

Leveraging Speaker Verification Models for PPG-Based Biometric Verification

Arthur Friler, Noah Lewkowicz, and Yair Moshe

Signal and Image Processing Laboratory (SIPL)

Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering

Technion – Israel Institute of Technology, Haifa, Israel

<https://sipl.ece.technion.ac.il>

Abstract—Photoplethysmogram (PPG) signals are a compelling biometric modality due to their physiological uniqueness and ubiquity in wearable technology. This study introduces a novel cross-domain framework for PPG-based biometric authentication by adapting architectures originally developed for automated speaker verification. Recognizing that both speech and PPG signals are quasi-periodic and shaped by individual physiological traits, we demonstrate that speaker verification models are exceptionally well-suited to capturing the subtle, dynamic temporal patterns inherent in cardiac cycles. Our methodology involves a preprocessing pipeline that aligns raw PPG data with the structural characteristics of acoustic signals, enabling the direct application of speaker verification models. Notably, this approach yields a highly generalizable framework that maintains robust performance across different physiological conditions and hardware configurations. Experimental results across multiple datasets demonstrate state-of-the-art performance, outperforming previously reported results in the literature.

Index Terms—Biometric authentication, biometrics, human verification, PPG (photoplethysmography), speaker recognition

I. INTRODUCTION

The objective of authentication is to verify the identity of an individual to safeguard access to sensitive information and digital infrastructures. Traditional security frameworks primarily rely on knowledge-based or possession-based factors; however, these approaches are increasingly compromised by credential theft, social engineering, and unauthorized sharing. These security gaps have motivated the exploration of biometric authentication, which leverages unique human physiological or behavioral characteristics to provide a more direct link between a user and their digital identity. Traditional modalities like face or iris recognition are vulnerable to presentation attacks and data leakage. In contrast, intrinsic physiological signals like cardiac activity are harder to replicate and provide inherent liveness detection as products of active biological functions [1], [2].

Photoplethysmography (PPG) is a non-invasive optical technique that measures blood volume changes in the microvascular bed of tissue by capturing the intensity of light reflected or transmitted to a photodetector. This process captures a pulsatile waveform generated by the cardiac cycle, the morphology and temporal behavior of which are shaped by a unique combination of physiological and anatomical factors, including vascular structure, arterial stiffness, blood volume distribution,

and cardiac dynamics. Because these cardiovascular characteristics are influenced by individual-specific traits such as vessel compliance, heart anatomy, and autonomic nervous system regulation, PPG signals serve as a robust biometric modality for identity verification. Furthermore, PPG-based authentication is particularly compelling due to its integration into the low-cost sensors already ubiquitous in consumer wearables, enabling a passive and continuous security layer that operates seamlessly without requiring explicit user interaction.

Prior work on PPG-based biometric verification has demonstrated encouraging results. Initial methodologies primarily focused on the extraction of handcrafted features, including fiducial points [3], [4], the autocorrelation of first and second PPG derivatives [5], Discrete Wavelet Transform (DWT) coefficients [6], and Mel Frequency Cepstral Coefficients (MFCC) [7]. These feature sets were typically paired with classical machine learning classifiers to establish identity. More recent approaches have shifted toward deep learning paradigms, employing convolutional architectures to learn identity-discriminative representations from either raw or minimally pre-processed PPG signals. A prevalent architectural configuration integrates Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) layers [8]–[10] or with attention mechanisms [11] for capturing temporal dependencies.

Despite these advancements, PPG-based authentication faces substantial obstacles to real-world adoption. PPG signals are characterized by high intra-subject variability, frequently exacerbated by motion artifacts, fluctuating physiological or emotional states, and inconsistencies in sensor placement. Moreover, the limited availability of comprehensive datasets limits the generalization capability of current deep learning models. Consequently, most current methodologies rely on controlled experimental conditions and require retraining for unseen subjects, thereby limiting the practical scalability of these systems [2]. Recent research has begun to address these constraints regarding enrollment efficiency and architectural robustness. For instance, Fusion-ID [12] utilizes a few-shot learning framework that facilitates user onboarding without requiring model retraining, needing only a few seconds of sample data to enroll new users. However, its reliance on a private dataset complicates comparative performance benchmarking. PulseID [13] introduces a multi-scale CNN. By maintaining

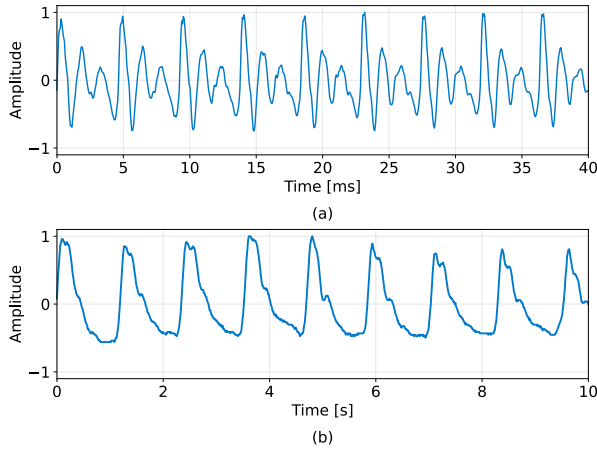


Fig. 1: Comparison of signal morphologies between (a) a voice waveform and (b) a PPG waveform. While both signals exhibit quasi-stationary periodic oscillation structures, they operate at distinct orders of magnitude in frequency.

high verification accuracy across several datasets, PulseID demonstrates significant progress toward scalable, hardware-agnostic PPG authentication.

In this work, we propose a novel framework that repurposes state-of-the-art architectures originally designed for speaker verification for PPG-based verification. This shift is motivated by the fundamental similarities between speech and PPG signals and by the fact that speaker verification models are significantly more mature and trained on massive diverse datasets. Our methodology yields a highly generalizable framework that extracts robust, identity-discriminative embeddings capable of operating across different PPG datasets and sensor configurations.

II. SPEAKER VERIFICATION FOR PPG-BASED VERIFICATION

Despite their disparate physiological origins, human voice and cardiac pulse signals share profound physical and morphological characteristics, as both may be modeled as outputs of a biological source-filter system driven by pressure-induced oscillation. In phonation, subglottal air pressure induces vocal fold oscillation to produce a source signal characterized by a fundamental frequency (F_0), or pitch, typically operating within ≈ 85 -255 Hz. Similarly, in PPG, arterial pressure fluctuations drive the rhythmic expansion and contraction of vessel walls, generating a quasi-periodic source operating at ≈ 1 -3 Hz. Although this operates two orders of magnitude below speech, it remains fundamentally linked by its underlying oscillatory mechanism. From a signal processing perspective, biometric uniqueness is encoded through the interaction of neural source control and the specific geometry of the biological filter. Just as the vocal tract's configuration defines resonance formants ($F_1, F_2, \dots, F_n$) that serve as an acoustic fingerprint, the microvascular branching and arterial elasticity shape the PPG waveform's filter characteristics. Within this framework, the high-energy systolic peak provides

a functional mapping to the first formant (F_1), while more subtle morphological nuances, such as the diastolic notch and diastolic peaks, provide the "timbre" analogous to higher-order formants (F_2, F_3) required for high-precision identity discrimination. Representative examples of these waveforms, illustrating these shared structural properties across disparate frequency domains, are shown in Figure 1.

Speaker verification aims to determine whether two audio samples originate from the same individual. Speaker verification systems rely on speaker embeddings, which are high-dimensional features representing a speaker's vocal tract characteristics and speaking style. Once created, these embeddings can be compared to determine if two audio samples come from the same person. A primary challenge in this field is the development of robust embeddings that remain consistent for a single speaker despite changes in background noise, emotional variance, or channel distortion, while being clearly distinct from other speakers [14]. Over the last decade, the superior feature extraction capabilities of Deep Neural Networks (DNNs) have established them as the dominant paradigm for speaker embedding. One of the foundational successes in speaker verification is the Time Delay Neural Network (TDNN) architecture, which processes input Mel-frequency cepstral coefficients (MFCCs) to produce robust embeddings known as x-vectors. A prominent state-of-the-art evolution of this approach is the Emphasized Channel Attention, Propagation, and Aggregation TDNN (ECAPA-TDNN) [15], which integrates 1D-Res2Net modules and Squeeze-and-Excitation (SE) layers to enhance multiscale temporal modeling and channel-wise feature recalibration. Parallel to these TDNN advancements, TitaNet [16] has emerged as a high-performance architecture. It utilizes a backbone of 1D depthwise separable convolutions to achieve a large receptive field with significantly lower computational complexity than standard convolutions. By incorporating global context through SE blocks and employing attentive statistics pooling, TitaNet effectively captures long-range dependencies and aggregates frame-level features into highly discriminative, fixed-dimensional embeddings.

Given the morphological parallels between PPG and voice signals, the methodologies established in speaker verification may be uniquely suited to capturing identity-relevant features from PPG signals, and vice versa. However, while this relationship is bidirectional, the speaker verification domain currently offers a superior level of maturity, marked by a large body of work, extensive datasets and proven generalizable architectures providing a valuable computational framework that can be exploited to advance PPG-based biometric identification. For example, TitaNet's specialized 1D convolutional structure and global context modeling allow it to robustly capture subtle physiological patterns of PPG signals, making it a powerful tool for identity verification on cardiac biometrics.

III. PROPOSED SOLUTION

In light of the motivation described in the previous section, the proposed biometric authentication framework adapts a

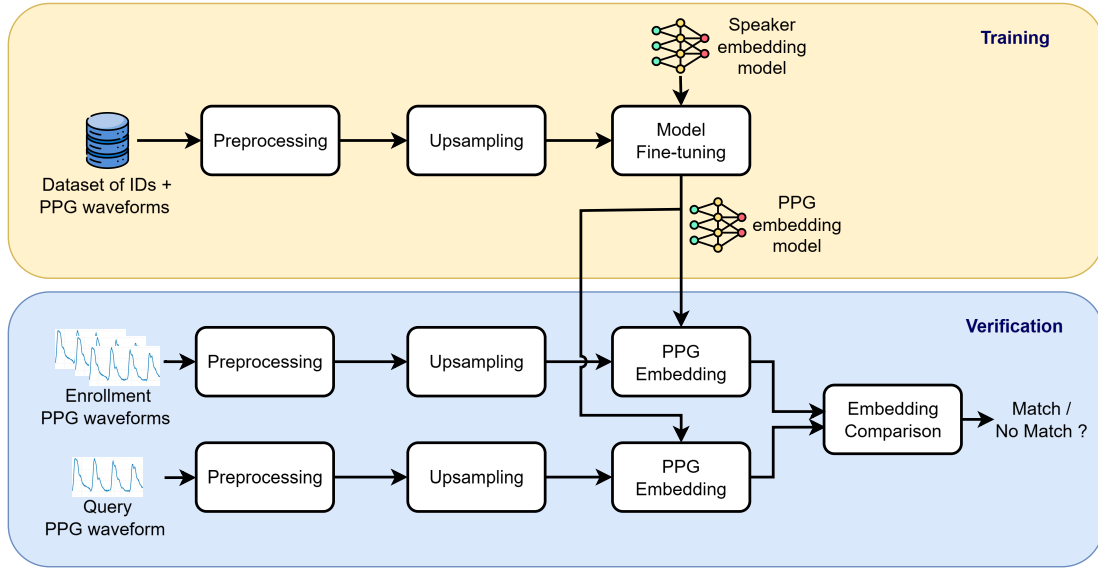


Fig. 2: Scheme of the proposed PPG-based human verification.

speaker embedding model, originally trained on a large-scale dataset of speech signals, for the task of individual identification via PPG signals. While such models are designed for vocal characteristics, previous research has demonstrated that cross-modal transfer learning is highly beneficial for improving performance in domains with limited data availability [17]. Consequently, we perform fine-tuning on a comprehensive dataset of PPG waveforms and subject IDs to specialize the embedding space for cardiac biometrics.

As shown in Figure 2, the architecture consists of a training phase and a verification phase. During training, the model is fine-tuned using a dataset where PPG waveforms are coupled with their corresponding person IDs, facilitating the transformation of the acoustic embedding architecture into a specialized PPG embedding model. We employ TitaNet-L [16], a high-capacity model featuring 25.3 million parameters originally trained on 3,373 hours of speech across 16,681 speakers, as the core acoustic embedding backbone. Both the original TitaNet-L and the resulting fine-tuned model produce an embedding represented as a vector of length 192. The backbone is fine-tuned on VitalDB [18], a large-scale open-access dataset comprising vital signs collected during surgical procedures. VitalDB contains recordings from 6,388 patients with continuous multi-hour PPG signals recorded at a high temporal resolution. However, as VitalDB contains segments where heavy noise often masks the underlying physiological properties of the signal, we implemented a noise-filtering protocol. Following this filtration, 5,107 distinct subjects remained. The fine-tuning process was subsequently performed using 30-second signal segments extracted exclusively from these remaining subjects. This dataset is significantly more extensive than those typically used in prior PPG-based biometric studies, enabling the development of a robust, generalized embedding model through large-scale transfer learning. The model is

fine-tuned as a supervised classification task using subject-disjoint splits to prevent identity leakage. Post-convergence, the classification head is removed, and the penultimate layer serves as a fixed embedding extractor.

During the verification stage, the proposed framework operates within either a one-shot or few-shot learning paradigm. Unlike most other approaches that require a fixed set of subjects and costly retraining to add new users, our method utilizes the fine-tuned model as a generalized feature extractor. Both enrollment and query waveforms are processed through the model to extract high-dimensional feature vectors, where the individuals being verified were entirely absent from the training phase. Unlike traditional closed-set classification protocols that require a fixed subject pool and costly retraining to onboard new users, our method is inherently scalable and capable of generalizing to an open population without additional parameter updates. This evaluation framework directly reflects a deployment-oriented verification pipeline, proving particularly effective when retraining is computationally impractical, enrollment time must be minimized for user convenience, or data access is limited. However, it should be noted that while maintaining scalability is advantageous, fine-tuning the model on a specific dataset typically yields superior performance by enhancing the discriminative power within the specific biometric embedding space.

The verification process quantifies the similarity between the query PPG embedding and each enrollment embedding using cosine distance. This metric is particularly well-suited for PPG-based biometrics as it prioritizes the angular orientation of feature vectors over their magnitude, offering inherent robustness against signal amplitude fluctuations caused by varying sensor pressure or physiological differences in skin tone. Following the computation of pairwise distances, a binary decision (match or non-match) is produced for each comparison based on a predefined threshold. In a configuration

involving multiple enrollment samples, the system selects the enrollment-query pair with the minimum distance (maximum similarity) before applying the threshold to determine the final identity verification.

Since PPG signals are inherently susceptible to motion artifacts and environmental interference, a rigorous pre-processing pipeline is essential to ensure that the extracted embeddings reflect physiological identity rather than noise. Signals are segmented and Z-score normalized to ensure invariance to sensor gain or contact pressure. A 4th-order Butterworth bandpass filter (0.5 – 32.5 Hz) is then applied to isolate the cardiac cycle from electronic noise and respiratory wander. Additionally, for the VitalDB dataset only, a two-stage quality assurance protocol is implemented to prune "noisy" subjects. In the first stage, the auto-correlation of the signal is checked against a predefined threshold to evaluate its periodicity. This stage is required to verify that the signal captures a rhythmic pulse rather than motion-induced artifacts. In the second stage, the remaining segments are validated against physiological constraints by ensuring the instantaneous heart rate remains between 50 and 200 BPM.

The integration of PPG signals into an audio-centric architecture is achieved through temporal upsampling to 16 kHz, bridging the domain gap between physiological and acoustic signals. This upsampling transposes the cardiac cycle into the audible frequency range, enabling the pre-trained backbone to extract identity-bearing features while maintaining strict compatibility with state-of-the-art speaker verification input requirements. By standardizing the input to 16 kHz, the framework treats the PPG segment as a high-resolution 1D temporal acoustic-like waveform, leveraging the model's established capacity for sequence modeling and latent feature discrimination to perform robust biometric identification.

IV. RESULTS

To evaluate the generalizability of the proposed approach, we conducted experiments across three distinct datasets: CapnoBase [19], BIDMC [20], and BioSec2 [9]. These datasets provide a diverse range of physiological conditions and acquisition parameters. CapnoBase comprises 8-minute recordings from 42 subjects sampled at 300 Hz. It includes 23 cases of spontaneous breathing and 19 cases of controlled ventilation, providing a robust test for algorithm performance across different respiratory patterns. The BIDMC dataset focuses on clinical environments, containing 53 recordings from adult intensive care unit patients. Each 8-minute segment in this dataset is sampled at 125 Hz. Finally, BioSec2 represents the largest cohort in this study, consisting of PPG signals from 100 subjects (48 males, 52 females) captured in relaxed states. BioSec2 captures longitudinal variability via two separate sessions per subject. Each session includes three individual trials, 1.5–1.9 minutes each, sampled at 100 Hz from the fingertip using infrared pulse sensors.

To ensure a fair comparison with previous work, we evaluated our approach by retraining the model independently for each dataset. This aligns with established protocols where

TABLE I: Performance comparison across three datasets (adapted from [13]). Best results are highlighted in **bold**, while second-best results are underlined.

Dataset	Method	ACC (%) $\uparrow$	EER (%) $\downarrow$
CapnoBase [19]	Ye2021 [8]	94.20	4.10
	Walia2019 [5]	57.30	8.80
	Siam2021 [7]	90.00	2.30
	Karimian2017 [6]	84.10	3.40
	Wang2022 [11]	<u>99.20</u>	<u>0.60</u>
	Hwang2020 [9]	95.50	2.90
	PulseID2024 [13]	97.20	3.10
	Ours	99.93	0.04
BIDMC [20]	Ye2021 [8]	78.90	9.50
	Walia2019 [5]	63.50	10.30
	Siam2021 [7]	70.30	7.90
	Karimian2017 [6]	69.60	9.10
	Zhang2023 [4]	97.20	2.50
	Wang2022 [11]	96.50	<u>2.10</u>
	Hwang2020 [9]	93.70	4.80
	PulseID2024 [13]	97.80	2.80
	Ours	98.43	1.75
BioSec2 [9]	Ye2021 [8]	36.20	23.50
	Walia2019 [5]	12.70	23.40
	Siam2021 [7]	25.90	14.00
	Karimian2017 [6]	21.50	13.50
	Wang2022 [11]	69.30	5.80
	Hwang2020 [9]	87.10	12.90
	PulseID2024 [13]	<u>94.30</u>	2.90
	Ours	96.64	<u>3.36</u>

models are trained and evaluated on the same dataset. For each dataset, subjects were partitioned into training (60%), validation (20%), and test (20%) sets. For the subjects in the test set, we employed a multi-embedding strategy to better capture intra-subject variability. For a given subject S , the available PPG recordings are partitioned into two disjoint subsets: an enrollment set E_S and a query set Q_S . The enrollment set consists of $K \in \{1, 5, 10\}$ PPG segments randomly sampled from that subject's recordings, while the remaining segments form the query set and are used for verification. It has been demonstrated empirically that PPG signals change over time [9]. The multi-embedding configurations (where $K > 1$) account for such variability by populating a more comprehensive embedding space, thereby capturing a broader representation of the subject's physiological characteristics.

Each PPG segment undergoes the preprocessing pipeline detailed in Section III. We adopted segment lengths of 4.8 seconds for BIDMC and 10 seconds for CapnoBase and BioSec2. These durations were selected to align with established existing literature and to accommodate the distinct physiological properties inherent to each dataset.

System performance is evaluated using two standard biometric metrics: Verification Accuracy (ACC) and the Equal Error Rate (EER). The EER represents the operational point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equivalent. A lower EER indicates superior system discriminability. To ensure a robust evaluation independent of fixed decision thresholds or dataset-specific biases, the threshold is systematically varied across the full range of

TABLE II: Multi-embedding verification performance on the CapnoBase dataset [19] with different number of embeddings (K). Best results are highlighted in **bold**.

K	ACC (%) $\uparrow$	EER (%) $\downarrow$
1	99.53	0.47
5	99.85	0.17
10	99.93	0.04

similarity scores. These metrics facilitate a direct comparison with state-of-the-art PPG-based authentication methods, as detailed in Table I. The experimental results demonstrate that our proposed method consistently outperforms prior work across diverse datasets. Notably, on the CapnoBase dataset, our model achieves a near-perfect accuracy of 99.93% and a remarkably low EER of 0.04%, significantly surpassing the previous best EER of 0.60% held by [11]. Similar trends are observed in the BIDMC and BioSec2 datasets, where our approach maintains the highest verification accuracy. While [13] shows a slightly lower EER on the BioSec2 dataset, our method remains the most robust overall, providing the highest classification accuracy (96.64%) and establishing a new benchmark for high-precision PPG authentication.

The impact of the multi-embedding strategy on system performance is further detailed in Table II, using the CapnoBase dataset [19]. The results demonstrate a clear positive correlation between the number of enrollment segments (K) and verification accuracy. This trend validates the hypothesis that utilizing multiple segments allows the model to better account for intra-subject physiological variability by populating a more comprehensive and representative embedding space for each individual.

V. CONCLUSIONS

This paper demonstrates that architectures originally developed for automated speaker verification can be effectively repurposed for PPG-based biometric authentication. By up-sampling PPG signals to the audible frequency range, we successfully bridged the domain gap between physiological and acoustic signals, enabling the use of high-capacity models like TitaNet-L [16]. This cross-domain transfer learning approach overcomes the limitations of small datasets and less mature architectures, achieving state-of-the-art results. These findings establish a robust, scalable framework that leverages mature acoustic modeling techniques to advance the precision of cardiac-based identity verification.

ACKNOWLEDGMENT

The authors would like to thank Itay Katsnelson and Idan Oren for their initial contributions, which were instrumental in the development of this work.

REFERENCES

- [1] L. Li, C. Chen, L. Pan, L. Y. Zhang, Z. Wang, J. Zhang, and Y. Xiang, "A survey of ppg's application in authentication," *Computers & Security*, vol. 135, p. 103488, 2023.
- [2] D. Aguilar, A. Martínez-Cruz, K. A. Ramírez-Gutierrez, and M. Morales-Sandoval, "Ppg-based biometric authentication: A review on architectures, datasets, attacks and security challenges," *Computational and Structural Biotechnology Journal*, vol. 28, pp. 511–528, 2025.
- [3] T. Zhao, Y. Wang, J. Liu, Y. Chen, J. Cheng, and J. Yu, "Trueheart: Continuous authentication on wrist-worn wearables using ppg-based biometrics," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 30–39.
- [4] L. Zhang, A. Li, S. Chen, W. Ren, and K.-K. R. Choo, "A secure, flexible, and ppg-based biometric scheme for healthy iot using homomorphic random forest," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 612–622, 2023.
- [5] A. Walia and A. Kaul, "Human recognition via ppg signal using temporal correlation," in *2019 5th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2019, pp. 144–147.
- [6] N. Karimian, Z. Guo, M. Tehranipoor, and D. Forte, "Human recognition from photoplethysmography (ppg) based on non-fiducial features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4636–4640.
- [7] A. I. Siam, A. A. Elazm, N. A. El-Bahnasawy, G. M. El Banby, and F. E. Abd El-Samie, "Ppg-based human identification using mel-frequency cepstral coefficients and neural networks," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26 001–26 019, 2021.
- [8] Y. Ye, G. Xiong, Z. Wan, T. Pan, and Z. Huang, "Ppg-based biometric identification: Discovering and identifying a new user," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1145–1148.
- [9] D.-Y. Hwang, B. Taha, D. S. Lee, and D. Hatzinakos, "Evaluation of the time stability and uniqueness in ppg-based biometric system," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 116–130, 2020.
- [10] D. Biswas, L. Everson, M. Liu, M. Panwar, B.-E. Verhoef, S. Patki, C. H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg *et al.*, "Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 2, pp. 282–291, 2019.
- [11] D. Wang, Q. Hu, and C. Yang, "Biometric recognition based on scalable end-to-end convolutional neural network using photoplethysmography: A comparative study," *Computers in Biology and Medicine*, vol. 147, p. 105654, 2022.
- [12] H. Kumar, H. S. Mousavi, and B. Shahsavari, "Fusion-id: A photoplethysmography and motion sensor fusion biometric authenticator with few-shot on-boarding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3983–3987.
- [13] R. Wei, X. Xu, Y. Li, Y. Zhang, J. Wang, and H. Chen, "Pulseid: Multi-scale photoplethysmographic identification using a deep convolutional neural network," *Biomedical Signal Processing and Control*, vol. 88, p. 105609, 2024.
- [14] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.
- [15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [16] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.
- [17] A. Bizzego, G. Gabrieli, and G. Esposito, "Deep neural networks and transfer learning on a multivariate physiological signal dataset," *Bioengineering*, vol. 8, no. 3, p. 35, 2021.
- [18] H.-C. Lee, Y. Park, S. Yoon, S.-M. Yang, D. Park, and C.-W. Jung, "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients," *Scientific Data*, vol. 9, no. 1, p. 279, 2022.
- [19] W. Karlen, M. Turner, E. Cooke, G. Dumont, and J. M. Ansermino, "CapnoBase: Signal database and tools to collect, share and annotate respiratory signals," in *2010 Annual meeting of the society for technology in anesthesia*. Society for Technology in Anesthesia, 2010, p. 27.
- [20] M. A. Pimentel, A. E. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2016.