Deployment-Oriented Quantization of an ECGWavePuzzle-Based Personalization Pipeline for Arrhythmia Classification

Guilherme Silva[1], **Arthur Negrão**[1], **Pedro Silva**[2], **Eduardo Luz**[2]

[1] Postgraduate Program in Computer Science – Federal University of Ouro Preto – Brazil
guilherme.lopes@aluno.ufop.edu.br

[2]Computing Department – Federal University of Ouro Preto – Brazil
eduluz@ufop.edu.br

Abstract. *This paper evaluates INT8 quantization and mixed-precision execution in an ECGWavePuzzle-based personalization pipeline for arrhythmia classification. The SSL framework follows prior work, and the contribution is a deployment-oriented analysis under reduced numerical precision. On MIT-BIH with ANSI/AAMI patient-wise evaluation,* `head_int8` *achieves* 0.9013 ± 0.0198 *Accuracy and* 0.5364 ± 0.0416 *Macro-F1 on H200. FP16 also improves Macro-F1 over BF16 on H200 (0.5252 vs. 0.4969) with slightly lower runtime (2920 s vs. 3003 s). However, Jetson Nano execution remains expensive, requiring about 60048 s. The results show that selective quantization is feasible, but online personalization remains the main bottleneck for embedded deployment.*

1. Introduction

Automatic arrhythmia classification from ECG signals remains challenging under realistic conditions due to inter-patient variability, signal noise, and the need to follow clinically meaningful evaluation protocols [Luz et al. 2016, Silva et al. 2025]. At the same time, recent advances in self-supervised learning (SSL) have made patient-specific ECG adaptation a promising approach, since unlabeled data can be used to improve representation learning and reduce dependence on manual annotation [Silva et al. 2024, Mehari and Strodthoff 2022, Phan et al. 2022]. In computing applied to healthcare, however, the relevance of such methods depends not only on predictive performance, but also on whether they remain feasible under the latency, memory, and energy constraints of edge and embedded platforms [Silva et al. 2025].

In this context, we investigate the feasibility of applying deployment-oriented numerical optimization to a *ECGWavePuzzle*-based personalization pipeline proposed by Silva *et al.* [Silva et al. 2024]. Specifically, this paper evaluates whether INT8 quantization-aware training (QAT) and mixed-precision execution can be incorporated into a *ECGWavePuzzle*-based pipeline without significantly harming the performance of the arrhythmia classification.

Therefore, the main contribution of this paper is the deployment-oriented evaluation of *ECGWavePuzzle* under reduced numerical precision. The main contributions are summarized as follows: (i) a mixed-precision analysis of the non-quantized *ECGWavePuzzle*-based pipeline, (ii) an evaluation of selective INT8/QAT coverage in the supervised branch, and (iii) a comparison of online personalization strategies across high-

performance and embedded-oriented platforms. Experiments are performed on the MIT-BIH Arrhythmia Database using the ANSI/AAMI protocol and patient-wise evaluation. The results show that selective quantization preserves the behavior of the *ECGWavePuzzle* pipeline with limited degradation, but online personalization remains the main bottleneck for embedded deployment.

2. Related Works

Recent ECG SSL studies have focused on representation learning and downstream transfer. Liu *et al.* [Liu et al. 2021] proposed ECG-specific self-supervised pretraining using large amounts of unlabeled data, and Mehari and Strodthoff [Mehari and Strodthoff 2022] reported systematic gains from SSL on clinical 12-lead ECG classification. Closer to the present setting, Silva *et al.* [Silva 2023] introduced *ECGWavePuzzle* and a human-in-the-loop personalization pipeline for arrhythmia classification. These works motivate the use of SSL for ECG analysis, but they do not address reduced-precision deployment.

From the systems perspective, recent work has emphasized efficient ECG inference and lightweight deployment. Liu *et al.* [Liu et al. 2023] reported an FPGA-based ECG accelerator with 43.08$\times$ speedup over ARM Cortex-A53 and energy efficiency of 63.48 GOPS/W. An *et al.* [An et al. 2024] proposed a lightweight single-lead wearable model based on knowledge distillation, reporting 96.32% accuracy and a 1242.58$\times$ compression ratio. Gupta *et al.* [Gupta et al. 2024] reviewed efficient AI models for ECG rhythm classification and highlighted the need for lightweight and real-time solutions. However, prior work still lacks a unified analysis connecting patient-specific SSL personalization, selective quantization, and deployment-oriented evaluation. This is the gap addressed in the present paper.

3. Method

3.1. *ECGWavePuzzle* pipeline

The current study evaluates the *ECGWavePuzzle*-based pipeline introduced by Silva *et al.* [Silva et al. 2024]. The ECG signal is segmented into heartbeat-centered windows by detecting R-peaks, extracting beat-centered segments, and resampling each beat to 300 samples. No explicit denoising is applied, preserving acquisition conditions closer to a real world scenario.

The pretext task is *ECGWavePuzzle*, in which each heartbeat is divided into three parts and their order is permuted to generate six possible combinations, as shown in Figure 1. The model is trained to identify the correct permutation, turning the pretext stage into a six-class classification problem. This task encourages the network to learn ECG morphological structure before downstream classification [Silva et al. 2024].

Following Silva *et al.* [Silva et al. 2024], the workflow has two stages: an offline and online stage. In the **offline stage**, the model is pretrained on normal beats using *ECGWavePuzzle* in a high performance computer. In the **online stage**, the pretrained model is adapted to a target patient using the beats extracted from a five-minute recording. After adaptation, the pretext head is replaced with a three-class classification head and the network is fine-tuned for arrhythmia classification. The human-in-the-loop adaptation protocol follows the prior formulation of a patient-specific embedded pipeline.

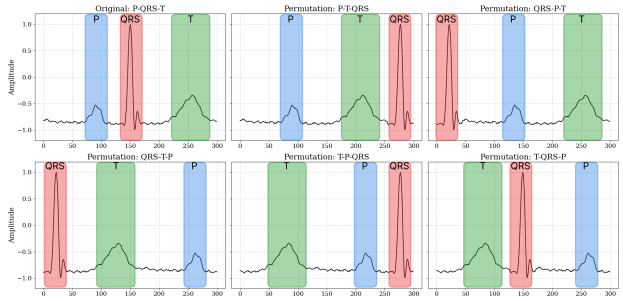

Figure 1. *ECGWavePuzzle* pretext task: the original sequence is split into three parts and rearranged into six possible permutations.

The backbone used throughout the pipeline is *ECGNet*, a one-dimensional convolutional neural network (1D-CNN) adopted from [Silva et al. 2024] and shared across both the *ECGWavePuzzle* pretraining stage and the downstream arrhythmia classification stage. The network receives heartbeat-centered ECG segments of length 300 as input. Its feature extractor is composed of five stacked 1D convolutional blocks. The first block contains 64 filters with kernel size 3, followed by a ReLU activation and a max-pooling operation. The second block contains 128 filters with kernel size 3, again followed by ReLU and max pooling. This pattern continues for three additional convolutional blocks with progressively larger numbers of filters, enabling the network to learn increasingly abstract morphological representations from the ECG waveform. After the convolutional stage, the feature maps are flattened and passed to a classifier composed of two fully connected layers: the first with 512 units followed by dropout with rate 0.5, and the second with 256 units and ReLU activation. The final output layer depends on the task: during *ECGWavePuzzle* pretraining, it contains 6 neurons corresponding to the six segment permutations, whereas during downstream learning it is replaced by a 3-neuron classification head for arrhythmia prediction.

3.2. Quantization and precision settings

Quantization is applied only in the **online stage**. This design is consistent with the intended deployment scenario: offline SSL pretraining can run once on high-performance hardware, whereas patient-specific adaptation and downstream inference is associated with edge execution. This separation allows the analysis to focus on the numerically constrained portion of the workflow that is most relevant to practical deployment. Three quantization configurations are then evaluated:

- **No quantization**: floating-point execution used as reference;
- **`head_int8`**: INT8/QAT applied only to the classification head;
- **`shallow_int8`**: INT8/QAT applied to the classification head and shallow backbone blocks.

Floating-point precision is analyzed separately from INT8 scope. In the non-quantized pipeline, we compare FP16 and BF16 to assess whether lower-precision floating-point execution changes the performance–runtime trade-off before introducing INT8 constraints.

4. Experimental Setup

All experiments use the MIT-BIH Arrhythmia Database [Goldberger et al. 2000], considering the modified Lead II. Evaluation follows the ANSI/AAMI EC57:1998/(R)2008

standard [ANSI/AAMI 2008] and the patient-wise split proposed by de Chazal *et al.* [De Chazal et al. 2004], ensuring that no patient overlap exists between training and testing sets and thus providing a more realistic assessment of inter-patient generalization. The offline stage is trained on DS1 with the *ECGWavePuzzle* pretext task, where the model learns the six possible segment permutations. The online stage is then carried out on DS2 on a per-patient basis, using each patient's heartbeats for *ECGWavePuzzle*-based personalization, followed by supervised fine-tuning on DS1 for arrhythmia classification. This protocol is consistent with the target scenario of adapting the model to unseen subjects before downstream diagnosis.

Two platforms are considered: **NVIDIA H200**, used as the high-performance reference, and **NVIDIA Jetson Nano**, used as the embedded-oriented platform. INT8/QAT experiments are reported only on H200, since the required conversion flow was not supported for the current pipeline on Jetson Nano. Results are reported as mean $\pm$ standard deviation over three seeds (7, 21, and 42).

5. Results and Discussion

5.1. RQ1: Mixed-Precision Effects

Table 1 presents the results of RQ1, which evaluates whether mixed-precision execution changes the non-quantized behavior of the *ECGWavePuzzle* pipeline. On H200, FP16 and BF16 yield nearly identical Accuracy (0.8904 vs. 0.8890), but FP16 improves Macro-F1 from 0.4969 to 0.5252 and slightly reduces runtime from 3003 s to 2920 s. This makes FP16 the best operating point among the tested floating-point settings. On Jetson Nano, FP16 preserves comparable predictive behavior, with Accuracy of 0.8941 and Macro-F1 of 0.5104, but runtime increases to about 60048 s, approximately $20\times$ slower than on H200. Therefore, mixed precision alone does not solve the deployment problem: model behavior is stable across platforms, but online execution remains expensive on embedded hardware.

Table 1. RQ1: mixed-precision results without INT8.

Platform	Precision	Accuracy	Macro-F1	Runtime (s)
H200	FP16	0.8904 ± 0.0211	$\mathbf{0.5252 \pm 0.0398}$	2920
H200	BF16	0.8890 ± 0.0094	0.4969 ± 0.0098	3003
Jetson Nano	FP16	$\mathbf{0.8941 \pm 0.0086}$	0.5104 ± 0.0119	60048

5.2. RQ2: Selective INT8/QAT Coverage

Table 2 presents the results of RQ2, which analyzes whether the *ECGWavePuzzle*-based pipeline remains effective when selective INT8/QAT is introduced into the supervised branch. The two quantization scopes produce very similar results for the SSL pipeline. The `head_int8` configuration reaches 0.9013 ± 0.0198 Accuracy and 0.5364 ± 0.0416 Macro-F1, whereas `shallow_int8` reaches 0.9003 ± 0.0248 Accuracy and 0.5258 ± 0.0408 Macro-F1. In practical terms, extending INT8/QAT from the head to the shallow backbone changes Accuracy by only 0.001 and Macro-F1 by about 0.011, indicating that selective INT8/QAT can be introduced without substantial degradation. This is the main result of the paper: the *ECGWavePuzzle* pipeline remains technically compatible with low-precision deployment.

Besides that, Table 2 also shows that the DS1-only baseline remains stronger than

Table 2. RQ2: selective INT8/QAT coverage on H200.

Pipeline	QAT scope	Accuracy	Macro-F1	Sensitivity
SSL + personalization	`head_int8`	0.9013 ± 0.0198	$\mathbf{0.5364 \pm 0.0416}$	$\mathbf{0.5434 \pm 0.0465}$
SSL + personalization	`shallow_int8`	0.9003 ± 0.0248	0.5258 ± 0.0408	0.5368 ± 0.0365
Baseline (DS1-only)	`head_int8`	0.9170 ± 0.0139	0.6261 ± 0.0307	0.6530 ± 0.0028
Baseline (DS1-only)	`shallow_int8`	$\mathbf{0.9194 \pm 0.0042}$	$\mathbf{0.6303 \pm 0.0165}$	$\mathbf{0.6542 \pm 0.0125}$

the personalized SSL pipeline under the current setting. The best baseline result, obtained with `shallow_int8`, reaches 0.9194 ± 0.0042 Accuracy and 0.6303 ± 0.0165 Macro-F1.

5.3. RQ3: Personalization Strategy Comparison

Table 3 presents the results of RQ3, which compares online personalization strategies before INT8 is introduced. The behavior of the personalization strategies is consistent across platforms. On H200, AdaBN improves Macro-F1 from 0.5272 to 0.5410 compared with head-only adaptation, but increases runtime from 2826.54 s to 3254.31 s and reduces Accuracy from 0.8911 to 0.8776. On Jetson Nano, the same pattern remains: AdaBN reaches the best Macro-F1 (0.5211), but runtime rises from 60288.91 s in head-only adaptation to 72776.07 s. Head-only adaptation yields the best or tied-best Accuracy on both platforms, while Adapters do not provide a competitive trade-off in the current setting.

Table 3. RQ3: personalization strategies on H200 and Jetson Nano.

Platform	Strategy	Accuracy	Macro-F1	Runtime (s)
H200	Head-only	$\mathbf{0.8911 \pm 0.0351}$	0.5272 ± 0.0250	**2826.54**
H200	AdaBN	0.8776 ± 0.0165	$\mathbf{0.5410 \pm 0.0172}$	3254.31
H200	Adapters	0.8455 ± 0.0365	0.5095 ± 0.0310	2828.55
H200	Last blocks	$\mathbf{0.8911 \pm 0.0351}$	0.5272 ± 0.0250	2831.71
Jetson Nano	Head-only	$\mathbf{0.8830 \pm 0.0249}$	0.4855 ± 0.0365	60288.91
Jetson Nano	AdaBN	0.8667 ± 0.0159	$\mathbf{0.5211 \pm 0.0151}$	72776.07
Jetson Nano	Adapters	0.8402 ± 0.0232	0.4705 ± 0.0161	**60003.61**

6. Conclusion

This paper evaluated the feasibility of applying selective INT8 quantization and mixed-precision execution to an *ECGWavePuzzle*-based personalization pipeline for arrhythmia classification. Using patient-wise experiments on MIT-BIH, we showed that the supervised branch of the pipeline is reasonably robust to selective INT8/QAT and that FP16 provides the best trade-off among the tested floating-point configurations.

The main contribution of this paper is therefore deployment-oriented evidence: *ECGWavePuzzle* can be translated to a reduced-precision setting without large predictive collapse, but practical edge use still depends on further optimization of the online adaptation stage. For computing applied to healthcare, this result is relevant because it indicates that hardware-aware simplifications can be introduced into SSL-based ECG analysis while preserving its core behavior. The evidence provided by this study also creates a basis for extending the analysis toward FPGA-based deployment and hardware acceleration, since it identifies which parts of the pipeline are compatible with reduced precision and which components still limit practical execution. As future work, we therefore intend to investigate FPGA-oriented implementations of the proposed methodology, with emphasis on latency, energy efficiency, and the feasibility of accelerating the online personalization stage in real-time personalized ECG monitoring scenarios.

Acknowledgments

The authors acknowledge the support of the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, project APQ-01768-24), the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), the *Universidade Federal de Ouro Preto* (PROPPI/UFOP), and its Graduate Program in Computer Science (PPGCC/UFOP).

References

An, X., Shi, S., Wang, Q., Yu, Y., and Liu, Q. (2024). Research on a lightweight arrhythmia classification model based on knowledge distillation for wearable single-lead ecg monitoring systems. *Sensors*, 24(24):7896.

ANSI/AAMI (2008). Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. American National Standards Institute, Inc. (ANSI), Association for the Advancement of Medical Instrumentation (AAMI). ANSI/AAMI/ISO EC57, 1998-(R)2008.

De Chazal, P., O'Dwyer, M., and Reilly, R. B. (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE transactions on biomedical engineering*, 51(7):1196–1206. Publisher: IEEE.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Gupta, U., Paluru, N., Nankani, D., Kulkarni, K., and Awasthi, N. (2024). A comprehensive review on efficient artificial intelligence models for classification of abnormal cardiac rhythms using electrocardiograms. *Heliyon*, 10(5):e26787.

Liu, H., Zhao, Z., and She, Q. (2021). Self-supervised ecg pre-training. *Biomedical Signal Processing and Control*, 70:103010.

Liu, W., Guo, Q., Chen, S., Chang, S., Wang, H., He, J., and Huang, Q. (2023). A fully-mapped and energy-efficient fpga accelerator for dual-function ai-based analysis of ecg. *Frontiers in Physiology*, 14:1079503.

Luz, E. J. d. S., Schwartz, W. R., Cámara-Chávez, G., and Menotti, D. (2016). ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer methods and programs in biomedicine*, 127:144–164. Publisher: Elsevier.

Mehari, T. and Strodthoff, N. (2022). Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114.

Phan, T., Le, D., Brijesh, P., Adjeroh, D., Wu, J., Jensen, M. O., and Le, N. (2022). Multimodality multi-lead ecg arrhythmia classification using self-supervised learning. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04. IEEE.

Silva, G., Silva, P., Moreira, G., Freitas, V., Gertrudes, J., and Luz, E. (2025). A systematic review of ecg arrhythmia classification: Adherence to standards, fair evaluation, and embedded feasibility. *arXiv preprint arXiv:2503.07276*.

Silva, G., Silva, P., Moreira, G., and Luz, E. (2024). Bridging the gap in ecg classification: Integrating self-supervised learning with human-in-the-loop amid medical equipment hardware constraints. In *International Symposium on Applied Reconfigurable Computing*, pages 63–74. Springer.

Silva, G. A. L. (2023). Self-supervised learning for arrhythmia classification. Thesis (master in computer science), Federal University of Ouro Preto, Ouro Preto.