



# Hypertension Management via Photoplethysmography: An Ensemble Learning-Based Approach for Classification of Blood Pressure Using Fourier Synchrosqueezed Transform

Noreddine Benaided<sup>1</sup>, Mohamed Hamza Meghraoui<sup>1\*</sup>, Zoubir Abdeslem Benselama<sup>2</sup>

<sup>1</sup> Electrotechnical and Automatic Engineering Department, GIDD Laboratory, University of Relizane, Relizane 48000, Algeria

<sup>2</sup> Electronics Department, LATSI Laboratory, University of Blida, Blida 09000, Algeria

Corresponding Author Email: [mohamedhamza.meghraoui@univ-relizane.dz](mailto:mohamedhamza.meghraoui@univ-relizane.dz)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410504>

## ABSTRACT

**Received:** 16 March 2024

**Revised:** 30 June 2024

**Accepted:** 20 August 2024

**Available online:** 31 October 2024

### Keywords:

*bagged trees (BT), blood pressure (BP), FSST, hypertension, photoplethysmography (PPG), TF analysis*

As the global prevalence of hypertension continues to rise, researchers have increasingly explored the potential of artificial intelligence (AI) for developing self-tracking blood pressure (BP) monitoring systems. An ideal approach would utilize photoplethysmography (PPG) signals, as they enable non-invasive wearable-based hypertension monitoring without reliance on cuff-based devices. This study investigated a PPG-based system for automated BP classification using an ensemble bagging technique with 200 decision trees. Given the nonstationary properties and motion artifact susceptibility of PPG signals, time-frequency (TF) analysis was conducted using Fourier Synchrosqueezed Transforms (FSST) to generate high-resolution TF representations. A set of 44 features were extracted from the transformed signals, revealing the dynamic statistical properties over time. Three experimental models were trained on datasets incorporating different FSST variables. Unlike prior studies using small datasets, the models were trained on a large dataset comprising 46,572 subject-segments across varied BP ranges, collected from the MIMIC-III intensive care database. This large dataset allowed boosting models accuracies and generalizability, achieving 100% training accuracy and 95.7% to 96.9% testing accuracy across the FSST experimental settings. The system also showed excellent results on three different classification tasks - normotension vs. hypertension, normotension vs. prehypertension, and non-hypertension vs. hypertension - with F1 scores reaching 99.1%. Moreover, the lightweight decision tree models enabled training in just minutes on this large dataset, indicating low computational complexity. Overall, this study presents an efficient PPG-based hypertension classification system. Results suggest potential for convenient clinical-grade BP monitoring beyond healthcare settings.

## 1. INTRODUCTION

Hypertension, a prevalent health condition defined by elevated BP, poses a significant global health burden, affecting more than 1.28 billion adults aged 30-79 years globally, as stated by the World Health Organization (WHO) [1]. Unfortunately, many hypertensive individuals remain unaware of their condition until they experience severe health complications, such as kidney failure, strokes, heart failure, and heart attacks [2]. Consequently, accurate and timely BP measurement and monitoring are imperative for effective hypertension management.

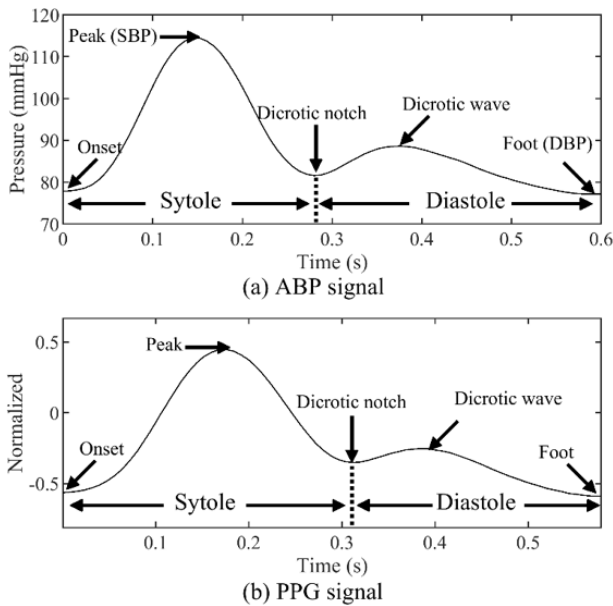
BP is a quantification of the force generated by the ventricular contraction in order to circulate blood throughout the cardiovascular system, is typically categorized into systolic and diastolic determinations. Systolic blood pressure (SBP) represents the peak arterial pressure attained during ventricular contraction while diastolic blood pressure (DBP) reflects the minimum arterial pressure registered just before the subsequent contraction [3]. Although invasive catheterization serves as the gold standard for accurate ABP measurement [4], it is primarily reserved for critically ill

patients due to its invasiveness and associated complications, such as arterial thrombosis, infection, nerve damage, and exsanguination [5].

To overcome the limitations of invasive techniques, non-invasive methods like cuff-based BP monitoring present safer and more cost-effective alternatives. Nevertheless, challenges persist, including discomfort during cuff use, the need for trained personnel to perform manual readings, and inaccuracies resulting from improperly fitting cuff sizes or unique arm dimensions [6-10]. Vascular unloading, which employs a variable finger cuff and PPG sensor, provides another non-invasive avenue for intra-arterial pressure estimation [11]. While some validated finger cuff devices have shown promise as alternatives for BP measurement in patients unfit for classical arm cuffs [12], prolonged use of finger cuffs can cause discomfort, numbness, and arterial congestion [13].

Alternatively, arterial tonometry (AT) offers a cuff-less approach for continuous ABP measurement using a pressure transducer to compress an artery. However, it may not be recommended for obese individuals due to the slowed propagation of pulse waves to the skin [13]. Considering the limitations of the aforementioned techniques, researchers have

turned their attention to explore the potential of AI in estimating ABP using PPG signals [14, 15], owing to the close correlation between these signals [16] and their similar shape characteristics (Figure 1).



**Figure 1.** ABP and PPG shape characteristics

PPG is a non-invasive technique that employ optical sensing to quantify variations in blood volume over time, providing a pulsatile waveform (PPG signal) for monitoring various physiological parameters related to blood circulation [17]. Various algorithms and models were suggested to estimate BP via these optical signals, with some studies meeting the guidelines established by the British Hypertension Society (BHS) and the Association for the Advancement of Medical Instrumentation (AAMI) for BP estimation [18, 19].

One approach involved employing a convolutional neural network (CNN) and transfer learning to estimate BP from PPG signals, attaining BHS grade A for diastolic and grade C for systolic estimates [20]. In another study, Maher et al. [21] proposed a calibration-free approach that attained Grade A ratings for both systolic and diastolic BP estimates based on the same BHS benchmarks. Additionally, El-hajj and Kyriacou [22] developed several PPG-based models applying recurrent neural networks with bidirectional connections and attention mechanisms, yielding results in line with the AAMI standards. In a separate study, the authors explored the potential of various PPG features and deep recurrent models for cuffless BP estimation, once more fulfilling the association's norms [23]. Furthermore, a hybrid neural network architecture was introduced by Qiu et al. [24], meeting both BHS and AAMI criteria for BP estimation.

Despite these promising advancements in non-invasive BP measurement techniques, the real-world implementation for hypertension management may face challenges due to poor knowledge among hypertensive individuals regarding acceptable BP values [25]. A study conducted in London revealed that over half of the participants were unable to correctly estimate an acceptable BP range [26], emphasizing the need for simplicity in designing BP monitoring systems in addition to non-invasiveness. An ideal monitoring method would involve providing patients with a system that notifies them promptly if their BP is abnormal. In this regard, classification methods using machine learning (ML) or deep

learning (DL) algorithms may be more suitable than regression approaches. Accordingly, the current study introduces a BP classification framework that utilizes PPG signals as inputs to an ensemble machine learning method known as bagged trees (BT). The study contributes to the existing literature in the following ways:

**Timely notification system:** The proposed system notifies the users instantly if their blood pressure is elevated, allowing for timely intervention and proactive management of hypertension.

**Enhanced dataset:** Unlike previous studies, we utilize a dataset containing a higher number of samples, which enhances the predictive capabilities of the proposed model, leading to more reliable and accurate BP classification.

**Addressing motion artefacts:** Despite the non-invasive benefit of PPG signals, they are susceptible to motion artefacts [27]. To overcome this challenge, a suitable feature extraction procedure is implemented. The study adopts Time-Frequency Reassignment (TFR), which offers more representative features and reduces the complexity of the extraction process. FSST is employed to leverage the sparsity in the time-varying oscillatory properties of PPG signals, resulting in a more concentrated TF representation compared to conventional techniques like Short-Time Fourier Transform (STFT) [28].

**Dataset diversity:** The reliability of a BP classification system heavily depends on acquiring diverse SBP and DBP values for each BP class level. During our data collection process, we ensure the inclusion of varied SBP and DBP values, providing the BT algorithm with a more representative dataset during the learning stage.

**Computational efficiency:** Dealing with large datasets conventionally prompts a preference for DL algorithms over ML due to their ability to handle complex and high-dimensional data. However, DL models are computationally intensive and time-consuming to train and infer. In contrast, the BT model offers faster and less computationally intensive performance, particularly when working with substantial datasets.

The remainder of this paper is structured as follows. Section 2 describes relevant studies and state of the art knowledge related to our study requirements. Section 3 provides a detailed description of the proposed methodology. Section 4 presents the experimental results and discusses the implications of the findings by identifying potential avenues for future research. Finally, Section 5 concludes the paper and summarizes the main contributions of this work.

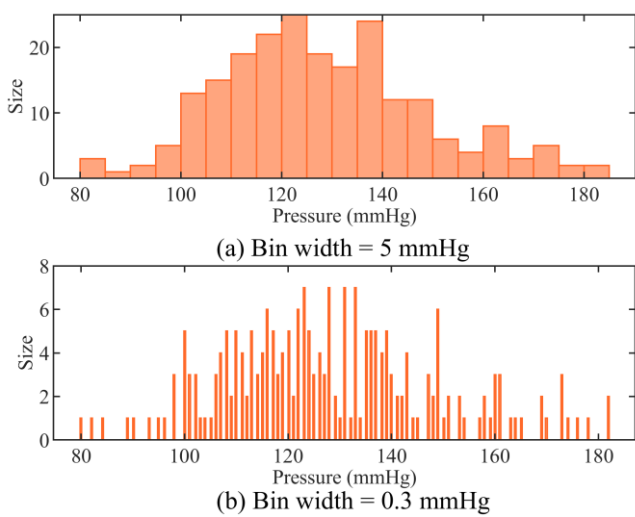
## 2. PRIOR RESEARCH AND FOUNDATIONAL CONCEPTS

### 2.1 PPG-based hypertension detection systems

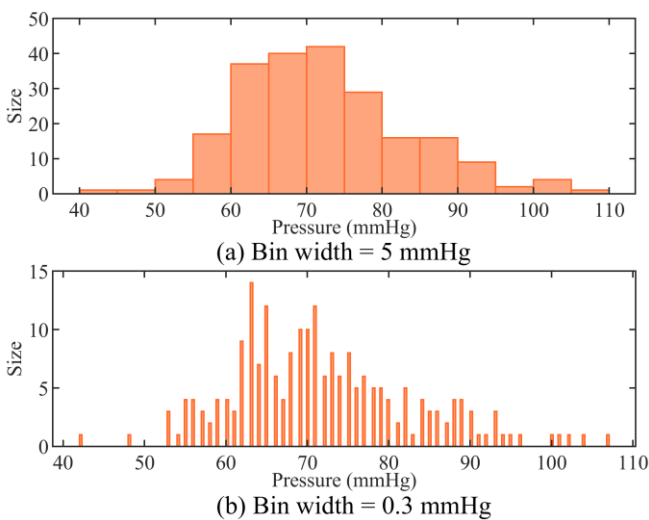
Several studies have investigated the use of ML and DL algorithms for predicting the risk of hypertension. For instance, Yen et al. [29] focused on enhancing hypertension classification accuracy using PPG signals from the PPG-BP figshare database [30]. They developed DL models designed with varying parameters and different architectures, including long short-term memory (LSTM), bidirectional long short-term memory (BLSTM), deep residual network convolutional neural network (ResNetCNN) and Extreme Inception (Xception). The best-performing model was a combination of Xception and BLSTM, achieving 48% precision, 45% recall,

and 76% accuracy. However, the results showed relatively low recall and precision values, implying a significant number of false negatives and false positives, which could lead to undetected hypertension cases and unnecessary interventions with higher healthcare costs.

On the other hand, Tjahjadi et al. [31] achieved better results in hypertension classification using the same dataset. They used TF-based PPG features and a BLSTM architecture, obtaining accuracy, recall, and specificity of 94.64%, 88.09%, and 98.57%, respectively. Nevertheless, the BLSTM required considerable training time (30 min and 26 secs) despite using a relatively small dataset (900 observations). In a different approach, Tjahjadi and Ramli [32] used the K-nearest neighbors (KNN) algorithm with 2100 PPG samples, obtaining accuracy, recall, and specificity of 86.77%, 74.28%, and 91.86%, respectively, for hypertension class level. While these studies [31, 32] outperformed Yen et al. [29], their datasets lacked diversity in SBP and DBP values as presented in Figure 2 and Figure 3, respectively.



**Figure 2.** Histograms of SBP values taken from the figshare database



**Figure 3.** Histograms of DBP values taken from the figshare database

In another study, Liang et al. [33] experimented with four ML models (AdaBoost, KNN, BT, and Logistic Regression) using various feature sets, including pulse arrival time (PAT),

PPG morphology features, and a combination of PPG and PAT features. They conducted three classification trials: normotension vs. hypertension (NT vs. HT), normotension vs. prehypertension (NT vs. PHT), and normotension plus prehypertension vs. hypertension (NT+PHT vs. HT). The KNN model with the combined features showed the best performance, achieving F1 scores of 84.34%, 94.84%, and 88.49% for the respective trials. However, using PAT as a feature is not practical for hypertension management, as it requires an additional sensor for electrocardiogram (ECG) signals. Besides, the features extraction process demands high-quality PPG signals [34], which is difficult due to their susceptibility to motion artefacts [27].

In a different study, the same researchers explored utilizing TF analysis via the continuous wavelet transform (CWT) method for transforming the signals into scalogram representations with three color channels. The resultant representations were then fed as RGB images into a pretrained CNN model [34]. The F1 score results for the classification trials were 92.55%, 80.52%, and 82.95% for NT vs. HT, NT vs. PHT, and NT+PHT vs. HT, respectively. While this approach improved the feature extraction process, the methodology was tested with a relatively small dataset, indicating the need for validation with a larger and more diverse dataset containing a wider range of SBP and DBP values.

## 2.2 FSST

PPG signals are non-stationary bio-signals characterized by time-varying properties resulting from dynamic physiological processes. Traditional frequency analysis techniques, such as the Fourier Transform (FT), face challenges when dealing with these signals due to their changing frequency components over time [35]. The FT operates under the supposition that the signal displays stationary behavior through an unchanging frequency constitution over its temporal evolution. This assumption is embedded in the fundamental definition of the FT, which provides a frequency-domain representation of a signal without considering time variations. The continuous FT of a signal  $s(t)$  is given by:

$$S(f) = \int_{-\infty}^{+\infty} s(t)e^{-2\pi ift} dt \quad (1)$$

where,  $S(f)$  is the Fourier Transform of  $s(t)$ , and  $f$  represents frequency. This integral transforms the entire signal  $s(t)$  into the frequency domain, assuming that the signal's frequency components are constant over the entire duration of the signal.

For non-stationary signals, such as PPG, where the frequency content changes over time, this assumption is invalid. The FT fails to capture these time-varying characteristics, as it provides only a global frequency representation, losing all time-related information. To address this issue, TF analysis has been employed to gain insights into signal behavior, facilitating the effective analysis of non-stationary signals by capturing the evolving frequency content throughout time. One of the widely used methods for TF analysis is the STFT. STFT attempts to address the challenge of non-stationarity by applying a window function to the signal, allowing the analysis of frequency content within small time segments [36]. The local frequency content of the signal is calculated using the equation:

$$V_S^D(t, \gamma) = \int_{-\infty}^{+\infty} S(\alpha) D(\alpha - t) e^{-2i\pi\gamma(\alpha - t)} d\alpha \quad (2)$$

where,  $S(\alpha)$  is the input signal,  $D(t)$  represents the window function,  $t$  denotes the time index,  $\gamma$  corresponds to frequency, and  $\alpha$  serves as a time-variable of integration.

However, STFT suffers from the classic trade-off between time and frequency resolution due to fixed window sizes [37]. A narrow window provides good time resolution but poor frequency resolution, while a wide window offers good frequency resolution but poor time resolution. This restricts STFT's ability to effectively capture rapid changes in both time and frequency that often appear in non-stationary signals like PPG.

To overcome these limitations, the FSST offers an innovative solution by applying a nonlinear post-processing mapping to the conventional STFT [38]. FSST aims to enhance the time-frequency representation by reassigning the signal's energy more accurately, resulting in better localization of frequency components [39]. The concept of reassignment in FSST facilitates signal interpretation through the redistribution of energy, and synchrosqueezing provides robust visualization and manipulation capabilities [28]. The FSST formula can be expressed as:

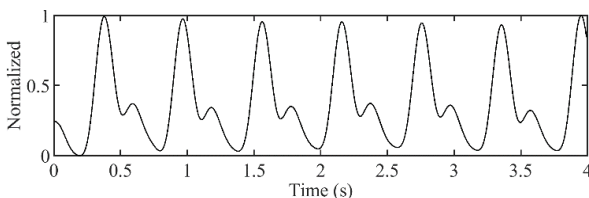
$$T_S^D(t, \omega) = \int_{-\infty}^{+\infty} V_S^D(t, \gamma) \delta(\omega - I_S(t, \gamma)) d\gamma \quad (3)$$

where,  $\delta(\omega)$  is the Dirac distribution, while  $I_S(t, \gamma)$  denotes the instantaneous frequency estimation at time  $t$  and frequency  $\gamma$  obtained using the following expression:

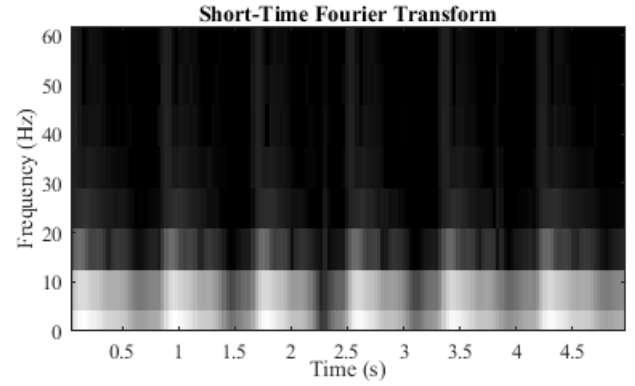
$$I_S(t, \gamma) = \frac{\partial \arg V_S^D(t, \gamma)}{\partial t} = \mathbb{R} \left\{ \frac{1}{2i\pi} \frac{\partial t V_S^D(t, \gamma)}{V_S^D(t, \gamma)} \right\} \quad (4)$$

such that  $V_S^D(t, \gamma) \neq 0$

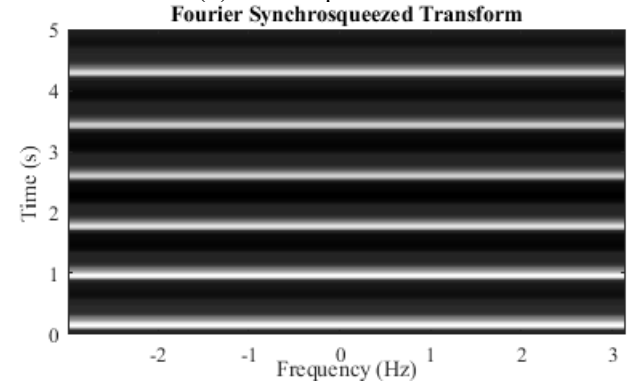
In FSST, the synchrosqueezing process significantly enhances the precision of frequency localization in the TF plane. This is achieved by reassigning the time-frequency representation  $V_S^D(t, \gamma)$  obtained from the STFT to concentrate the signal's energy more precisely around the true instantaneous frequency trajectories  $I_S(t, \gamma)$ . By dynamically adjusting the representation of the signal  $s(t)$ , FSST mitigates the trade-off between time  $t$  and frequency  $\gamma$  resolution. The signal's energy, initially spread over a broader area in the TF plane due to the fixed window function  $D(t)$  in STFT, is more accurately focused around  $\omega$  using the Dirac distribution  $\delta(\omega - I_S(t, \gamma))$ . This process allows  $T_S^D(t, \omega)$  to better capture the non-stationary characteristics of PPG signals by following the rapid changes in frequency components  $\gamma$  over time  $t$ . Consequently, FSST provides improved time  $t$  and frequency  $\gamma$  resolution, facilitating more detailed and accurate feature extraction from PPG signals for applications such as BP classification. A visual comparison between the resolution of FSST and STFT for a PPG signal is presented in Figure 4.



(a) Normalized PPG signal



(b) STFT representation



(c) FSST representation

Figure 4. TF representations of a PPG signal

### 3. PROPOSED METHODOLOGY

This section presents our proposed method for BP classification using TF analysis. The main steps involved in designing the system are visually depicted in Figure 5.

The process begins with the acquisition of a suitable dataset that satisfies the requirements of our approach, ensuring an adequate sample size and diversity in both SBP and DBP values to enhance the model's generalizability.

Next, we proceed to segment the PPG signals into discrete windows, each assigned to its corresponding BP label. This step allows us to isolate relevant temporal segments of the PPG signals for further analysis.

Following the segmentation, a set of features is extracted from each TF component of a given PPG segment.

In the final step, we carefully organize the dataset, taking into consideration factors such as potential bias and dependency of the test set on the training set. Once the dataset is prepared, we feed it into the BT algorithm for the actual classification process.

#### 3.1 Data source

The dataset utilized in the current research originates from the MIMIC-III (Medical Information Mart for Intensive Care III) database, an openly accessible storage repository holding anonymized medical information from individuals to intensive care unit (ICU) [40]. MIMIC-III is widely recognized for its extensive size and widespread adoption within the medical research community, making it a valuable resource for research and analysis. Within the MIMIC-III database, various data types are available, including clinical records, demographic information, laboratory results, medication details, and waveform data. In particular, the MIMIC-III

waveform database comprises over 3 million hours of waveform recordings across 67,830 records [41]. This includes simultaneously recorded ABP and PPG signals from thousands of ICU patients, sampled at 125 Hz.

### 3.2 Research subjects

To ensure diversity in BP values, a process was implemented during data collection whereby each acquisition

of an ABP record was inspected for SBP and DBP values. This process continued until a wide feasible range of BP values was achieved. Each 1-minute ABP signal was collected with its corresponding 1-minute PPG signal, both containing 7500 data points. This process yielded ABP signals with SBP spanning 69-216 mmHg and DBP spanning 34-115 mmHg. Figure 6 provides a visual comparison of the BP values in this dataset versus those utilized in previous related works [29-32].

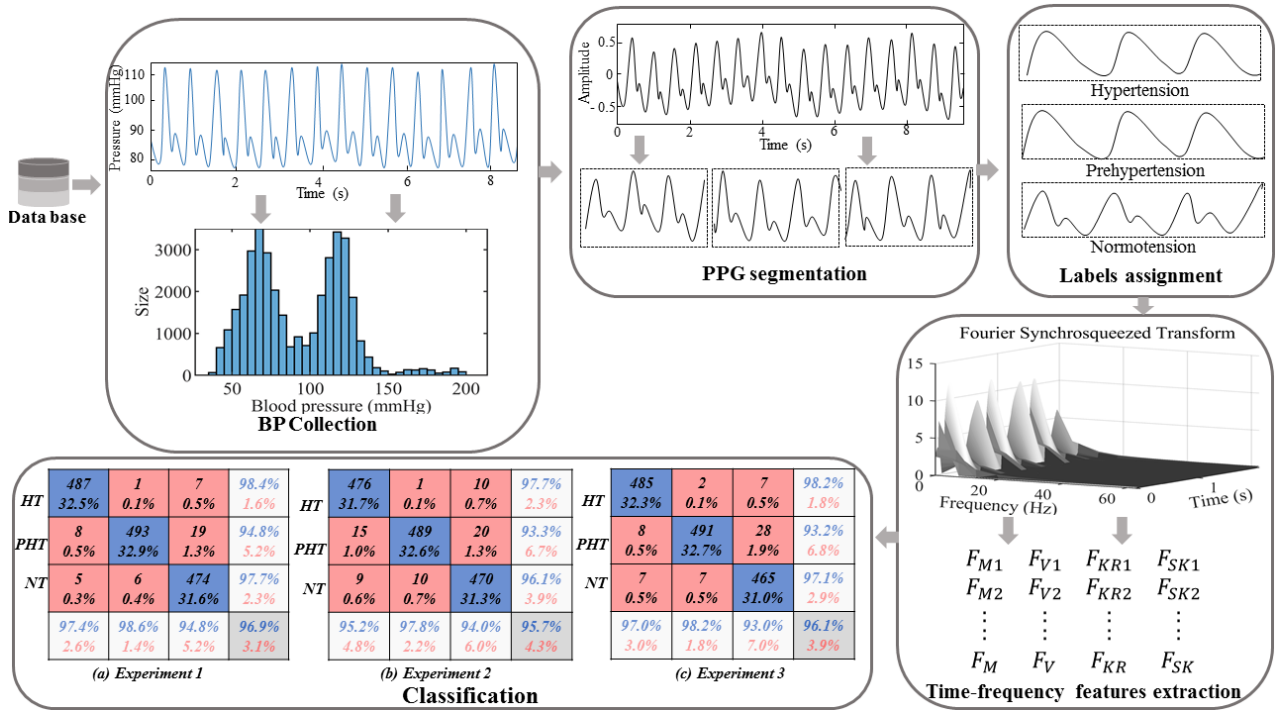


Figure 5. Methodological steps

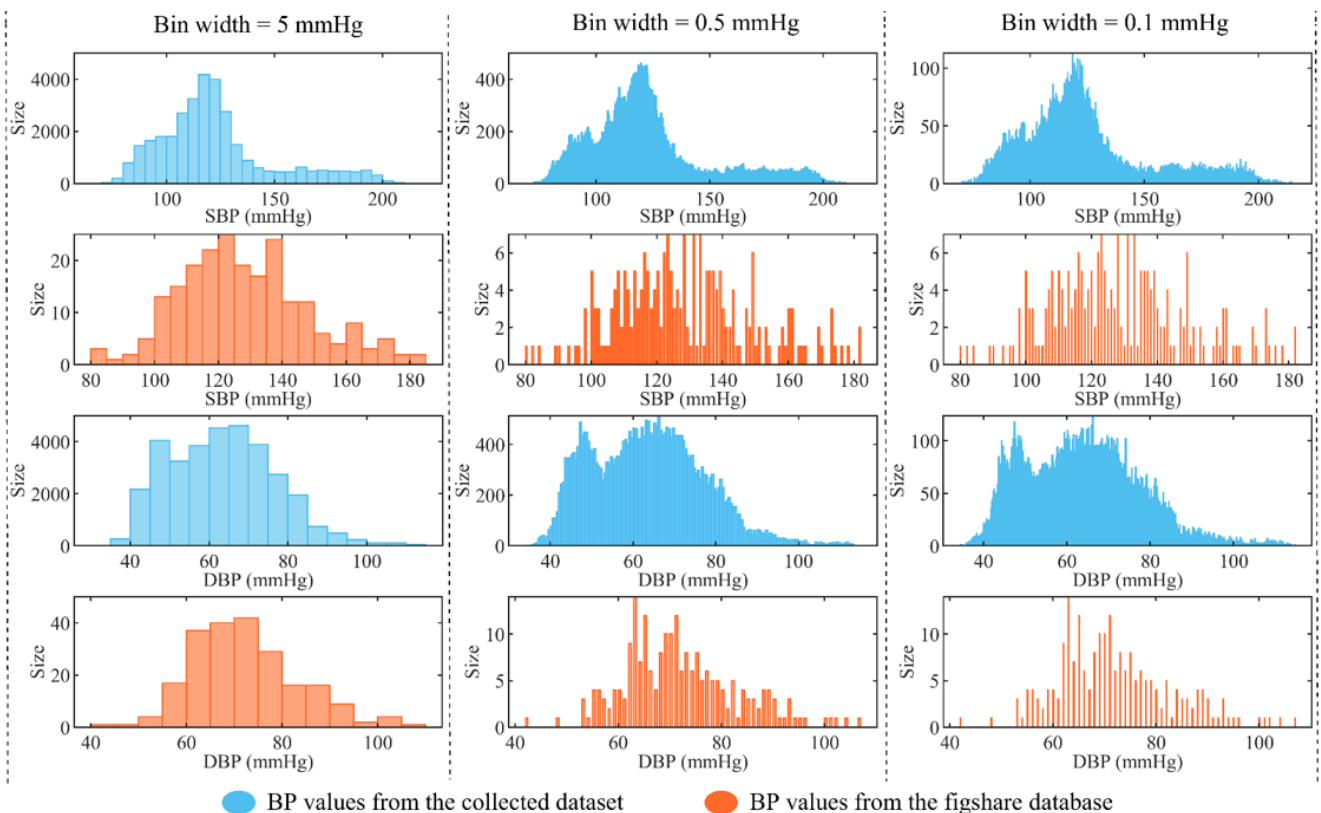
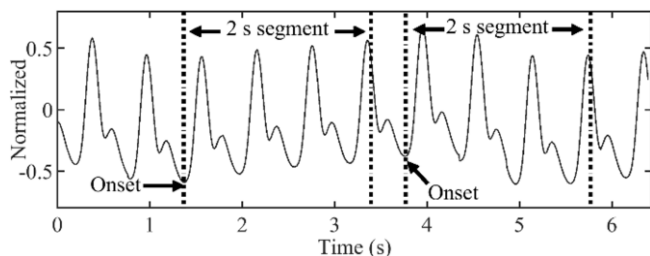


Figure 6. BP distribution comparison



**Figure 7.** PPG signal segmentation

**Table 1.** JNC 7 BP classification [42]

BP level	SBP (mmHg)		DBP (mmHg)
NT	< 120	and	< 80
PHT	120 to 140	or	80 to 90
HT	> 140	or	> 90

Next, the acquired signals underwent processing to generate input subjects for the classification system. Specifically, the PPG signals were segmented into 2-second windows (250 samples each) initiating at the onset of each pulse wave, as detailed in Figure 7. The resulting segments were then categorized into three groups, each corresponding to a specific BP class label: HT, PHT, and NT, following the Joint National Committee 7 (JNC7) guidelines [42] (Table 1).

However, it is important to note that a single BP reading is insufficient for a hypertension diagnosis, which requires multiple readings due to the fluctuation of BP over time, as emphasized in clinical literatures [43-45]. Therefore, the average SBP and DBP values ((SBP), (DBP)) were calculated

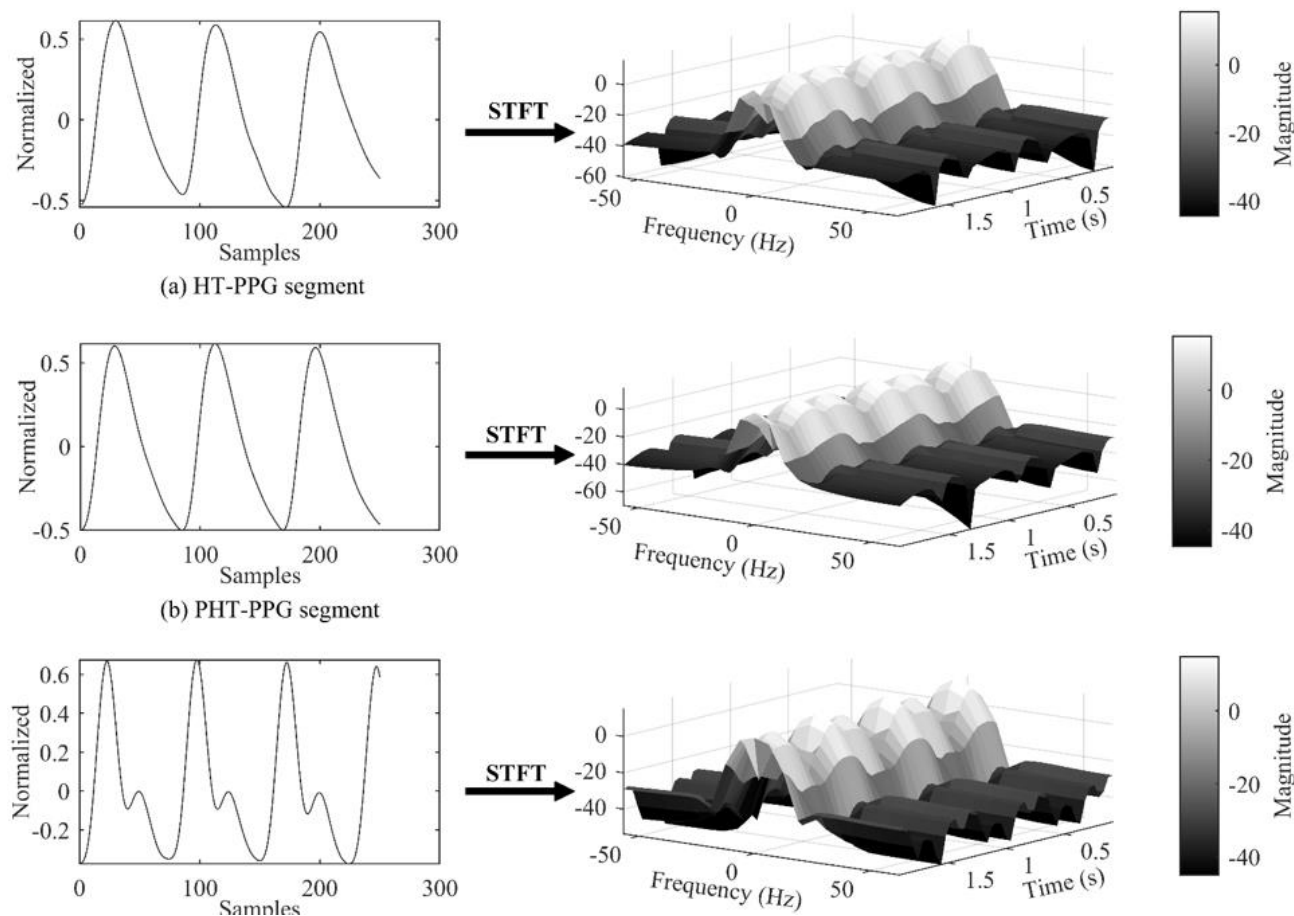
within each 1-minute ABP record. The BP class labels were then determined by comparing the (SBP) and (DBP) values to the JNC7 criteria. As a result, the final dataset comprises 6,068 hypertensive subjects, 9,010 prehypertensive subjects, and 16,024 normotensive subjects.

### 3.3 Time-frequency PPG features

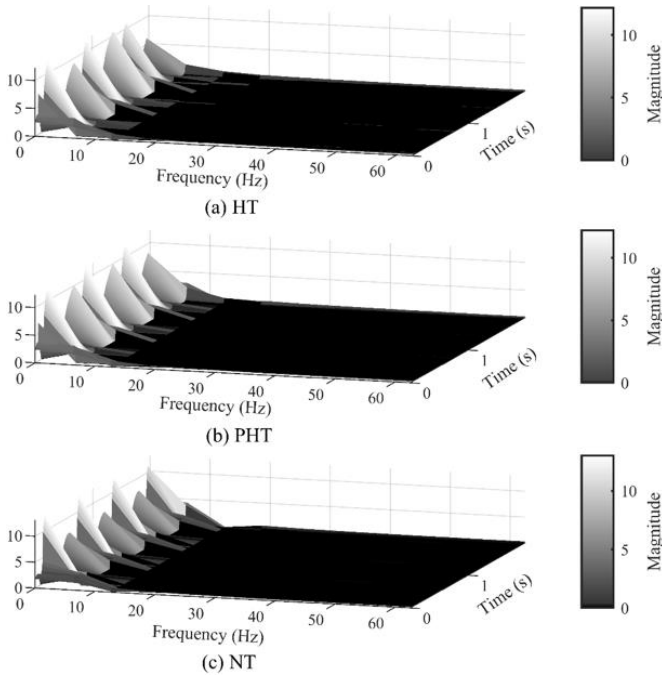
The feature extraction process aims to reduce the input dimensions per subject while retaining their predictive capabilities for BP classification. It involves two main steps: representing the PPG signal's time-varying spectrum using the FSST and reducing the dimensionality of the FSST output matrix.

The process begins by computing the STFT of a given 2-second PPG signal. Specifically, the signal is divided into 20 short segments, each windowed using a 20-sample Hamming window. Overlapping windows with 19-sample overlap are used for smooth spectral tracking over time. Within each segment, the fast Fourier transform (FFT) is applied to capture the frequency information. This generates a TF representation, showing how spectral content evolves over time. Figure 8 displays the 3D views of TF planes representing signals from each BP level group.

Next, the Fourier coefficients from the STFT are reassigned to new time locations based on instantaneous frequency estimates using Eq. (3), creating the FSST (Figure 9). This squeezing process sharpens the PPG signal's time-varying spectrum and provides a more concentrated view in the TF plane compared to the smeared STFT in Figure 8.



**Figure 8.** STFT 3D-representation of labeled PPG signals



**Figure 9.** FSST 3D-representation of labeled PPG signals

The FSST outputs a  $250 \times 11$  complex-valued matrix, with rows representing reassigned time instances and columns representing frequency bins. The 250 rows correspond to the original 250 signal samples, indicating continuous time after reassignment. The 11 columns represent the unique frequency components needed to characterize the real input signal's spectrum. Specifically, due to conjugate symmetry of the FT of real signals, the full spectrum can be compactly represented using only 11 non-redundant frequency bins, rather than the full FFT length.

Finally, to reduce dimensionality, each frequency bin column is processed to derive four statistical features, instead of using all 250 complex time values. This compresses the FSST matrix into  $4 \times 11$ , while retaining predictive information about the behavior of the reassigned time instances in each frequency bin. As a result, a set of 44 input features are used for BP classification per subject. Furthermore, three classification experiments are adopted within this study, each consisting of a different reference variable to compute the input features, including the magnitude (absolute value), real and imaginary parts of the complex values. Let's consider  $X_r$  a sequence of the reassigned coefficients denoted as:

$$X_r = \{x_1, x_2, x_n, \dots, x_N\} \quad (4)$$

where,  $r$  is the frequency index,  $N$  is the sequence length and  $x$  is the expected complex-value from a given sample index  $n$  and defined as:

$$x_n = a_n + jb_n \quad (5)$$

where,  $a_n$  is the real part,  $b_n$  is the imaginary part and  $j$  denotes the imaginary unit. The magnitude of  $x_n$  becomes:

$$|x_n| = \sqrt{a_n^2 + b_n^2} \quad (6)$$

The first feature provides an overall sense of the central tendency of the values in the frequency bin and is defined as

the mean, or average:

$$F_M = \frac{1}{N} \sum_{n=1}^N C_n^{ex} \quad (7)$$

$$\text{Such that } \begin{cases} C_n^{ex} = a_n & \text{when } ex = 1 \\ C_n^{ex} = b_n & \text{when } ex = 2 \\ C_n^{ex} = |x_n| & \text{when } ex = 3 \end{cases} \quad (8)$$

The second feature measures how far the values spread out from the mean. It is calculated using the variance equation:

$$F_V = \frac{\sum_{n=1}^N (C_n^{ex} - F_M)^2}{N} \quad (9)$$

The third feature examines the asymmetry of the distribution using skewness. Skewness quantifies the extent that values are unevenly distributed to one side versus the other side of the data's mean. A positive evaluation suggests the values tail off more to the right, whereas a negative evaluation indicates they tail off more to the left [46]. The skewness formula can be expressed as:

$$F_{SK} = \sqrt{\frac{\sum_{n=1}^N (C_n^{ex} - F_M)^3}{N \sqrt{F_V}}} \quad (10)$$

The fourth feature is derived using the kurtosis, which measures the heaviness of the tails of the values relative to the normal distribution [47] and is defined as:

$$F_{KR} = \frac{\frac{1}{N} \sum_{n=1}^N (C_n^{ex} - F_M)^4}{\left(\frac{1}{N} \sum_{n=1}^N (C_n^{ex} - F_M)^2\right)^2} \quad (11)$$

### 3.4 Data management

In this study, we generated three classification models, each trained on an experimental dataset created using one of the FSST variables ( $C_n^1, C_n^2, C_n^3$ ), as detailed in the features extraction section. Prior to model development, the datasets were partitioned into separate training and test sets to allow for proper evaluation on new data. The training sets included 5,568 HT subjects, 8,510 PHT subjects, and 15,524 NT subjects. The test sets contained 500 subjects from each class.

To account for class imbalance, the minority classes (HT and PHT) were oversampled in the training sets until all classes had equal representation. Oversampling was done after partitioning to avoid test set contamination which can artificially inflate the classification performance. As a result, the total dataset contained 48,072 subjects with the training set comprising 15,524 subjects for each class and the test set comprising 500 subjects for each class.

### 3.5 Ensemble learning classifier

In this study, an ensemble bagging approach was implemented for classification, with decision trees (DTs) as base learners. Ensemble methods combine multiple base models to improve predictive performance compared to single models [48]. Bagging involves training each base model on a bootstrap replica of the training data randomly drawn with

replacement [49]. This decorrelates the learners and reduces overfitting compared to a single model trained on all data [50, 51].

In total, 200 decision trees were trained using bootstrap sampling with replacement. At each split, a random subset of features was selected to enhance diversity [52]. Out-of-bag estimation was also used to evaluate model error without needing a separate validation set. Table 2 summarizes the key parameters used to design the bagged trees ensemble.

**Table 2.** Bagged trees setting components

Parameters	Detail
Samples (N)	46,572
Predictors (P)	44
Learner	Decision Tree
Maximum number of splits	N - 1
Number of learners	200
Number of predictors per split	$\sqrt{P}$
Resampling	On
Resampling fraction	100%
Replacement	On
Minimum leaf size	1
Learning rate	1

To validate the performance of the BT model, we conducted empirical comparisons against several state-of-the-art ensemble learning-based classification algorithms for BP classification using PPG signals. Specifically, we compared our proposed BT model to:

**Boosted Trees (BOT):** An ensemble algorithm that builds a series of weak learners (typically DTs) sequentially [53], where each tree is trained to correct the errors of its predecessor, enhancing model performance through iterative refinement.

**Random Under-sampling Boosted Trees (RUSBT):** An adaptation of BOT that incorporates random under-sampling during training to handle class imbalance [54]. RUS selectively removes majority class instances to balance the class distributions, which can enhance the model's ability to learn from minority class examples.

**Subspace KNN (SKNN):** An ensemble technique that leverages multiple KNN classifiers, each trained on a different random subspace of the original feature set [55].

**Subspace Discriminant (SD):** Implements multiple linear discriminant analysis (LDA) models, each trained on a different random subspace of the input features.

### 3.6 Evaluation metrics

To evaluate model performance, we utilized several standard classification metrics, including specificity, sensitivity, accuracy, precision and F1 score. These metrics provide important insights into different aspects of a classifier's predictive capabilities:

-Sensitivity or recall reflects the ability to identify true positive rate by measuring the percentage of positive instances correctly detected:

$$SP = \frac{TP}{TP + FN} \quad (12)$$

High sensitivity indicates the model is effective at detecting the positive class.

-Specificity reflects the ability to identify true negative rate

by measuring the percentage of negative instances correctly detected:

$$SE = \frac{TN}{TN + FP} \quad (13)$$

High specificity indicates the model is effective at excluding the negative class.

-Precision quantifies the proportion of positive predictions that are correct:

$$PR = \frac{TP}{TP + FP} \quad (14)$$

High precision denotes low false positive rate.

-Accuracy measures overall correctness of predictions:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \quad (15)$$

F1 Score combines precision and recall as the harmonic mean:

$$F1 = \frac{2(SE \times PR)}{SE + PR} \quad (16)$$

In the equations above, TP (True Positives) represents cases correctly predicted as positive. FP (False Positives) represents incorrect positive predictions. TN (True Negatives) represents cases correctly predicted as negative. FN (False Negatives) represents incorrect negative predictions.

## 4. RESULTS AND DISCUSSION

### 4.1 Classification results

We designed an ensemble learning-based classification system using MATLAB (version R2020a) to non-invasively determine BP levels from PPG signals. To validate the effectiveness of the proposed BT model, we conducted empirical comparisons against several state-of-the-art ensemble classifiers for this task. Specifically, we evaluated the BT model against BOT, SKNN, RUSBT, and SD methods. All classifiers were implemented as integrated in the MATLAB's Classification Learner application interface, where each model consisted of 30 base learners. To evaluate the models' sensitivity to variations in data size, we trained them on three different proportions of the full dataset: 10%, 30%, and 50%. Tables 3-5 report the classification results achieved by each algorithm over three trials for the 10%, 30%, and 50% data subsamples, respectively. This systematic validation framework allowed for an impartial assessment of the BT model's performance relative to other prominent ensemble techniques, under varying data availability.

To validate and compare the classification performance of models, we utilized a standardized evaluation framework. Specifically, we employed an 80:20 train-test split on the 3 subsets of the dataset created using the  $C_n^3$  variable, which encapsulates both real and imaginary components of the FSST features. To ensure robust and unbiased estimates, we adopted a k-fold cross-validation approach. Specifically, we randomly partitioned each  $C_n^3$  sub-datasets into 5 equal mutually exclusive subsets (folds). Then, each model was trained on 4 folds and directly evaluated on the held-out fold, repeating this



process 10 times such that each fold was used once for validation. The average classification accuracy across the 10 iterations was then computed to provide a reliable aggregate performance metric for each model, minimizing variability that could arise from a single train-test split.

**Table 3.** Models performance using 10% of the data

Results	BT	BOT	SKNN	RUSBT	SD
Training accuracy (%)	96.7	81.9	94.5	76.2	73.9
Training time (sec)	16	32	40	25	14
Prediction speed (obs/sec)	6500	6300	450	8400	5500
Test accuracy	91.5	79.2	87.4	76.7	74.2

Notes: 1. obs: observation. 2. sec: second.

**Table 4.** Models performance using 30% of the data

Results	BT	BOT	SKNN	RUSBT	SD
Training accuracy (%)	98.5	82.3	96.9	76.8	74.9
Training time (sec)	46	111	661	109	25
Prediction speed (obs/sec)	16000	20000	110	23000	9400
Test accuracy	95.3	78.8	90	67.4	71.2

**Table 5.** Models performance using 50% of the data

Results	BT	BOT	SKNN	RUSBT	SD
Training accuracy (%)	98.5	80.8	97.1	75.1	74.1
Training time (sec)	89	218	3342	221	43
Prediction speed (obs/sec)	21000	31000	42	34000	9100
Test accuracy	96.2	77.8	91.6	64.4	71.9

The results across Tables 3-5 indicate that increasing the proportion of data generally improved performance of the BT and SKNN models. Specifically, the SKNN algorithm achieved training accuracies ranging from 94.5% to 97.1% and test accuracies ranging from 87.4% to 91.6% as the set size increased from 10% to 50% of the full dataset. The BT model demonstrated relatively superior and more consistent classification ability, with training accuracies varying between 96.7-98.5% and test accuracies from 91.5-96.2% over the different data sizes. However, the BOT, RUSBT and SD algorithms did not exhibit substantial gains in accuracy despite being provided with larger subsets. This suggests they may have been less capable of effectively learning the underlying patterns from the FSST-based feature space, likely attributable to inherent limitations in their respective architectures for this particular classification task and dataset.

For instance, the BOT model achieved a peak training accuracy of 82.3% and test accuracy of 79.2%. This suboptimal performance can be attributed to the fact that boosting algorithms are highly sensitive to noise and overfitting, especially in datasets with complex, non-linear relationships. The sequential nature of BOT means that any noise in the training data can be amplified [56], leading to poor generalization on the test set. Similarly, the RUSBT model attained a maximum training accuracy of 76.8% and validation accuracy of 76.7%. RUSBT combines random undersampling with boosting to handle class imbalance. However, the random

undersampling process can lead to the loss of important information, especially in small or moderately sized datasets. This loss of information, coupled with the boosting process's sensitivity to noise, likely hindered the RUSBT model's ability to learn effectively.

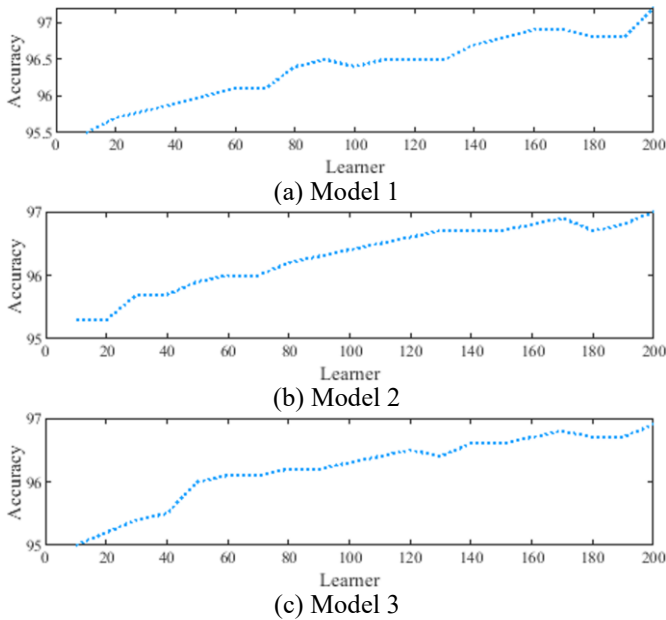
The SD model achieved a maximum training accuracy of 74.9% and test accuracy of 74.2%. The SD approach employs LDA on random subspaces of the feature space. However, LDA rests on the stringent assumption that data is linearly separable between classes [57]. PPG signals exhibit inherent nonlinearity and non-stationarity due to their complex physiological origin. Thus, the linear decision boundaries used by LDA are insufficiently expressive to accurately capture the intricate, nonlinear relationships embedded within the PPG feature space mapped by the FSST. Ultimately, while the BT and SKNN models demonstrated strong performance and robustness across different dataset sizes, the BOT, RUSBT, and SD models struggled due to their sensitivity to noise, loss of information during undersampling, and assumptions of linear separability, respectively. This analysis underscores the importance of selecting appropriate models that align with the data's characteristics to achieve optimal classification performance.

In terms of computational efficiency, the BT and SD models exhibited the most favorable training time scaling as the proportion of data was increased. Specifically, BT required 16, 46, and 89 seconds to train on 10%, 30%, and 50% of the full dataset, respectively. Comparatively, SD took 14, 25, and 43 seconds over the same training sizes. However, during model prediction the BT approach significantly outperformed SD in terms of speed, with observed throughput of up to 21,000 obs/sec versus 9,400 obs/sec for SD. This discrepancy can be attributed to inherent differences in their algorithmic architectures. BT leverages the parallelizable and efficient tree-based structure to rapidly classify new samples. In contrast, SD's classification mechanism involves more computationally intensive Linear Discriminant Analysis calculations during each prediction.

The remaining models exhibited less favorable scaling of training time with increasing data proportions compared to BT and SD. Notably, despite using a modest ensemble of only 30 base learners, the SKNN algorithm was significantly slower during model fitting. Specifically, SKNN required 3342 seconds to train on 50% of the full dataset, two orders of magnitude longer than BT or SD, with a correspondingly low prediction throughput of 42 obs/sec. This inefficiency stems from KNN's underlying classification mechanism, which involves calculating the distance between each new sample and all examples in the training set [58]. As data volume grows, this distance computation becomes prohibitively expensive from a complexity standpoint. While SKNN achieved good prediction accuracy, its quadratic computational complexity severely limits the algorithm's potential for further tuning by increasing hyperparameters like ensemble size or training data proportions. Any such adjustments would excessively prolong training time at the detriment of runtime performance required for real-time applications.

Training time also scaled less favorably for the BOT and RUSBT models as dataset proportions increased. BOT's sequential approach, whereby each additional tree is fit to emphasize mistaken examples from previous trees, intrinsically demands greater computational burden relative to a purely parallel ensemble method like BT. Moreover, RUSBT

compounds this complexity by incorporating random undersampling during training. The added steps of subsampling and sequential boosting accumulation elongate its training relative to simpler bootstrapping-based techniques.



**Figure 10.** BT tuning process

Ultimately, BT emerged as the preferable algorithmic choice for this real-time BP classification application owing to its balanced optimization of accuracy, efficiency, and scalability. Its parallelizable tree-based framework facilitates rapid model fitting and prediction even on larger datasets through sample-wise decomposition and averaging. This makes BT well-suited for applications requiring timely inference, such as continuous hemodynamic monitoring using PPG sensors. In contrast, approaches like BOT and RUSBT that impose sequential dependencies exhibited poorer calibration of computational demand to growing data volumes. Therefore, BT's combination of robust performance and favorable runtime properties support its suitability for the proposed real-time blood pressure estimation system.

To gain further insights into the BT model, we conducted parameter optimization by systematically increasing the number of base learner trees from 10 to 200. This tuning process employed a held-out validation subset comprising 20% of the original data. Figure 10 depicts the classification accuracy of BT models trained on datasets constructed from different FSST features, as the number of constitutive decision trees is incremented.

Across all feature representations, classification accuracy increases rapidly with additional trees, suggesting the BT ensembles converge near 200 trees. The tuning process thereby provides further empirical evidence confirming BT as an appropriate modeling approach for these nonlinear challenges involving large, complex biomedical datasets. Table 6 displays the results of the tuned model using the 48,072 size dataset, with the training set comprising 15,524 subjects for each class and the test set comprising 500 subjects for each class. Training required approximately 4 minutes per model, a reasonable duration given the large data volume. Performance was evaluated based on accuracy, confusion matrices, and F1 score using specific classification tasks as detailed below.

**Table 6.** The BT models' performances

Results	Model 1	Model 2	Model 3
Training accuracy (%)	100	100	100
Prediction speed (obs/s)	6600	4700	4600
Training time (s)	248	237	226
Test accuracy (%)	96.9	95.7	96.1

As depicted in Table 6, the proposed methodology exhibited good performance, achieving 100% training accuracy for each experiment. Testing accuracies were 96.9%, 95.7% and 96.1% for Experiments 1, 2 and 3, respectively. It is evident that model 1 performed best among the three, although all three models proved to be effective predictors of BP levels.

The training and testing confusion matrices (Figures 11 and 12) provide insight into model performance. The matrices outline relationships between true versus predicted classes, with diagonal cells (blue) representing TPs/TNs and off-diagonal cells (red) representing FPs/FNs. Columns indicate predicted classes (output axis) while rows indicate actual classes (target axis). Grey cells denote positive predictive values (column-wise) and true positive rates (row-wise).

The training matrices confirmed 100% accuracy distinguishing actual cases within each class. It indicates also that of all predicted cases 100% were correctly classified. However, evaluating the reliability of the experimental models relies on analyzing their test performance. Specifically, experiment 1 correctly classified 97.4%, 94.8%, and 98.6% of actual HT, PHT, and NT cases, respectively. Further, experiment 1 indicates also that of all HT, PHT and NT predictions, 98.4%, 97.7% and 94.8% were correctly classified, respectively.

Similarly, experiment 2 correctly classified 95.2%, 94%, and 97.8% of actual HT, PHT, and NT cases, respectively, while indicating also that of all HT, PHT and NT predictions, 97.7%, 96.1% and 93.3% were correctly classified, respectively. Furthermore, experiment 3 correctly classified 97%, 93%, and 98.2% of actual HT, PHT, and NT cases, respectively, while indicating also that of all HT, PHT and NT predictions, 98.2%, 97.1% and 93.2% were correctly classified, respectively.

The models underwent three classification trials: non-HT (NT + PHT) vs HT, NT vs HT, and NT vs PHT, in line with previous literature [31-34]. Tables 7-9 overview performance metrics - F1 score, precision, specificity, sensitivity, and accuracy - for model 1-3 across the trials. Model 3 attained the best F1 score (99%) for NT vs HT, while Models 1-2 scored 97.5-98%. Model 1 outperformed for NT+PHT vs HT (99%) and NT vs PHT (99.1%) tasks, while models 2-3 scored 98.9-96.6% and 98.8-98.8%, respectively. Overall, the models demonstrated comparable and consistently high performance across trials, indicating capabilities for BP level detection.

## 4.2 Comparative analysis

This section evaluates the system's classification performance relative to prior studies. Table 10 compares F1 scores from our work and others [31-34] for the (NT + PHT) vs HT, NT vs PHT, and NT vs HT classification tasks. According to the comparison results, our proposed models demonstrate superior robustness and efficiency. A major reason is that our models were trained on a much larger and more diverse dataset comprising a wider BP value range, an aspect not considered in earlier work.

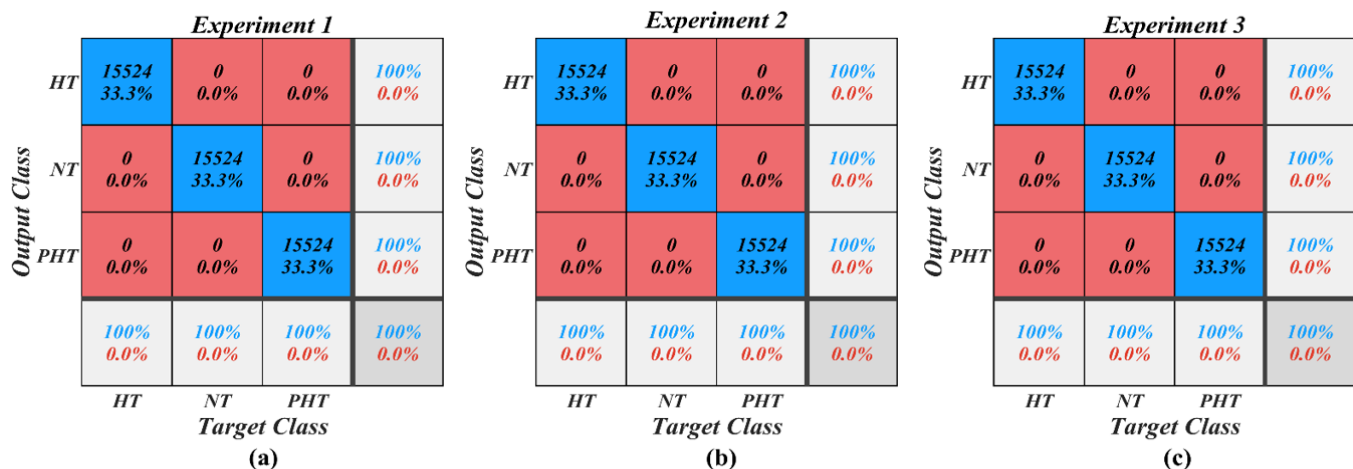


Figure 11. Confusion matrices describing the training performance of the three experimental models

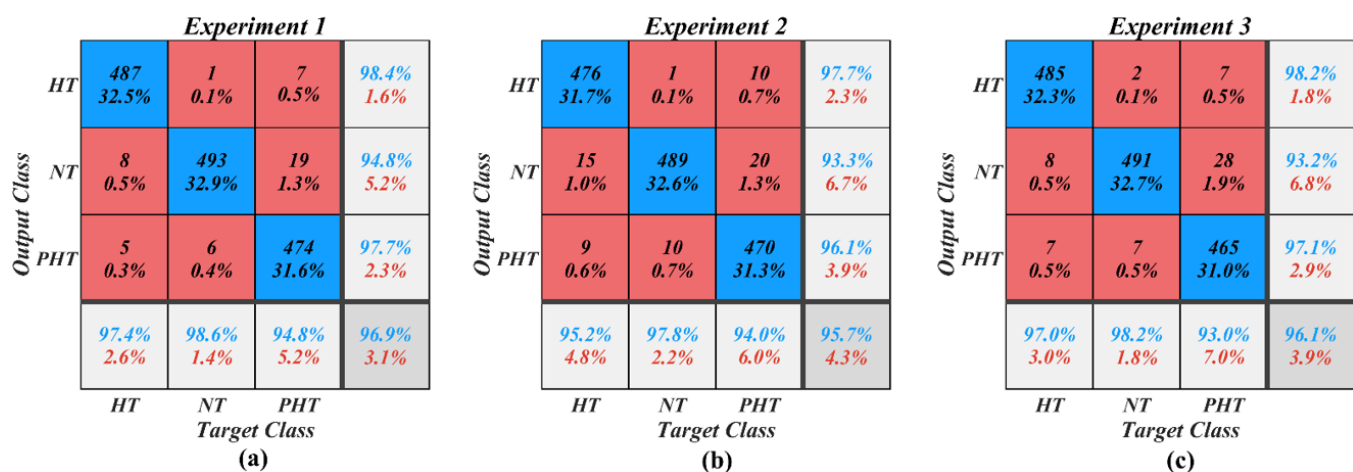


Figure 12. Confusion matrices describing the testing performance of the three experimental models

Table 7. The models' performances using the first experimental dataset (real variable)

Trial	TP	TN	FP	FN	Specificity (%)	Precision (%)	Sensitivity (%)	Accuracy (%)	F1 Score (%)
PHT	474	980	11	26	98.9	97.7	94.8	97.5	96.2
HT	487	967	8	13	99.2	98.4	97.4	98.6	97.9
NT	493	961	27	7	97.3	94.8	98.6	97.7	96.7
NT vs PHT	493	474	1	8	99.8	99.8	98.4	99.1	99.1
NT vs HT	493	487	19	6	96.3	96.3	98.8	97.5	97.5
(NT + PHT) vs HT	992	487	13	8	97.4	98.7	99.2	98.6	99

Table 8. The models' performances using the second experimental dataset (imaginary variable)

Trial	TP	TN	FP	FN	Specificity (%)	Precision (%)	Sensitivity (%)	Accuracy (%)	F1 Score (%)
PHT	470	965	19	30	98.1	96.1	94	96.7	95
HT	476	959	11	24	98.9	97.7	95.2	97.6	96.4
NT	489	946	35	11	96.4	93.3	97.8	96.9	95.5
NT vs PHT	489	470	20	10	96	99.8	98	97	98.9
NT vs HT	489	476	15	1	97	96.3	99.8	98.4	98
(NT + PHT) vs HT	989	476	24	11	97.1	98.6	98.9	97.7	98.8

Table 9. The models' performances using the third experimental dataset (absolute variable)

Trial	TP	TN	FP	FN	Specificity (%)	Precision (%)	Sensitivity (%)	Accuracy (%)	F1 Score (%)
PHT	465	976	14	35	98.6	97.1	93	96.7	95
HT	485	956	9	15	99.1	98.2	97	98.4	97.6
NT	491	950	36	9	96.4	93.2	98.2	97	95.6
NT vs PHT	491	465	28	7	94.3	94.6	98.6	96.5	96.6
NT vs HT	491	485	8	2	98.4	98.4	99.6	99	99
(NT + PHT) vs HT	991	485	15	9	97	98.5	99.1	98.4	98.8

**Table 10.** Results comparison relative to prior studies

Approach	Classification Task	Features	Research Subjects	Classifier	F1 Score
PAT features [33] (PPG and ECG signals)	NT (46 TS) vs HT (34 TS)	PAT and 10 PPG features	121 subjects	KNN	84.34%
	NT (46 TS) vs PHT (41 TS)				94.84%
	(NT + PHT) (87 TS) vs HT (34 TS)				88.49%
PAT features [33] (PPG and ECG signals)	NT (46 TS) vs HT (34 TS)	PAT and 10 PPG features	121 subjects	BT	94.13%
	NT (46 TS) vs PHT (41 TS)				83.88%
	(NT + PHT) (87 TS) vs HT (34 TS)				88.22%
PAT features [33] (ECG signals)	NT (46 TS) vs HT (34 TS)	PAT features	121 subjects	BT	68.10%
	NT (46 TS) vs PHT (41 TS)				66.95%
	(NT + PHT) (87 TS) vs HT (34 TS)				53.19%
PAT features [33] (only PPG signals)	NT (46 TS) vs HT (34 TS)	10 PPG features	121 subjects	BT	84.98%
	NT (46 TS) vs PHT (41 TS)				78.48%
	(NT + PHT) (87 TS) vs HT (34 TS)				75.32%
PPG features [33] (only PPG signals)	NT (46 TS) vs HT (34 TS)	10 PPG features	121 subjects	KNN	78.62%
	NT (46 TS) vs PHT (41 TS)				86.94%
	(NT + PHT) (87 TS) vs HT (34 TS)				78.44%
Raw PPG Signal [34]	NT (46 TS) vs HT (34 TS)	RGB images using CWT (scalogram)	2904 subjects (images)	CNN	80.52%
	NT (46 TS) vs PHT (41 TS)				92.55%
	(NT + PHT) (87 TS) vs HT (34 TS)				82.95%
Raw PPG Signal [32]	NT (46 TS) vs HT (34 TS)	2100 PPG features points	900 subjects	KNN	100%
	NT (46 TS) vs PHT (41 TS)				100%
	(NT + PHT) (87 TS) vs HT (34 TS)				90.80%
Raw PPG Signal [31]	NT (38 TS) vs HT (38 TS)	TF-features using STFT	900 subjects	BLSTM	97.29%
	NT (38 TS) vs PHT (38 TS)				97.39%
	(NT + PHT) (76 TS) vs HT (38 TS)				93.93%
This study (experiment 1)	NT (500 TS) vs HT (500 TS)	44 TF-features using FSST	48,072 subjects	BT	97.5%
	NT (500 TS) vs PHT (500 TS)				99.1%
	(NT + PHT) (1000 TS) vs HT (500 TS)				99%
This study (experiment 2)	NT (500 TS) vs HT (500 TS)	44 TF-features using FSST	48,072 subjects	BT	98%
	NT (500 TS) vs PHT (500 TS)				98.9%
	(NT + PHT) (1000 TS) vs HT (500 TS)				98.8%
This study (experiment 3)	NT (500 TS) vs HT (500 TS)	44 TF-features using FSST	48,072 subjects	BT	99%
	NT (500 TS) vs PHT (500 TS)				96.6%
	(NT + PHT) (1000 TS) vs HT (500 TS)				98.8%

Note: TS: Test Subjects.

The proposed BT models achieved F1 scores of 99%, 99.1%, and 99% for the NT vs. HT, NT vs. PHT, and non-HT vs. HT classification tasks, respectively. This exceeded prior published method performance, aside from study [32] which reported F1 scores of 100% for NT vs. HT and NT vs. PHT using a KNN model. However, their approach attained a lower F1 of 90.8% for the non-HT vs. HT task compared to our BT models' consistent F1 scores exceeding 98.8%. Meanwhile, Tjahjadi et al. [31] achieved competitive yet slightly lower F1 scores of 97.29%, 97.39%, and 93.93% across the same tasks using a smaller 786 sample dataset. However, both studies [31, 32] relied on more limited data lacking the diversity in BP ranges observed in our sample collection (Figures 2, 3, 6). Additionally, our BT models were evaluated on a larger test set of 500 subject-segments per class, providing a more robust evaluation. Moreover, training each BT model in our study required only approximately 4 minutes to calibrate, significantly faster training times than previous works - for example over 30 minutes for the BLSTM model in reference [31] using a smaller 786 sample set, and over 350 minutes for the CNN approach in reference [34] using 2323 samples.

For a fair comparison with previous studies that employed similar modeling techniques, Table 10 presents results from Liang et al. [33] who used various feature extraction methods as inputs for a BT classifier. In their study, a combination of PAT and 10 PPG morphological features delivered the highest F1 scores of 94.13%, 83.88%, and 88.22% for NT vs. HT, NT vs. PHT, and non-HT vs. HT classifications, respectively. In contrast, when only FSST features of the PPG signal were used

as input, our proposed BT model demonstrated significantly enhanced performance across all tasks, with F1 scores exceeding 96%. Although Liang et al. [33] achieved competitive results using a KNN model with PPG+PAT features, their method required an additional sensor beyond just PPG. Moreover, morphological feature extraction is sensitive to signal quality [34] and susceptible to motion artifacts, which may limit its practical applicability [27].

When solely using PPG signals without additional sensors, Liang et al. [33] reported markedly lower performance from their BT model, achieving F1 scores of 84.98%, 78.48%, and 75.32% for NT vs. HT, NT vs. PHT, and non-HT vs. HT classifications, respectively. Their KNN classifier yielded nearly identical results. In contrast, our BT models attained significantly higher F1 scores exceeding 96% across all tasks using only 44 features directly extracted from the raw PPG waveform via FSST, without requiring special pre-processing steps to isolate morphological characteristics. Notably, our proposed framework maintained this consistent, superior classification ability while overcoming limitations of prior work such as smaller, less diverse datasets [31, 32], computationally expensive training procedures [31, 34], and reliance on additional sensors or susceptibility to noise inherent in complex feature extraction [33, 34]. Overall, these comparisons demonstrate the proposed models establish new state-of-the-art performance for non-invasive BP monitoring using exclusively PPG data, represented by efficiently extracted Fourier features, without pre-processing vulnerabilities or constraints of earlier efforts.

### 4.3 System design and clinical perspectives

PPG provides a convenient and noninvasive means of measuring pulsatile blood volume changes, with the periodic waveform conveying insights into cardiovascular health [17]. However, deriving meaningful features from PPG signals is confounded by the nonstationary nature of the signals, whereby the statistical properties evolve dynamically over time. To address this, time-frequency decomposition was performed using FSST to generate high-resolution spectrograms revealing transient cardiovascular variabilities related to BP regulation. Spectrographic analysis enabled engineering statistical features including mean, variance, skewness and kurtosis, all together represent pressure-dependent physiological changes missed by basic morphological features. This set of TF signal features formed the input vectors for the BP classifier, providing a robust representation of the non-stationary dynamics within the PPG data.

For the classification architecture, an ensemble approach was pursued using 200 bagged decision trees, improving accuracy and robustness over individual models by reducing variance and overfitting [50]. Bagging provides an efficient means to leverage large training datasets as the models can be fit in parallel on data subsets. Specifically, the model was trained on a sizeable dataset spanning normotensive, prehypertensive, and hypertensive BP ranges collected from the MIMIC-III database [41]. This diversity is expected to improve real-world generalizability and reduce biases compared to models trained on a narrow BP distribution. Low computation complexity enabled training in just minutes per experiment on a dataset of this scale. Additionally, testing on a balanced dataset better assesses performance across the spectrum of the population.

The approach demonstrated potential as inexpensive assistant tool in clinical setting, as the experimental models achieved F1 scores ranging from 96.6% to 99.1% for non-HT vs. HT, NT vs. HT, and NT vs. PHT classifications. Trials that play a significant role for routine prehypertension and hypertension detection, which in turn contribute for the early hypertension diagnosis and management [34]. Additionally, these results justify employing the proposed system in real-world scenarios where various sources of disturbance could interfere, owing to synchrosqueezing stability in handling perturbed signals [59]. Which is reasonable as the PPG signals were captured from patients in ICU wards [40].

Furthermore, by requiring only raw PPG input, the technique provides a convenient means for wearable-based hypertension monitoring without reliance on cuff devices. The lightweight tree models allow implementation without cloud computing, enabled by low memory footprint compared to deep neural networks. The key advantages in using bagged trees over DL models in wearable monitoring devices are summarized as follows:

-Bagged decision tree models are highly suitable for implementation on wearable devices due to their lower complexity and minimal computational requirements. Storing the parameters for decision tree models requires very little memory on the wearable hardware. This allows the entire bagged ensemble to be housed on the device, rather than relying on cloud connectivity for model storage and predictions. This reduces latency, costs, and privacy concerns by keeping data localized. In contrast, deep neural networks have substantial storage needs due to the large number of

weight parameters distributed across many layers, making cloud storage more practical.

-For prediction, bagged trees run efficiently on cheaper, lower-power CPUs by avoiding expensive high-RAM GPUs required by deep nets to process their computationally intensive activation layers in parallel. By avoiding such specialized hardware, tree ensembles enable cost savings on the device processors.

-Bagged trees support continuous incremental learning by continuously adding new models to the ensemble on-device as more training data becomes available [48]. Retraining deep nets to update their weights is much more computationally expensive. This makes regular model updating prohibitive on wearable processors.

-For clinical applications, model explanations matter for regulatory approval, accountability, and transparency around AI assisted medical decision making. Accordingly, decision trees have clear logical structures that are easy to understand and visualize making them more interpretable. In contrast, deep neural networks operate like a black box, thus, obscuring the basis for their predictions.

By enabling affordable BP monitoring directly through consumer wearables, this approach promotes accessible screening without expensive medical devices. The lightweight system allows self-tracking globally, even in resource-constrained settings, supporting prevention through timely personalized notifications. This addresses many patients' lack of knowledge regarding acceptable BP levels [25, 26], which currently relies on infrequent clinic measurements. Widespread deployment can enhance early diagnosis, convenient management, and improved outcomes by averting target organ damage and associated healthcare costs from untreated hypertension.

## 5. CONCLUSIONS

This research presented a novel technique for classifying BP levels by employing only PPG signals. TF analysis was employed to reveal subtle cardiovascular dynamics in PPG waveforms associated with BP changes. FSST generated high-resolution TF representations from which 44 statistical features were extracted, capturing transient physiological variations related to BP regulation missed by basic PPG morphology.

An ensemble classifier comprising 200 bagged decision trees was trained on a large dataset encompassing 46,572 subjects across varied BP ranges. This boosted accuracy and generalizability across three different experimental datasets, as evidenced by the achieved training accuracies of 100% and testing accuracies ranging from 95.7% to 96.9% across the experimental models. Furthermore, the developed system demonstrated exceptional performance across three different classification trials, including NT vs. HT, NT vs. PHT and non-HT vs. HT, with F1 scores up to 99.1%.

Unlike prior works limited by smaller datasets, this research leveraged substantial variability in BP levels, demonstrating the technique's efficiency and versatility. The approach requires only single-lead PPG data, avoiding reliance on multiple monitoring devices. Additionally, the low-complexity tree models enable direct on-device implementation without cloud connectivity, overcoming barriers of DL models.

Overall, the methodology shows considerable promise as an

accessible, affordable self-screening tool to promote early hypertension detection through consumer wearables. By enabling convenient BP tracking anytime, anywhere, the approach can help improve outcomes globally through timely notifications and personalized hypertension management. Further research should focus on evaluating real-world accuracy across diverse populations and age groups, including detailed reporting on the demographic and health characteristics of the study participants. With refinement, the technique could empower patients and providers with convenient tools for early diagnosis and prevention of silent cardiovascular risks.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support provided through our institutional affiliations during the completion of this research project.

## REFERENCES

- [1] Lim, S.S., Vos, T., Flaxman, A.D., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859): 2224-2260. [https://doi.org/10.1016/S0140-6736\(12\)61766-8](https://doi.org/10.1016/S0140-6736(12)61766-8)
- [2] World Health Organization. (2013). A global brief on hypertension: Silent killer, global public health crisis: World Health Day 2013. <https://apps.who.int/iris/handle/10665/79059>.
- [3] Brzezinski, W.A. (1990). Blood Pressure. In *Clinical Methods: The History, Physical, and Laboratory Examinations* (3rd Edition).
- [4] Sharma, M., Barbosa, K., Ho, V., Griggs, D., Ghirmai, T., Krishnan, S.K., Hsiai, T.K., Chiao, J.C., Cao, H. (2017). Cuff-less and continuous blood pressure monitoring: A methodological review. *Technologies*, 5(2): 21. <https://doi.org/10.3390/technologies5020021>
- [5] Alexander, B., Cannesson, M., Quill, T.J. (2013). Blood pressure monitoring. In *Anesthesia Equipment*, pp. 273-282.
- [6] Schoot, T.S., Weenk, M., Van De Belt, T.H., Engelen, L.J., Van Goor, H., Bredie, S.J. (2016). A new cuffless device for measuring blood pressure: A real-life validation study. *Journal of Medical Internet Research*, 18(5): e85. <https://doi.org/10.2196/jmir.5414>
- [7] Nachman, D., Gilan, A., Goldstein, N., et al. (2021). Twenty-four-hour ambulatory blood pressure measurement using a novel noninvasive, cuffless, wireless device. *American Journal of Hypertension*, 34(11): 1171-1180. <https://doi.org/10.1093/ajh/hpab095>
- [8] Pan, F., He, P., Wang, H., et al. (2021). Development and validation of a deep learning-based automatic auscultatory blood pressure measurement method. *Biomedical Signal Processing and Control*, 68: 102742. <https://doi.org/10.1016/j.bspc.2021.102742>
- [9] Muntner, P., Shimbo, D., Carey, R.M., et al. (2019). Measurement of blood pressure in humans: A scientific statement from the American Heart Association. *Hypertension*, 73(5): e35-e66. <https://doi.org/10.1161/HYP.0000000000000087>
- [10] Yüksel, S., Altun-Uğraş, G., Altınok, N., Demir, N. (2020). The effect of cuff size on blood pressure measurement in obese surgical patients: A prospective crossover clinical trial. *Florence Nightingale Journal of Nursing*, 28(2): 205. <https://doi.org/10.5152/FNJV.2020.19119>
- [11] Penaz, J. (1973). Photoelectric measurement of blood pressure, volume and flow in the finger. In *Digest of the 10th International Conference on Medical and Biological Engineering*, Dresden, Germany.
- [12] Athaya, T., Choi, S. (2022). A review of noninvasive methodologies to estimate the blood pressure waveform. *Sensors*, 22(10): 3953. <https://doi.org/10.3390/s22103953>
- [13] Lim, M.J., Tan, C.W., Tan, H.S., Sultana, R., Eley, V., Sng, B.L. (2020). Correlation of patient characteristics with arm and finger measurements in Asian parturients: A preliminary study. *BMC Anesthesiology*, 20: 218. <https://doi.org/10.1186/s12871-020-01131-6>
- [14] Athaya, T., Choi, S. (2021). An estimation method of continuous non-invasive arterial blood pressure waveform using photoplethysmography: A U-Net architecture-based approach. *Sensors*, 21(5): 1867. <https://doi.org/10.3390/s21051867>
- [15] Harfiya, L.N., Chang, C.C., Li, Y.H. (2021). Continuous blood pressure estimation using exclusively photoplethysmography by LSTM-based signal-to-signal translation. *Sensors*, 21(9): 2952. <https://doi.org/10.3390/s21092952>
- [16] Martínez, G., Howard, N., Abbott, D., Lim, K., Ward, R., Elgendi, M. (2018). Can photoplethysmography replace arterial blood pressure in the assessment of blood pressure? *Journal of Clinical Medicine*, 7(10): 316. <https://doi.org/10.3390/jcm7100316>
- [17] Tamura, T., Maeda, Y., Sekine, M., Yoshida, M. (2014). Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2): 282-302. <https://doi.org/10.3390/electronics3020282>
- [18] O'Brien, E., Petrie, J., Littler, W., et al. (1990). The British hypertension society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems. *Journal of Hypertension*, 8(7): 607-619.
- [19] White, W.B., Berson, A.S., Robbins, C., Jamieson, M.J., Prisant, L.M., Roccella, E., Sheps, S.G. (1993). National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers. *Hypertension*, 21(4): 504-509. <https://doi.org/10.1161/01.HYP.21.4.504>
- [20] Wang, W., Zhu, L., Marefat, F., Mohseni, P., Kilgore, K., Najafizadeh, L. (2020). Photoplethysmography-based blood pressure estimation using deep learning. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, pp. 945-949. <https://doi.org/10.1109/IEEECONF51394.2020.9443447>
- [21] Maher, N., Elsheikh, G.A., Anis, W.R., Emara, T. (2021). Enhancement of blood pressure estimation method via machine learning. *Alexandria Engineering Journal*, 60(6): 5779-5796. <https://doi.org/10.1016/j.aej.2021.04.035>
- [22] El-Hajj, C., Kyriacou, P.A. (2021). Deep learning models for cuffless blood pressure monitoring from PPG signals

- using attention mechanism. *Biomedical Signal Processing and Control*, 65: 102301. <https://doi.org/10.1016/j.bspc.2020.102301>
- [23] El-Hajj, C., Kyriacou, P.A. (2021). Cuffless blood pressure estimation from PPG signals and its derivatives using deep learning models. *Biomedical Signal Processing and Control*, 70: 102984. <https://doi.org/10.1016/j.bspc.2021.102984>
- [24] Qiu, Y., Liu, D., Yang, G., et al. (2021). Cuffless blood pressure estimation based on composite neural network and graphics information. *Biomedical Signal Processing and Control*, 70: 103001. <https://doi.org/10.1016/j.bspc.2021.103001>
- [25] Slark, J., Khan, M.S., Bentley, P., Sharma, P. (2014). Knowledge of blood pressure in a UK general public population. *Journal of Human Hypertension*, 28(8): 500-503. <https://doi.org/10.1038/jhh.2013.136>
- [26] Wright-Nunes, J.A., Luther, J.M., Ikizler, T.A., Cavanaugh, K.L. (2012). Patient knowledge of blood pressure target is associated with improved blood pressure control in chronic kidney disease. *Patient Education and Counseling*, 88(2): 184-188. <https://doi.org/10.1016/j.pec.2012.02.015>
- [27] Seok, D., Lee, S., Kim, M., Cho, J., Kim, C. (2021). Motion artifact removal techniques for wearable EEG and PPG sensor systems. *Frontiers in Electronics*, 2: 685513. <https://doi.org/10.3389/felec.2021.685513>
- [28] Auger, F., Flandrin, P., Lin, Y.T., McLaughlin, S., Meignen, S., Oberlin, T., Wu, H.T. (2013). Time-frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30(6): 32-41. <https://doi.org/10.1109/MSP.2013.2265316>
- [29] Yen, C.T., Chang, S.N., Liao, C.H. (2021). Deep learning algorithm evaluation of hypertension classification in less photoplethysmography signals conditions. *Measurement and Control*, 54(3-4): 439-445. <https://doi.org/10.1177/00202940211001904>
- [30] Liang, Y., Chen, Z., Liu, G., Elgendi, M. (2018). A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China. *Scientific Data*, 5(1): 180020. <https://doi.org/10.1038/sdata.2018.20>
- [31] Tjahjadi, H., Ramli, K., Murfi, H. (2020). Noninvasive classification of blood pressure based on photoplethysmography signals using bidirectional long short-term memory and time-frequency analysis. *IEEE Access*, 8: 0735-20748. <https://doi.org/10.1109/ACCESS.2020.2968967>
- [32] Tjahjadi, H., Ramli, K. (2020). Noninvasive blood pressure classification based on photoplethysmography using k-nearest neighbors algorithm: A feasibility study. *Information*, 11(2): 93. <https://doi.org/10.3390/info11020093>
- [33] Liang, Y., Chen, Z., Ward, R., Elgendi, M. (2018). Hypertension assessment via ECG and PPG signals: An evaluation using MIMIC database. *Diagnostics*, 8(3): 65. <https://doi.org/10.3390/diagnostics8030065>
- [34] Liang, Y., Chen, Z., Ward, R., Elgendi, M. (2018). Photoplethysmography and deep learning: Enhancing hypertension risk stratification. *Biosensors*, 8(4): 101. <https://doi.org/10.3390/bios8040101>
- [35] Meignen, S., Oberlin, T., Pham, D.H. (2019). Synchrosqueezing transforms: From low-to highfrequency modulations and perspectives. *Comptes Rendus Physique*, 20(5): 449-460. <https://doi.org/10.1016/j.crhy.2019.07.001>
- [36] Mateo, C., Talavera, J.A. (2020). Bridging the gap between the short-time Fourier transform (STFT), wavelets, the constant-Q transform and multi-resolution STFT. *Signal, Image and Video Processing*, 14(8): 1535-1543. <https://doi.org/10.1007/s11760-020-01701-8>
- [37] Flandrin, P., Auger, F., Chassande-Mottin, E. (2018). Time-frequency reassignment: From principles to algorithms. In *Applications in Time-Frequency Signal Processing*, pp. 179-204. <https://doi.org/10.1201/9781315220017-5>
- [38] Thakur, G. (2015). The synchrosqueezing transform for instantaneous spectral analysis. In *Excursions in Harmonic Analysis, Volume 4. Applied and Numerical Harmonic Analysis*. Birkhäuser, Cham. [https://doi.org/10.1007/978-3-319-20188-7\\_15](https://doi.org/10.1007/978-3-319-20188-7_15)
- [39] Degirmenci, D., Yalcin, M., Ozdemir, M.A., Akan, A. (2020). Synchrosqueezing Transform in Biomedical Applications: A mini review. In *2020 Medical Technologies Congress (TIPTEKNO)*, Antalya, Turkey, pp. 1-5. <https://doi.org/10.1109/TIPTEKNO50054.2020.9299225>
- [40] Johnson, A.E., Pollard, T.J., Shen, L., et al. (2016). Data descriptor: MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3: 160035. <https://doi.org/10.1038/sdata.2016.35>
- [41] Moody, B., Moody, G., Villarroel, M., Clifford, G.D., Silva, I. (2017). MIMIC-III waveform database. *PhysioNet*. <https://doi.org/10.13026/C2607M>
- [42] Chobanian, A.V., Bakris, G.L., Black, H.R., et al. (2003). Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension*, 42(6): 1206-1252. <https://doi.org/10.1161/01.HYP.0000107251.49515.c2>
- [43] O'Shea, J.C., Califf, R.M. (2006). 24-Hour ambulatory blood pressure monitoring. *American Heart Journal*, 151(5): 962-968. <https://doi.org/10.1016/j.ahj.2005.03.020>
- [44] Raymaekers, V., Brenard, C., Hermans, L., Frederix, I., Staessen, J.A., Dendale, P. (2019). How to reliably diagnose arterial hypertension: lessons from 24 h blood pressure monitoring. *Blood Pressure*, 28(2): 93-98. <https://doi.org/10.1080/08037051.2018.1557508>
- [45] Parati, G., Schumacher, H. (2014). Blood pressure variability over 24 h: Prognostic implications and treatment perspectives. An assessment using the smoothness index with telmisartan-amlodipine monotherapy and combination. *Hypertension Research*, 37(3): 187-193. <https://doi.org/10.1038/hr.2013.145>
- [46] Singh, A.K., Gewali, L.P., Khatiwada, J. (2019). New measures of skewness of a probability distribution. *Open Journal of Statistics*, 9(5): 601-621. <https://doi.org/10.4236/ojs.2019.95039>
- [47] Cohen, J.E., Davis, R.A., Samorodnitsky, G. (2020). Heavy-tailed distributions, correlations, kurtosis and Taylor's Law of fluctuation scaling. *Proceedings of the Royal Society A*, 476(2244): 20200610. <https://doi.org/10.1098/rspa.2020.0610>
- [48] Karanikola, A., Karlos, S., Kazllarof, V., Kotsiantis, S. (2018). An incrementally updateable ensemble learner. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*, Athens Greece, pp. 243-248. <https://doi.org/10.1145/3291533.3291536>

- [49] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24: 123-140. <https://doi.org/10.1007/BF00058655>
- [50] Liu, H., Gegov, A., Cocca, M. (2016). Collaborative rule generation: An ensemble learning approach. *Journal of Intelligent & Fuzzy Systems*, 30(4): 2277-2287. <https://doi.org/10.3233/IFS-151997>
- [51] Hu, Q., Yu, D., Xie, Z., Li, X. (2007). EROS: Ensemble rough subspaces. *Pattern Recognition*, 40(12): 3728-3739. <https://doi.org/10.1016/j.patcog.2007.04.022>
- [52] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5-32. <https://doi.org/10.1023/A:1010933404324>
- [53] Bartlett, P., Freund, Y., Lee, W.S., Schapire, R.E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5): 1651-1686. <https://doi.org/10.1214/aos/1024691352>
- [54] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1): 185-197. <https://doi.org/10.1109/TSMCA.2009.2029559>
- [55] Ho, T.K. (1998). Nearest neighbors in random subspaces. In *Advances in Pattern Recognition: Joint IAPR International Workshops SSPR'98 and SPR'98 Sydney, Australia*, pp. 640-648. <https://doi.org/10.1007/BFb0033288>
- [56] Ponti, M.A., Oliveira, L.D.A., Román, J.M., Argerich, L. (2022). Improving Data Quality with Training Dynamics of Gradient Boosting Decision Trees. *arXiv preprint arXiv:2210.11327*. <https://doi.org/10.48550/ARXIV.2210.11327>
- [57] Elkhailil, K., Kammoun, A., Couillet, R., Al-Naffouri, T.Y., Alouini, M.S. (2020). A large dimensional study of regularized discriminant analysis. *IEEE Transactions on Signal Processing*, 68: 2464-2479. <https://doi.org/10.1109/TSP.2020.2984160>
- [58] Hamza, M.M., Noredine, B., Abdeslem, B.Z., Benyssaad, Y., Benselama, S.I. (2024). Classifying abnormal arterial pulse patterns in cardiovascular diseases: A photoplethysmography and machine learning approach. *Traitement du Signal*, 41(2): 543-562. <https://doi.org/10.18280/ts.410201>
- [59] Thakur, G., Brevdo, E., Fučkar, N.S., Wu, H.T. (2013). The synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications. *Signal Processing*, 93(5): 1079-1094. <https://doi.org/10.1016/j.sigpro.2012.11.029>