DOI: 10.2507/36th.daaam.proceedings.033

DUAL-VIEW EXPLAINABLE AI FOR 12-LEAD ECG WITH OPENMP-ACCELERATED POSTPROCESSING AND EXTERNAL CALIBRATION

Bilgin Demir & Ervin Domazet

This Publication has to be referred as: Demir, B[ilgin] & Domazet, E[rvin] (2025). Dual-View Explainable AI for 12-Lead ECG With OpenMP-Accelerated Postprocessing and External Calibration, Proceedings of the 36th DAAAM International Symposium, pp.0244-0248, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-47-1, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/36th.daaam.proceedings.033

Abstract

Accurate electrocardiogram (ECG) classifiers are increasingly deployed, but practical use requires interpretable outputs and CPU-class latency. This study presents a lightweight pipeline that couples a compact 1D-CNN with dual-view explanations-Grad-CAM for time-localised saliency and Integrated Gradients (IG) reduced to per-lead importance and accelerates explanation postprocessing via a C++/OpenMP module. Internal benchmarking adopts the official PTB-XL split. External assessment uses the St. Petersburg INCART database with an adapt → calibrate (isotonic) → test protocol. On PTB-XL, the model attains AUC = 0.808 at the window level and AUC = 0.833 at the record level on the official test fold. On INCART, uncalibrated scores thresholded using the calibration subset yield ROC-AUC $\approx$ 0.84 and PR-AUC $\approx$ 0.85 on a small test cohort; isotonic calibration shifts the operating point toward higher accuracy ($\approx$ 0.70) with ROC-AUC $\approx$ 0.70 and PR-AUC $\approx$ 0.68. The OpenMP postprocessing reproduces NumPy exactly (max $|\Delta| = 0$) while reducing CPU runtime. Results indicate that accuracy, interpretable rationales, and CPU-feasible explanation latency can be achieved simultaneously with a disciplined internal/external evaluation design.

Keywords: Explainable AI; Grad-CAM; 12-lead ECG; OpenMP; Integrated Gradients

1. Introduction

Electrocardiography (ECG) remains a first-line diagnostic for cardiac disorders, yet clinical deployment of learning systems still wrestles with three practical requirements: accurate discrimination on public benchmarks, interpretable rationales that clinicians can audit, and CPU-class latencies suitable for bedside devices. Public 12-lead corpora such as PTB-XL[1] have standardised internal evaluation through patient-stratified folds and SCP-ECG–based labels, enabling like-for-like comparisons across studies [2]. At the same time, post hoc explanation methods adapted from computer vision-Grad-CAM and Integrated Gradients (IG)-offer complementary views of model evidence; Grad-CAM localises salient time segments, while IG attributes importance in an axiomatic manner. However, external validity and calibrated operation remain unevenly addressed; deployment often requires probability calibration (e.g., isotonic regression) and threshold selection on a dedicated cohort before testing on a never-touched set.

Finally, most reports optimise classification throughput but rarely quantify the runtime cost of explanations themselves, despite many monitors running on CPUs. This study presents a compact 1D-CNN with dual-view explanations and an OpenMP postprocessing path, evaluated on PTB-XL (internal benchmark) and INCART [3] (external protocol).

2. Related Work

Recent work has converged on PTB-XL as the reference benchmark for 12-lead ECG modelling because its official patient-stratified folds (1-8/9/10) and SCP-ECG–based labels enable reproducible, like-for-like comparisons across architectures [2]. PTB-XL+ extends this ecosystem with harmonised feature sets for the same records, supporting hybrid modelling and ablation studies while reinforcing standardised evaluation practice [4]. In parallel, explainability has progressed from image-centric tools to time-series ECG. Studies demonstrate that Grad-CAM can highlight diagnostically relevant intervals on multi-lead signals for myocardial infarction detection, and ECG-tailored variants such as ECGradCAM generate attention maps adapted to 12-lead morphology [5], [6]. At the same time, reviews emphasise that visually persuasive saliency is not automatically reliable: many cardiac-AI papers lack quantitative checks of explanation stability or consistency with preprocessing, motivating paired use of complementary methods and simple, auditable aggregation steps [7]. Clinical-facing work increasingly stresses external validation and probability calibration; for example, the AIRE platform is evaluated across sites with attention to actionable, reliable risk estimation from a single ECG [8], while statistical guidance formalises isotonic regression and related post hoc calibration to correct probabilities and transport thresholds under class imbalance or prevalence shift, with explicit separation of calibration and test cohorts [9]. A complementary strand targets efficiency for edge/CPU deployment: compact or knowledge-distilled CNNs reduce latency and power draw and convincingly optimise classification throughput, but most reports do not measure the runtime cost of explanations themselves (e.g., saliency smoothing and per-lead aggregation) on CPUs, where many bedside systems actually run [10], [11]. High-performance ECG filtering demonstrates an emphasis on deployable, CPU-class signal-processing pipelines [12]. Saliency-based interpretability has also been explored in applied control settings, underscoring practical demands for transparent models [13]. Unlike prior work by the authors focusing on classification speed/accuracy [14], the present study integrates dual-view explainability, calibrated external testing, and CPU-feasible XAI postprocessing. Against this backdrop, this study adds value by integrating dual-view ECG-XAI (time-localised Grad-CAM with per-lead attributions), an explicit adapt → calibrate (isotonic) → test protocol on a different database, and a CPU-practical OpenMP implementation for explanation postprocessing, thereby closing the gap between interpretability, external validity, and deployable runtime that prior work typically treats in isolation.

3. Methods

Here we evaluate a lightweight, interpretable ECG pipeline under an internal–external design: internal benchmarking on the PTB-XL official split and external assessment on INCART[3]. Signals are standardised and windowed; a compact 1D-CNN produces window scores that are averaged to records, and explanations are dual-view Grad-CAM for time-localised saliency and Integrated Gradients aggregated into per-lead importance. To keep latency CPU-feasible, explanation postprocessing is implemented in C++/OpenMP with numerical parity to NumPy. External operation uses isotonic regression on a dedicated calibration split with a fixed threshold transported to the held-out test split.

3.1. Datasets and splits

Internal benchmarking used PTB-XL (12-lead ECG, 10 s, 500 Hz) with the official patient-stratified folds 1–8/9/10 for training/validation/test to ensure comparability with the literature [2]. External assessment used the St. Petersburg INCART database (12-lead ECG with beat annotations). Records were partitioned into three disjoint sets: adaptation (for distributional normalisation choices), calibration (for probability calibration and threshold selection), and test (held-out evaluation). The binary task on PTB-XL followed the dataset's diagnostic mapping: normal vs abnormal (any diagnostic class other than purely normal) [1].

3.2 Preprocessing and windowing

Signals were read via WFDB, reordered to a fixed 12-lead layout (I, II, III, aVR, aVL, aVF, V1–V6), band-pass filtered (0.5–40 Hz), and standardised by per-record z-scoring. Each record was segmented into overlapping windows of 512 samples with 256-sample stride. This corresponds to ≈1.02 s per window at 500 Hz (PTB-XL) and ≈1.99 s at 257 Hz (INCART). Inference operated at the window level; record-level scores were obtained by averaging window probabilities.

3.3. Model Architecture and Training

A compact 1D-CNN with residual connections was used: Conv1D-BatchNorm-ReLU blocks with skip connections, intermittent max-pooling, global average pooling, dropout (0.25), and a sigmoid output.

Optimisation used Adam (1×10^{-3}) with binary cross-entropy and class weighting computed on the training fold. Early stopping monitored validation ROC-AUC. Models were saved and reloaded using Keras' native serialisation for exact reproducibility.

3.4. Explainability

Two complementary post hoc methods were applied. Grad-CAM was computed from the final convolutional block to obtain time-localised saliency on 1-D signals. Integrated Gradients (IG) used a zero baseline with 64-128 integration steps to obtain attribution scores with axiomatic guarantees.

3.5. Parallel Postprocessing (OpenMP)

The module was compiled with -O3 -fopenmp (aggressive optimisation; OpenMP support) and run with OMP_PROC_BIND=true, OMP_PLACES=cores, and schedule(static) (core pinning and even work sharing to minimise overhead). The C++ outputs were verified to be numerically identical to the NumPy reference (maximum absolute difference = 0). Design choices for OpenMP parallelisation followed prior real-time ECG acceleration work by the authors and were adapted here to the XAI postprocessing stage.

3.6. Calibration and thresholding (external)

For INCART, raw scores were transformed using isotonic regression fit on the calibration split only. The decision threshold was selected on that same calibration split (F1 maximisation) and then applied unchanged to the held-out test split. Uncalibrated metrics at the calibration-chosen threshold were also reported for comparison.

3.7. Evaluation

Internal PTB-XL reporting included ROC-AUC and accuracy at the window level and after record-level averaging. External INCART reporting included ROC-AUC, PR-AUC, accuracy, precision, recall, and F1. To reflect the limited size of the external test cohort, bootstrap 95% confidence intervals were computed at the record level. All metrics were computed with scikit-learn under fixed seeds and batched inference settings. Isotonic regression (a monotonic mapping) was fit at the record level on the calibration split and applied to record-level scores on test.

4. Results

On the PTB-XL official split (training on folds 1-8, validation on fold 9, and testing on fold 10), the compact 1D-CNN achieved strong discrimination internally. At the window level, the model reached an AUC of 0.829 on validation and 0.808 on test, with corresponding accuracies of 0.744 and 0.732. Aggregating decisions to the record level by averaging window probabilities further improved performance: validation AUC rose to 0.854 and test AUC to 0.833, with accuracies of 0.764 and 0.759, respectively. Score polarity was consistently verified throughout the analysis, with no instances of inversion being necessary during the final evaluation. Each evaluation metric was computed on the officially established data folds, thereby maintaining full comparability with earlier studies conducted on the PTB-XL dataset.

For external evaluation, we assessed the frozen PTB-XL model on the INCART database using a three-way record split-adaptation, calibration, and test-to emulate deployment without re-training. Uncalibrated scores, thresholded using the calibration split, yielded a test ROC-AUC of approximately 0.840 and PR-AUC of 0.853, with accuracy around 0.50 on ten held-out records. Applying isotonic regression on the calibration split adjusted probability estimates and operating points: on the same test split the ROC-AUC was ~0.700 and PR-AUC ~0.683, with accuracy ~0.70 at the calibration-selected threshold. Bootstrap analysis at the record level (≈2000 resamples) reflected the small test set: the uncalibrated AUC had a mean of ~0.844 with a 95% CI [0.50, 1.00], and the isotonic-calibrated AUC had a mean of ~0.704 with a 95% CI [0.375, 1.00].

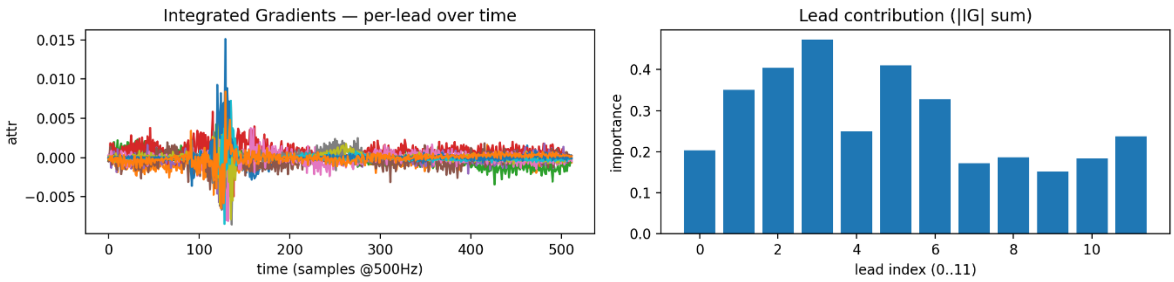

Fig. 1. Integrated Gradients (IG) on a representative record. Left: per-lead attributions over time after smoothing (kernel = 9); Right: per-lead contribution summary (IG summed across time). IG uses a zero baseline and 128 integration steps.

The explanation analysis combined Integrated Gradients (IG) and Grad-CAM to localise model evidence in time and across leads. IG bar plots of per-lead importance indicated lead-specific contributions that varied by case, while Grad-CAM heatmaps overlaid on the raw ECG highlighted short temporal segments around detected events; representative true-positive and false-positive examples illustrate these patterns. These views provide complementary insight-IG offers a global attribution per lead, and Grad-CAM pinpoints salient temporal intervals-without making clinical claims about causality. An example Integrated Gradients attribution is shown in Figure 1, with per-lead importance (right) and the time course of attributions (left). Complementary Grad-CAM temporal saliency is illustrated in Figure 2, highlighting short intervals that drove the decision.

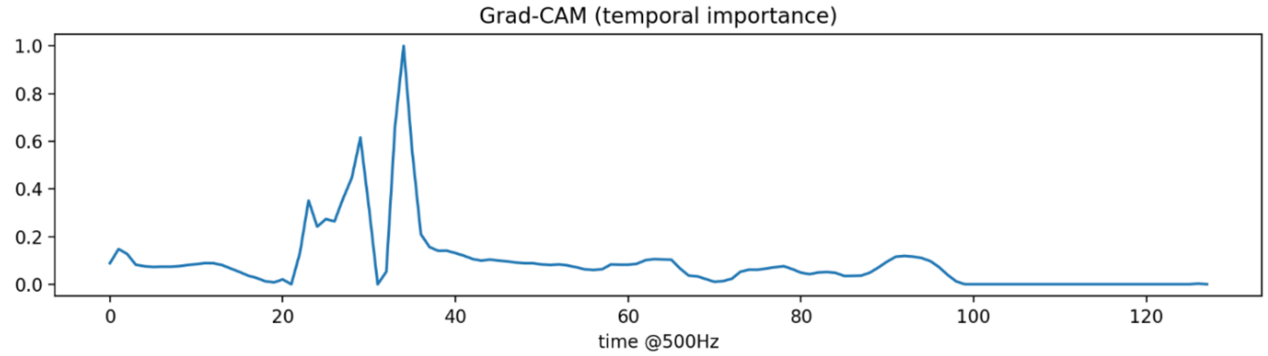

Fig. 2. Grad-CAM temporal importance for the same case as Figure 1. Saliency is computed from the final convolutional block and normalised to [0,1]; peaks align with short, decision-driving segments.

The OpenMP module reproduced NumPy exactly (max |Δ| = 0) and met the sub-30 ms latency target on CPU. For the timing batch used here, single-thread execution was fastest; adding threads increased overhead and did not improve latency (Figure 3). Nevertheless, C++ delivered large gains over pure-Python loops (×49 at 1 thread; Figure 4). This indicates the explanation workload is small enough that parallel start-up and scheduling costs dominate; parallel benefits are expected only for larger batches.

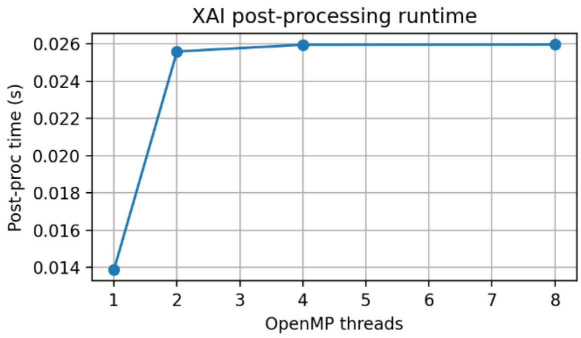

Fig. 3. Postprocessing latency vs OpenMP thread count (median over repeated runs; fixed affinity). Latency is minimised at 1 thread for this workload.

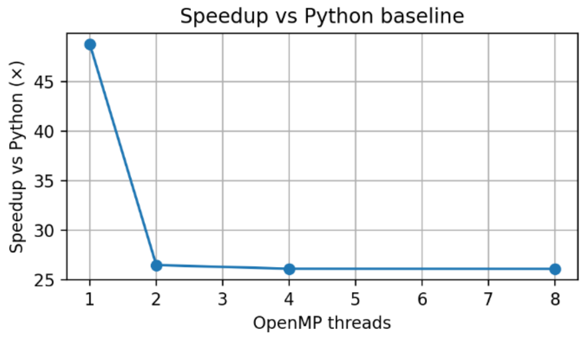

Fig. 4. Speedup relative to a pure-Python baseline. Single-threaded C++ already provides ×49; multithreading does not help at this batch size

5. Discussion

The internal PTB-XL evaluation shows that a compact 1D-CNN, trained with standard preprocessing and record-level aggregation, achieves strong discrimination on a widely used benchmark. On the external INCART cohort, calibration and strict cohort separation were essential. Uncalibrated scores paired with a threshold chosen on the calibration split yielded modest accuracy; isotonic calibration shifted the operating point toward a more balanced precision–recall trade-off. The dual-view explanation design provides complementary insight. Grad-CAM highlights short, decision-driving time segments, while Integrated Gradients-after simple smoothing-yields per-lead importance vectors, making both “when” and “which lead” auditable. From an engineering perspective, the OpenMP postprocessing reproduces the NumPy reference exactly (max |Δ| = 0) and meets the latency target on a single CPU core. Limitations include a small external cohort, a binary PTB-XL task (normal vs abnormal) rather than multi-label diagnostics, and qualitative assessment of explanations; expanding external datasets, moving to multi-label targets, and conducting stability and reader studies are natural next steps.

6. Conclusion

This study consolidates three capabilities, typically approached in isolation for ECG machine learning: in situ transparent benchmarking, reasoned interpretation with explicit traceability, and deployment-ready execution within compact runtime constraints. Evaluated on the standardised PTB-XL division, our compact one-dimensional convolutional network exhibited robust class separation; decision sharpening through record-level averaging yielded marked stability against variances in window-based outputs. On INCART, strict cohort separation (adapt → calibrate → test) and isotonic calibration illustrated how probability mapping and threshold transport affect operating points under class imbalance-an essential step for deployment beyond the development dataset. Limitations include the small size of the external test cohort (wide uncertainty) and the binary task formulation on PTB-XL. Future work should expand to additional external databases, adopt multi-label diagnostic targets, quantify explanation stability (e.g., perturbation and test–retest analyses), and assess human factors through reader studies; exploring SIMD/AVX or GPU variants could further broaden deployment options. Overall, the results support a practical pathway wherein accuracy, calibrated external validity, and auditable explanations are jointly attainable on CPU-class hardware.

7. References

[1] PTB-XL Database, PhysioNet, (2020). Available from: https://physionet.org/content/ptb-xl/ Accessed: 2025-09-06

[2] Wagner, P.; Strodthoff, N.; Bousseljot, R.; et al. (2020). PTB-XL: A large publicly available electrocardiography dataset. Scientific Data, Vol. 7, No. 1, (May 2020) Article 154, ISSN 2052-4463, DOI:10.1038/s41597-020-0495-6

[3] INCART Database, St Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database, PhysioNet, (2008). Available from: https://physionet.org/content/incartdb/ Accessed: 2025-09-06

[4] Strodthoff, N.; Wagner, P.; Schaeffter, T.; et al. (2023). PTB-XL+: A comprehensive electrocardiographic feature dataset. Scientific Data, Vol. 10, No. 1, (May 2023) Article 279, ISSN 2052-4463, DOI:10.1038/s41597-023-02153-8

[5] Jahmunah, V.; Oh, S. L.; San, T. R.; et al. (2022). Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Computers in Biology and Medicine, Vol. 146, (July 2022) Article 105550, ISSN 1879-0534, DOI:10.1016/j.compbiomed.2022.105550

[6] Hicks, S. A.; Isaksen, J. L.; Thambawita, V.; et al. (2021). Explaining deep neural networks for knowledge discovery in electrocardiogram analysis (ECGradCAM). Scientific Reports, Vol. 11, No. 1, (May 2021) Article 10949, ISSN 2045-2322, DOI:10.1038/s41598-021-90285-5

[7] Salih, A. M.; Lundervold, A.; Oehme, A.; et al. (2024). A review of evaluation approaches for explainable AI with applications in cardiology. Artificial Intelligence Review, Vol. 57, No. 9, (September 2024) Article 240, ISSN 1573-7462, DOI:10.1007/s10462-024-10852-w

[8] Sau, A.; Pastika, L.; Sieliwonczyk, E.; et al. (2024). Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: A model development and validation study. The Lancet Digital Health, Vol. 6, No. 11, (November 2024) e791–e802, ISSN 2589-7500, DOI:10.1016/S2589-7500(24)00172-9

[9] Ojeda, F. M.; Kappert, P.; Schulte-Steinberg, S.; et al. (2023). Calibrating machine-learning approaches for probability estimation: A comprehensive comparison. Statistics in Medicine, Vol. 42, No. 29, (December 2023) pp. 5451–5478, ISSN 1097-0258, DOI:10.1002/sim.9921

[10] An, X.; Shi, S.; Wang, Q.; et al. (2024). Research on a lightweight arrhythmia classification model based on knowledge distillation for wearable single-lead ECG monitoring systems. Sensors, Vol. 24, No. 24, (December 2024) Article 7896, ISSN 1424-8220, DOI:10.3390/s24247896

[11] Lee, K.; Lee, J.; Shin, M. (2024). Lightweight beat score map method for electrocardiogram-based arrhythmia classification. Biocybernetics and Biomedical Engineering, Vol. 44, No. 4, (October 2024) pp. 844–857, ISSN 0208-5216, DOI:10.1016/j.bbe.2024.11.002

[12] Domazet, E.; Gusev, M. & Ristov, S. (2016). Optimizing High-Performance CUDA DSP Filter for ECG Signals, Proceedings of the 27th International Symposium, pp. 0623–0631. DOI: 10.2507/27th.daaam.proceedings.091

[13] Schwaiger, S.; Aburaia, M.; Aburaia, A. & Woeber, W. (2021). Explainable Artificial Intelligence for Robot Arm Control, Proceedings of the 32nd International Symposium. DOI: 10.2507/32nd.daaam.proceedings.090

[14] Demir, B. & Tafa, Z. (2025). Contribution to the development of intelligent systems for multi-lead ECG signal classification using convolutional neural networks. Proceedings of the 14th Mediterranean Conference on Embedded Computing (MECO), June 2025, Budva, Montenegro, ISBN 979-8-3315-1341-2, pp. 1–5, IEEE, DOI:10.1109/MECO66322.2025.11049271