Spectral Transformation-Based PPG Representation for Improved Blood Pressure Estimation Using Transformer Models

Shyamala Subramanian[1], Sashikala Mishra[1], Ranjeet Bidwe[1]*

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune 412115, Maharashtra, India.
shyamalamathi123@gmail.com, shashikala.mishra@sitpune.edu.in, *ranjeet.bidwe@sitpune.edu.in

Abstract. The measurement of blood pressure (BP) accurately and non-invasively is still a very important challenge in monitoring cardiovascular health. This exploratory study evaluates multiple spectral conversion methods, Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), Fast Fourier Transform (FFT), and Cepstrum, for deep learning BP prediction using photoplethysmogram (PPG) and electrocardiogram (ECG) signals. This study aims to address the need for a systematic comparison of spectral conversion methods towards facilitating the development of cuffless blood pressure estimation methods that can support early diagnosis and continuous health monitoring in clinical and remote settings. STFT provided the best trade-off between predictive performance and computational cost. For our final ensemble model we deployed (an ensemble based upon a weighted geometric mean of the six best performing STFT based transformer configurations), the final ensemble achieved MAE values of 3.91 mmHg (DBP) and 4.97 mmHg (SBP), with standard deviations of 6.96 mmHg and 7.88 mmHg, respectively. Pearson correlation coefficients were 0.9220 (DBP) and 0.9391 (SBP). Both STFT models in terms of MAE and their respective standard deviations satisfy the clinical acceptability criteria (as defined by the Association for the Advancement of Medical Instrumentation (AAMI). Overall, these results indicate strong potential for reliable and clinically acceptable cuffless BP estimation using STFT-based features with transformer architectures.

Keywords: Acute Myeloid Leukemia, Extracellular vesicles, Microvesicles, Exosomes, Mesenchymal stromal cells.

1 Introduction

The continuous and noninvasive monitoring of blood pressure is increasingly important for the identification and management of cardiovascular diseases such as hypertension, stroke, and heart failure. Hypertension is a chronic condition that affects greater than a billion people worldwide and is a leading cause of heart-related illness and death [1]. Cuff-based sphygmomanometers are the gold standard and provide fairly accurate assessments of blood pressure in the clinic; however, cuff-based designs are not

A. Vaidya et al. (eds.), *Proceedings of the Conference on Bioengineering for Global Health (SYMRESEARCH 2.0 2025)*, Advances in Biological Sciences Research 55,
https://doi.org/10.2991/978-94-6239-756-9_21

appropriate for continuous measurements as they are bulky and uncomfortable to wear, thus rendering them incompatible with wearable or mobile technologies [2], [3].

There has been growing interest in photoplethysmography (PPG) and electrocardiography (ECG) for non-invasive blood pressure estimation without a cuff. These signals are non-invasive, relatively easy to extract through wearable devices, and have a large amount of physiological information [4]. Together, PPG and ECG signals provide different information on cardiovascular behavior [5], which are complementary. Consequently, the combination of PPG and ECG signals may improve the accuracy of blood pressure estimates in real-world environments.

Deep learning has recently opened up new opportunities to analyze these signals in a way that is more accurate and nuanced compared with traditional techniques. When considering traditional approaches like linear regression or the handcrafted extraction of features, neural networks are more beneficial in that they can better learn complex nonlinear relationships that cannot be modeled as usefully with traditional techniques [6], [7]. Even with advances in modeling, signal preprocessing is still necessary to determine performance of the model. Time-domain features are appealing because they are intuitive and follow the shape of the original signal. However, they can be sensitive to noise and motion in real-world environments and particularly when using wearable data, and these factors lower the value of using the time-series signal. Both a frequency and time-frequency approach will generally provide the best performance as they provide a more reliable representation of changes in signals by capturing the timing and spectral content of changes in a signal.

In this work, we systematically compare four spectral transformations (STFT, CWT, FFT, and Cepstrum) within a common transformer-based framework, considering both prediction performance and computational cost. The contributions made to this study are summarized as follows to highlight the key findings and advancements in methodologies used:

An innovative transformer-based blood pressure estimation framework is proposed, which incorporates spectral features from multimodal PPG and ECG inputs while eliminating the need to manually engineer the feature set used to estimate the blood pressure.

The study compares four signal transformation techniques, STFT, CWT, FFT, and Cepstrum by systematically analyzing their impact on accuracy, robustness, computational demand, and scalability.

In order to increase stability and generalize one's prediction results, several transformer models are combined into an ensemble method for their predictions by performing a geometric mean aggregation over each model's output.

The comparative evaluation highlights the relative trade-offs between accuracy and computational efficiency among the investigated spectral transformations.

The remainder of the paper will consist of: Section 2 will review existing literature associated with spectral transform methods & cuffless BP estimation; Section 3 will discuss the methodology used to conduct this research, including a description of the datasets used, preprocessing, processing methods, and model construction. Section 4 explains the experimental setup, while Section 5 presents results and discussion. Section 6 concludes the study.

2 Related work

The advancement of cuffless BP estimation has been rapid in recent years due to the rise of deep learning methods on physiological signals like PPG and ECG signals. A critical part of such pipelines is the conversion of raw signals to informative representations. Research has demonstrated the utility of converting information to different signal domains: time, frequency, and time-frequency, with each domain having its specific features, strengths, and weaknesses.

Time-domain features (e.g., pulse arrival time (PAT), pulse transit time (PTT), systolic time intervals) are popular because of their relation to physiology, low computational costs, and relative ease of extraction. Kachuee et al. [8] developed a CNN-based regression model to perform cuffless BP estimation using a combination of pulse transit time (PTT) and a set of handcrafted features, but only reached moderate accuracy. Though time-domain metrics such as those above are usable, they are highly prone to noise and motion artifacts, which may limit the practicality of these estimates in reality [3]. Researchers have also investigated frequency domain methods for blood pressure estimation using Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT). For instance, Xing and Sun [9] extracted amplitude and phase features from FFT transformed PPG signals and fed them into a neural network. Similarly, Nawaz et al. [10] used DCT-based frequency domain learning in a transformer model to synthesize ABP waveforms and estimate BP. Time-frequency methods can address tradeoffs between time-domain and frequency-domain methods. The STFT and CWT are two examples. Omer et al. [11] applied wavelet scattering transforms with an LSTM model, validated against several lengths of time series, highlighting the benefit of being more resilient to distortions in the signal. Moreover, Wu et al. [12] applied CWT within a deep learning pipeline, finding it better able to capture transience in cardiovascular features compared to traditional methods. While CWT has been shown to perform very well, its computational costs limit its scalability. STFT, which has the benefit of constant time-frequency resolution, has gained attention due to its computational efficiency and feasibility for real-time implementations. Recently, several studies have utilized STFT features within deep learning models for classification tasks using PPG [13], [14], [15]. However, there has been no comparative analysis of STFT, CWT, FFT, and cepstral techniques within a common regression framework for blood pressure estimates. Cepstral analysis is one more approach that has not been fully explored. Cepstral analysis provides periodic representations in the frequency spectrum, which serves as a reliable account of signal behaviour from a different angle. Cepstral analysis has been explored in areas like human identification and blood glucose estimation using PPG signals with Mel-frequency cepstral coefficients (MFCCs) [16], [17], but this has not been done within blood pressure predictions. Transformer-based architectures have earned attention due to their ability to learn long-range dependencies via self-attention mechanisms. Nawaz et al. [10] employed a transformer encoder–decoder model with frequency-domain learning for arterial BP waveform synthesis from PPG, achieving clinical-grade accuracy. Arjomand et al. [7] introduced TransfoRhythm, a transformer framework using only PPG signals for cuff-less BP estimation, trained on the MIMIC-IV dataset.

Most existing cuffless BP studies evaluate only one representation or transformation at a time, which limits direct comparison of accuracy–efficiency trade-offs across methods. Moreover, relatively few studies benchmark results against clinical acceptability criteria (e.g., MAE < 5 mmHg and SD < 8 mmHg for SBP and DBP), which is essential for assessing real-world wearable or real-time applicability.
This work addresses this gap by evaluating STFT, CWT, FFT, and cepstral representations using the same transformer-based pipeline across multiple dataset sizes. The prediction accuracy, computational efficiency, clinical benchmarks (AAMI) were assessed, and the impact of ensemble modeling on prediction stability. This study specifically investigates the utility of the short-time Fourier transform (STFT) as a robust and scalable feature representation for blood pressure estimation focusing on its utility for models of different complexity and datasets of differing sizes. It will examine the model's performance in regard to both the AAMI-required systolic and diastolic BP and the impact of ensemble modeling techniques on model-level prediction stability by using a broader set of more detailed metrics for each configuration.

3 Methodology

This section describes the dataset details, data processing pipeline, spectral transformation techniques, model architecture, ensemble strategy, and evaluation metrics that were used to examine how multimodal spectral features influence the transformer-based BP model (Fig. 1).

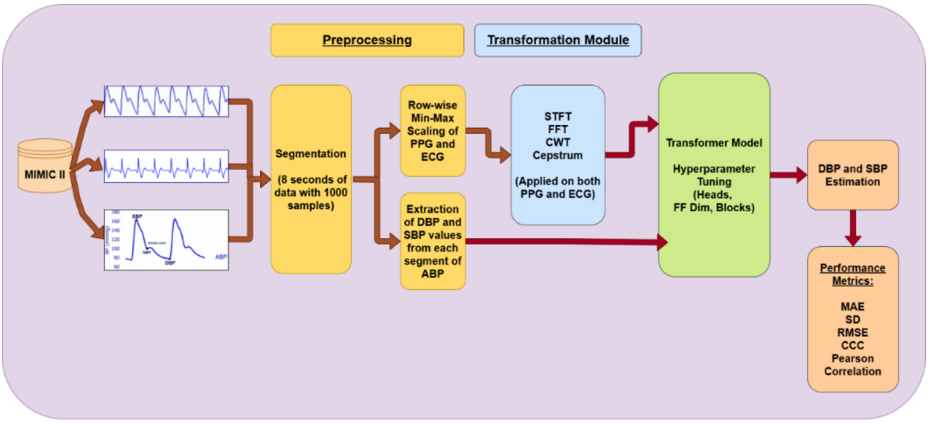

Fig. 1. Transformer-based blood pressure estimation framework.

The pipeline starts with the acquisition of both the PPG and ECG signals synchronized, followed by the transformation into the frequency domain using a transformation technique such as the STFT, CWT, FFT, or Cepstrum. These frequency features will be used as the input features to the transformers-based regression models used for predicting BP. The output of the models can be evaluated with the standard evaluation metrics and the ensemble methods can be used to add variation to the prediction.

3.1 Dataset

To provide diversity in the data used to develop and evaluate the models, the MIMIC III physiological signal database was used in this study [18]. The MIMIC III database consists of high-quality, synchronized PPG, ECG, and invasive arterial BP annotation recordings. MIMIC III is a large-scale critical care dataset collected from ICU patients and contains a very rich and varied set of physiological signals from a wide range of patients. Additionally, the dataset contains continuous recordings of vital signals and reference BP values obtained through arterial lines and, therefore, is well-suited to develop and validate cuffless BP estimation models. By using MIMIC III, the models developed will be generalizable and robust across different clinical environments. The recordings used in this study were accessed via PhysioNet [19].

3.2 Data Processing

In addition, PPG and ECG signals will be utilized for BP estimation due to the different ways they represent cardiovascular activity. For this research, all data were extracted from the MIMIC III database and processed to ensure that they were consistent and could be used in deep learning models. Both PPG and ECG signals were divided into non-overlapping segments of 1000 samples each, which corresponded to 8 seconds of activity at the original sample rate of 125 Hz. The arterial blood pressure (ABP) waveform was similarly segmented ensuring alignment with the chosen input signals. For any ABP segment, systolic (SBP) and diastolic (DBP) blood pressure values were estimated by locating the local peaks and troughs in the waveform. The values typically spanned approximately eight cardiac cycles which were stages and averaged into the final SBP and DBP for that segment. To keep signal range consistent to improve training stability min–max normalization was applied to each PPG and ECG segment separately. Basic artifact filtering was also applied to enhance signal quality during preprocessing when necessary.

3.3 Transformation Techniques

To evaluate the effects of spectral representations on BP estimation utilizing transformer-based architectures, four established transformations were applied to both PPG and ECG signals: STFT, FFT, CWT, and creating a cepstrum with Mel-frequency cepstral coefficients (MFCC). These were chosen to represent a comprehensive range of transformation type, contrasting its adopted usage in the literature on BP and/or heart rate variability, these four techniques span time and frequency-domain to time-frequency domains to invoke some of the possible advantages of these different

common representations (Fig. 2) that each can provide information regarding the structure and dynamics of the cardiovascular signal.

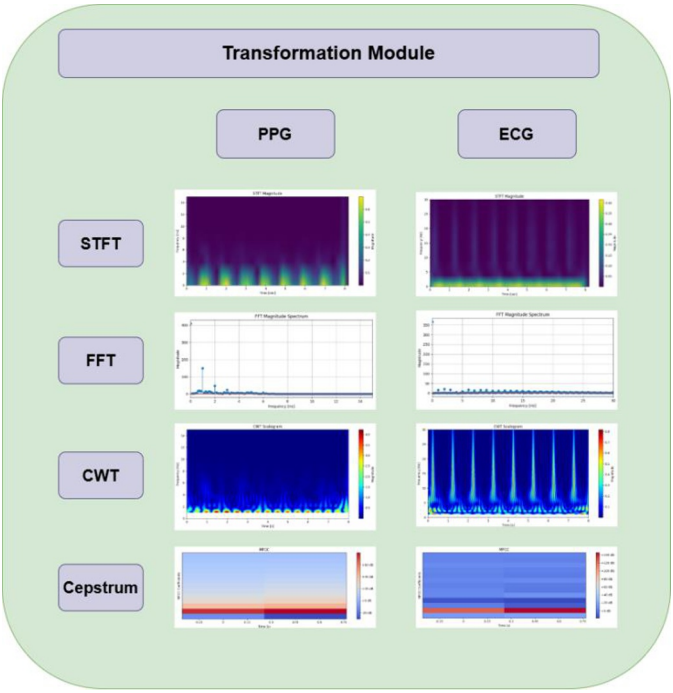

Fig. 2. Visualization of transformation techniques applied to PPG and ECG signals, showing STFT, FFT, CWT, and MFCC representations.

The Short-Time Fourier Transform (STFT) (Equation 1) allows for localized frequency analysis by computing the Fourier transform over short, overlapping windows. It is expressed as:

$$STFT\{x(t)\}(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j\omega\tau}d\tau \qquad (1)$$

Where x(t) denotes the input PPG or ECG signal, $w(\tau - t)$ is the window function, and ω denotes frequency. In the study, a Hamming window of 64 samples (nperseg=64) with 32 sample overlap (noverlap=32) was used, providing a balanced trade-off between time and frequency resolution suitable for the 125 Hz sampling rate. The Short Time Fourier Transform (STFT) is an important tool for detecting time-varying spectral characteristics [13] that characterize non-steady-state physiological signals. The Fast Fourier Transform (FFT) also provides a worldwide frequency domain representation via a method to calculate the Discrete Fourier Transform with great computational efficiency (Equation 2). It is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\left(\frac{2\pi}{N}\right)kn} \qquad (2)$$

The discrete-time PPG or ECG signal is represented by x(n) and was sampled at a frequency of 125 Hz, producing a signal of length N = 1000. The corresponding spectral coefficients of this signal are represented as X(k). To identify the dominant periodic components (heart rate and breathing) in the recorded signal, an FFT was performed on each non-overlapping segment of the signal, although the use of FFT does not provide temporal resolution. Using CWT (Equation [3]), it is possible to generate a multi-resolution time-frequency representation of the signal. It is described by:

$$CWT_{x(t)}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(\tau) \varphi^* \left(\frac{\tau - b}{a}\right) d\tau \qquad (3)$$

Where φ is the mother wavelet, a is the scale, and b is the time shift. CWT was implemented using the Morlet wavelet over scales 1 to 32, allowing dynamic frequency resolution across a 0-30 Hz band. The transformation excels at isolating transient cardiovascular features [11]. Lastly, cepstral analysis was performed using Mel-frequency cepstral coefficients (MFCCs), traditionally employed in speech processing but increasingly applied to biomedical signals. The MFCCs (Equation 4) are derived by applying a discrete cosine transform (DCT) on the log-scaled outputs of a Mel-filtered power spectrum.

$$MFCC_n\big(x(t)\big) = \sum_{m=1}^{M} \log(S_m) \cos\left[\frac{n\pi}{M}(m - 0.5)\right] \qquad (4)$$

Where S_m represents the energy in the mth Mel filter, and n indexes the coefficient. In this study, 13 mel-frequency cepstral coefficients (MFCCs) were extracted from each 1000-sample segment of signals sampled at 125 Hz. MFCCs provide representations that are useful in capturing the periodic constructs of the same signal in the frequency domain [17], giving the frequency spectrum of the signal in consideration of regularity and harmonics. In combination, the MFCCs and the previous transformations provide distinct measurements to characterize the temporal and spectral properties of the cardiovascular signals. These representations are evaluated according to prediction accuracy, computational efficiency, and clinical significance.

3.4 Model Architecture

A transformer-based framework was designed to predict SBP and diastolic DBP blood pressure values from spectral-based data of PPG and ECG signals. The model architecture was established to receive separate spectral-based representations of the input from both modalities, to merge the unique capabilities of these modalities into one suitable neural network for allowing both temporal and spectral properties and information from each modality to proceed simultaneously (Fig. 3).

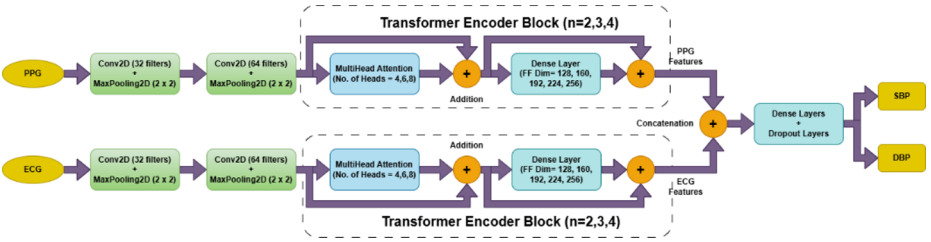

Fig. 3. Transformer-based architecture for SBP and DBP estimation

The overall structure of the model has two independent branches, one for each input modality, where one branch processes PPG signals and the other process ECG signals. The branches use two-dimensional convolutional layers (Conv2D) and a max pooling operation (MaxPooling2D) to extract the spatial features from the spectral inputs. The incorporation of convolutional layers reduces the dimensionality of the inputs while preserving local information. The outputs from the convolutional blocks were then reshaped to a sequence and passed through the transformer layers. However, prior to being input into the transformer layers, the spectral features would allow the model to learn temporal structure.

In addition to using multi-head self-attention and feed-forward networks in the transformer encoder layers, it allowed the model to establish dependencies over longer timeframes and identify complex relationships between the series of spectral features. Furthermore, using multi-head attention facilitated learning multiple relationships from the input sequence at once, increasing the model's ability to detect patterns related to blood pressure dynamics.

Feature fusion was performed by concatenating the outputs from the PPG (photoplethysmography) and ECG (electrocardiography) branches after passing through the transformer encoders. The resulting fused set of features was then input into a fully-connected (dense) layer to combine modality-specific information to create a single representative feature set. Systematic hyperparameter tuning was applied to uncover the most successful architectural combinations. Some of the main hyperparameters that were altered during the tuning process were: {number of attention heads: 4, 6, 8}; {feed-forward neural network size: 128, 160, 192, 224, 256}; {number of transformer encoder blocks: 2, 3, 4}. Numerous architecture configurations were explored through cross-validation to ensure consistent performance across all parts of the dataset.

3.5 Ensemble Strategy

Identify applicable funding agency here. If none, delete this text box.

An approach was then implemented that utilized the "ensemble" strategy to improve the robustness of estimates and reduce the risk of overfitting. The ensemble aggregated the predictions from the three best-featured transformer configurations, according to cross-validated performance metrics for each Transformer variant (STFT, CWT, FFT,

and Cepstrum). Using these models to produce a set of final systolic/dyastolic BP estimates.

Three aggregation strategies were examined to accomplish this ensemble (i.e., simple mean, weighted mean, and geometric mean). Across the validation sets, geometric mean aggregation had better aggregate performance than either of the two other methods. Consequently, the final ensemble approach was determined to be geometric mean aggregation, given that it allowed for the complementary model strengths to produce better BP estimates across subjects, while at the same time, reducing the impact from larger-than-expected value outliers.

3.6 Evaluation Metrics

Identify applicable funding agency here. If none, delete this text box.

The performance of the models was assessed by combining error metrics, correlation metrics and cost metrics. These three evaluations gave a complete analysis of how well a model predicted BP values, and this data has been organised in Table I. All metrics were computed by the authors using standard formulations.

Table.1 Performance Metrics

Metrics	Definition	Formula
Mean Absolute Error (MAE)	Measures the average absolute difference between predicted and reference BP values	$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left\lvert \widehat{\mathrm{BP}_{\mathrm{i}}} - \mathrm{BP}_{\mathrm{i}} \right\rvert$ Where BP_i and $\widehat{BP_i}$represent the reference and predicted SBP or DBP values for the i[th] sample.
Root Mean Square Error (RMSE)	Provides a measure that emphasizes larger errors	$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\widehat{\mathrm{BP}_{\mathrm{i}}} - \mathrm{BP}_{\mathrm{i}}\right)^2}$
Standard Deviation of Error (SD)	Represents the spread of error	$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\left(\widehat{\mathrm{BP}_{\mathrm{i}}} - \mathrm{BP}_{\mathrm{i}}\right) - \mathrm{ME}_{\mathrm{BP}}\right)^2}$ $ME_{BP} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\widehat{\mathrm{BP}_{\mathrm{i}}} - \mathrm{BP}_{\mathrm{i}}\right)$
Concordance Correlation	Assesses agreement between predictions and references	$CCC = \frac{2\rho_{BP}\sigma_{\widehat{BP}}\sigma_{BP}}{\sigma_{BP}^2 + \sigma_{\widehat{BP}}^2 + (\mu_{\widehat{BP}} - \mu_{BP})^2}$

Coefficient (CCC)		Where ρ_{BP} is the Pearson correlation coefficient, μ and σ denote the mean and standard deviation of predicted or reference BP values.
Pearson Correlation Coefficient (ρ_{BP})	Measure linear correlation	$\rho_{BP} = \frac{\sum_{i=1}^{N}(\widehat{BP_i} - \mu_{\widehat{BP}})(BP_i - \mu_{BP})}{(N-1)\sigma_{\widehat{BP}}\sigma_{BP}}$ $R^2 = 1 - \frac{\sum_{i=1}^{N}(BP_i - \widehat{BP_i})^2}{\sum_{i=1}^{N}(BP_i - \mu_{BP})^2}$

Computational Cost: Efficiency was used to assess how efficient it would be to use the proposed model in real time, and this evaluation took into consideration both inference time and memory used by the model. The inference time (in milliseconds) is the average time it takes the model to estimate BP for one sample, while the peak memory consumption (in megabytes) is the maximum amount of memory used to perform the estimate and show that the proposed model will work for real time applications in both wearable and resource constrained systems.

4. Experimental Setup

4.1 Hardware and Software Environment

The experiments were performed using the Kaggle Platform and two NVIDIA Tesla T4 GPUs. The models were built using Python with TensorFlow as the primary deep-learning framework and Keras as the application programmer's interface to build, train, evaluate, etc. the models. Other libraries used include scikit-learn for data splitting and evaluation, SciPy for signal transformations, and additional packages such as NumPy, pickle, and time for data manipulation and performance monitoring. The specific software versions corresponded to Kaggle's standard Python environment at the time of experimentation.

4.2 Model Training Details

The models were trained using a batch size of 32 and the Adam optimizer, with an initial learning rate of 1×10^{-4}. Learning rate (LR) reduction was performed using ReduceLROnPlateau, decreasing the rate by a factor of 0.5 upon plateauing validation loss (patience of 5 epochs; minimum LR 1×10^{-6}. Early stopping was used with a patience of 5 epochs, restoring the best model weights. Dropout regularization was used with a rate of 0.3 in the final combined layers.

A 5-fold cross-validation strategy was adopted. For each fold, 80% of the data was allocated for training and 20% for validation. Splits were generated using scikit-learn's KFold, resulting in random partitions across folds without stratification (Table II).

Table 2. Summary of Experimental Configuration

1 Component	Configuration / Value
Hardware	Kaggle, 2 × NVIDIA Tesla T4 GPUs
Programming language	Python (Kaggle standard)
Frameworks/ Libraries	TensorFlow (Keras), scikit-learn, SciPy, NumPy
Frameworks/ Libraries	TensorFlow (Keras), scikit-learn, SciPy, NumPy
Batch size	32
Optimizer	Adam (1×10^{-4} initial learning rate)
Learning rate scheduler	ReduceLROnPlateau (factor 0.5, patience 5, min lr 1×10^{-6}
Early stopping	Patience 5 epochs, restore best weights
Dropout	0.3 in final layers
Cross-validation	5-fold, 80/20 split per fold, random

5 Results

A comparative evaluation of STFT, CWT, FFT, and Cepstrum was conducted across multiple dataset sizes and transformer configurations.

5.1 Performance Across Spectral Transformations

As shown in Table III, STFT and CWT achieved lower MAE and higher CCC values across dataset sizes compared to FFT and Cepstrum.

Table 3. Comparative performance of spectral transformations for BP estimation across metrics and dataset sizes.

Transform-ation	Records	MAE DBP	MAE SBP	RMSE DBP	RMSE SBP	SD DBP	SD SBP	CCC DBP	CCC SBP	R^2 DBP	R^2 SBP	Pearson DBP	Pearson SBP	Training Time (s)
STFT	37476	4.49	7.70	6.26	10.3	6.26	10.3	0.85	0.88	0.74	0.78	0.86	0.88	385.
T	4	45	63	75	794	75	794	22	53	38	71	25	79	33
STFT	74948	4.38	7.39	6.16	9.93	6.16	9.93	0.85	0.89	0.75	0.80	0.87	0.89	355.
T	8	09	57	61	56	61	56	71	27	54	72	01	87	08
STFT	149896	3.76	6.10	5.25	8.36	5.20	8.36	0.89	0.92	0.82	0.86	0.91	0.92	452.
T	96	23	71	49	34	58	18	63	57	24	35	13	94	4
CW	60008	4.67	7.88	6.76	10.9	6.96	10.9	0.76	0.70	0.65	0.55	0.81	0.75	402.
T	58	39	75	899	61	553	51	3	17	62	65	62	36
CW	100059	4.56	7.87	6.69	10.6	6.89	10.5	0.81	0.76	0.65	0.62	0.81	0.79	470.
T	0	59	97	56	226	39	682	09	01	62	04	95	07	41
CW	150001	4.53	7.79	6.44	10.5	6.64	10.4	0.83	0.81	0.65	0.67	0.81	0.83	581.
T	0	75	01	51	413	18	609	23	91	76	78	61	81	56

	CW	2000	4.32	7.58	6.41	10.2	6.42	10.3	0.84	0.89	0.67	0.69	0.83	0.82	638.
T		0	64	67	86	468	91	996	98	68	02	33	23	34	43
	CW	2500	4.21	7.46	6.21	9.77	6.39	9.60	0.85	0.89	0.68	0.69	0.83	0.83	624.
T		0	77	01	3	85	67	47	62	69	01	36	71	56	2
	FFT	3747	9.49	18.3	12.3	22.5	12.3	22.4	0.00	0.00	0.00	0.00	0.05	0.09	109.
		4	33	551	771	441	757	508	19	43	14	45	93	26	72
	FFT	7494	8.53	16.9	11.1	20.8	11.1	20.7	0.31	0.25	0.20	0.15	0.45	0.39	669.
		8	55	998	459	238	454	774	77	69	39	45	4	93	93
	FFT	1498	7.13	12.9	9.47	16.9	9.43	16.9	0.58	0.60	0.42	0.44	0.65	0.66	895.
		96	45	652	33	248	4	214	29	3	28	11	51	48	64
	Ceps	3747	6.13	10.8	8.41	14.6	8.38	14.5	0.69	0.73	0.53	0.57	0.73	0.76	213.
trum		4	6	316	68	135	32	997	54	8	82	79	64	1	33
	Ceps	7494	5.69	9.98	7.74	13.3	7.73	13.2	0.75	0.78	0.61	0.65	0.78	0.81	293.
trum		8	53	22	55	843	43	688	57	71	56	07	58	05	4
	Ceps	1498	5.06	8.79	7.07	12.0	7.06	12.0	0.80	0.83	0.67	0.71	0.82	0.84	579.
trum		96	13	64	21	994	25	831	02	25	83	44	54	57	75

The chart (Fig. 4) shows how STFT yields lower MAE and SD as sample size increases, while FFT remains significantly higher. CWT and Cepstrum show moderate results with CWT fluctuating more.

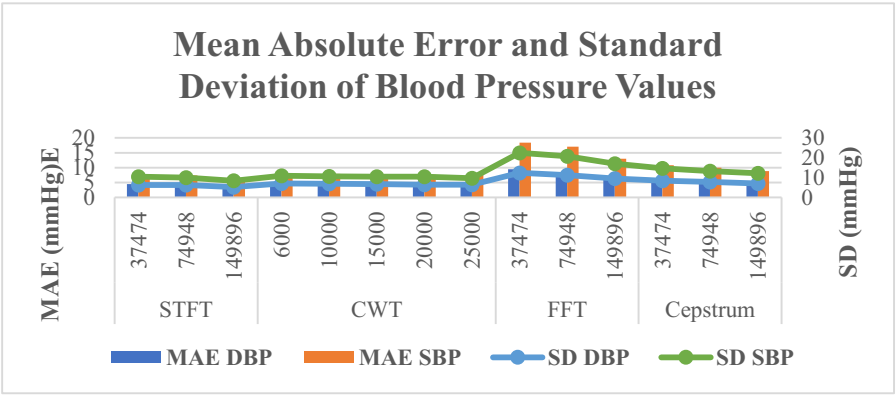

Fig. 4. MAE and SD of BP Predictions Across Transformations and Sample Sizes

As shown in Fig. 5, STFT has a highly correlated result with a manageable training time. CWT will yield a similar R^2 value to STFT; however, the cost of CWT will increase as more and more samples are added to the training set. Cepstrum provides the most efficient training per sample, while FFT is significantly behind STFT in terms of correlation metric performance.

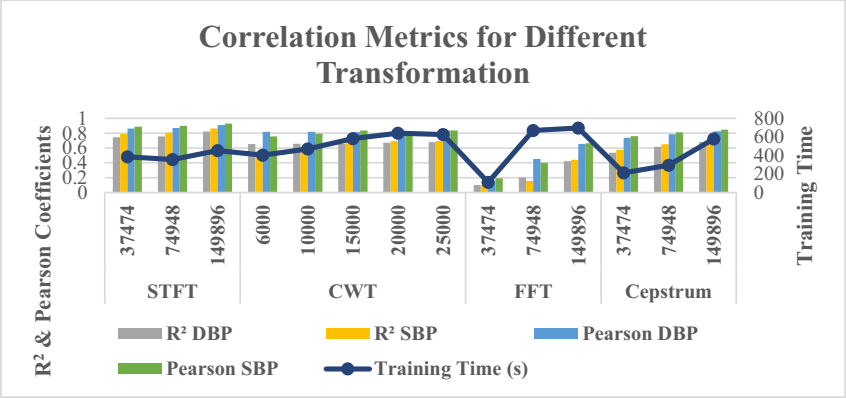

Fig. 5. Correlation Metrics and Training Time across STFT, CWT, FFT, and Cepstrum transformations.

The results of Cepstrum demonstrated consistent and moderate performance, demonstrating that it is a valid option when there are limited resources as a result of lower computational requirements. The results of FFT were inferior when using the overall metrics and sizes tested, with consistently high values for error and low values for the concordance correlation coefficient. Only after testing the largest dataset size did there start to be marginal improvement. The findings shown in the CCC (Fig. 6) further concur with those shown in the MAE and correlation results. STFT performed consistently well in alignment with reference BP, with CWT second and Cepstrum third. Across all sizes, FFT performed poorly.

Overall, STFT demonstrated the most consistent performance across accuracy and computational metrics.

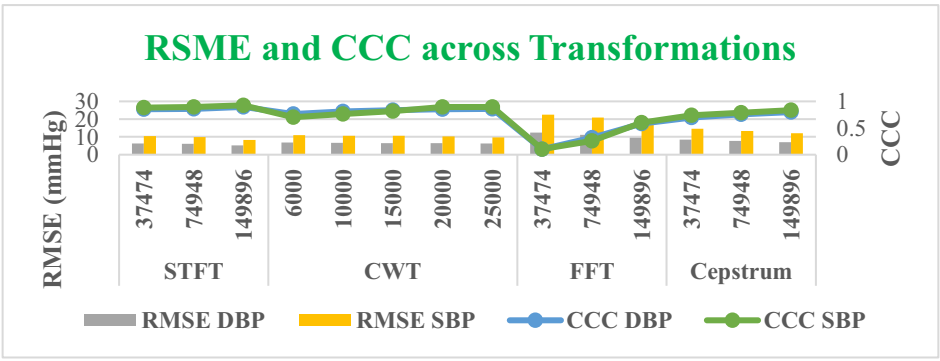

Fig. 6. RMSE and CCC Across Transformations

5.2 STFT-Based Modeling and Selection of Top Models

Based on the performance metrics reported in Table III, STFT was selected for further architectural evaluation. This process utilized 37,474 records that had been converted to STFT for exploring different configurations of transformer models. Configurations of transformer models consisted of changes to number of attention heads, feedforward size, and transformer block structure. The top six configurations found through performance metrics on the validation portion of the dataset are summarized in Table IV.

Table 4. Performance of Top Six STFT-Based Transformer Configurations

Heads / FF / Blocks	Loss Formula	MAE DBP	MAE SBP	SD DBP	SD SBP	CCC DBP	CCC SBP	R^2 DBP	R^2 SBP	Corr DBP	Corr SBP	Training Time (s)
6/192/2	mse + mae + 0.005 × (1.0 − corr)	4.4945	7.706	6.2675	10.3794	0.8522	0.8853	0.7438	0.7871	0.8625	0.8879	3085.33
6/224/3	0.5 × mse + 0.5 × mae + 0.01 × (1.0 − corr)	4.6196	7.638	6.4281	10.1549	0.8378	0.8822	0.7244	0.7936	0.8551	0.8927	3053.19
6/192/3	0.5 × mse + 0.5 × mae + 0.01 × (1.0 − corr)	4.6554	7.721	6.4202	10.4075	0.8483	0.8863	0.7245	0.7847	0.8563	0.8883	3007.64
6/192/4	mse + mae + 0.005 × (1.0 − corr)	4.7155	7.898	6.4569	10.6915	0.8433	0.8729	0.7274	0.7706	0.8534	0.8799	3050.53
6/256/3	mse + mae + 0.005 × (1.0 − corr)	4.7253	7.715	6.4668	10.3252	0.834	0.8788	0.7269	0.7891	0.8544	0.8889	3061.61
6/192/3	0.7 × mse + 0.3 × mae + 0.01 × (1.0 − corr)	4.8162	8.028	6.5297	10.5431	0.8396	0.8751	0.7183	0.7725	0.8499	0.8835	310.7

5.3 Ensemble Performance of STFT Models

To enhance prediction stability and reduce variance, multiple ensemble aggregation strategies were evaluated using the three best-performing STFT transformer configurations. The investigated aggregation methods included weighted geometric mean, harmonic mean, arithmetic mean, weighted mean, and median. Each approach was assessed based on MAE, CCC, and R^2 metrics. Among the evaluated strategies, the weighted geometric mean achieved the lowest MAE values and the highest correlation metrics across both SBP and DBP (Table V), indicating superior overall ensemble performance compared to the other aggregation methods.

Table 5. Ensemble Performance of Top STFT Models

Aggregation Method	MAE DBP	MAE SBP	SD DBP	SD SBP	CCC DBP	CCC SBP	R^2 DBP	R^2 SBP	Corr DBP	Corr SBP
Weighted Geometric Mean	3.9056	4.9707	6.9633	7.8837	0.922	0.9391	0.8655	0.8884	0.9302	0.9453
Weighted Mean	4.117	5.7665	5.8211	9.1148	0.8866	0.9158	0.8121	0.8492	0.8991	0.9258
Median	4.0918	6.051	5.7378	8.981	0.8889	0.9186	0.8187	0.8546	0.9012	0.9279
Arithmetic Mean	4.1296	6.5336	5.8453	9.3652	0.8826	0.9093	0.8079	0.8381	0.8945	0.9201
Harmonic Mean	4.2375	6.9589	5.927	9.3672	0.8801	0.9116	0.8013	0.8367	0.891	0.918

Ensemble aggregation improved prediction stability and reduced variance compared to individual models.

5.4 Benchmarking Against Existing Methods

To better illustrate how well the STFT-based transformer architectures have performed, a comparison against similar studies in the literature that have used other transformation techniques and methods was conducted. Table VI shows that there were significantly lower MAE and SD values for both of the BP ranges when using the STFT-based transformer architecture, along with being compliant with the AAMI clinical standards (i.e., MAE < 5 mmHg and SD < 8 mmHg). The proposed STFT-based ensemble achieved MAE < 5 mmHg and SD < 8 mmHg for both SBP and DBP, satisfying AAMI clinical criteria.

Table 6. Comparative performance of the proposed method against prior studies.

Study	Transform-ation	Domain	Model Type	MAE (mmHg)	ME (mmHg)	SD (mmHg)	Remarks
Kachuee et al. [8]	PAT, PTT	Time	CNN + SVR	DBP=7.45 SBP=11.2		—	Requires calibration

Xing et al. [9]	FFT	Frequency	ANN		DBP=0.01 SBP=0.06	DBP=4.66 SBP=7.08	Loss of temporal resolution
Omer et al. [11]	Wavelet Scattering	Time-Frequency	LSTM	DBP=5.09 SBP=10.8		—	Robust to signal distortion
Ours (STFT+Ensemble)	STFT	Time-Frequency	Transformer (Ensemble)	DBP=3.91 SBP=4.97		DBP=6.96 SBP=7.88	Meets AAMI criteria

6 DISCUSSION

Although CWT shows competitive accuracy, especially when using smaller datasets, the use of CWT results in significant computation overhead. When using very large datasets, particularly with over 15,000 records on the Kaggle free-tier GPU, the high dimensionality of the CWT-transformed inputs caused memory issues that inhibited the ability to grow the model and resulted in incomplete training of the models on the full training set. In contrast, STFT demonstrated an optimal balance of predictive accuracy and training efficiency across all evaluated sizes of datasets, making STFT the most scalable and reliable transformation.

The achieved error margins remain within AAMI-recommended thresholds, supporting the clinical feasibility of the proposed framework. These outcomes further demonstrate that STFT-based spectral representations can serve as reliable predictors of clinically-significant, accurate blood-pressure values when applied within deep-learning models. The outcome measure also allowed for a meaningful discussion regarding the interactions between various spectral transformation techniques, model-development strategies, and deep-learning-based blood-pressure prediction performance. These topics are outlined below:

STFT demonstrated the most consistent performance across all considered metrics, including MAE, CCC, and R^2, and at all data scales. The ability of STFT to capture both time and frequency components of cardiovascular signals made it a particularly effective feature representation for transformer-based models. In addition to providing high levels of prediction accuracy, the STFT has also been associated with faster training times, making them ideal for large scale learning, even in hardware-constrained environments like free-tier services.

Model performance improved with increasing data size, affirming the benefits of learning from more complex data. As the dataset increased in size from approximately 37k samples to approximately 150k samples, both STFT based models and Cepstrum based models demonstrated consistent decreases in mean absolute error (MAE) and increases in concordance correlation coefficient (CCC). This trend indicates improved generalization with increasing dataset size.

Ensemble methods improved stability and robustness of predictions, particularly when combining the six best STFT transformer models. Among the tested aggregation schemes, the weighted geometric mean achieved the best overall error and correlation metrics. In addition to reducing variance between their predictions; ensembles had high agreement with their ground truth values.

Cepstrum provided moderate accuracy with lower computational cost, making it suitable for resource-constrained settings. Although the cepstrum did not perform as well as the STFT; it was still an attractive compromise between efficiency and accuracy, especially in cases where there were limitations in memory or processing power.

The FFT method did not perform well on all sample sizes, because it had a high error rate and low CCC value, particularly when using smaller sample sizes. Also, while there was more data for modeling, there was not much improvement, indicating that the FFT method does not do as well with the dynamic and non-stationary characteristics of physiological signals, like PPG.

Although CWT achieved competitive accuracy, its high memory demand created bottlenecks as dataset size increased. The high output dimensionality often resulted in out of memory errors when using standard GPUs and limited the scalability of the CWT method. The CWT method appeared to perform well with smaller datasets, but the computational limitations will ultimately prevent it from being used for larger applications.

Overall, these findings emphasize the importance of selecting spectral representations that balance predictive accuracy and computational feasibility.

7 Conclusion

The primary objective of this research was to assess the different spectral transformation techniques (STFT, CWT, FFT, and Cepstrum) for transformer-based BP estimation using PPG data. A rigorous evaluation across the various dataset sizes and all model configurations showed that the Short-Time Fourier Transform (STFT) consistently performed better than the other spectral transformation techniques, allowing the best performance with a good trade-off between accuracy and computational cost.

When using the STFT features as the input into the transformer architectures and developing the final ensemble models from the weighted geometric mean of the top six performing configurations, the performances achieved an overall MAE absolute error of 3.91 mmHg for DBP and 4.97 mmHg for SBP. The corresponding standard deviations were 6.96 mmHg (DBP) and 7.88 mmHg (SBP), with Pearson correlation coefficients of 0.93 and 0.95, respectively.

These findings demonstrate strong generalisation and are also within the clinically acceptable ranges recommended by AAMI, which recommends less than 5 mmHg mean absolute error, and less than 8 mmHg for standard deviation for SBP and DBP.

The results highlight the importance of selecting appropriate spectral representations for transformer-based BP estimation. These results support the use of STFT-based features in transformer models for cuffless BP estimation.

Conflict of Interest: The authors declare that they have no conflict of interest.

Funding Declaration: This research did not receive any specific grant from funding agencies in the public, commercial or not-for profit sectors.

Ethics Declaration: Not applicable.

Clinical Trials: Not applicable.

Data Availability Statement: Data will be available upon reasonable request.

References

1. World Health Organization, "Hypertension," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension
2. R. Mukkamala *et al.*, "Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 8, pp. 1879–1901, Aug. 2015.
3. G. Slapničar, N. Mlakar, and M. Luštrek, "Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network," *Sensors*, vol. 19, no. 15, p. 3420, 2019.
4. Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 8, pp. 1902–1910, Aug. 2015.
5. Z. B. Zhou *et al.*, "Wearable continuous blood pressure monitoring devices based on pulse wave transit time and pulse arrival time: A review," *Materials*, vol. 16, no. 6, p. 2133, 2023.
6. N. Jean Effil and R. Rajeswari, "Wavelet scattering transform and long short-term memory network-based noninvasive blood pressure estimation from photoplethysmograph signals," *Signal Image Video Process.*, vol. 16, no. 1, pp. 1–9, 2022.
7. A. Arjomand *et al.*, "TransfoRhythm: A transformer architecture conductive to blood pressure estimation via solo PPG signal capturing," 2024. [Online]. Available: arXiv:2404.15352.
8. M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuffless blood pressure estimation algorithms for continuous health-care monitoring," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 4, pp. 859–869, Apr. 2017.
9. X. Xing and M. Sun, "Optical blood pressure estimation with photoplethysmography and FFT-based neural networks," *Biomed. Opt. Express*, vol. 7, no. 8, pp. 3007–3020, 2016.
10. M. W. Nawaz *et al.*, "Cuff-less arterial blood pressure waveform synthesis from single-site PPG using transformer and frequency-domain learning," 2024. [Online]. Available: arXiv:2401.05452.
11. O. A. Omer *et al.*, "Blood pressure estimation from photoplethysmography using hybrid scattering–LSTM networks," *BioMedInformatics*, vol. 4, no. 1, pp. 139–157, 2024.
12. J. Wu *et al.*, "Improving the accuracy in classification of blood pressure from photoplethysmography using continuous wavelet transform and deep learning," *Int. J. Hypertens.*, vol. 2021, Art. no. 9938584, 2021.
13. H. Tjahjadi, K. Ramli, and H. Murfi, "Noninvasive classification of blood pressure based on photoplethysmography signals using bidirectional long short-term memory and time-frequency analysis," *IEEE Access*, vol. 8, pp. 20735–20748, 2020.
14. N. Nasir *et al.*, "Deep learning classification of photoplethysmogram signal for hypertension levels," 2024. [Online]. Available: arXiv:2405.14556.
15. J. Chen *et al.*, "Signal quality assessment of PPG signals using STFT time-frequency spectra and deep learning approaches," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2021, pp. 1153–1156.
16. A. I. Siam *et al.*, "PPG-based human identification using Mel-frequency cepstral coefficients and neural networks," *Multimed. Tools Appl.*, vol. 80, no. 17, pp. 26001–26019, 2021.

17. A. Prabha *et al.*, "Intelligent estimation of blood glucose level using wristband PPG signal and physiological parameters," *Biomed. Signal Process. Control*, vol. 78, p. 103876, 2022.
18. A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, p. 160035, 2016.
19. A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.