Fusion of Handcrafted and Deep-Learned Cardiorespiratory Features for Breathing Pattern Classification

Manuel Lage Cañellas
CMVS, Infotech Oulu
University of Oulu, Finland
0000-0002-4917-340X

Constantino Álvarez Casado
CMVS, Infotech Oulu
University of Oulu, Finland
0000-0002-3052-4759

Miika Malin
Biomimetics and Intelligent Systems Group
University of Oulu, Finland
0009-0004-7243-4219

Nicoletta Prencipe
Center for Ubiquitous Computing
University of Oulu, Finland
0000-0001-5981-8016

Sasan Sharifipour
CMVS, Infotech Oulu
University of Oulu, Finland
0000-0001-6670-4252

Miguel Bordallo López
CMVS, Infotech Oulu
University of Oulu, Finland
0000-0002-5707-9085

***Abstract*—Convolutional networks have shown strong performance in physiological signal classification, generating robust temporal representations without expert knowledge. In contrast, handcrafted features, are able to capture statistical, fractal, complexity-based, cardiac, and respiratory characteristics at the cost of feature engineering. This paper explores a fusion framework combining physiological one-dimensional features from deep convolutional networks and handcrafted parameters extracted from signals obtained from mmWave radar, RGB and depth cameras. A cross-attention module guides the deep feature extraction process, allowing convolutional features to refine their representations based on handcrafted descriptors. The fusion approach improves accuracy across all physiological classification tasks, achieving 15% improvement in pose estimation, 24% in breathing pattern classification. These results highlight the advantages of hybrid feature integration for remote health monitoring, particularly in elderly care and physiological assessment.**

***Index Terms*—Biometric analysis, multimodal data fusion, breathing pattern**

I. Introduction

Non-invasive monitoring of physiological signals, such as respiration and cardiovascular activity, has gained significant attention in healthcare applications. Radar [1] and camera-based [2] systems capture data that can be treated as one-dimensional time-series data while preserving privacy and remaining unaffected by varying light conditions. The resulting signals provide valuable data for classification tasks such as breathing pattern analysis and physiological state monitoring used in remote health assessment and elderly care [3]. The signals can be analyzed using two primary approaches. One approach relies on handcrafted feature-based models, which capture non-linear characteristics such as frequency-domain representations but require domain expertise to extract specialized features from cardiac or respiratory signals. Alternatively, deep learning models automatically learn features by focusing on spatio-temporal patterns, eliminating the need for manual feature engineering while leveraging data-driven representation learning. Multimodal sensor fusion frameworks have explored integrating different sensors at various levels (e.g., sensor, model, and decision) to leverage multiple modalities. However, limited efforts have been made to develop models that effectively fuse one-dimensional handcrafted features with deep-learned features across different modalities and sensors for physiological analysis. In this work, we propose a novel architecture that fuses handcrafted features, obtained through feature engineering, with convolutional features extracted from a self-supervised model. These features are derived from four signals representing breathing and heart activity, captured using an mmWave radar and an RGB-D camera. The motivation behind this fusion arises from the hypothesis that combining linear, temporal convolutional features with non-linear, frequency-domain, and complex-domain features can enhance model performance by leveraging the complementary nature of these representations. Our model introduces a multimodal fusion framework at two levels, with key contributions summarized as follows:

- First, we propose an intermodality feature fusion approach that integrates handcrafted features from the frequency, complex, and non-linear domains with convolutional temporal features extracted from a self-supervised model, guided by a cross-attention mechanism.
- Second, we implement post-fusion modality by combining four intermodality fused streams. These modalities are derived from heart and respiration waveforms captured by RGB-D and mmWave radar signals.
- Finally, we evaluate the framework on classification tasks, including breathing pattern estimation and biometric information, such as sex classification. Additionally, we conduct an ablation study to assess the improvements introduced by our fusion model.

II. Related Work

Physiological studies have traditionally derived features from signal analysis, extracting handcrafted features from biological time series using specialized libraries such as Heartpy [4] and Neurokit [5]. This approach leverages the non-linearity of handcrafted features and extends beyond temporal components by analyzing frequency, fractal-based descriptors, and purely statistical features. These features are then used in classification tasks to characterize activities, and identify abnormal respiration patterns [6]. Despite demonstrating good generalization [7], handcrafted features-based models require expert knowledge and manual feature engineering to capture meaningful characteristics.

A different approach uses deep learning (DL) models, particularly convolutional neural networks (CNNs), which have demonstrated strong performance in learning feature representations directly from raw signals [8]. CNNs effectively capture local spatio-temporal characteristics, eliminating the need for predefined feature extraction [9]. However, CNNs primarily focus on local structures model linearity and temporality in data, which limits their ability to capture diverse domain-specific features, such as frequency information. Although DL provides advantages in certain approaches, the choice between DL and handcrafted features should be carefully considered for each specific study [10].

The limited exploration of fusing handcrafted and DL features has led to investigations into their complementarity, aiming to enhance classification performance when integrated [11]. In physiological signal analysis, hybrid approaches have been explored to improve robustness by leveraging handcrafted descriptors alongside DL features [12]. This has motivated the development of fusion strategies that effectively integrate both representations, ensuring improved interpretability and classification accuracy in biosignal-based health monitoring. Although experiments on fusing handcrafted and DL features have been conducted for image processing [13] and facial expression recognition [14], to the best of our knowledge, no studies have explored this approach for one-dimensional physiological data using different sensors, such as cameras and radar, across multiple modalities, including breathing and cardiac waveforms.

In this work, we propose a framework for studying the fusion of features extracted from biosignals using both DL models and handcrafted feature extraction to classify breathing pattern activities and biometric signals.

III. Methodology

A. Dataset

The heart and respiration waveforms used in this work are obtained from the OMuSense-23 dataset [15]. This dataset consists of labeled data collected from an RGB-D camera (Intel RealSense D435) and a millimeter-wave radar (Texas Instruments IWR1443), capturing 50 users in three different poses (standing, sitting, and lying down). Each user performs four 30-second activities representing specific breathing patterns (normal breathing, reading, guided breathing, and apnea) in three recordings, resulting in a total of 150 videos. The four selected signals from the dataset used in our experiments are the following, with a sample shown in Figure 1.

- **Heart activity waveform.** Obtained from the mmWave radar signal and filtered within the common human heart frequency band from 0.8 to 4 Hz [16].
- **Heart activity waveform.** Derived from the RGB stream using a rPPG signal, processed through the Face2PPG [17] framework with a chrominance-based method (CHROM) [18].
- **Respiration activity waveform.** Obtained from the radar signal, filtered within the common respiration frequency band from 0.1 to 0.6 Hz. [16].
- **Respiration activity waveform.** Derived from the depth stream, where chest movement is calculated by averaging a selected RGB patch data into a scalar value.

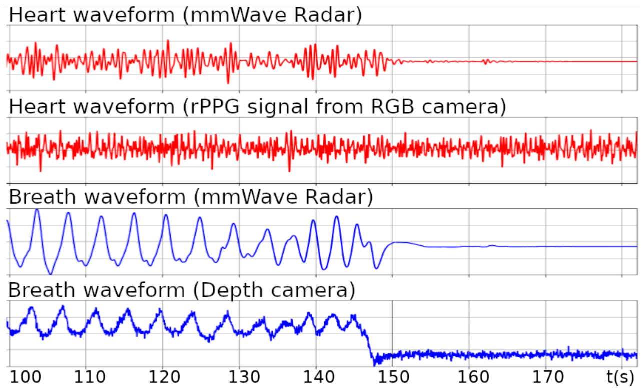

Fig. 1. Sample of the four signals extracted from mmWave Radar and RGBD depicting two breathing patterns (guided breathing and apnea).

We follow the data preprocessing of previous work [15], standardizing signals per video to reduce setup-specific variations, resampling to 20 Hz, and segmenting into 10-second windows with a 1-second sliding interval.

B. Handcraft feature extraction module

The handcrafted extraction module derives a subset of features from different domains. These features are calculated for all the signals contained in every window.

1) Statistical features: Statistical features describe the distribution and variability of the signal over time. The extracted features include *mean, standard deviation, maximum, minimum, percentiles, and four interquartile ranges*. Furthermore, the *dynamic range* and the *mean crossing rate* are computed.

2) Fractal features: Biological signals, due to their inherent complexity and non-linear nature, exhibit fractal properties that can reflect changes in sympathetic activity. [19]. Self-similarity and scaling properties are measured by *Katz and Higuchi* [20]. Fractal dimensions are also computed to quantify irregularities in the signal over time. *Detrended Fluctuation Analysis (DFA)* is applied to assess long-range correlations in the data while *Permutation entropy* measures the randomness or predictability of a time series. *Hjorth* parameters are calculated to provide additional insights into rapid signal fluctuation.

3) Complex features: We use the NeuroKit2 [5] library to explore complex features that capture non-linear relationships within the signal. *Hurst exponent*, identifies and self-similarity. *Multiscale Entropy (MSE)* assesses signal complexity across multiple time scales, while *Approximate Entropy (ApEn)* and *Sample Entropy (SampEn)* quantify the regularity and unpredictability of the signal.

4) Cardiac features: From the two heart-related signals, a subset of physiological features is extracted using HeartPy library [4]. Time-domain features measure *heart rate variability* (HRV) and inter-beat intervals (IBI), which represent the average time between consecutive heartbeats. Poincaré features are used to reflect short and long term *HRV* components. *Cardiac Sympathetic Index (CSI)* and the *Cardiac Vagal Index (CVI)*, provide insights into autonomic nervous system regulation. Frequency-domain such as *High-Frequency Normalized Units (HNU)* are also extracted.

5) Respiratory features: The two breathing signals are analyzed to extract temporal, amplitude, and respiratory rate characteristics. The *exhalation-inhalation transition (NN)* is evaluated in terms of its *mean*, *variability*, and *average duration*. Amplitude from respiratory peaks and Respiratory rate are also added. The total number of features per window is distributed as: 23 statistical, 10 fractal, 4 complex, 25 heart-related, and 4 breath-related features. The resulting handcrafted feature vector is named H_{enc}.

C. Deep convolutional feature extraction module

Deep convolutional feature extraction is performed using an SSL multimodal framework [21]. This framework encodes four physiological signals into distinct feature vectors using a ResNet-based 1-D encoder. A temporal contrastive module is then applied to all pairwise combinations of these encodings, leveraging contrastive loss to capture temporal dependencies. Next, a contextual contrastive module employs a Normalized Temperature-scaled Cross Entropy (NT-Xent) loss to minimize the distance between representations from the same time window while maximizing the distance between those from different windows [22]. The total self-supervised loss is computed as a weighted sum of the temporal and contextual contrastive modules, averaged across all modality pairs, treating them as augmented views of the same physiological phenomenon [23].

Training is conducted without labels, enabling the system to learn robust representations of cardiovascular activity from multiple signals at a given moment. Once training is complete, these learned features serve as a strong foundation for fully supervised downstream tasks using a small percentage of labeled data. At this stage, the temporal and contextual contrastive modules are removed, leaving only the encoders for the extraction of features and the convolutional features named D_{enc}.

D. Fusion Pipeline Framework

Both deep convolutional and handcrafted features are projected into smaller subspaces, denoted as D_{proj} and H_{proj}, before being fed into the fusion pipeline. The proposed fusion framework uses a cross-attention mechanism, where deep convolutional features are modified by attending to specific parts of the handcrafted features. An attention weight matrix S is computed, capturing information from both deep convolutional and handcrafted features, given by $S = D_{\text{proj}} H_{\text{proj}}^{\top}$. To generate a fusion encoding vector, the Hadamard product is applied between the original deep convolutional features and the normalized attention score matrix S using a softmax function. This ensures that D_{enc} is refined while retaining information from both the D_{proj} and H_{proj} feature vectors. Since H_{proj} represents high-level handcrafted features, it serves as a guide for convolutional features to learn complementary information. This operation is defined in Equation 1.

$$F_{\text{fused}} = (D_{\text{enc}} \odot \text{softmax}(S)) \tag{1}$$

This process is repeated across the four different modalities within each window. The pipeline is illustrated in Figure 2. The fused features are subsequently aggregated in a final post-fusion modality, calculating the mean of the four modalities predictions. which is then fed into a classifier. The classifier consists of two fully connected layers separated by a layer normalization and a ReLU activation.

E. Training and Evaluation

To evaluate the model, we propose three physiological classification tasks: breathing pattern, pose, and a combination of both, referred to as PosAct. Additionally, sex classification is performed as biometric analysis. The classification tasks are defined as follows:

- Breathing pattern (four classes): *Normal Breathing, Reading, Guided Breathing* and *Apnea*.
- Pose (three classes): *Standing, Sitting* and Lying *Down*.
- Sex (two classes): based on responses provided in a user filled form, *Female* and *Male*.
- PosAct (12 classes): all possible combinations of breathing patterns and poses.

OmuSense-23 dataset is divided into five folds, each containing data from 10 users with an equal sex distribution. The classifier performance is evaluated using accuracy within a train-validation-test split, where training is performed on three folds, validation on one, and testing on another. Results are also reported separately for each modality to analyze the contribution of individual sensor information. This approach allows us to assess the role of each feature type and evaluate the effectiveness of the fusion model. To ensure a balanced and comprehensive evaluation, we also report results using five-fold cross-validation. All classification models are trained in a supervised manner using a cross-entropy loss function. Additionally, we conduct an ablation study where each feature stream (handcrafted and deep features) is independently evaluated across the same classification tasks. This analysis provides insight into the contribution of the fusion system to overall accuracy. The performance of handcrafted features alone is assessed using an XGBoost classifier trained on a concatenation of all extracted features. It is also evaluated within the fusion pipeline without the convolutional encoder

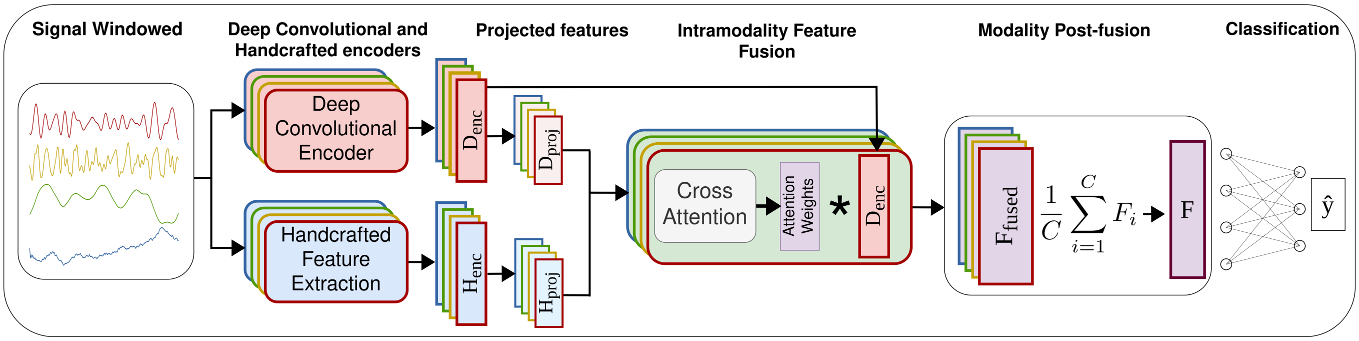

Fig. 2. The fusion pipeline for handcrafted and deep convolutional features is repeated along all different modalities (C) within each window

module and without the intramodality feature fusion component. Automatic features are evaluated both as a standalone model in their SSL framework and within the fusion pipeline where the feature extraction module and intramodality feature fusion are removed to assess their impact.

IV. Results and Discussion

The fusion results from the fusion pipeline are presented in Table I. The model achieves 85% accuracy in activity classification and 97% accuracy in pose classification. For sex classification, performance to 63% and for PosAct the accuracy is 79%. The results reveal interesting insights. Notably, the fusion approach outperforms both single-feature classifications, demonstrating its effectiveness in integrating information from multiple feature types. The radar-based breathing modality consistently outperforms the other modalities, indicating that contactless radar sensors may capture subtle respiratory dynamics with greater fidelity than visual methods. The five-fold cross-validation results, Table II, demonstrate consistency across different folds, suggesting the strong generalizability of the fusion pipeline.

TABLE I
Accuracy for fusion models with handcrafted and convolutional features

	Activity	Pose	Sex	PosAct
Fusion pipeline (F)	**85**	**97**	**63**	**79**
Rad-heart	55	68	50	43
Cam-heart	44	55	59	24
Rad-breath	74	83	47	56
Cam-breath	73	79	57	60
Random Guess	25	33	50	08

TABLE II
Accuracy for fusion models using five-fold cross-validation

	Activity	Pose	Sex	PosAct
Fold I	**85**	**97**	**63**	**79**
Fold II	89	93	60	80
Fold III	87	92	53	82
Fold IV	88	85	50	70
Fold V	87	96	56	85
Average	87.2	92.6	56.4	79.2

We also conducted an ablation study for each feature stream independently. The ablation study for handcrafted features alone (H) is conducted using a train-validation-test approach. Experiments are performed with an XGBoost classifier and the fusion pipeline, excluding both the deep convolutional encoder and the intramodality feature fusion module. Interestingly, both models yield similar performance, suggesting that they effectively capture all relevant information from handcrafted features. Since neither model applies convolutional operations, their results remain comparable. Modality-specific results indicate that the radar-based breathing modality contributes the most to classification performance, a trend already seen in the fusion model. Nevertheless, sex classification performance is close to random guessing, suggesting that biometric information is largely absent from handcrafted features. The baseline results for handcrafted features are summarized in Table III.

TABLE III
Accuracy for models with only handcrafted features

	Activity	Pose	Sex	PosAct
XGBoost Baseline	81	86	50	68
Fusion pipeline (H)	83	79	49	69
Rad-heart	49	57	53	32
Cam-heart	40	58	56	24
Rad-breath	63	68	52	46
Cam-breath	60	76	51	47

The ablation study using only deep convolutional features (D) was conducted in two settings: first, within the original SSL downstream pipeline, and second, in the fusion pipeline without the handcrafted feature extraction module and the intramodality feature fusion. Similar to the handcrafted feature results, the consistency across both methods suggests that the models effectively capture the most meaningful information from convolutional features, regardless of the architecture as shown in Table IV. While handcrafted models achieved higher accuracy in physiological classification tasks (81% vs. 74% for activity classification and 86% vs. 78% for pose classification), convolutional models performed better in biometric sex classification (61% vs 50%). The fusion results indicate that integrating handcrafted and convolutional features could enhance overall performance by leveraging the strengths of both approaches for physiological and biometric classification. Moreover the fusion of handcrafted and deep convolutional features not only boosts general classification accuracy, but also captures critical physiological nuances for breathing and pose classification, while deep features enhance biometric tasks such as sex classification. These findings challenge conventional assumptions about the sufficiency of deep learning alone in physiological monitoring and highlight the potential of hybrid models to reveal more robust representations.

TABLE IV
ACCURACY FOR MODELS WITH ONLY CONVOLUTIONAL FEATURES

	Activity	Pose	Sex	PosAct
SSL Baseline	74	78	61	63
Fusion pipeline (D)	69	80	60	62
Rad-heart	55	60	46	37
Cam-heart	39	59	62	24
Rad-breath	59	61	52	40
Cam-breath	62	50	52	34

V. Conclusion

This research presented a fusion pipeline based on cross-attention between handcrafted and deep convolutional features extracted from cardiac and respiratory waveforms using an RGB-D camera and mmWave radar. Our results demonstrated that feature fusion outperformed models using handcrafted or deep convolutional features independently, suggesting an information complementarity between the two feature sets. This aligns with findings in other domains, where handcrafted and deep representations have been shown to provide complementary information, improving classification performance [11]. Specifically, the fusion model achieved an improvement of 15% in breathing pattern classification, 24% in pose estimation, and 25% in a 12-class classifier that mix both tasks, highlighting its effectiveness across different classification tasks. By incorporating a late fusion strategy across multiple modalities, we also introduced a framework for evaluating the impact of each modality individually. This is particularly useful in real-world scenarios where certain modalities may be unavailable due to occlusions or privacy concerns, such as when cameras cannot be used. For future work, this pipeline could be further validated on additional datasets to deepen the analysis of feature complementarity between deep convolutional models and handcrafted features.

In conclusion, this paper has demonstrated the effectiveness of our fusion pipeline in integrating features from different domains such as temporal, amplitude, fractal, and complexity-based features with purely convolutional features. Our findings also suggest that convolutional models alone struggle to capture the full complexity of certain frequency-domain features, highlighting the value of incorporating handcrafted representations in such tasks.

References

[1] S. M. M. Islam, "Radar-based remote physiological sensing: Progress, challenges, and opportunities," *Frontiers in Physiology*, vol. 13, oct 11 2022.

[2] N. Molinaro, E. Schena, S. Silvestri, F. Bonotti, D. Aguzzi, E. Viola, F. Buccolini, and C. Massaroni, "Contactless Vital Signs Monitoring From Videos Recorded With Digital Cameras: An Overview," *Frontiers in Physiology*, vol. 13, feb 18 2022.

[3] E. J. Argüello-Prada, M. A. D. Cantín, and J. C. Victoria, "A photoplethysmography-based system for talking detection in bedridden patients," *Biomedical Signal Processing and Control*, vol. 81, p. 104477, 2023.

[4] P. Van Gent, H. Farah, N. Nes, and B. van Arem, "Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data," in *Proceedings of the 6th HUMANIST Conference*, 2018, pp. 173–178.

[5] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021.

[6] L. Nguyen, C. Á. Casado, O. Silvén, and M. B. López, "Identification, activity, and biometric classification using radar-based sensing," in *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2022, pp. 1–8.

[7] N. Bento, J. Rebelo, M. Barandas, A. V. Carreiro, A. Campagner, F. Cabitza, and H. Gamboa, "Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition," *Sensors*, vol. 22, no. 19, p. 7324, 2022.

[8] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 1–13, 7 2018.

[9] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction Using Deep Learning Methods in Healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–29, jan 16 2023.

[10] L. Rundo and C. Militello, "Image biomarkers and explainable ai: handcrafted features versus deep learned features," *European Radiology Experimental*, 2024.

[11] R. M. Pereira, Y. M. Costa, R. L. Aguiar, A. S. Britto, L. E. Oliveira, and C. N. Silla, "Representation learning vs. handcrafted features for music genre classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[12] C. Á. Casado, M. L. Cañellas, and M. B. López, "Depression recognition using remote photoplethysmography from facial videos," *IEEE Transactions on Affective Computing*, vol. 14, 2023.

[13] S. R. Bhadra, S. P. Maity, and J. Sil, "Enhancement of alzheimer's disease detection technique by fusing features of deep learning and handcrafted methods on smri images," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*. IEEE, 2024, pp. 1079–1090.

[14] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[15] M. Lage Canellas, L. Nguyen, A. Mukherjee, C. Á. Casado, X. Wu, P. Susarla, S. Sharifipour, D. B. Jayagopi, and M. B. López, "Omusense-23: A multimodal dataset for contactless breathing pattern recognition and biometric analysis," *arXiv preprint arXiv:2407.06137*, 2024.

[16] T. Instruments, *TI mmWave Labs Vital Signs Measurement*, 2017.

[17] C. A. Casado and M. B. López, "Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[18] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[19] R. Acharya, P. S. Bhat, N. Kannathal, A. Rao, and C. M. Lim, "Analysis of cardiac health using fractal dimension and wavelet transformation," *Itbm-Rbm*, vol. 26, no. 2, pp. 133–139, 2005.

[20] B. Raghavendra and D. N. Dutt, "A note on fractal dimensions of biomedical waveforms," *Computers in Biology and Medicine*, vol. 39, no. 11, pp. 1006–1012, 2009.

[21] M. Lage Canellas, C. Alvarez Casado, L. Nguyen, and M. Bordallo Lopez, "A self-supervised multimodal framework for 1d physiological data fusion in remote health monitoring," *preprint SSRN 5044405*, 2024.

[22] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2021.

[23] S. Dixon, L. Yao, and R. Davidson, "Modality aware contrastive learning for multimodal human activity recognition," *Concurrency and Computation: Practice and Experience*, p. e8020, 2024.