Assessing the Quality of Photoplethysmograms via Gramian Angular Fields and Vision Transformer

Pedro Garcia Freitas*, Rafael G. de Lima†, Giovani D. Lucafo‡, and Otávio A. B. Penatti§
*Department of Computer Science, University of Brasília, Brasília, Brazil
†‡§Samsung R&D Institute Brazil (SRBR), Campinas, São Paulo, Brazil

***Abstract*—Nowadays, the field of healthcare delivery is undergoing a revolutionary transformation through the implementation of real-time health monitoring. This innovative approach utilizes everyday environments, particularly leveraging the latest wearable health devices, to enable the continuous monitoring of individuals irrespective of their location or time. This breakthrough allows for the timely identification and prevention of numerous diseases. Among the array of technologies integrated into wearable devices to facilitate ongoing health monitoring, one of the most significant techniques is Photoplethysmography (PPG). PPG is a non-intrusive, cost-effective, easily implementable, and thus convenient method for tracking physiological signals, including oxygen saturation in the bloodstream, heart rate variability, respiration rate, and more. Due to these merits, PPG has gained extensive usage across various health applications, especially in commercially available wearable devices. Nevertheless, despite its advantages, PPG suffers from a primary drawback of being highly vulnerable to motion artifacts and environmental disturbances. These factors significantly hinder the efficacy of PPG-based applications, particularly when recording PPG signals through wearable devices. Consequently, in order to ensure reliable measurements, it is crucial to assess the quality of the signals and discard unreliable ones. Signal quality assessment emerges as the foremost priority in this context. This paper presents an innovative method for evaluating the quality of PPG signals, accomplished through a fusion of Gramian Angular Fields (GAF) and Visual Transformer (ViT) techniques. The results demonstrate that the proposed approach achieves a competitive accuracy in predicting signal quality when compared to the current state-of-the-art methods.**

***Index Terms*—photoplethysmography, digital health, Gramian angular fields, signal quality, vision transformers**

I. Introduction

The integration of wearables and Internet of Things (IoT) systems has significantly propelled the progress of remote health monitoring. These systems employ specialized infrastructure for sensing, communication, and computation to ensure continuous monitoring of individuals' health and prompt decision-making. Crucially, wearable devices like smartwatches and smart rings play a pivotal role in this process, enabling the continuous acquisition of biomedical signals such as Electrocardiogram (ECG) and PPG [1]. PPG, an optical technique, non-invasively captures variations in blood volume by utilizing the properties of light transmission and reflection through a light sensor conventionally positioned on the finger or wrist [1]. The changes observed in these collected biosignals provide diverse physiological information, including cardiac oscillation, oxygen saturation, and respiration rate, facilitating various clinical applications such as pain assessment, vascular stiffness evaluation, and cuffless blood pressure estimation [2]. Due to its simplicity of implementation and wide range of applications, PPG is extensively incorporated into diverse wearable devices [3]. However, the primary limitation of PPG lies in its susceptibility to noise, particularly arising from motion artifacts that obscure or distort the information encoded within the signal. This vulnerability can lead to erroneous predictions of health status, misdiagnosis, and other issues in monitoring, which are unacceptable in the context of healthcare applications. Consequently, the development of a reliable PPG quality assessment method is crucial to discern between dependable and unreliable biosignals and prevent any misinterpretation.

The automatic assessment of quality using objective metrics is an extensively explored field across a diverse range of signal types [4]–[7]. In the particular scope of this paper (i.e., the assessment of 1D physiological signals), the primary objective of a quality assessment method is to automatically classify signals as either "good" or "bad". To achieve this classification task, various methods have been proposed in the literature, including waveform morphological analysis (referred to as template matching [2] or rules [8]) and Machine Learning (ML) techniques. Rule-based approaches involve extracting diverse statistical features from the peaks and valleys in the signals, such as pulse amplitude, width, slope, entropy, power, skewness, and kurtosis. These features are used to assess signal quality based on specific criteria. These methods compare each pulse waveform to a template representing a high-quality signal and establish a correlation to differentiate between "good" or "bad" waveforms. Thresholds are employed to distinguish between the waveforms. Rule-based approaches offer the advantage of computational simplicity and speed. However, they often suffer from inaccuracies and difficulties in generalization across different implementations.

Learning-based methods for assessing signal quality center around the utilization of data and classification algorithms to construct predictive models. Machine learning (ML) algorithms encompass a broad spectrum, ranging from support vector machines [9] to deep neural networks [10]. More recently, novel learning-based approaches have emerged that leverage the spatiotemporal information embedded within the PPG signal by transforming the one-dimensional waveform into a two-dimensional representation. For instance, Roh &

Shin [2] propose the transformation of one-dimensional PPG signals into a two-dimensional image using recurrence plots, enabling the characterization of the system's behavior in phase space. Similarly, Chatterjee et al. [11] convert one-dimensional signal segments into image files, which are then fed as input to a two-dimensional Convolutional Neural Network (CNN) architecture, thereby leveraging the strengths of image classification techniques.

In this manuscript, we present a novel technique adopting the aforementioned strategy. It assesses the quality of PPG signal by expanding the one-dimension PPG into a two-dimensional GAF to build a deep learning model to achieve an efficient assessment of signal quality. Using the GAF as input, the method learns to differentiate 'good' and 'bad' signals by a ViT on top of it. The proposed method is able to efficiently recognize the waveforms that most approximately match the ground-truth signals with proper reliable/unreliable labels.

II. Proposed Method

Figure 1 illustrates the proposed pipeline which is divided into three steps: (1) GAF generation, (2) patchfication, and (3) image classification that employs a Transformer-like architecture over patches of the generated GAF to perform quality index estimation. In the following subsections, we provide a detailed description of each individual step illustrated in this diagram.

A. Gramian Angular Field Generation

Given a PPG signal $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ of n samples so that all values are in the interval $x_i \in \mathbb{R}\,[-1, 1]$. We can represent this signal $\boldsymbol{x}$ using polar coordinates by encoding the value of signal amplitude as the angular cosine and the timestamp as the radius as follows:

$$\begin{cases} \phi_i = \arccos(x_i), & -1 \leq x_i \leq 1 \\ r = \dfrac{t_i}{N}, & t_i \in \mathbb{N}, \end{cases} \tag{1}$$

where t_i is the timestamp and N is a regularization factor to stretch over the polar coordinate system. Using this polar coordinate-based representation, we can encode the 1D signal into an encoding map that is bijective as $\cos(\phi)$ and monotonic when $\phi \in [0, \pi]$. It means that, for a given 1D signal, the transformed map produces one and only one result in the polar coordinate system with a unique inverse map. Moreover, contrary to Cartesian coordinates, polar coordinates representations retain temporal relations. After this transformation, we can model the GAF using the sum or the difference between each point to describe the temporal correlation within various time intervals. Therefore, in summary, two GAF methods can be derived, namely Gramian Difference Angular Field (GAF_D) and Gramian Summation Angular Field (GAF_S), that are defined as follows:

$$\begin{aligned} GAF_D &= \{\sin(\phi_i - \phi_j)\}_{i,j} \\ &= \left(\sqrt{\mathbb{1} - \boldsymbol{x}^2}\right)^{\intercal} \cdot \boldsymbol{x} - \boldsymbol{x}^{\intercal} \cdot \sqrt{\mathbb{1} - \boldsymbol{x}^2} \end{aligned} \tag{2}$$

and

$$\begin{aligned} GAF_S &= \{\cos(\phi_i + \phi_j)\}_{i,j} \\ &= \boldsymbol{x}^{\intercal} \cdot \boldsymbol{x} - \left(\sqrt{\mathbb{1} - \boldsymbol{x}^2}\right)^{\intercal} \cdot \sqrt{\mathbb{1} - \boldsymbol{x}^2}, \end{aligned} \tag{3}$$

where $\mathbb{1}$ is the unit row vector $[1, 1, \cdots, 1]$ with the same length of $\boldsymbol{x}$. After the transformation of the 1D signal into the polar coordinate system, two types of GAFs can be defined by the inner products $\langle x, y\rangle = x \cdot y - \sqrt{1 - x^2} \cdot \sqrt{1 - y^2}$ and $\langle x, y\rangle = \sqrt{1 - x^2} \cdot y - x \cdot \sqrt{1 - y^2}$. These products are quasi-Gramian matrices since the defined functions $\langle x, y\rangle$ do not satisfy the property of linearity in the inner-product space.

GAFs have multiple benefits. They provide a form of preserving temporal dependency since time increments as the position moves from top-left to bottom-right. They include temporal correlation since their elements represent the corresponding correlation by superposition of directions concerning a time interval. The main diagonal is a special case that comprises the original angular information. From this main diagonal, it is possible to reconstruct the 1D signal from high-level features learned by a neural network.

Figure 2 depicts how the 1D signal waveform relates to the 2D GAF representations. The first column of this figure (i.e., Figure 2-(a)) illustrates the waveforms of a 'good' and a 'bad' signal. Figure 2-(b) illustrates the corresponding two-dimensional projections of the GAF for each of them. From these images, we can observe a noticeable visual regularity in the projected version of 'good' signals. On the other hand, for 'bad' signals, the GAF projections present more irregularity and chaotic visual patterns. The regularity characteristic becomes more apparent when we examine the similarity between the various patches, as illustrated in Figures 2-(c), (d), and (e). Hence, the distinction between the quality of the PPG signals labeled as "good" or "bad" can be achieved by either comparing their two-dimensional GAF representation and by inter-correlating their intra-patches. Moreover, the motivation for patchfication is also an essential part of ViT algorithm and is more detailed in the next section.

B. Patchfication

In the proposed approach, images in Vision Transformers (ViT) are represented as sequences, enabling the independent learning of visual structures. The images are treated as a series of patches, with each patch flattened into a single vector by concatenating the channels of all pixels and linearly projecting it to the desired embedding dimension. For the PPG signals in this study, they are divided into windows of 3 seconds with 75 samples each and an overlap of 5 samples. As a result, the generated 2D GAF representations correspond to images with dimensions of 75×75 pixels. To ensure a perfect division of the 75×75 image, patch sizes (PS) of 25×25, 15×15, and 5×5 were considered, as shown in Figure 2-(c), Figure 2-(d), and Figure 2-(e) respectively.

C. Transformer-based Image Classification

The ViT model represents an input image as a series of patches, like the series of word embeddings used when using

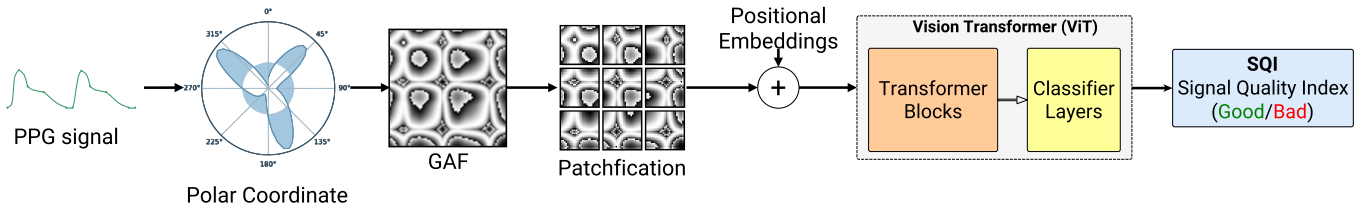

Fig. 1. Depiction of the proposed Photoplethysmogram Signal Quality Assessment method using Gramian Angular Fields and Vision Transformers.

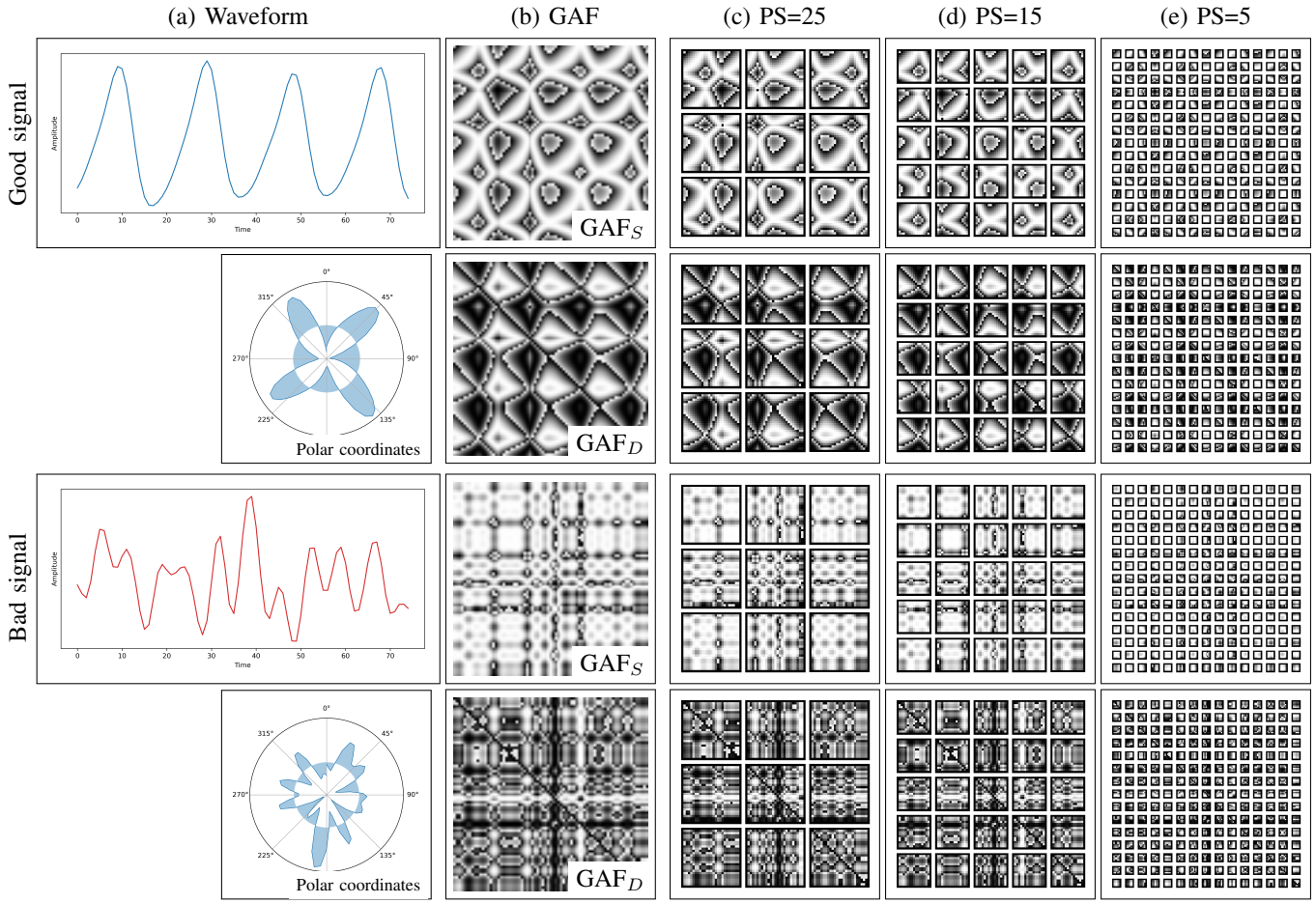

Fig. 2. Example of signals and their corresponding GAF encoding maps. A 'good' signal (a) produces symmetric encoding maps with a more redundant pattern. On the other hand, a 'bad' signal (b) generates asymmetric encoding maps with less redundancy and higher visual variation. The last three columns depict the effect of patchfication as function of different patch sizes (PS).

transformers to text, and directly predicts class labels for the image. This model applies the Transformer [12] to the image classification task as proposed by Dosovitskiy *et al.* [13]. The model architecture is almost the same as the original Transformer but with a twist to allow images to be treated as input, just like natural language processing.

Figure 3 depicts an overview of the ViT model. Its core, the Transformer Encoder, is the same as the standard Transformer proposed by Vaswani *et al.* [12]. This means that Transformer Encoder receives as input a 1D sequence of token embeddings and extracts features from the image, passing these processed features into a Multiplayer Percecptron (MLP) head model for classification. In order to handle 2D GAF images, the patches are reshaped, flattened, and sent through a single Feed Forward (FF) layer to get a linear patch projection so each patch can be treated as a token, which can be input to the Transformer.

To assist with the classification bit, Dosovitskiy *et al.* [13] took insights from Kenton *et al.* [14] by concatenating a learnable embedding with the other patch projections (i.e., the original BERT's `[class]` token, represented as '*' in Figure 3). Finally, the outputs of the Transformer Encoder are then sent into a MLP to perform the quality classification.

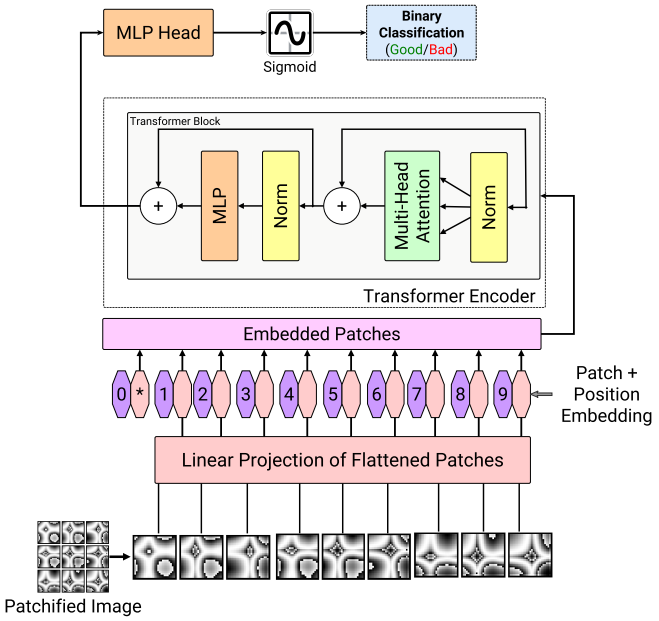

Fig. 3. ViT model overview. A GAF image is split into fixed-size patches, linearly embedded, and position embeddings are added. The resulting sequence of vectors feeds an original Transformer Encoder (the same used by Vaswani *et al.* for NLP purposes [12]). To perform the binary classification, we modified the original Softmax activation layer by replacing it with a Sigmoid.

III. Experimental Results

We used a dataset referred to as 'ICON' that was originally developed for Inter-Beat Interval (IBI) detection purposes [8]. This dataset was originally collected with 46 volunteer subjects. These subjects include 9 volunteers with permanent Atrial Fibrillation (AF), 16 volunteers with Normal Sinus Rhythm (NSR), and 21 volunteers with other non-specified arrhythmias. The majority of these volunteers are older than 60 years, with a mean age of 66 years old and a median age of 70 years. The dataset used throughout this study was collected using a Samsung Galaxy Watch Active 2 at 25Hz lasting 45-60 minutes per subject. Due to the purpose of the ICON dataset, the quality of the signals in this dataset is manually labeled by experts according to their waveforms, i.e., a 'good' signal can provide reliable IBI measurements.

We used Optuna [15] to find the optimized combination of hyperparameters. Since the PPG signals are split in windows of 75 samples in the ICON dataset, the search space included the Patch Sizes (PSs) of 25×25, 15×15, and 5×5, as illustrated in Figure 2-(c), Figure 2-(d), and Figure 2-(e), respectively. Additionally, the ViT's 'number of encoding dimensions' ($\mathcal{D}$), 'patch size´ ($\mathcal{PS}$,), and 'number of transformer blocks' ($\mathcal{NB}$) are also included in the search space. We implemented our ViT model using Keras [16] and Einops [17] for a faster patchfying execution.

During the Optuna execution, we saved the top-performing results found in the optimization process using the F1-score as the optimization criterion. Table I depicts the top-15 performance metrics achieved in this process as well as the corresponding hyperparameters used to achieve these metrics.

TABLE I
Top performing metrics obtained using GAF and the corresponding ViT parameters used to achieve these metrics. $\mathcal{D}$, $\mathcal{PS}$, and $\mathcal{NB}$ represent the number of encoding dimensions, patch size, and number of transformer blocks, respectively. The best scores and selected parameters are **boldfaced**.

Trial	Accuracy	Recall	Precision	F-Score	$\mathcal{D}$	$\mathcal{PS}$	$\mathcal{NB}$
1	0.89534	0.92326	0.92007	0.92166	75	15	10
2	0.90597	0.92894	0.92997	0.92946	25	5	3
3	**0.92212**	0.93383	**0.94857**	**0.94114**	**32**	**5**	**1**
4	0.88756	0.91837	0.91347	0.91591	64	15	6
5	0.89066	0.90796	0.92660	0.91718	64	15	3
6	0.89525	0.91337	0.92838	0.92082	32	5	6
7	0.92162	0.93429	0.94744	0.94082	32	5	1
8	0.90075	0.91684	0.93315	0.92492	64	5	10
9	0.90645	0.91321	0.94465	0.92867	25	15	3
10	0.90747	0.92646	0.93423	0.93033	75	15	1
11	0.90007	0.90571	0.94219	0.92359	32	25	1
12	0.92121	0.93324	0.94779	0.94046	32	5	1
13	0.92093	0.93405	0.94666	0.94032	32	5	1
14	0.90293	0.91464	0.93823	0.92628	32	25	1
15	0.92180	**0.93664**	0.94557	0.94108	32	5	1

From this table, the best results achieved are highlighted. These results correspond to 32, 5, and 1 for $\mathcal{D}$, $\mathcal{PS}$, and $\mathcal{NB}$, respectively.

These best hyperparameters produce an accuracy of 0.92212, a recall equal to 0.93383, a precision of 0.94857, and an F1 score of 0.94114 using the proposed method. These results are also reported in Table II This table also presents the results obtained using three state-of-the-art methods, expressed as 'Lucafo$_1$', 'Lucafo$_2$', and 'Hao & Bo'. All results in this table were obtained using the ICON dataset with the same train-test split and other experimental conditions. Based on the results presented in Table II, it is noticeable the advantage of the proposed approach.

TABLE II
Performance comparison of the proposed and state-of-the-art methods in the ICON dataset. The best results are **boldfaced**.

Method	Reference	Accuracy	Recall	Precision	F1
Lucafo$_1$	[8]	0.8980	0.8860	0.9590	0.9190
Lucafo$_2$	[8]	0.8990	0.8750	**0.9710**	0.9200
Hao & Bo	[18]	0.8060	0.7570	0.9050	0.8240
Proposed	∅	**0.9221**	**0.9338**	0.9485	**0.9411**

Table III compares the results of previous studies with our proposed approaches. From these results, Sukor *et al.* [19] and Naeini *et al.* [20] tie with a classification accuracy of 0.8300, which is much lower than the models created from the proposed approach, which achieves an accuracy score of 0.9221. Moreover, in the studies of Sukor *et al.* [19], Selveraj *et al.* [21], Liu *et al.* [10], Li and Clifford [22], and Naeini *et al.* [20], the number of subjects was small, which imply in lower reliability of results. A small number of subjects may induce bias and low variability in data, which can impair the trustworthiness of these high accuracy scores reported in Table III for these works. Considering this point, Roh and

Shin [2], Fisher *et al.* [23], and this present study report their results using a dataset with a large number of subjects, which enable more reliable results. From these results, Roh's and Fisher's present a higher performance than ours. However, Fisher's research presumes the detection of waveform features, therefore it is dependent on the feature detector in realistic usage. Roh's method has high computational complexity and its trade-off between performance and convenience needs to be considered. As a result, these approaches may not be proper to perform the signal quality assessment in some contexts. On the other hand, our method can be applied in more diverse scenarios since it uses only a GAF projection without any extraordinary preprocessing of the original signal.

TABLE III
SIGNAL QUALITY ASSESSMENT PERFORMANCE COMPARED TO PREVIOUS STUDIES (N IS THE NUMBER OF SUBJECTS).

Method	Reference	Input	N	Accuracy
Fischer et al.	[23]	Detected Features	69	0.9780
Sukor et al.	[19]		13	0.8300
Selvaraj et al.	[21]		10	0.9480
Li and Clifford	[22]		13	0.9520
Liu et al.	[24]		10	0.8300
Roh and Shin	[2]	Raw Signal	76	0.9750
Liu et al.	[10]		14	0.9500
Naeini et al.	[20]		1	0.8301
Proposed	∅		**56**	**0.9221**

IV. CONCLUSION

In this paper, we proposed a novel method of assessing the quality of PPG signals that are based on the GAF projections and ViT. We verified that the performance of the proposed model is comparable to the results of a state-of-the-art, showing a competitive level of performance without separated complex pre-processing and feature detection steps. However, in this study, the quality of the waveform is based on the single projection method (i.e., the GAF algorithm). Therefore, in future research, it is necessary to investigate whether GAF can be used in combination with other 1D-to-2D projection techniques, such as recurrence plots or Markov transition fields. In addition, this study should be extended to other types of 1D signals and time series.

REFERENCES

[1] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.

[2] D. Roh and H. Shin, "Recurrence plot and machine learning for signal quality assessment of photoplethysmogram in mobile environment," *Sensors*, vol. 21, no. 6, p. 2188, 2021.

[3] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 1, p. 130, 2017.

[4] P. G. Freitas, W. Y. Akamine, and M. C. Farias, "No-reference image quality assessment based on statistics of local ternary pattern," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[5] R. Diniz, P. G. Freitas, and M. C. Q. Farias, "Towards a point cloud quality assessment model using local binary patterns," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[6] P. G. Freitas, R. Diniz, and M. C. Farias, "Point cloud quality assessment: unifying projection, geometry, and texture similarity," *The Visual Computer*, vol. 39, no. 5, pp. 1907–1914, 2023.

[7] R. Diniz, P. G. Freitas, and M. C. Q. Farias, "Color and geometry texture descriptors for point-cloud quality assessment," *IEEE Signal Processing Letters*, vol. 28, pp. 1150–1154, 2021.

[8] G. Lucafo, P. G. Freitas, R. Lima, G. Luz, R. Bispo, P. Rodrigues, F. Cabello, and O. Penatti, "Signal quality assessment of photoplethysmogram signals using hybrid rule- and learning-based models," in *XIX Congresso Brasileiro de Informática em Saúde (CBIS-2022)*. Brazilian Health Informatics Association (SBIS), 2022.

[9] T. Pereira, K. Gadhoumi, M. Ma, X. Liu, R. Xiao, R. A. Colorado, K. J. Keenan, K. Meisel, and X. Hu, "A supervised approach to robust photoplethysmography quality assessment," *IEEE journal of biomedical and health informatics*, vol. 24, no. 3, pp. 649–657, 2019.

[10] S.-H. Liu, R.-X. Li, J.-J. Wang, W. Chen, and C.-H. Su, "Classification of photoplethysmographic signal quality with deep convolution neural networks for accurate measurement of cardiac stroke volume," *Applied Sciences*, vol. 10, no. 13, p. 4612, 2020.

[11] T. Chatterjee, A. Ghosh, and S. Sarkar, "Signal quality assessment of photoplethysmogram signals using quantum pattern recognition technique and lightweight cnn module," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 3382–3386.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[14] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[16] F. Chollet *et al.* (2015) Keras. [Online]. Available: https://github.com/fchollet/keras

[17] A. Rogozhnikov, "Einops: Clear and reliable tensor manipulations with einstein-like notation," in *International Conference on Learning Representations*, 2022.

[18] J. Hao and G. Bo, "A quality assessment system for ppg waveform," in *2021 IEEE 3rd International Conference on Circuits and Systems (ICCS)*, 2021, pp. 170–175.

[19] J. A. Sukor, S. Redmond, and N. Lovell, "Signal quality measures for pulse oximetry through waveform morphology analysis," *Physiological measurement*, vol. 32, no. 3, p. 369, 2011.

[20] E. K. Naeini, I. Azimi, A. M. Rahmani, P. Liljeberg, and N. Dutt, "A real-time ppg quality assessment approach for healthcare internet-of-things," *Procedia Computer Science*, vol. 151, pp. 551–558, 2019.

[21] N. Selvaraj, Y. Mendelson, K. H. Shelley, D. G. Silverman, and K. H. Chon, "Statistical approach for the detection of motion/noise artifacts in photoplethysmogram," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4972–4975.

[22] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiological measurement*, vol. 33, no. 9, p. 1491, 2012.

[23] C. Fischer, M. Glos, T. Penzel, and I. Fietze, "Extended algorithm for real-time pulse waveform segmentation and artifact detection in photoplethysmograms," *Somnologie*, vol. 21, no. 2, pp. 110–120, 2017.

[24] S.-H. Liu, J.-J. Wang, W. Chen, K.-L. Pan, and C.-H. Su, "Classification of photoplethysmographic signal quality with fuzzy neural network for improvement of stroke volume measurement," *Applied Sciences*, vol. 10, no. 4, p. 1476, 2020.