# A machine learning approach for hypertension detection based on photoplethysmography and clinical data

**Article** *in* Computers in Biology and Medicine · April 2022

DOI: 10.1016/j.compbiomed.2022.105479

CITATIONS
40

READS
456

**3 authors:**

Erick Axel Martinez-Ríos
Tecnológico de Monterrey
**17** PUBLICATIONS **300** CITATIONS
SEE PROFILE

Luis Montesinos
Tecnológico de Monterrey
**48** PUBLICATIONS **1,373** CITATIONS
SEE PROFILE

Mariel Alfaro Ponce
Tecnológico de Monterrey
**63** PUBLICATIONS **595** CITATIONS
SEE PROFILE

Computers in Biology and Medicine 145 (2022) 105479

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# A machine learning approach for hypertension detection based on photoplethysmography and clinical data

Erick Martinez-Ríos, Luis Montesinos, Mariel Alfaro-Ponce *

*Tecnologico de Monterrey, School of Engineering and Sciences, Mexico City, 14380, Mexico*

ARTICLE INFO

*Keywords:*
Blood pressure
Classification
Machine learning
Risk stratification
Wavelet transform

ABSTRACT

High blood pressure early screening remains a challenge due to the lack of symptoms associated with it. Accordingly, noninvasive methods based on photoplethysmography (PPG) or clinical data analysis and the training of machine learning techniques for hypertension detection have been proposed in the literature. Nevertheless, several challenges arise when analyzing PPG signals, such as the need for high-quality signals for morphological feature extraction from PPG related to high blood pressure. On the other hand, another popular approach is to use deep learning techniques to avoid the feature extraction process. Nonetheless, this method requires high computational power and behaves as a black-box approach, which impedes application in a medical context. In addition, considering only the socio-demographic and clinical data of the subject does not allow constant monitoring. This work proposes to use the wavelet scattering transform as a feature extraction technique to obtain features from PPG data and combine it with clinical data to detect early hypertension stages by applying Early and Late Fusion. This analysis showed that the PPG features derived from the wavelet scattering transform combined with a support vector machine can classify normotension and prehypertension with an accuracy of 71.42% and an F1-score of 76%. However, classifying normotension and prehypertension by considering both the features extracted from PPG signals through wavelet scattering transform and clinical variables such as age, body mass index, and heart rate by either Late Fusion or Early Fusion did not provide better performance than considering each data type separately in terms of accuracy and F1-score.

## 1. Introduction

Hypertension or high blood pressure (BP) is a global public health concern that is associated with diseases such as heart attack [1], heart failure [2], diabetes [3], and strokes [4]. Besides, based on the data provided by the World Health Organization, hypertension influences 1.1 billion individuals worldwide, with more than a half living in low-income nations [5]. Additionally, overseeing this health condition is demanding, and expensive [6]. One of the main challenges that hypertension imposes is an early diagnosis [5]. The above occurs since high BP does not generate noticeable symptoms until it has reached a higher stage; that is why it is known as the silent killer [7]. The common method used to diagnose and monitor high BP is a sphygmomanometer. Based on the readings of this device, a patient can be classified into four stages of hypertension. This classification can be appreciated in Table 1. Nevertheless, cuff-based monitoring systems only take a snapshot of BP, which requires taking several readings and averaging them to obtain a more reliable value [8]. In addition, they generate arterial compression, which can affect the quality of life of users [9]. Moreover, self-measurements could provide a wrong diagnosis if the user does not follow an appropriate methodology to use cuff-based devices [10,11].

On the other hand, other technologies have been developed to provide a less invasive and precise system to monitor high BP. For example, photoplethysmography (PPG) waveforms are used to develop cuffless BP estimation or high BP risk stratification systems [13]. One common technique is to calculate the Pulse Arrival Time (PAT) based on PPG and electrocardiography (ECG) waveforms or extract morphological characteristics from the PPG signals and their first and second derivatives [14,15]. Then, these characteristics are used as inputs into regression or classification machine learning (ML) techniques to monitor or detect high BP. In addition, other authors have used deep learning (DL) techniques to avoid feature extraction approaches, in which multilayer perceptrons (MLPs) [16], convolutional neural networks (CNNs) [17], and recurrent neural networks (RNNs) [18] are the common types of architectures that have been employed. Finally, socio-demographic and clinical variables such as age, sex, and body mass index (BMI) have been

* Corresponding author.
*E-mail address:* marielalfa@gmail.com (M. Alfaro-Ponce).

https://doi.org/10.1016/j.compbiomed.2022.105479
Received 12 January 2022; Received in revised form 1 March 2022; Accepted 31 March 2022
Available online 2 April 2022

**Table 1**
High BP classification [12].

| BP classification | Systolic BP (mmHg) | Diastolic BP (mmHg) |
|---|---|---|
| Normal | <120 | <80 |
| Prehypertension | 120–139 | 80–89 |
| Stage 1 Hypertension | 140–159 | 90–99 |
| Stage 2 Hypertension | ≥ 160 | ≥ 100 |

used to develop ML-based high BP risk stratification systems [19].

However, several problems need consideration while developing systems for BP estimation or hypertension risk stratification. First, PPG morphological characteristics required a high-quality waveform collected with a high sampling rate to extract them effectively [17]. Moreover, morphological characteristics are susceptible to drifts, artifacts, and noise, which makes its extraction difficult [20]. On the other hand, DL requires a large sample size and high computational power to provide a high-performance model. Additionally, the setting of DL is more like an art since a precise methodology to set parameters such as the number of neurons, layers, or the learning rate is not available and is highly dependable on the particular task for which they are being used [21]. The above has made DL approaches black-box techniques that have great complexity but no interpretability, an impediment that limits ML or DL for medical applications [22]. Finally, considering only the clinical or socio-demographic information of the patient does not allow constant monitoring of cardiovascular health.

This work proposes the Wavelet Scattering Transform (WST) by Stéphane Mallat [23] as a signal representation technique for the classification of early stages of high BP applied to PPG waveforms to avoid the problems associated with DL techniques and morphological feature extraction from PPG waveforms. In contrast to other transforms such as the Fourier Transform (FT) and Wavelet Transform (WT), the WST generates a translation-invariant representation that is stable to small time-warping deformations to have a suitable signal representation for classification tasks. Moreover, ML fusion strategies to handle multimodal data, specifically Early and Late Fusion, are also explored to consider the effect of socio-demographic and clinical variables in combination with physiological signals for high BP early risk stratification. A schematic representation of the methodology proposed in this work is illustrated in Fig. 1.

In summary, the main contributions of the present research are the following:

- The WST is used to extract features from PPG signals for normotension (NT) and prehypertension (PHT) detection based on ML techniques.
- An analysis of the clinical variables (i.e., age, heart rate, sex, and BMI) associated with high BP was performed through ML techniques and feature selection using the Gini Importance.
- An evaluation of multimodal techniques (i.e., Early Fusion and Late Fusion) for NT and PHT detection through the use of clinical data (i.e., age, BMI, and heart rate) and PPG features derived from the WST.
- A comparison between unimodal and multimodal classifiers for detecting NT and PHT classes based on ML techniques.

This work is structured as follows. First, the related research is presented. Subsequently, the materials, background, and methodology are explained. Section 4 shows the results obtained through the proposed methods, while Section 5 shows the analysis of the results and discussion of the present study. Finally, Section 6 presents the conclusions and future work.

## 2. Related works

Previous studies have developed models based on ML techniques for high BP risk stratification, using clinical data or physiological waveforms as inputs. In this regard, the models are trained with variables such as age, BMI, sex, and heart rate, or the raw PPG waveforms or the characteristics extracted from PPG by using time analysis, frequency analysis, time-frequency analysis, or chaotic analysis [24]. Moreover, morphological features from PPG signals have also been studied and can be viewed inside the scope of time analysis. In addition, DL techniques have been applied to avoid the process of handcrafting characteristics from the PPG waveforms [24]. An overview of the approaches for high BP detection is presented in Fig. 2. Furthermore, a brief literature review and analysis are provided below.

### 2.1. High BP detection based on clinical data

Some examples of works that have used clinical data for high BP detection are described in this section. Lopez et al. [25] proposed a model to evaluate the relationship between sex, race, BMI, age, smoking status, and risk of hypertension using a logistic regression (LR) based on data taken from the USA' National Health Nutrition Examination Survey (NHANES) from 2007 to 2016. The fitted model's sensitivity was 77%, the specificity was 68%, and the area under the receiver operating characteristic curve (AUC) was 73%. In a different work, Lopez et al.

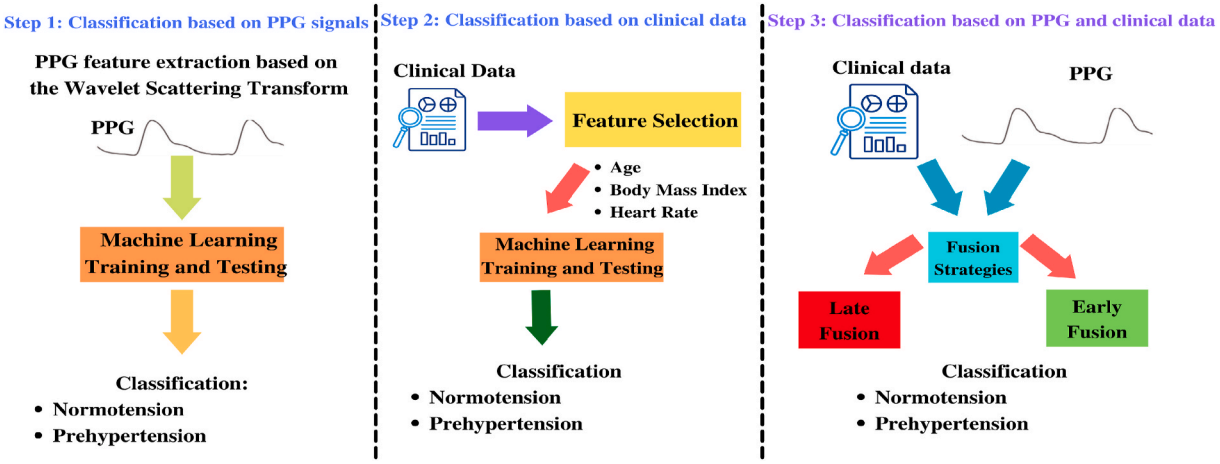


**Fig. 1.** The overall methodology applied in this work. The first step consists of classifying NT and PHT classes based on features from PPG signals derived from the WST. The second step considers clinical data and feature selection to classify NT and PHT classes. Finally, Late Fusion and Early Fusion are tested to assess both clinical data and PPG features for classifying NT and PHT classes.

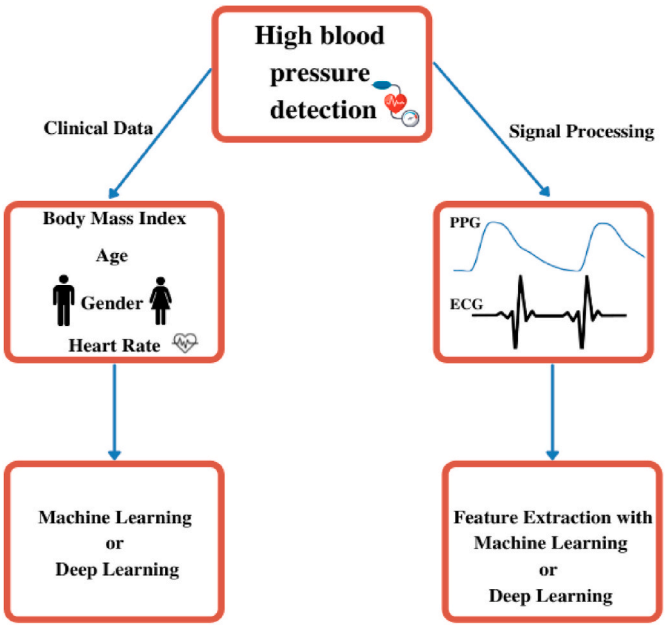


**Fig. 2.** Overview of approaches for high BP detection based on clinical data or physiological signals through ML techniques.

[19] proposed an artificial neural network (ANN) in contrast to his original work in which a LR was fitted by employing the NHANES dataset again. The reported results showed a sensitivity of 40%, a specificity of 87%, and an AUC of 77%. Despite the improvement in the AUC, the complexity of the model increased considerably when using three hidden layers in the proposed ANN compared to a LR. LaFreniere et al. [26] also opted for a DL approach by employing an ANN based on data taken from Canadian Primary Care Sentinel Surveillance Network (CPCSSN). The above work reported an accuracy of 82%, considering as input data the age, sex, BMI, lipoproteins, triglycerides of each subject. More sophisticated approaches like CNNs have also been studied. For example, Luo et al. [27] employed a CNN with a 1-dimensional kernel size by analyzing the Medical Information Mart for Intensive Care (MIMIC) II Waveform Database Matched Subset [28]. The reported results demonstrated an accuracy of 89.95%.

### 2.2. High BP detection based on physiological data

Related to how ECG and PPG waveforms have been used for hypertension risk stratification, the following proposals are described. Liang et al. [29] studied the relationship between the morphological features of the PPG signal and its first and second derivatives with systolic BP. The PPG-BP dataset [30] was studied in this work. A total of 125 morphological features were defined and studied. Nevertheless, based on different feature selection approaches, only ten features were considered to train a Linear Discriminant Analysis (LDA), weighted k-Nearest Neighbor (KNN), LR, and cubic Support Vector Machine (SVM). The authors reported an F1-score of 72.97% for NT versus PHT, 81.82% for NT, and PHT versus hypertension, and 92.31% for NT versus hypertension. In another study, Liang et al. [31] studied the potential use of PPG morphological features and PAT based on both the ECG and PPG signals available in the MIMIC dataset. The F1-score for the classification of NT versus PHT was 84.34%, 94.84% for NT versus hypertension, and 88.49% for NT plus PHT versus hypertension.

One of the disadvantages of extracting morphological features is the need for a high-quality PPG waveform. Nevertheless, this becomes difficult since the PPG and its derivatives are distorted due to artifacts and noise. In addition, calculating the PAT from ECG and PPG is complicated due to the lack of synchronization of both waveforms, the need for stable and high-quality signals, and it is more invasive to the user [24,32,33]. Yao et al. [34] proposed to filter the PPG signal with a technique called Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and wavelet threshold analysis to overcome the problems of noise and artifacts. Nonetheless, the methodology or algorithm to extract the features is not extensively explained, and it only explored 20 morphological features in contrast to Liang's work [29] where 125 features were studied. Moreover, in the work of Liang et al. [31], and Yao et al. [34], the features used for the classification tasks differ a little due to the importance that each of the feature selection techniques assigns to the computed features.

Other works have opted to use pre-trained CNNs combined with the time-frequency representations of PPG waveforms to avoid the feature extraction approach. For example, in Ref. [17] the Continuous Wavelet Transform (CWT) was applied to the PPG signal by considering a Morse Wavelet. The produced CWT scalograms were used to fine-tune the GoogLeNet pre-trained CNN. The data used for this study was the MIMIC dataset. The results in terms of the F1-score showed a value of 80.52% for NT versus PHT, 92.55% for NT versus hypertension, and 82.95% for NT plus PHT versus hypertension. Likewise, Sun et al. [20] proposed the Hilbert-Huang Transform (HHT) to apply it to the PPG signals and its derivatives to combine the spectrogram of each waveform and produce RGB images to be used as input to fine-tune the AlexNet pre-trained architecture. The F1-scores of this method showed 85.80% for NT versus PHT, 98.90% for NT versus hypertension, and 93.54% for NT plus PHT versus hypertension. One of the critical disadvantages of applying a Transfer Learning approach using pre-trained CNNs is that they require high computational power, a drawback also mentioned in Ref. [17]. Furthermore, the interpretability is also affected due to the use of deep CNNs like AlexNet and GoogLeNet. One of the advantages of transfer learning is that it can be used in situations with a lack of training examples. Nevertheless, a specific number of training examples while applying transfer learning is not specified, and different authors that have used this approach have reported sample sizes of 582 [20], 510 [35], and 121 [17].

Other tools explored for feature extraction are statistical features in the time domain and features computed in the frequency domain. An example of this approach is the work of Aydemir et al. [36], in which the WT, chirp z-transform (CZT), the total band power (TBP), autoregressive model parameters (ARMP), zero-crossing rate (ZCR), and the standard deviation (SD) of the first derivative of the PPG were used to obtain features from the PPG waveform. According to the authors, by combining features from the CZT, TBP, ARMP, ZCR, and the SD of the first PPG derivative and training an SVM, an accuracy of 77.52% was obtained. One of the problems of this approach is that it is necessary to compute many PPG signal representations and extract features from those representations to train and fit the model. In addition, a priori, there is no certainty that those representations will be helpful to discriminate between the hypertension classes. Finally, few works have studied the use of both physiological signals and clinical data for hypertension risk stratification. Of the works above, only Yao et al. [34] considered the effect of both types of data.

## 3. Materials and methods

This section presents the dataset used for this work and the background related to the techniques used to develop this study. Moreover, the procedure used for this study is also explained. The methods are divided principally into three parts, classification of high BP stages based on PPG waveforms and WST, classification of high BP stages based on socio-demographic and clinical data, and classification based on multimodal data fusion strategies.

### 3.1. Dataset

The dataset used for this work was collected at the Guilin People's

Hospital in Guilin, China [30]. There are a total of 219 samples from patients admitted to the hospital. The sample has 37% of subjects diagnosed with NT, 38% with PHT, and 25% with Stage 1 and Stage 2 hypertension. Moreover, age, sex, BMI, heart rate, height, and weight were collected for each patient. In addition to this socio-demographic and clinical data, the patient's systolic and diastolic BP and three PPG signal segments per subject were collected with a sampling frequency of 1 kHz. The segments were recorded for 2.1 s, which led to segment lengths of 2100 samples. Three PPG waveform segments per subject were collected to have a variety of PPG segments and select the ones with the lowest amount of noise, drift, and artifacts. The waveform selection was based on a Signal Quality Index (SQI) measured by computing the skewness of the segment. The authors provided the values of skewness for all the subject's PPG segments [30]. In this case, a high-quality segment will be the one with the highest skewness value. It is worth mentioning that the Perfusion Index is the gold standard for measuring the quality of PPG waveforms [37]. However, based on the analysis presented by Krishnan et al. [38], and Elgendi et al. [37] and the recommendation given in the dataset documentation [30] the skewness was the selected metric. The skewness can be computed with equation (1). Where $x_i$ is a single sample of the signal segment $x$, $\mu_x$ is the mean of the signal, $N$ is the signal's segment length, and $\sigma$ is the SD of the signal segment.

$$SQI = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\mu_x)^3}{\sigma^3} \tag{1}$$

Only the NT and PHT classes were compared for this analysis since they have similar proportions in the dataset of 37% and 38%, respectively. The above was decided since ML techniques tend to classify better the majority class, but they give the wrong impression of a highly accurate model [39]. The above led to analyzing only 165 PPG signal segments and samples related to age, sex, BMI, heart rate, weight, and height. Moreover, it was necessary to filter out the signals to discard the high-frequency noise. An Infinite Impulse Response Butterworth low-pass filter was used to remove the signals' high-frequency content. The low-pass filter characteristics were set to a cut-off frequency of 25 Hz of order six, as suggested in the work of Chowdhury et al. [40].

### 3.2. Wavelet scattering transform

The WST was first introduced by Stéphane Mallat [23] and further developed by Anden et al. [41] and Bruna et al. [42]. Its main objective is to generate a representation of signals or images used for classification tasks. The above is achieved by creating a translation-invariant representation that is stable to small time-warping deformations. As stated in Ref. [42] the magnitude of the FT is invariant to translations but not stable to small-time warping deformations, and the WT is translation covariant. The WST employs convolution with wavelets, nonlinearity applied through modulus operations, and averaging through scaling functions or low-pass filters to generate stability to deformations.

The WST is constructed with the help of a scaling function $\phi(t)$ and a family of dilated wavelets $\psi_\lambda(t)$, where $\lambda > 0$. The scaling function is a low-pass filter that computes the invariant representation by averaging the scattering coefficients at a particular size defined by $2^J$. The zero-order scattering coefficients ($S_0x(t)$) are computed by performing the convolution of the input signal $x(t)$ and the scaling function $\phi(t)$. This process is expressed in equation (2).

$$S_0x(t) = x(t)\star\phi(t) \tag{2}$$

On the other hand, $\psi_\lambda(t)$ are wavelets dilated at different frequencies. For any $\lambda > 0$, a dilated wavelet of center frequency $\lambda$ is obtained. These wavelets work as band-pass filters and are expressed in the time and frequency domain as shown in equations (3) and (4), respectively [41].

$$\psi_\lambda(t) = \lambda\psi(\lambda t) \tag{3}$$

$$\widehat{\psi_\lambda}(\omega) = \widehat{\psi}\left(\frac{\omega}{\lambda}\right) \tag{4}$$

These wavelets are used to recover the high-frequency information discarded from the signal when computing the zero-order scattering coefficients $S_0x(t)$, the frequency resolution of this family of wavelets is controlled by a factor $Q$ that defines the number of wavelets per octave or in other words the number of filters per octave. The center frequency of the wavelets is normalized to 1. Considering that the $Q$ factor represents the number of wavelets per octave, lambda is equivalent to $\lambda = 2^{\left(\frac{k}{Q}\right)}$ for $k\varepsilon\mathbb{Z}$. Besides, the bandwidth of $\widehat{\psi}$ is in the order of $Q^{-1}$ these band-pass wavelet filters span the entire frequency axis [41].

The first-order coefficients ($S_1x(t, \lambda_1)$) are obtained by first convolving the family of wavelets dilated at different octaves controlled by the factor $Q$ with the input signal, later the modulus is calculated, and then the result is convolved with the scaling function $\phi(t)$. This process is described in the expression below.

$$S_1x(t,\lambda_1) = |x(t)\star\psi_{\lambda_1}|\star\phi(t) \tag{5}$$

Moreover, to obtain the second-order coefficients ($S_2x(t, \lambda_1, \lambda_2)$), the coefficients obtained by convolving the family of wavelets with the input signal using the first family of wavelets $\psi_{\lambda_1}$ and its modulus are convolved again with a second family of wavelets $\psi_{\lambda_2}$, and again the modulus is calculated. Finally, the result of the previous operations is convolved with the scaling function $\phi(t)$. There is only one wavelet per octave for this second layer, which implies that $Q = 1$. These operations are described in the expression below [41].

$$S_2x(t,\lambda_1,\lambda_2) = ||x(t)\star\psi_{\lambda_1}|\star\psi_{\lambda_2}|\star\phi(t) \tag{6}$$

The WST can be extended to further layers. Nevertheless, with more layers, the signal's energy starts to dissipate. According to Bruna et al. [42], and Anden et al. [41] for most applications, two layers are sufficient since the energy of the signals starts dissipating with more layers. In Fig. 3 it is shown the scattering transform operations in a graphical representation. Notice that the tree representation of the scattering transform also receives the name of Deep Scattering Spectrum [41,42].

Since the WT is a contractive operator as well as the complex modulus, the whole transform is a contractive operator. This property reduces the variance of the representation but also makes it stable to additive noise. Moreover, this method can be compared as an analogy to CNNs, since the input signal $x(t)$ is first processed through a filtering operation, then a nonlinearity is applied, which refers to the complex modulus operation, which is analogous to the activation functions used in CNNs, and finally, the convolution with the scaling function $\phi(t)$ which computes an average of every coefficient could be compared with the pooling operation of CNNs. The main differences are that in the WST, the filters are fixed, while in the CNN, the filters are learned. Furthermore, each layer in the scattering transform produces the features used

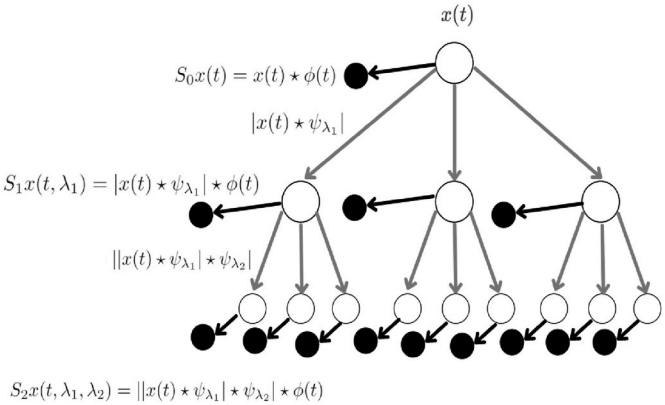


**Fig. 3.** Schematic representation of the WST.

for classification, while in the CNN, the final layers produce the final features or representation for classification. Besides, each layer can be presented and understood mathematically. At the same time, in the CNNs, the analysis becomes more complex due to the nonlinearities, hyperparameters, and the deepness of the network [41]. Finally, the WST has the following properties with an adequate selection of wavelets [41]:

- Contracting operator
- Preserves the energy
- Stable to small time-warping deformations
- Fast Computation

### 3.3. Fusion approaches for multimodal data based classification

The literature has opted to use model-agnostic approaches to handle multimodal data in the context of ML. These approaches can be divided into Early Fusion (feature-based), Late Fusion (decision-based), and Hybrid Fusion. Early Fusion combines the features immediately after they have been extracted; this is commonly done by concatenating them. Otherwise, Late Fusion applies the integration after each modality has made a decision either in regression or classification tasks. Finally, Hybrid Fusion merges the outputs from Early Fusion and individual unimodal predictors. One of the key advantages of model-agnostic approaches is that they can be applied to any unimodal classifier or regressor [43].

#### 3.3.1. Early fusion (feature level)

Early Fusion can be interpreted as an attempt by researchers to generate multimodal representation learning, as it can learn to use the correlation and interactions between low-level features of each data modality. Besides, it only requires the training of a single model, compared to Hybrid or Late Fusion, making the final model more simple [43,44].

#### 3.3.2. Late Fusion (decision level)

Late Fusion allows the use of different models for each of the data modalities. Besides, this technique makes it easier to make predictions when one or more modalities are missing. Nevertheless, Late Fusion does not consider low-level interactions between modalities as Early Fusion. Instead, Late Fusion merges the results of each modality through averaging, voting mechanism, weighting based on channel noise, signal variance, or a learned model [43,44].

Only Early and Late Fusion strategies were employed for this study since these strategies do not require a meticulous design process. In addition, Hybrid Fusion or Joint Fusion strategies have been developed in the context of DL, which can affect the interpretability of the final model [44].

### 3.4. Machine learning techniques

This section provides a brief overview of the ML techniques used in this work. Classical ML techniques were selected since they can deal with small datasets instead of DL techniques or random forest. In addition, the selection of the ML algorithms was determined by considering the simplicity of the generated model. For instance, LR and LDA produce linear decision boundaries, which contributes to the simplicity of the generated model and consequently to its interpretability [45]. Besides, decision trees make classifications based on rules which can be easily understood by humans [45]. Furthermore, the KNN algorithm was selected since it has shown good performance for similar classification tasks like in the work of Liang et al. [33]. Finally, SVM was chosen since it can produce linear decision boundaries and does not require a large sample size to be trained [46].

All ML techniques tested in this work were trained using the scikit-learn library of Python [47]. The computational resources employed to train the models were 25.46 GB of RAM and an Intel(R) Xeon(R) CPU of 2.30 GHz.

#### 3.4.1. Logistic regression

LR is a technique used to model binary classification problems. To map a set of inputs $X$ to a value between 0 and 1, the LR uses the logistic function representation that can be appreciated in equation (7) [48].

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{7}$$

Where the coefficients $\beta_0$ and $\beta_1$ should be adjusted based on the given training data. One of the methods used to fit the LR model is through maximum likelihood [48].

#### 3.4.2. Linear Discriminant Analysis

LDA, also known as Fisher's discriminant, is a linear classification technique that tries to maximize the distance between the means of two classes and minimize the variance inside the classes. LDA provides class discrimination by generating a decision region between the different classes, assuming that the data follows a normal distribution. The decision region is decided by maximizing the between-class variance and the within-class variance ratio. LDA can be applied to classification, and dimensionality reduction tasks [48].

#### 3.4.3. Decision trees

Decision or classification trees are sequential models. When a test set arrives, the decision tree compares a numeric characteristic to a threshold value or a nominal characteristic to a range of possibilities. A decision tree classifies a new data point as belonging to the most frequent class in a partitioned region when it falls within that region. Different algorithms have been developed to construct decision trees, like C4.5, CART, and SPRINT. The main objective of decision trees is to generate the best partitioning. This partitioning process is performed based on impurity. If only one class belongs to a subset, it is established as pure; otherwise, it is considered impure. The purer a split is, the better it is. There are different techniques to measure how to perform the splitting process; among them are Information Gain, Gini Value, and Gain Ratio [49].

#### 3.4.4. K-nearest neighbor

In the KNN algorithm, the learning process is accomplished by comparing a given test set to similar training sets. There are $n$ characteristics that characterize the training sets. When a new test example appears, KNN searches for the $k$ training points in the n-dimensional space closer to the unknown test set. A distance metric, commonly the Euclidean distance, is used to define the closeness of the test data point to the training set. A majority voting mechanism defines the most prevalent class among the test set. For instance, if $k = 1$, the test set is assigned to the training set class in the n-dimensional pattern space closest to it. The KNN algorithm is considered a lazy learner because it memorizes the training dataset rather than learning a discriminative function from the training data [50].

#### 3.4.5. Support vector machines

SVM is one of the most powerful ML techniques, in which the classification problem is solved by constructing hyperplanes that maximize the margin between the classes and training points, as shown in Fig. 4. The training data points on the margins receive the name of support vectors. SVM penalizes the distance of points on the opposite side of their margin when an overlap appears between the classes. The above allows a limited number of uncorrected classifications to be accepted close to the margin. Another characteristic of SVMs is kernel functions to convert non-linearly separable classes into a higher-dimensional space where they can be separated. One of the most common kernels that are used are radial basis functions (RBFs) [46].

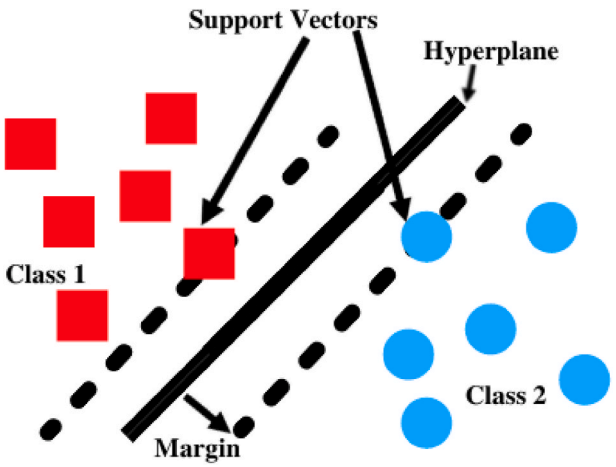


**Fig. 4.** Separation of two classes using optimal hyperplanes.

### 3.5. Classification of high BP based on PPG signals and WST

For this work, the WST was performed using the Python implementation through the Kymatio library [51]. It is necessary to establish a set of wavelets per octave ($Q$) to compute the first-order scattering coefficients that allow controlling the resolution in the frequency domain and the size of the invariant $J$, which controls the time-averaging window. Moreover, the length of the signals needs to be a power of two. Therefore, the PPG segments need to be cut into segments of 2048 samples that are equivalent to $2^{11}$. Besides, the number of wavelets per octave was established as $Q = 1$, and the size of the invariant was defined as $J = 5$. Otherwise, for the second-order scattering coefficients, it is not possible to manipulate the number of wavelets per octave as it is established in the work of Bruna et al. [42], Anden et al. [41], and in the Kymatio documentation [51]. Therefore, only one wavelet per octave is allowed. It is essential to mention that although a larger value of wavelets per octave can provide a better frequency resolution, the number of features obtained will be larger and could lead to overfitting problems and provide a more complex model. Therefore, to obtain a sparse representation, only one wavelet per octave was defined in this work.

After applying the WST and obtaining the corresponding WST coefficients, a two-dimensional array is obtained. One dimension represents the scales or frequency resolution of the scattering transform (y-axis) and the second dimension represents the discrete-time of the signal (x-axis). The scalogram representations shown in Fig. 5 are averaged along the time dimension to obtain features from this representation. Following the suggestion of the Kymatio documentation, the coefficients of the WST were transformed with the natural logarithm; this is called the log-scattering transform. Consequently, the WST coefficients were averaged along time. A comparison of a PHT subject with a normal subject after averaging along time the coefficients can be appreciated in Fig. 6. After averaging along time, a completely time-invariant representation is obtained with no time resolution. In this case, the representation that is obtained is similar to a frequency representation since each value in the y-axis of the scalograms shown in Fig. 5 represents a frequency scale of the WST.

In Fig. 6, the zero-frequency scale represents the average along time of the zero-order WST, the frequency scales from 1 to 6 are the average along time coefficients of the first-order WST, and the frequency scales from 7 to 18 are the average along time coefficients of the second-order WST. Each of the values of each frequency scale was used as input features to the ML algorithms. A total of 19 features were obtained by applying the above procedures.

### 3.6. Classification of high BP based on clinical and socio-demographic variables

Another analysis that was performed was to classify the NT, and PHT subjects through the use of only the socio-demographic and clinical variables available in the PPG-BP dataset [30] and applying only classic ML classifiers to maintain the interpretability and simplicity of the models. The first step consisted in considering all variables available in the dataset to fit the models. In this case, the techniques were trained using the BMI, age, heart rate, weight, height, and sex, which constitutes six variables. Subsequently, by employing feature selection, the model was fitted again to reduce the dimensionality and compare the reduced model with the original one that uses all variables of the dataset. The above is done to generate the simplest model that explains the data [52].

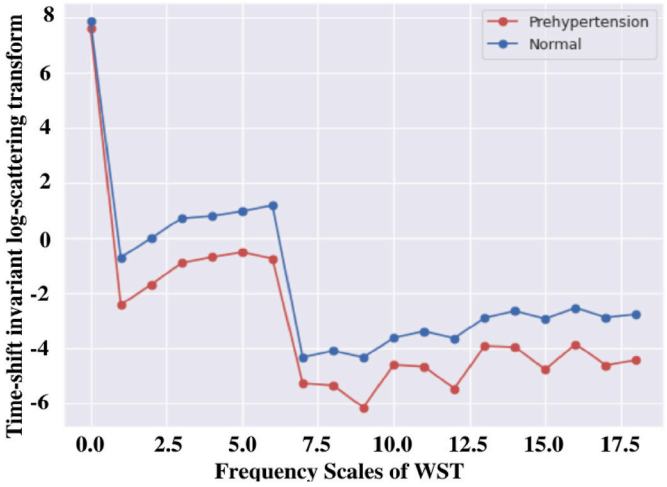


**Fig. 6.** Time-shift invariant representation after averaging along time the log-scattering transform. A comparison between a normal (blue) and PHT (red) subject is provided.

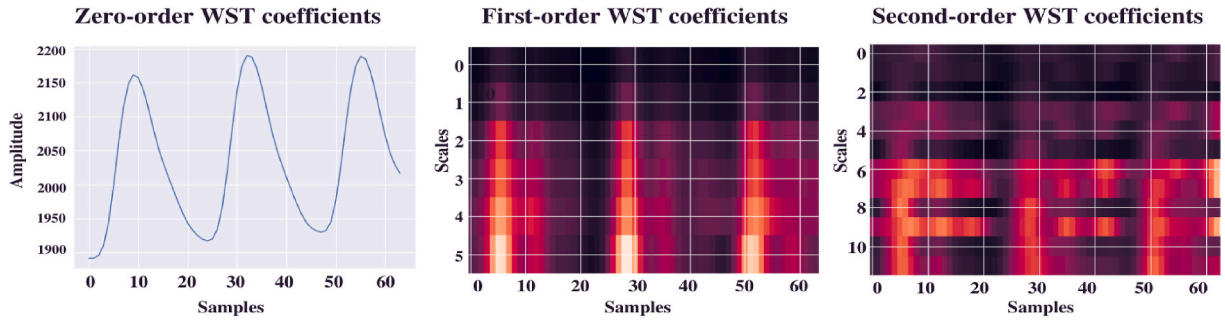


**Fig. 5.** Example of the representation generated by computing the WST transform by considering $Q = 1$ for the first-order WST. From left to right it is presented the zero, first, and second-order WST coefficients.

The feature selection process selected for this analysis was through the mean decrease of impurity, also known as Gini Importance. This index estimates the probability that a specific feature is classified incorrectly when selected randomly. The expression of the Gini Index(GI) is shown in equation (8), where $p_i$ is the relative frequency of class ($c$) in ($D$) [49].

$$GI(D) = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{8}$$

If all the samples in a feature are associated with a single class, the feature is pure. If the equation outputs a value near zero, it is concluded that the feature is pure. Otherwise, if the value is near one, it is concluded that the feature is impure. The relative feature importance was computed by training a decision tree through the scikit-learn package of Python. The results of the feature importance can be appreciated in Fig. 7.

This analysis shows that BMI and age have the highest relative feature importance between NT and PHT classes. This is in accordance with the variables other authors have used in the literature [19,25]. In the case of weight and height, they are associated with BMI; therefore, to avoid redundancy, both variables will be excluded to reduce the model. Finally, sex had the lowest relative feature importance. Consequently, it was decided to exclude it from this analysis. The above process led to considering only three variables (i.e., BMI, age, and heart rate) from the original six.

### 3.7. Classification of high BP based on PPG signals and clinical data

A fusion strategy was applied to consider the predictive power of multimodal data coming from the PPG waveforms and clinical data. The fusion strategies were implemented without any deep neural network to avoid computational complexity and maintain the final model's interpretability. The first method that was tested is based on Early Fusion, in this case, the features extracted from the PPG waveforms through the WST were concatenated with the variables selected through the Gini Impurity Index to produce a single feature vector this is graphically shown in Fig. 8.

On the other hand, the Late Fusion approach was implemented by merging the results of the best classifiers by majority voting. The schematic representation of the Late Fusion approach can be appreciated in Fig. 8. The above implies that every classifier votes for a class, and the class with the most votes wins. Since a majority voting was selected, it is necessary to select an odd number of classifiers to be trained. Only three classifiers were selected to maintain the simplicity of the final model.

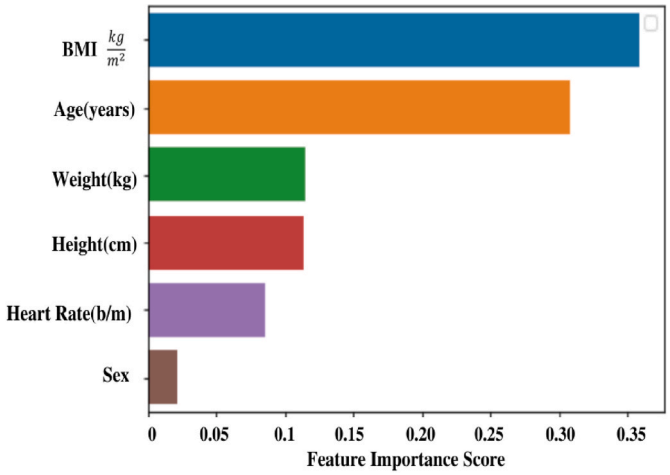


**Fig. 7.** Relative Feature Importance based on the Gini Impurity Index for the clinical data.

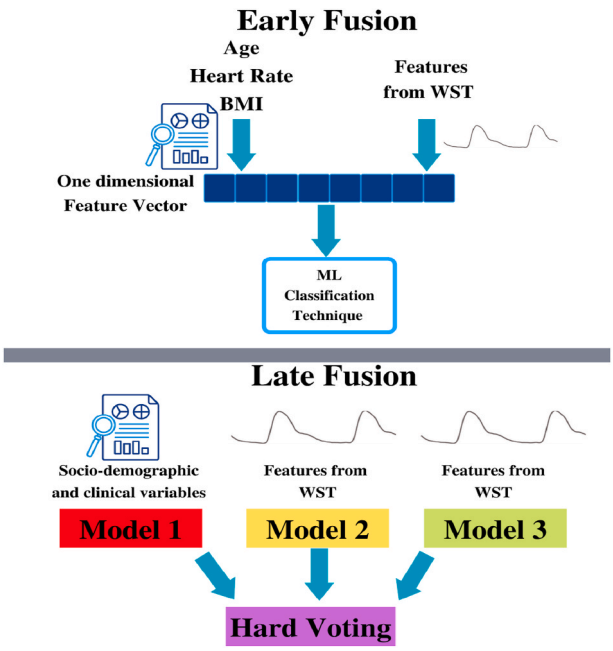


**Fig. 8.** Schematic representation of the Early and Late Fusion approaches for multimodal data.

## 4. Results

For this work, the data was divided into 75% for training and 25% for testing. Moreover, for each of the classification stages, the variables and features from the WST were standardized with a z-score according to the expression shown in equation (9), where $x$ represents a sample, $\mu_x$ is the mean of the training examples $x$, and $\sigma_x$ is the SD of $x$. This process allows having a mean of zero and an SD of one across all features. Standardizing the data is crucial when there are different ranges of values across all features or variables since distance-based techniques may give a higher weight to variables with a broader range.

$$z = \frac{x - \mu_x}{\sigma_x} \tag{9}$$

Moreover, the metrics used to measure the performance of the fitted models were the Accuracy (equation (10)), Recall/Sensitivity (equation (11)), Precision (equation (12)), and F1-score (equation (13)). In these equations, TP is the True Positives, TN is the True Negatives, FP is the False Positives, and FN is the False Negatives. These metrics were calculated for both NT and PHT subjects. The F1-score, precision, and recall were reported exclusively for the testing set, and the accuracy was reported for both the training and testing sets to evaluate the overfitting of each ML technique.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F1 - score = \frac{2(Precision)(Recall)}{Precision + Recall} \tag{13}$$

### 4.1. Results of the ML training with WST features

The results of employing the WST as a feature vector can be appreciated in Table 2. Moreover, the behavior of the accuracy for the training

**Table 2**
Results of applying the WST and ML Classifiers in the PPG waveforms. In bold is shown the best classifier.

| ML Method | Hyperparameters | Class | Training Accuracy% | Testing Accuracy% | Precision % | Recall % | F1- Score % |
|---|---|---|---|---|---|---|---|
| **SVM** | **C**:100, **Kernel**: RBF **Gamma**: 0.01 | NT | **72.35** | **71.42** | **84.61** | **52.38** | **64.70** |
| | | PHT | | | **65.51** | **90.47** | **76.00** |
| LR | **C**: 10 **Penalty**: L2 | NT | 66.66 | 64.29 | 65.00 | 61.90 | 63.41 |
| | | PHT | | | 63.64 | 66.67 | 65.12 |
| LDA | **Solver**: Singular Value Decomposition (SVD) | NT | 70.73 | 59.52 | 58.33 | 66.67 | 62.22 |
| | | PHT | | | 61.11 | 52.38 | 56.41 |
| KNN | **Nearest Neighbors**: 6 | NT | 65.85 | 64.29 | 62.50 | 71.43 | 66.67 |
| | | PHT | | | 66.67 | 57.14 | 61.54 |
| Decision Tree | **Criterion**: Gini Impurity **Max depth**: 6 | NT | 82.92 | 57.14 | 55.56 | 71.43 | 62.50 |
| | | PHT | | | 60.00 | 42.86 | 50.00 |

and testing sets by changing the value of the hyperparameters can be appreciated for all the ML techniques in Fig. 9 except for the LDA, since the selected solver (singular value decomposition) does not allow controlling a shrinkage parameter according to the scikit-learn documentation for the LDA technique [47]. The candidate set of the regularization hyperparameters ($C$) for the SVM were $C$ = [0.1, 1, 10, 100, 500, 1000], a fix gamma of 0.01, and a RBF kernel. For the LR, the candidate set of the regularization hyperparameters ($C$) by applying an $L2$ regularization or ridge regression were $C$ = [0.01, 0.1, 1, 10, 50, 100]. The regularization hyperparameters were set exponentially as suggested by the scikit-learn documentation related to the LR and SVM [47]. The above was done for all the experiments of this work. For the KNN, the values were chosen to be integer values ranging from 1 to 9 neighbors. For the decision tree, the depth was varied with integer values ranging from 1 to 9.

Additionally, Fig. 10 shows the box plots of the features of the WST compared between NT versus PHT subjects. It is possible to appreciate a difference in the means between both classes across all features. This difference suggests a difference in the energy of the PPG signals between both classes.

The observed difference in the means was tested through a two-sided Welch's T-Test, with a significance level of $\alpha = 0.05$. The significance level of 0.05 was chosen since it is the most common in practice [53,54]. The results can be appreciated in Table 3. The Welch's T-Test was chosen since the sample size and the variance of each class are unequal, as appreciated in Table 3 through the SD values reported for each feature and class. Moreover, this type of test provides better control of Type 1 error when the premise of homogeneity of variance is not satisfied, as explained by Delacre et al. [55]. The two-sided version of Welch's T-Test was chosen since there is insufficient evidence to state that the mean value of the WST features for the NT class is greater than the PHT class or vice versa. Therefore, it was decided to use the two-sided version of Welch's T-Test instead of a one-sided Welch's T-Test to consider the effect in both directions. The normality requirement of Welch's T-Test

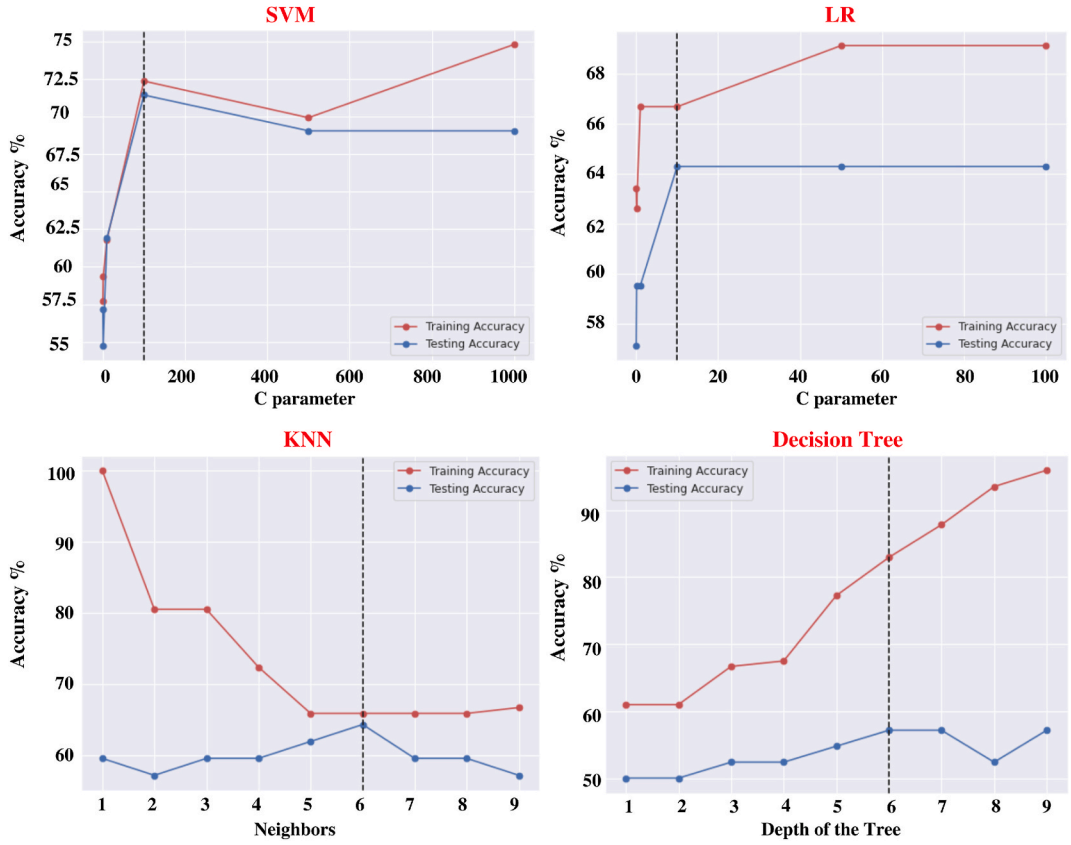


**Fig. 9.** Training (red) and testing (blue) set accuracies while varying the hyperparameters of the respective ML techniques trained with the features of PPG signals derived from the WST for NT and PHT detection. The vertical dash line shows the hyperparameter value that produces the best trade-off between the training and testing sets' accuracies. The top graphs show the results of the SVM and LR; the bottom graphs show the results of KNN and the decision tree.

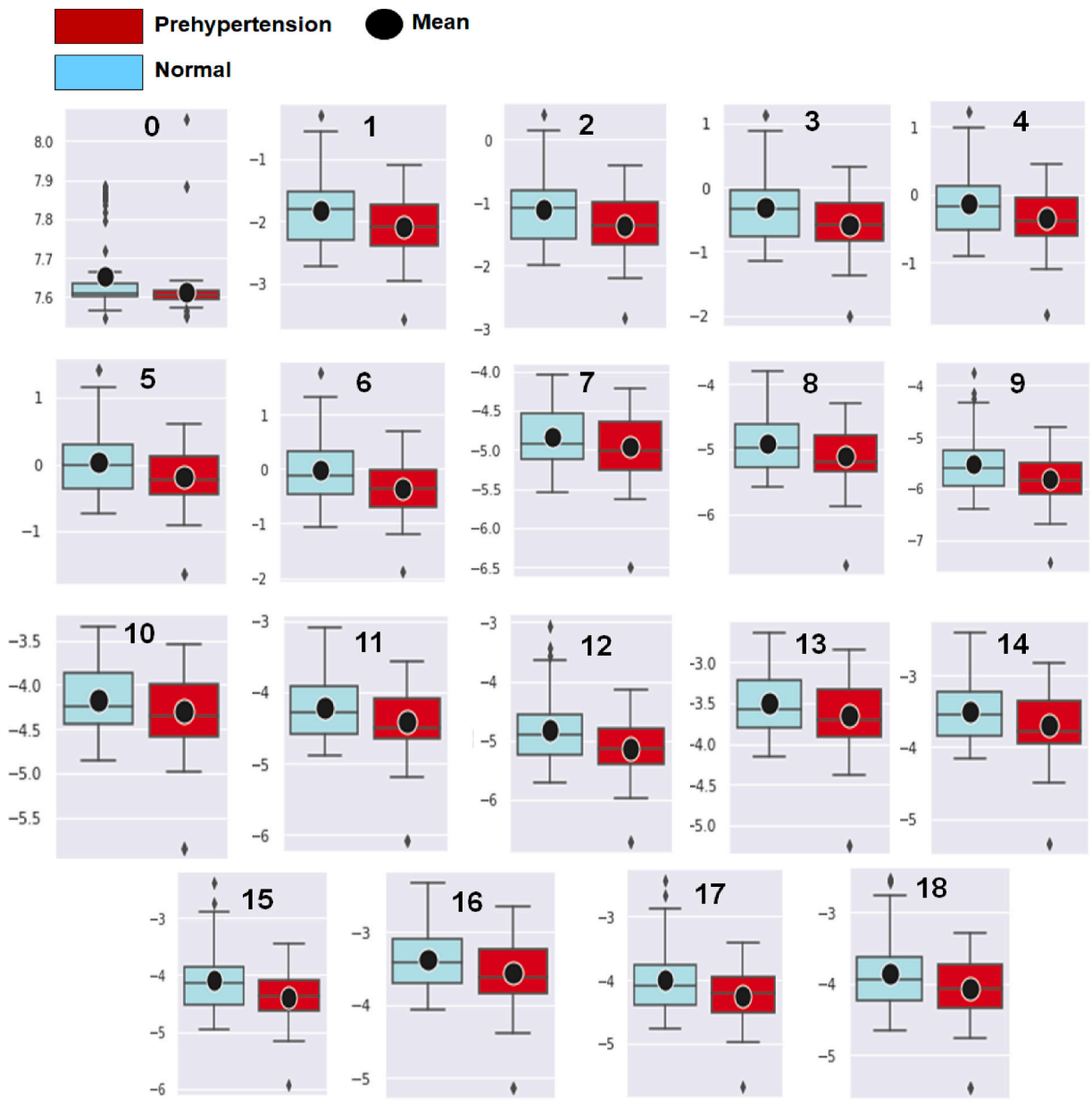


**Fig. 10.** Comparison of the box plot of each frequency scale obtained through the WST after averaging them along time between NT and PHT subjects.

was assumed following the central limit theorem, which states that as the sample size increases, the distribution of sample means approaches a normal distribution, independent of the population's distribution [56]. For the central limit theorem to hold, sample sizes of 30 or more are frequently regarded as sufficient [54]. In this case, the sample size of the NT class is 80 and for the PHT class is 85. For most features, it is possible to appreciate that the difference is statistically significant based on the reported $p - values$, except for features 7 and 10. However, there is an overlapping between the spreads of each feature between NT versus PHT. The above also explains why, despite a degree of classification between both classes, the values are below the 72% of accuracy when employing linear classifiers (e.g., SVM, LDA, and LR).

### 4.2. Results of the ML training with socio-demographic and clinical data

When only the clinical and socio-demographic data are considered to train the models, the results presented in Table 4 are obtained. Moreover, the accuracy for the training and testing sets by varying the hyperparameters is shown in Fig. 11. The candidate set of hyperparameters for the SVM were $C = [0.1, 1, 10, 100, 1000, 10000]$, a fix gamma of 0.00001, and a RBF kernel. For the LR, the candidate set of regularization hyperparameters were $C = [0.1, 1, 5, 10, 100]$ by considering an $L2$ regularization. For the KNN, the values were chosen to be integer values ranging from 1 to 9 neighbors. For the decision tree, the depth was varied with integer values starting from 1 and ending with 9. For this first test, weight, height, BMI, sex, age, and heart rate were used as input variables to train each model.

Furthermore, after applying feature selection and considering only age, BMI, and heart rate to train the models, the results are presented in Table 5. Similar to the other experiments of this work, the training and testing set accuracies are shown in Fig. 12 for each ML technique for different hyperparameters values. The candidate set of regularization hyperparameters for the SVM were $C = [0.1, 1, 10, 100, 500, 1000, 10000]$, a fix gamma of 0.001, and a RBF kernel. For the LR, the candidate set of regularization hyperparameters were $C = [0.1, 1, 5, 10, 100]$ by considering an $L2$ regularization. For the KNN, the values were chosen to be integer values ranging from 1 to 9 neighbors. In the case of the LDA, there are no hyperparameters that can be controlled. Finally,

**Table 3**
Welch's T-Test to determine the difference of mean for the obtained coefficients by employing the WST between NT and PHT. In bold is shown the non-significant features.

| Two-sided Welch's T-Test | | | | |
|---|---|---|---|---|
| $\alpha = 0.05$ $n_{NT} = 80$, $n_{PHT} = 85$ | | | | |
| WST Scales | NT Mean ± SD | PHT Mean ± SD | Test Statistic | p-value |
| 0 | 7.6542 ± 0.0995 | 7.6116 ± 0.0596 | 3.2559 | 0.0013 |
| 1 | −1.8228 ± 0.5722 | −2.1005 ± 0.4567 | 3.4111 | 0.0008 |
| 2 | −1.0980 ± 0.5627 | −1.3696 ± 0.4494 | 3.3923 | 0.0008 |
| 3 | −0.3184 ± 0.5391 | −0.5748 ± 0.4321 | 3.3363 | 0.0010 |
| 4 | −0.1273 ± 0.4995 | −0.3566 ± 0.4067 | 3.2032 | 0.0016 |
| 5 | 0.0452 ± 0.4987 | −0.1862 ± 0.3993 | 3.2581 | 0.0013 |
| 6 | −0.0223 ± 0.6284 | −0.3510 ± 0.4711 | 3.7605 | 0.0002 |
| **7** | **−4.8378 ± 0.3803** | **−4.9621 ± 0.3938** | **2.0491** | **0.0420** |
| 8 | −4.9219 ± 0.4455 | −5.1063 ± 0.4193 | 2.7173 | 0.0073 |
| 9 | −5.5236 ± 0.5983 | −5.8293 ± 0.4610 | 3.6378 | 0.0003 |
| **10** | **−4.1657 ± 0.3842** | **−4.2970 ± 0.3950** | **2.1506** | **0.0329** |
| 11 | −4.2257 ± 0.4481 | −4.4118 ± 0.4224 | 2.7236 | 0.0071 |
| 12 | −4.8252 ± 0.5952 | −5.1264 ± 0.4574 | 3.6062 | 0.0004 |
| 13 | −3.5027 ± 0.3945 | −3.6482 ± 0.4041 | 2.3257 | 0.0212 |
| 14 | −3.5008 ± 0.4495 | −3.6876 ± 0.4290 | 2.7092 | 0.0074 |
| 15 | −4.1057 ± 0.5812 | −4.3906 ± 0.4450 | 3.4976 | 0.0006 |
| 16 | −3.3778 ± 0.4428 | −3.5606 ± 0.4362 | 2.6529 | 0.0087 |
| 17 | −3.9986 ± 0.5384 | −4.2433 ± 0.4186 | 3.2257 | 0.0015 |
| 18 | −3.8503 ± 0.5082 | −4.0556 ± 0.4066 | 2.8376 | 0.0051 |

the depth of the decision tree ranged from 1 to 9.

Interestingly, an improvement in the performance of the SVM and LDA is observed. Differently, performance remains unchanged for the KNN and decision tree models in terms of accuracy. In general, it is possible to observe that the overall performance obtained is below 70% for the test accuracy by only considering clinical data.

### 4.3. Results of early and Late Fusion based on WST features and clinical data

Finally, the results of applying the Early and Late Fusion can be appreciated in Tables 6 and 7, respectively. In the Early Fusion approach, the 19 features obtained from PPG signals using WST were concatenated with age, BMI, and heart rate into a single feature vector, generating 22 input features. Besides, for the Early Fusion approach, Fig. 13 shows the trade-off between training and testing set accuracies for each ML technique for different values of their respective hyperparameters. The candidate set of hyperparameters for the SVM were $C =$ [0.01, 0.1, 1, 10, 100, 1000], a fix gamma of 0.01, and a RBF kernel. For the LR, the candidate set of hyperparameters were $C =$ [0.01, 0.1, 1, 10, 50, 100] by considering a $L2$ regularization. For the KNN, the hyperparameter values were chosen to be integer values ranging from 1 to 9 neighbors. For the decision tree, the depth was varied with integer values beginning from 1 and ending with 9. There are no hyperparameters that can be adjusted with the LDA.

On the other hand, the best models obtained for each data modality were selected and merged through a majority or hard voting approach in the Late Fusion approach, which results can be appreciated in Table 7. The selected models trained with the WST features derived from PPG signals were SVM and LR shown in Table 2 with their respective hyperparameters, while each of the models based on clinical and socio-demographic variables trained through feature selection presented in Table 5 with their respective hyperparameters were tested. In addition, the trained models with their respective data modality, hyperparameters, and training accuracies used for the Late Fusion approach are shown in Table 8.

## 5. Discussion

### 5.1. High BP detection based on PPG signals

By looking at Table 2, it is possible to observe that the best model was the SVM in terms of test set accuracy and F1-score for the PHT class, and the worst model was the decision tree and LDA in terms of testing accuracy. The LR and KNN have similar performances in terms of testing accuracy. However, the KNN is a lazy learner (i.e., no model is generated) and requires higher computational power than the LR in which a linear model is generated. Besides, Fig. 9 shows the trade-off between the training and testing sets accuracies while varying their respective hyperparameters. It can be observed that the decision tree is highly overfitted as the depth increases; contrary, the KNN is less overfitted as the number of neighbors increases. In the case of the SVM and LR, the performance increases while increasing the regularization parameter until the model starts overfitting. However, the performance of all ML techniques is low. The above can be attributed to the distribution of the analyzed sample. As shown in Fig. 6, the features obtained from the WST represent the average energy over the frequency scales. However, despite the difference in the mean of each group, their data range overlaps (see Fig. 10), making classification difficult for the trained models. This similarity between both classes is expected for some cases since the PPG does not measure BP. The factors that may contribute to not observing an association between BP and PPG were also discussed by Xing et al. [57]. Several factors can affect this association, such as the subject's finger size [58], poor circulation or cold temperature (which can also affect the peripheral pulsation) [59,60], and elevated blood viscosity (which can slow down the blood and alter the PPG signal). Another assumption is that the blood volume measured through the PPG is proportional to the total hemoglobin. However, this assumption can be affected if the subject has anemia or edema [57,61]. Finally, arrhythmia, diabetes, and pregnancy are factors that were not considered in this study or during the data collection process of the collected sampled that could influence the PPG signal [57]. Therefore, a perfect separation between the features obtained from PPG waveforms between NT versus PHT as shown in Fig. 6 is complicated for all cases.

**Table 4**
Results of applying ML Classifiers using the clinical and socio-demographic data without feature selection. In bold is shown the best classifier.

| ML Method | Hyperparameters | Class | Training Accuracy% | Testing Accuracy% | Precision % | Recall % | F1- Score % |
|---|---|---|---|---|---|---|---|
| **SVM** | **C**:1000, **Kernel**: RBF, **Gamma**: 0.00001 | NT | **64.22** | **64.29** | **68.75** | **52.38** | **59.46** |
| | | PHT | | | **61.54** | **76.19** | **68.09** |
| LR | **C:** 10, **Penalty**: L2 | NT | 63.41 | 61.90 | 61.90 | 61.90 | 61.90 |
| | | PHT | | | 61.90 | 61.90 | 61.90 |
| LDA | **Solver**: SVD | NT | 65.04 | 57.14 | 57.14 | 57.14 | 57.14 |
| | | PHT | | | 57.14 | 57.14 | 57.14 |
| KNN | **Nearest Neighbors**: 5 | NT | 73.98 | 61.90 | 63.16 | 57.14 | 60.00 |
| | | PHT | | | 60.87 | 66.67 | 63.64 |
| Decision Tree | **Criterion**: Gini Impurity, **Max depth**: 1 | NT | 65.04 | 54.76 | 53.57 | 71.43 | 61.22 |
| | | PHT | | | 57.14 | 38.10 | 45.71 |

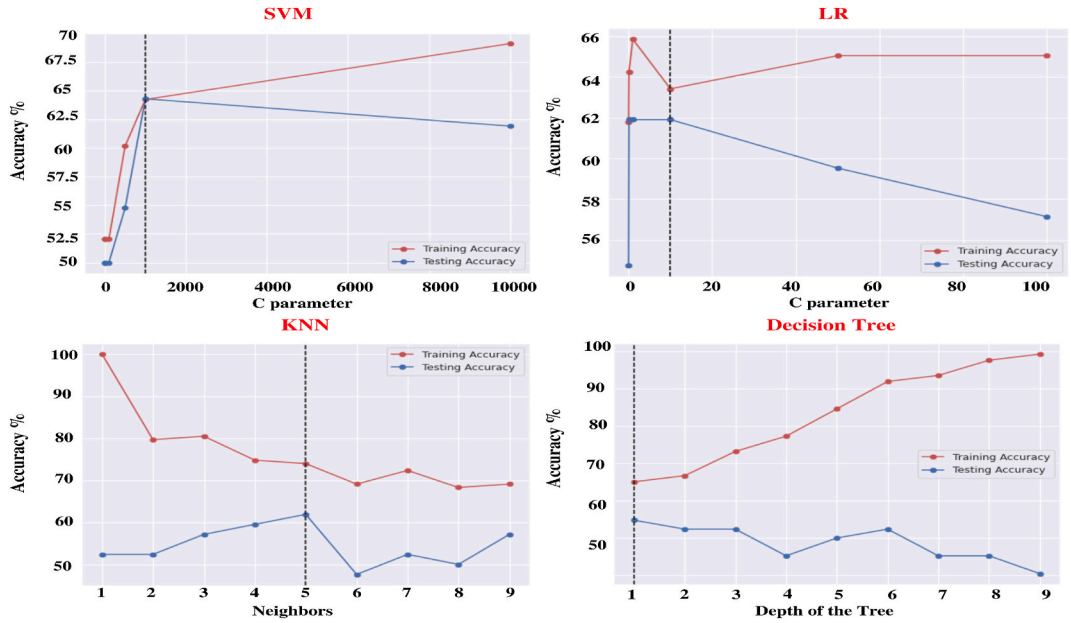


**Fig. 11.** Training (red) and testing (blue) sets' accuracy while varying the ML techniques' hyperparameters trained with the age, sex, BMI, heart rate, height, and weight for NT and PHT detection. The vertical dash line shows the hyperparameter value that produces the best trade-off between the training and testing set accuracies. The top graphs show the results of the SVM and LR; the bottom graphs show the results of KNN and the decision tree.

**Table 5**
Results of applying ML classifiers to the clinical and socio-demographic information after applying Feature Selection. In bold is shown the best classifier.

| ML Method | Hyperparameters | Class | Training Accuracy% | Testing Accuracy% | Precision % | Recall % | F1- Score % |
|---|---|---|---|---|---|---|---|
| **SVM** | **C**:1000, **Kernel**: RBF, **Gamma**: 0.001 | NT | **69.10** | **66.67** | **70.59** | **57.14** | **63.16** |
| | | PHT | | | **64.00** | **76.19** | **69.57** |
| LR | **C**: 1, **Penalty**: L2 | NT | 63.41 | 61.90 | 61.90 | 61.90 | 61.90 |
| | | PHT | | | 61.90 | 61.90 | 61.90 |
| LDA | **Solver**: SVD | NT | 63.41 | 61.90 | 61.90 | 61.90 | 61.90 |
| | | PHT | | | 61.90 | 61.90 | 61.90 |
| KNN | **Nearest Neighbors**: 8 | NT | 72.35 | 61.90 | 61.90 | 61.90 | 61.90 |
| | | PHT | | | 61.90 | 61.90 | 61.90 |
| Decision Tree | **Criterion**: Gini Impurity, **Max depth**: 1 | NT | 65.04 | 54.76 | 53.57 | 71.43 | 61.22 |
| | | PHT | | | 57.14 | 38.10 | 45.71 |

### 5.2. High BP detection based on clinical data

The results of applying ML techniques to the clinical and socio-demographic variables as inputs for NT and PHT detection can be appreciated in Table 4 without feature selection. In addition, by inspecting Fig. 11 it can be noticed that the decision tree suffers from overfitting as the depth of the tree increases. The KNN has less overfitting, considering five nearest neighbors. In the case of the SVM and LR, they had similar behavior in their bias and variance trade-off while increasing the regularization parameter *C*. Nevertheless, the SVM had better training and testing accuracies.

In addition, Table 5 presents the results after applying feature selection and Fig. 12 presents the trade-off between training and testing accuracies for different hyperparameter values for each ML technique. Similar to the experiments without feature selection, the decision tree has overfitting problems while the depth of the fitted tree increases. For the KNN, there is less difference between the training and testing accuracies as the number of neighbors increases. Nevertheless, the techniques that show the less difference between the training and testing accuracies are the LR and SVM, where the SVM showed a higher value and less difference between the training and testing accuracies.

It can be appreciated that, in general, the testing accuracy is below 70% with and without feature selection. Moreover, applying the feature selection based on the Gini Impurity to reduce the number of variables slightly increases the testing accuracy and F1-score of each class of the SVM. Differently, the accuracy remains unchanged for the KNN, LR, and decision tree. However, there is some variation in the F1-score for the KNN, while all performance metrics remain unchanged for the LR model. Additionally, the decision tree still shows poor performance with and without feature selection, and the LDA shows the same performance with feature selection as the LR. The above can be attributed to both the LR and LDA being linear classifiers, and the decision regions coincide in both cases. Moreover, it is essential to mention that although the variables have a certain degree of association with BP as stated in Ref. [24], the discriminative power of the variables highly depends on the distribution of the data that is studied, which explains the variability in the results reported in other works based principally on clinical data as in Refs. [19,25–27] and not entirely on the chosen classification technique.

### 5.3. High BP detection using a multimodal data approach

By looking at the results presented in Tables 6 and 7, it is possible to appreciate the performance of applying Early and Late Fusion, respectively. The Late Fusion approach provides a better classification than Early Fusion for most models in terms of the testing accuracy, except for the KNN of the Early Fusion approach. Nevertheless, the KNN is a lazy learner and requires the whole training data to make the classification, which increases the computational cost. On the other hand, the trade-off between training and testing accuracies for the Early Fusion approach presented in Fig. 13 shows that the SVM, LR, and decision tree are

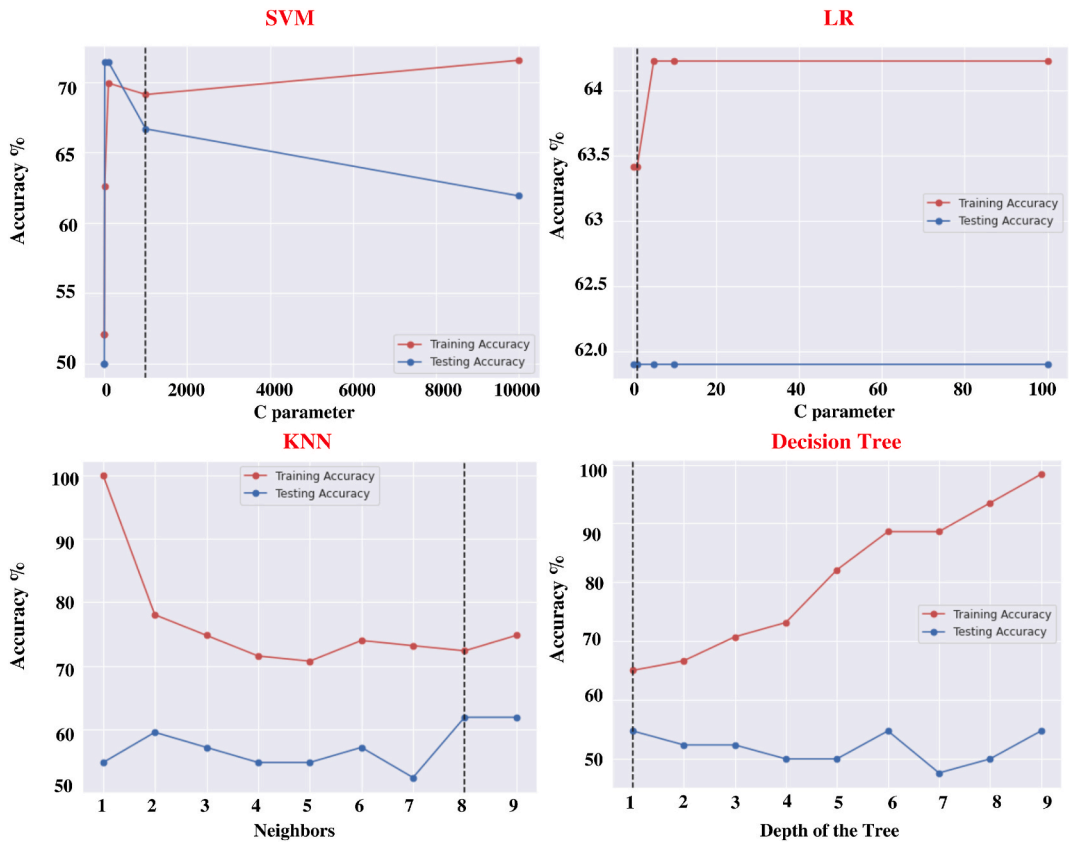


**Fig. 12.** Training (red) and testing (blue) sets' accuracies while varying the hyperparameters of the respective ML techniques trained with the age, BMI, and heart rate selected through the Gini Importance for NT and PHT detection. The vertical dash line shows the hyperparameter value that produces the best trade-off between the training and testing set accuracies. The top graphs show the results of the SVM and LR; the bottom graphs show the results of KNN and the decision tree.

**Table 6**
Results of applying Early Fusion and ML Classifiers. In bold is shown the best classifier.

| ML Method | Hyperparameters | Class | Training Accuracy% | Testing Accuracy% | Precision % | Recall % | F1- Score % |
|---|---|---|---|---|---|---|---|
| SVM | **C**:100,**Kernel**: RBF, **Gamma**: 0.01 | NT | 82.11 | 61.90 | 63.16 | 57.14 | 60.00 |
| | | PHT | | | 60.87 | 66.67 | 63.64 |
| LR | **C**: 0.001 **Penalty**: L2 | NT | 73.17 | 64.29 | 66.67 | 57.14 | 61.54 |
| | | PHT | | | 62.50 | 71.43 | 66.67 |
| LDA | **Solver**: SVD | NT | 73.17 | 61.90 | 61.90 | 61.90 | 61.90 |
| | | PHT | | | 61.90 | 61.90 | 61.90 |
| **KNN** | **Nearest Neighbors:**5 | NT | **72.35** | **69.05** | **70.00** | **66.67** | **68.29** |
| Decision Tree | **Criterion**: Gini Impurity, **Max Depth**: 3 | NT | 73.17 | 57.14 | 57.89 | 52.38 | 55.00 |
| | | PHT | | | 56.52 | 61.90 | 59.09 |

**Table 7**
Models, features, and test results of each of the classifiers after applying Late Fusion. In bold is shown the best classifier.

| PPG Models | Clinical/Socio-demographic models | Features | Class | Testing Accuracy % | Precision % | Recall % | F1-Score % |
|---|---|---|---|---|---|---|---|
| **(SVM + LR)** | **SVM** | 19 WST features + Age, BMI, and Heart Rate | NT | **69.05** | **75.00** | **57.14** | **64.86** |
| | | | PHT | | **65.38** | **80.95** | **72.34** |
| | LR | | NT | 66.67 | 70.59 | 57.14 | 63.16 |
| | | | PHT | | 64.00 | 76.19 | 69.57 |
| | LDA | | NT | 66.67 | 70.59 | 57.14 | 63.16 |
| | | | PHT | | 64.00 | 76.16 | 69.57 |
| | KNN | | NT | 66.67 | 73.33 | 52.38 | 61.11 |
| | | | PHT | | 62.96 | 80.95 | 70.83 |
| | Decision Tree | | NT | 69.05 | 72.22 | 61.90 | 66.67 |
| | | | PHT | | 66.67 | 76.19 | 71.11 |

overfitted for most of the candidate hyperparameters. Furthermore, the performance of the fusion techniques that were tested compared to considering each data modality separately does not improve the performance of the classification task in terms of testing accuracy. This lack of improvement can be related to the absence of differences in age, BMI, and heart rate between the NT and PHT subjects. This can also be observed in the performance of the models trained only with socio-demographic and clinical data, as shown in Tables 4 and 5.

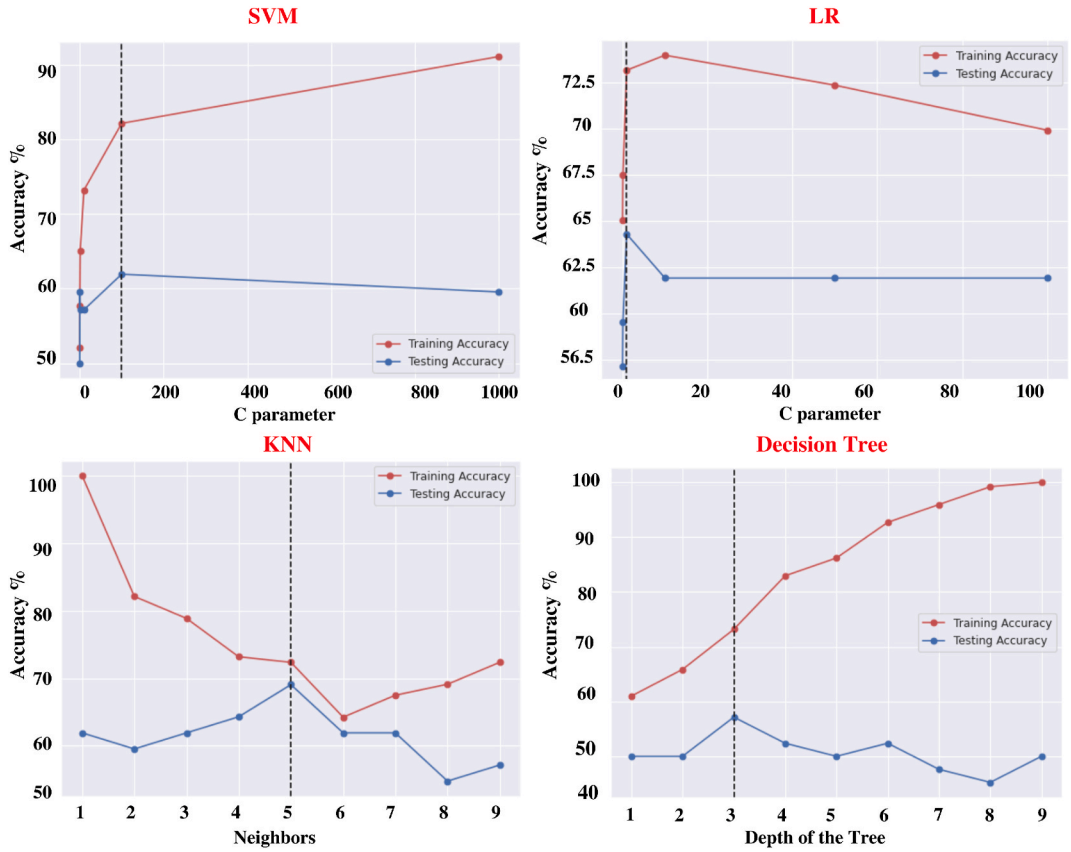


**Fig. 13.** Training (red) and testing (blue) sets' accuracy while varying the hyperparameters of the ML techniques trained through the concatenated feature vector generated from 19 features from the WST plus age, BMI, and heart rate for NT and PHT detection. The vertical dash line shows the hyperparameter value that produces the best trade-off between the training and testing set accuracies. The top graphs show the results of the SVM and LR; the bottom graphs show the results of KNN and the decision tree.

**Table 8**
Chosen classifiers for each data modality to perform the Late Fusion approach based on majority voting.

| ML Method | Data Modality | Hyperparameters | Training Accuracy% |
|---|---|---|---|
| SVM | 19 WST Features | **C**:100, **Kernel**: RBF **Gamma**: 0.01 | 72.35 |
| LR | 19 WST Features | **C**:10 **Penalty**:L2 | 66.66 |
| SVM | Age, BMI, Heart Rate | **C**:1000, **Kernel**: RBF **Gamma**: 0.001 | 69.10 |
| LR | Age, BMI, Heart Rate | **C**:1 **Penalty**:L2 | 63.41 |
| LDA | Age, BMI, Heart Rate | **Solver**:SVD | 63.41 |
| KNN | Age, BMI, Heart Rate | **Nearest Neighbors**: 8 | 72.35 |
| Decision Tree | Age, BMI, Heart Rate | **Criterion**:Gini Impurity **Max Depth**:1 | 65.04 |

Comparing the fusion techniques that were tested, the following points can be discussed. The problem with using the Late Fusion approach is that it requires the separate training of each classifier, which increases the computing time and the complexity of the final model. Besides, the required number of classifiers depends on the designer's criteria. In this work, only three classifiers were used based on the individual performance to handle a particular data modality. In contrast, the Early Fusion approach only requires a single classifier, but the dimensionality of the model increases, leading to overfitting as shown in Fig. 13 for the SVM, LR, and decision tree. The above produce lower performance in the testing set accuracy as appreciated in Table 6 and Fig. 13. One potential solution to this overfitting problem seen in Early Fusion is to reduce the dimensionality of the concatenated feature vector that considers the PPG features and clinical variables (e.g., through principal component analysis) or apply feature selection [62–64]. Finally, one advantage of both Early and Late Fusion is that the design process is less complex than Joint or Hybrid Fusion, leading to a more interpretable classifier [44].

### 5.4. Comparison with the literature

Table 9 shows a comparison between the results obtained in this work with other works available in the literature. It is possible to observe that in general, when the authors compare NT versus PHT classes, the maximum F1-score for the PHT class is 85.80% according to the results presented by Ref. [20] while the minimum value obtained is 65% based on the results presented by Ref. [34]. This shows the difficulty that is to discriminate between NT versus PHT subjects has likewise pointed out by Sannino et al. [66]. Nevertheless, a fair comparison of each of these works with the present study is difficult since the data on which these algorithms were trained is different. For instance, some authors used the MIMIC dataset, while others used the PPG-BP. Nonetheless, even among the works that employed the same dataset, a comparison is still difficult since every author used different subsets of the respective datasets or specific samples that are not specified, preventing reproducibility. In addition, the authors reported different metrics or only reported the metrics for a particular class. As shown in Table 9 the F1-score is the most frequent metric that is reported, and the F1-score for the PHT class obtained through the proposed methods is inside the maximum and minimum values that are reported in the literature.

**Table 9**
Comparison of the models trained for this study with the works available in the literature for NT and PHT detection.

| Author | Dataset | Features | ML Method | Metrics % |
|---|---|---|---|---|
| Liang et al. [17] | MIMIC | CWT (Morse Wavelet) | Pre-trained GoogLeNet | **F1-score: 80.52** |
| Liang et al. [31] | MIMIC | PAT plus 10 PPG Features | KNN | **F1-score: 84.34** Sensitivity: 83.92 Specificity: 84.76 |
| Liang et al. [29] | PPG-BP | PPG morphological features | Weight KNN | **F1-score: 72.97** Sensitivity: 65.85 Precision: 81.82 |
| Sannino et al. [65] | MIMIC | PPG waveform no feature engineer | Random Forest | **F1-score: 85.7** |
| Yao et al. [34] | PPG-BP | Socio-demographic and PPG morphological features | SVM | **F1-score: 65.00** Precision: 67.00 Recall: 63.00 Specificity: 69.00 |
| Sun et al. [20] | MIMIC | PPG Derivatives and HHT | Pre-trained AlexNet | **F1-score: 85.80** Recall: 95.26 Specificity: 71.88 |
| This Study | PPG-BP | 19 WST Features | SVM | **F1-score: 76.00** Precision: 65.51 Recall: 90.47 Specificity:52.38 |
| This Study | PPG-BP | Age, BMI, and Heart Rate | SVM | **F1-score: 69.57** Precision: 64.00 Recall: 76.19 Specificity: 57.14 |
| This Study | PPG-BP | Age, BMI, Heart Rate, plus 19 WST Features | Early Fusion KNN | **F1-score: 69.77** Precision: 68.18 Recall: 71.43 Specificity: 66.67 |
| This Study | PPG-BP | Age, BMI, Heart Rate, plus 19 WST Features | Late Fusion (SVM + SVM + LR) | **F1-score: 72.34** Precision: 65.38 Recall: 80.95 Specificity: 57.14 |

On the other hand, certain advantages of the present study can be further explored. For instance, the work of Liang et al. [17] showed that despite transfer learning could achieve a relatively good performance, the computational cost to retrain the CNNs and, consequently, the training time is a drawback that must take into account. Although the computational cost can be countered with the use of a graphics processing unit, the costs of the specialized hardware must be considered. Moreover, the interpretability of the model is reduced since pre-trained architectures like GoogleNet and AlexNet are very deep in their structure and have many parameters. On the other hand, when pre-trained CNNs are used, a time-frequency representation is often required as input. In this case, the results could depend on how the parameters of this time-frequency representation are selected. For instance, Sun et al. [20] seems to have a better F1-score compared to Liang et al. [17] by considering the HHT instead of employing the CWT. Nevertheless, a black-box approach is still required to achieve the reported results, and the computational cost is still high. Otherwise, the WST used in this work is less computational expensive [23,41,42,51] and all algorithms used in this work were executed in a CPU. Besides, the features obtained are an energy representation of the PPG signals as shown in Figs. 5, Fig. 6, and Fig. 10. Therefore, all features can be interpreted contrary to transfer learning.

In another study, Liang et al. [31] employed the PAT computed through the ECG and PPG waveforms using the MIMIC dataset. Nevertheless, this dataset has a drawback related to the lack of synchronization in time, a problem also noted in Ref. [13], and it is a more invasive method since the ECG waveform is required. Furthermore, the KNN is a lazy classifier since a model is not inferred from the data, requiring high memory and computational cost. In a related study, Liang et al. [29] studied 125 morphological features extracted from PPG and its derivatives to analyzed its relationship with high BP. Nevertheless, extracting morphological features is complicated since they require a high-quality PPG waveform to be extracted correctly. Finally, in the case of the work of Sannino et al. [65] since no feature extraction was reported, the computational cost of using as input the raw PPG waveform was high, and the best model was a random forest that is also viewed as a black-box classifier [67]. On the other hand, the WST does not require the computation of the PPG derivatives, and it is not affected by the distortions induced by noise, drifts, and artifacts [41].

Another aspect to consider is that most previous works analyzed the MIMIC dataset. This dataset was collected from intensive care units, which can affect the quality of the sample for hypertension detection since BP could have been affected by the medication of each participant. In addition, the sampling frequency was of 125 Hz, which could affect the representation of the PPG waveform and consequently produce abnormalities in the PPG morphology. The above factors were also pointed out by Yao et al. [34]. In this case, the PPG-BP dataset is more suitable for studying high BP detection since it was collected outside of intensive care units. However, the PPG-BP dataset collection was under the subjects' resting position, limiting its use under other factors or conditions such as body movements.

Otherwise, the work of Aydemir et al. [36] used the same PPG-BP dataset as the present work. Nonetheless, one drawback of this study is that the classes that were compared are not specified, which makes it difficult to compare the reported results. Moreover, the authors only reported the accuracy and omitted other metrics such as precision, recall, sensitivity, and F1-score. The problem of reporting only the accuracy is that it gives the wrong impression of a very accurate classifier, but the classifier's performance for each class is overlooked. The above could lead to biased results if the dataset is unbalanced, as is the case for the PPG-BP dataset. Besides, several PPG waveform representations or features are required to obtain the reported results, such as the CZT, TBP, ARMP, ZCR, and SD of the PPG derivative. The above hinders feature extraction, and compared to the WST used in this work; only a single framework is required to extract useful features for classification as demonstrated in Fig. 10 and Table 3.

Finally, the work of Yao et al. [34] performs a feature extraction process in which morphological features are computed from the PPG waveform and its derivatives. Nevertheless, from the 20 features computed, only 11 had a difference according to an Analysis of Variance (ANOVA) test. Contrastingly, in the present approach, from the 19 features obtained through the WST, only two were not different according to the performed Welch's T-Test. Moreover, in the work of Yao et al. [34] it is necessary to compute a new representation of the original PPG waveform through its derivatives and find specific points in those representations that are not necessarily associated with high levels of BP. Nevertheless, comparing the performance with Yao's work is complicated since, for the classification between NT and PHT, only a subset of the PPG-BP dataset was used but did not specify which samples were used.

## 6. Conclusion

This work proposed to use the WST as a feature extraction technique from PPG signals and training ML techniques for NT and PHT detection, where methods such as SVM and LR achieved the highest performance in terms of F1-score for PHT detection. In addition, the feature selection of clinical variables like age, BMI, and heart rate showed a better performance than considering all clinical variables (i.e., age, BMI, heart rate, weight, height, and sex) for detecting NT and PHT by training an SVM.

Moreover, the influence of considering PPG and clinical data by implementing both Early and Late Fusion was also analyzed. It was observed that comparing NT versus PHT subjects and considering the effect of clinical data in combination with PPG features derived from the WST does not improve the classification task's performance compared to considering each data modality separately. Of the two fusion approaches tested, Late Fusion applied through hard voting provides better performance than Early Fusion for detecting NT and PHT in terms of F1-score for the PHT class. Future work can explore the influence of clinical data and PPG data by stratifying the fitted models based on age, or BMI ranges through a multilevel model instead of considering a single model.

## Declaration of competing interest

None Declared.

## Acknowledgement

The work of E. Martinez-Ríos was supported by a scholarship awarded by Tecnologico de Monterrey and Consejo Nacional de Ciencia y Tecnologia (CVU: 1010770).

## References

[1] F.D. Fuchs, P.K. Whelton, High blood pressure and cardiovascular disease, Hypertension 75 (2) (2020) 285–292.
[2] A. Kannan, R. Janardhanan, Hypertension as a risk factor for heart failure, Curr. Hypertens. Rep. 16 (7) (2014) 447.
[3] D. Sun, T. Zhou, Y. Heianza, X. Li, M. Fan, V.A. Fonseca, L. Qi, Type 2 diabetes and hypertension: a study on bidirectional causality, Circ. Res. 124 (6) (2019) 930–937.
[4] B.B. Johansson, Hypertension mechanisms causing stroke, Clin. Exp. Pharmacol. Physiol. 26 (7) (1999) 563–565.
[5] W.H. Organization, et al., A Global Brief on Hypertension: Silent Killer, Global Public Health Crisis: World Health Day 2013, World Health Organization, 2013. Tech. rep.
[6] E. Wierzejska, B. Giernaś, A. Lipiak, M. Karasiewicz, M. Cofta, R. Staszewski, A global perspective on the costs of hypertension: a systematic review, Arch. Med. Sci.: AMS 16 (5) (2020) 1078.
[7] R. Rapport, Hypertension. silent killer, N. J. Med.: J. Med. Soc. N. J. 96 (3) (1999) 41–43.
[8] J.A. Pandit, D. Batlle, Snapshot hemodynamics and clinical outcomes in hypertension: precision in the measurements is key, Hypertension 67 (2) (2016) 270–271.
[9] J. Lee, S. Yang, S. Lee, H.C. Kim, Analysis of pulse arrival time as an indicator of blood pressure in a large surgical biosignal database: recommendations for developing ubiquitous blood pressure monitoring methods, J. Clin. Med. 8 (11) (2019) 1773.
[10] J.A. Pandit, E. Lores, D. Batlle, Cuffless blood pressure monitoring: promises and challenges, Clin. J. Am. Soc. Nephrol. 15 (10) (2020) 1531–1538.
[11] D. Shimbo, M. Abdalla, L. Falzon, R.R. Townsend, P. Muntner, Role of ambulatory and home blood pressure monitoring in clinical practice: a narrative review, Ann. Intern. Med. 163 (9) (2015) 691–700.
[12] A.V. Chobanian, G.L. Bakris, H.R. Black, W.C. Cushman, L.A. Green, J.L. Izzo Jr., D. W. Jones, B.J. Materson, S. Oparil, J.T. Wright Jr., et al., Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure, Hypertension 42 (6) (2003) 1206–1252.
[13] M. Elgendi, R. Fletcher, Y. Liang, N. Howard, N.H. Lovell, D. Abbott, K. Lim, R. Ward, The use of photoplethysmography for assessing hypertension, NPJ Digit. Med. 2 (1) (2019) 1–11.
[14] M. Kachuee, M.M. Kiani, H. Mohammadzade, M. Shabany, Cuffless blood pressure estimation algorithms for continuous health-care monitoring, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (4) (2016) 859–869.
[15] M. Elgendi, Y. Liang, R. Ward, Toward generating more diagnostic features from photoplethysmogram waveforms, Diseases 6 (1) (2018) 20.
[16] L. Wang, W. Zhou, Y. Xing, X. Zhou, A novel neural network model for blood pressure estimation using photoplethesmography without electrocardiogram, J. Healthc. Eng. 2018 (2018) 7804243.
[17] Y. Liang, Z. Chen, R. Ward, M. Elgendi, Photoplethysmography and deep learning: enhancing hypertension risk stratification, Biosensors 8 (4) (2018) 101.
[18] U. Senturk, K. Polat, I. Yucedag, A non-invasive continuous cuffless blood pressure estimation using dynamic recurrent neural networks, Appl. Acoust. 170 (2020), 107534.
[19] F. López-Martínez, E.R. Núñez-Valdez, R.G. Crespo, V. García-Díaz, An artificial neural network approach for predicting hypertension using nhanes data, Sci. Rep. 10 (1) (2020) 1–14.
[20] X. Sun, L. Zhou, S. Chang, Z. Liu, Using cnn and hht to predict blood pressure level based on photoplethysmography and its derivatives, Biosensors 11 (4) (2021) 120.
[21] G. Panchal, A. Ganatra, Y. Kosta, D. Panchal, Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers, Int. J. Comput. Theor. Eng. 3 (2) (2011) 332–337.
[22] J. Petch, S. Di, W. Nelson, Opening the black box: The promise and limitations of explainable machine learning in cardiology, Can. J. Cardiol. 38 (2) (2022) 204–213.
[23] S. Mallat, Group invariant scattering, Commun. Pure Appl. Math. 65 (10) (2012) 1331–1398.
[24] E. Martinez-Ríos, L. Montesinos, M. Alfaro-Ponce, L. Pecchia, A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data, Biomed. Signal Process Control 68 (2021), 102813.
[25] F. Lopez-Martinez, A. Schwarcz, E.R. Núñez-Valdez, V. Garcia-Diaz, Machine learning classification analysis for a hypertensive population as a function of several risk factors, Expert Syst. Appl. 110 (2018) 206–215.
[26] D. LaFreniere, F. Zulkernine, D. Barber, K. Martin, Using machine learning to predict hypertension from a clinical dataset, in: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2016, pp. 1–7.
[27] Y. Luo, Y. Li, Y. Lu, S. Lin, X. Liu, The prediction of hypertension based on convolution neural network, in: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), IEEE, 2018, pp. 2122–2127.
[28] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, Mimic-iii, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[29] Y. Liang, Z. Chen, R. Ward, M. Elgendi, Hypertension assessment using photoplethysmography: a risk stratification approach, J. Clin. Med. 8 (1) (2019) 12.
[30] Y. Liang, Z. Chen, G. Liu, M. Elgendi, A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China, Sci. Data 5 (1) (2018) 1–7.
[31] Y. Liang, Z. Chen, R. Ward, M. Elgendi, Hypertension assessment via ecg and ppg signals: an evaluation using mimic database, Diagnostics 8 (3) (2018) 65.
[32] H. Koshimizu, R. Kojima, Y. Okuno, Future possibilities for artificial intelligence in the practical management of hypertension, Hypertens. Res. 43 (12) (2020) 1327–1337.
[33] Y. Liang, D. Abbott, N. Howard, K. Lim, R. Ward, M. Elgendi, How effective is pulse arrival time for evaluating blood pressure? challenges and recommendations from a study using the mimic database, J. Clin. Med. 8 (3) (2019) 337.
[34] L.-P. Yao, W.-Z. Liu, Hypertension assessment based on feature extraction using a photoplethysmography signal and its derivatives, Physiol. Meas. 42 (6) (2021), 065001.
[35] G. Slapničar, N. Mlakar, M. Luštrek, Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network, Sensors 19 (15) (2019) 3420.
[36] T. Aydemir, M. Sahin, O. Aydemir, Determination of hypertension disease using chirp z-transform and statistical features of optimal band-pass filtered short-time photoplethysmography signals, Biomed. Phys. Eng. Expr. 6 (6) (2020) 65033.
[37] M. Elgendi, Optimal signal quality index for photoplethysmogram signals, Bioengineering 3 (4) (2016) 21.
[38] R. Krishnan, B. Natarajan, S. Warren, Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data, IEEE Trans. Biomed. Eng. 57 (8) (2010) 1867–1876.
[39] L. Wang, M. Han, X. Li, N. Zhang, H. Cheng, Review of classification methods on unbalanced data sets, IEEE Access 9 (2021) 64606–64628.
[40] M.H. Chowdhury, M.N.I. Shuzan, M.E. Chowdhury, Z.B. Mahbub, M.M. Uddin, A. Khandakar, M.B.I. Reaz, Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques, Sensors 20 (11) (2020) 3127.
[41] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Trans. Signal Process. 62 (16) (2014) 4114–4128.
[42] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886.
[43] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[44] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit. Med. 3 (1) (2020) 1–9.
[45] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Commun. ACM 63 (1) (2019) 68–77.
[46] S. Mishra, A. Datta-Gupta, Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences, Elsevier, 2017.
[47] O. Kramer, Scikit-learn, in: Machine Learning for Evolution Strategies, Springer, 2016, pp. 45–53.
[48] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, vol. 112, Springer, 2013.
[49] S.B. Kotsiantis, Decision trees: a recent overview, Artif. Intell. Rev. 39 (4) (2013) 261–283.
[50] A. Subasi, K. Khateeb, T. Brahimi, A. Sarirete, Human activity recognition using machine learning methods in a smart healthcare environment, in: Innovation in Health Informatics, Elsevier, 2020, pp. 123–144.
[51] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. AndÃ © n, E. Belilovsky, J. Bruna, V. Lostanlen, M. Chaudhary, M.J. Hirn, E. Oyallon, S. Zhang, C. Cella, M. Eickenberg, Kymatio: scattering transforms in python, J. Mach. Learn. Res. 21 (60) (2020) 1–6.
[52] F. Avellaneda, Efficient inference of optimal decision trees, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 3195–3202.
[53] R.H. Riffenburgh, Statistics in Medicine, Elsevier, 2011.
[54] S.M. Ross, Introductory Statistics, Academic Press, 2017.

[55] M. Delacre, D. Lakens, C. Leys, Why psychologists should by default use welch's t-test instead of student's t-test, Int. Rev. Social Psychol. 30 (1).
[56] S.G. Kwak, J.H. Kim, Central limit theorem: the cornerstone of modern statistics, Kor. J. Anesthesiol. 70 (2) (2017) 144.
[57] X. Xing, Z. Ma, M. Zhang, Y. Zhou, W. Dong, M. Song, An unobtrusive and calibration-free blood pressure estimation method using photoplethysmography and biometrics, Sci. Rep. 9 (1) (2019) 1–8.
[58] A. Grabovskis, Z. Marcinkevics, U. Rubins, E. Kviesis-Kipge, Effect of probe contact pressure on the photoplethysmographic assessment of conduit artery stiffness, J. Biomed. Opt. 18 (2) (2013) 27004.
[59] H. Hsiu, C. Hsu, C. Chen, W. Hsu, H. Hu, F. Chen, Correlation of harmonic components between the blood pressure and photoplethysmography waveforms following local-heating stimulation, Int. J. Biosci. Biochem. Bioinf. 2 (4) (2012) 248.
[60] H. Hsiu, S.-M. Huang, C.-L. Hsu, S.-F. Hu, H.-W. Lin, Effects of cold stimulation on the harmonic structure of the blood pressure and photoplethysmography waveforms, Photomed. Laser Surg. 30 (2) (2012) 77–84.
[61] T. Aoyagi, Pulse oximetry: its invention, theory, and future, J. Anesth. 17 (4) (2003) 259–266.
[62] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (1) (2004) 1–12.
[63] A.I. Pratiwi, et al., On the feature selection and classification based on information gain for document sentiment analysis, Appl. Comput. Intell. Soft Comput. 2018 (2018) 1407817.
[64] X. Ying, An overview of overfitting and its solutions, in: Journal of Physics: Conference Series, IOP Publishing, 2019, p. 22022.
[65] G. Sannino, I. De Falco, G. De Pietro, Non-invasive risk stratification of hypertension: a systematic comparison of machine learning algorithms, J. Sens. Actuator Netw. 9 (3) (2020) 34.
[66] G. Sannino, I. De Falco, G. De Pietro, Photoplethysmography and machine learning for the hypertension risk stratification, in: 2020 IEEE Globecom Workshops, GC Wkshps, IEEE, 2020, pp. 1–6.
[67] H. Zhang, M. Wang, Search for the smallest random forest, Stat. Interface 2 (3) (2009) 381.