Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TensorECG: Adaptive-Rank Temporal–Spatial Tensor Compression of Multi-lead ECG Signals with Wavelet and low bit residual refinement

Ashish Pratap

ashish.pratapphd@nituk.ac.in

National Institute of Technology

Hariharan Muthusamy

National Institute of Technology

S. D. Sudarsan

C-DAC

Ramesh Naidu Laveti

C-DAC

Research Article

Keywords: ECG compression, tensor-train decomposition, multi-lead electrocardiography, wavelet transform, rate-distortion, adaptive-rank compression

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11396766/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TensorECG: Adaptive-Rank Temporal–Spatial Tensor Compression of Multi-lead ECG Signals with Wavelet and low bit residual refinement

Ashish Pratap[1], Hariharan Muthusamy[2*], S. D. Sudarsan[3], Ramesh Naidu Laveti[4]

[1,2]Department of Electronics Engineering, National Institute of Technology, Uttarakhand, India.

[1,3,4]Centre for Development of Advanced Computing, C-DAC, Bengaluru, India.

*Corresponding author(s). E-mail(s): hariharanm@nituk.ac.in;
Contributing authors: ashish.pratapphd@nituk.ac.in; sds@cdac.in; rameshl@cdac.in;

Abstract

The rapid growth of long-duration ECG recordings from wearable and ambulatory monitors has created a compression bottleneck for wireless transmission and storage. Existing block- and transform-based methods largely exploit redundancy along only one or two dimensions, offering limited joint modeling of inter-beat and inter-lead correlations. TensorECG addresses this gap by introducing a three-way tensor representation for multi-lead ECG that extends conventional matrix-based approaches to three physiologically meaningful axes: a temporal axis capturing self-similarity across successive heartbeats, a spatial axis capturing inter-lead correlations arising from different projections of cardiac electrical activity, and a morphological axis representing individual beat waveform shapes. A discrete wavelet transform is applied to each beat prior to tensorization, concentrating signal energy into few significant coefficients and reducing the effective dimensionality of the morphological axis. These three dimensions are jointly represented in a unified tensor rather than modeled separately, and the resulting coefficient tensor is compressed via rank-r Tensor-Train SVD, jointly capturing beat and lead correlations. TT-core coefficients are entropy coded, with an optional low-bit residual-refinement layer entropy-coding reconstruction error for additional rate–fidelity control. TT-rank, the key sensitive hyperparameter, is tuned once on a 35-record MIT-BIH training split and fixed for testing on a held-out split and three PhysioNet databases (2–15 leads, 250–1000 Hz). An adaptive-rank DWT+TT variant achieves compression ratios of 18.9–51.6 across 4–12% PRD targets and clinical validation showing sub-millivolt amplitude and sub-millisecond interval errors

Keywords: ECG compression, tensor-train decomposition, multi-lead electrocardiography, wavelet transform, rate-distortion, adaptive-rank compression

1 Introduction

Modern cardiac monitoring is expanding along two axes simultaneously: the number of multiple leads captured per device and the duration of recording, which now expand from minutes to weeks. This dual growth has driven an exponential rise in the volume of ECG data generated and transmitted, evident across Holter monitors, implantable loop recorders, ICU telemetry systems, and consumer wearables. A contemporary 12-lead Holter system, sampling at 500 Hz with 11–12 bit resolution over a 24-hour period, generates approximately 750 MB to 1 GB of raw signal per patient per day. When this data is scaled across an entire monitored population, and further multiplied by multi-day or multi-week recording protocols common in long-term ambulatory monitoring, the cumulative data volume grows into the tens of gigabytes per patient — placing significant strain on wireless transmission bandwidth, storage infrastructure, and the battery life of power-constrained wearable devices [1, 2]. As a result, ECG compression has progressed from direct time-domain predictors (AZTEC, turning-point encoding) to transform coders based on the discrete cosine transform (DCT), the discrete wavelet transform (DWT), and wavelet-tree coders such as Set Partitioning In Hierarchical Trees (SPIHT) and Embedded Zerotree Wavelet (EZW), all of which exploit redundancy between successive beats

and, to a lesser extent, between leads [3, 4]. What these studies have in common is that they compress one lead at a time. This leaves two additional redundant axes unexploited in a multi-lead recording: (i) beat-to-beat redundancy, which arises from the highly self-similar morphology and timing of subsequent heartbeats from the same patient, and (ii) lead-to-lead redundancy, which arises from the twelve (or fewer) leads of a standard ECG being distinct linear projections of the same underlying cardiac dipole and therefore strongly correlated. Several approaches stack consecutive beats into a 2D matrix, applying 2D DCT or wavelet transforms (analogous to JPEG compression) to exploit inter-beat redundancy [5, 6]. A smaller number of studies have used principal component analysis or independent component analysis along the lead axis to exploit inter-lead redundancy [7]. The mathematical tool for this problem is tensor decomposition, which represents a multi-dimensional array (in this case, beats $\times$ leads $\times$ samples) as a small set of factors, one per axis, whose sizes control how much redundancy along each axis is retained. The Tensor-Train (TT) decomposition [8] is especially well suited to signal-compression techniques among the family of tensor factorizations because (i) its storage cost in the tensor order is linear rather than exponential, (ii) it can be computed by a numerically stable sequence of SVDs (TT-SVD), and (iii) its ranks have a direct operational meaning as the number of "modes" retained along each transition between axes, giving a natural distortion knob. Tensor techniques have previously proven effective for feature extraction and classification in biomedical applications; however, their systematic application to multi-lead ECG signal compression remains largely unexplored [9, 10]. Previous work reports strong compression alongside incidental denoising, but generally lacks a rank-selection protocol with held-out validation, a rate-distortion characterization against classical (wavelet/DCT) baselines, and an evaluation of clinical fidelity beyond visual waveform inspection. The primary contributions of this study are as follows.

(i) **A hybrid wavelet–tensor-train compression pipeline (TensorECG)** that arranges a block of consecutive, multi-lead heartbeats into a three-way tensor of per-lead DWT coefficients applies TT decomposition jointly across the beat and lead axes, and entropy encodes the resulting TT core values.

(ii) **An adaptive-rank variant** that meets a target distortion directly by searching a rank grid at run time, and a **residual-refinement stage** that adds a small uniformly quantized correction on top of the DWT+TT reconstruction, both of which we show substantially improve the achievable rate–distortion–complexity trade-off.

(iii) **An extensive, statistically rigorous evaluation** comprising a component-wise ablation that separates the multiplicative contribution of the wavelet front, the TT core, and entropy coding; a per-record interpolated rate-distortion study against four baselines (DCT+Huffman, DWT+Huffman, block-2-D-DCT baseline, and TT-only); and a clinically focused evaluation of R-peak amplitude, QRS duration, ST-segment deviation, T-wave amplitude/location, and QT-interval preservation, with paired statistical tests on each metric.

(iv) **A patient-disjoint protocol for selecting the TT-rank**. The rank, identified as the pipeline's primary sensitive hyperparameter, is grid-searched only on the training split of 35 MIT-BIH records and then fixed for evaluation, with no further tuning or patient overlap, on a held-out test split of 13 MIT-BIH records and three independent PhysioNet databases.

The remainder of the paper is organized as follows. Section 2 reviews classical and tensor-based ECG compression. Section 3 introduces the notation and mathematical preliminaries for the DWT and the TT decomposition. Section 4 describes the TensorECG pipeline, the adaptive-rank variant, and the residual-refinement stage. Section 5 details the datasets, the train/test protocol, the metrics, and the baselines. Section 6 reports and discusses the experimental results, Section 7 discusses the findings, and Section 8 discusses limitations and future work before Section 9 concludes.

2 Related Work

2.1 Classical ECG compression

The turning-point algorithm, coordinate reduction time encoding (CORTES), and Amplitude Zone Time Epoch Coding (AZTEC) [11] are examples of direct (time-domain) ECG compression techniques that predict each sample from its predecessors and encode only the prediction residual. They are computationally inexpensive but achieve relatively modest compression ratios and are known to distort diagnostic features. Prior work is categorized into direct, transform, and parameter-extraction families [12], a taxonomy that remains a useful benchmark for evaluating a unified approach to ECG compression. Transform coders decorrelate the signal across an orthogonal basis, then discard or coarsely quantize small coefficients. The Discrete Cosine Transform (DCT) is particularly useful for ECG signal compression via optimized coefficient quantization [13]. The discrete wavelet transform (DWT), which is based on multiresolution theory [14] and compactly

supported wavelet families [15], has become important for ECG because its basis matches the signal's structure, such as sharp QRS transients and smooth ST/T morphology. Subsequent research expanded on these methods by introducing explicit rate-control mechanisms [16] to meet bandwidth limits and incorporating quality-guarantee protocols [17–19] to ensure that diagnostic ECG features remain uncompromised after reconstruction and ensuring quality Another study[20] proposed a weighted-diagnostic-distortion measure and an associated PRD-based quality taxonomy which is used for TT-rank selection criterion — percentage root-mean-square difference (PRD) emerged as the predominant distortion metric in the literature. Finally, lossless compression is achieved via entropy coding, grounded in Shannon's source-coding theorem [21] [22].

2.2 Tensor decompositions for biomedical signals

In biomedical signal analysis, low-rank tensor formats—such as Tucker decomposition [23], Higher-Order SVD [24], CANDECOMP/PARAFAC (CP), and Tensor-Train (TT)—are widely used for multidimensional factorization. These representations are particularly effective for tasks like source separation, artifact removal, and feature extraction. Consequently, evaluating their numerical trade-offs is essential for selecting the optimal tensor format. The Tensor-Train format is particularly attractive for *compression* because TT-SVD constructs the decomposition through a sequence of truncated SVDs, allowing the approximation error to be controlled through the selected TT-ranks and the associated discarded singular values. Moreover, for fixed or bounded TT-ranks, its storage complexity scales linearly with tensor order, unlike the Tucker representation, whose core size can grow exponentially with tensor order. Despite its suitability, the systematic application of TT and related tensor decompositions to lossy, rate-controlled multi-lead ECG compression — with quantitative comparison against standard per-lead and block-2D baselines under matched compression ratios and a principled, leakage-free rank-selection procedure. To address this gap, we propose TensorECG, with the TT-rank selected exclusively on a disjoint training split and fixed before evaluation on held-out records and independent databases.

3 Preliminaries

3.1 Notation

We denote a block of B consecutive heartbeats, each with L leads and S time samples, as a three-way tensor $\mathcal{X} \in \mathbb{R}^{B\times L\times S}$, with entries $\mathcal{X}(b,\ell,s)$. We use $\hat{\mathcal{X}}$ for a reconstruction of $\mathcal{X}$ after lossy compression.

3.2 Discrete Wavelet Transform

For a single beat/lead signal $x \in \mathbb{R}^S$, an ℓ-level DWT with a compactly supported orthogonal wavelet (we use Daubechies-4, "db4") produces one approximation band and ℓ detail bands,

$$c = \mathcal{W}_\ell(x) = [c_{A_\ell}, c_{D_1}, \ldots, c_{D_\ell}] \in \mathbb{R}^K, \tag{1}$$

with inverse $x = \mathcal{W}_\ell^{-1}(c)$ recovering the signal exactly when no coefficients are discarded. Classical wavelet ECG coders threshold this coefficient vector c, retaining only its $k = \lceil \rho K \rceil$ largest-magnitude entries (ρ being the keep ratio), and entropy-code the survivors; this is the DWT+Huffman baseline used in Section 5.

3.3 Tensor-Train Decomposition

The Tensor-Train (TT) format represents a d-way tensor $\mathcal{Y} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ as a chain of third-order core tensors $\mathcal{G}_k \in \mathbb{R}^{r_{k-1}\times n_k\times r_k}$, $k = 1,\ldots,d$, with boundary ranks $r_0 = r_d = 1$, such that

$$\mathcal{Y}(i_1,\ldots,i_d) \;=\; \mathcal{G}_1[i_1]\,\mathcal{G}_2[i_2]\cdots\mathcal{G}_d[i_d], \tag{2}$$

where $\mathcal{G}_k[i_k] \in \mathbb{R}^{r_{k-1}\times r_k}$ is the i_k-th (matrix) slice of $\mathcal{G}_k$ and the right-hand side is a chain of matrix products collapsing to a scalar. The tuple $\mathbf{r} = (r_0, r_1, \ldots, r_d)$ is the TT-rank. Exact representation requires r_k up to $\min\big(\prod_{j\le k} n_j,\ \prod_{j>k} n_j\big)$; truncating each r_k to a smaller value yields a controllable, best-in-Frobenius-norm-per-unfolding approximation computable by a sequence of $d-1$ truncated SVDs (TT-SVD), with total storage

$$\mathrm{numel}(\mathcal{G}_1,\ldots,\mathcal{G}_d) \;=\; \sum_{k=1}^{d} r_{k-1}\,n_k\,r_k \;=\; O\big(d\,n\,r^2\big), \tag{3}$$

for $n = \max_k n_k$ and $r = \max_k r_k$ — linear in d and n, in sharp contrast to the $O(n^d)$ cost of the raw tensor. In our setting $d = 3$ (beats, leads, wavelet-coefficient index), and the rank tuple reduces to $\mathbf{r} = (1, r_1, r_2, 1)$; we tie $r_1 = r_2 = r$ throughout, so a single scalar rank r controls both internal transitions — this is what we mean below by "TT-rank r."

4 Proposed Method: TensorECG

4.1 Overview

In the TensorECG pipeline, raw multi-lead ECG is band-pass filtered (0.5–40 Hz, 4th-order Butterworth, zero-phase), segmented into fixed-length beats around detected/annotated R-peaks, and grouped into blocks of B consecutive beats as illustrated in Figure 1. Each beat in a block is wavelet-transformed per lead (Eq. 1), producing a coefficient tensor $\mathcal{C} \in \mathbb{R}^{B\times L\times S}$. Performing TT decomposition after the wavelet front end lets the low-rank structure act on an already-sparsified, decorrelated representation rather than on raw time samples. The TT-rank r is selected by sweeping candidate ranks on a held-out validation set and tracing the resulting compression-ratio–distortion trade-off curve; the chosen rank is then fixed across all test records rather than re-tuned per record, so that the reported performance reflects a deployable, leakage-free operating point. The resulting TT cores are uniformly scalar-quantized, with the quantization step size Δ chosen to balance the rate–distortion trade-off, and the quantized coefficient stream is losslessly compressed with Huffman coding to produce the final bitstream for transmission or storage.

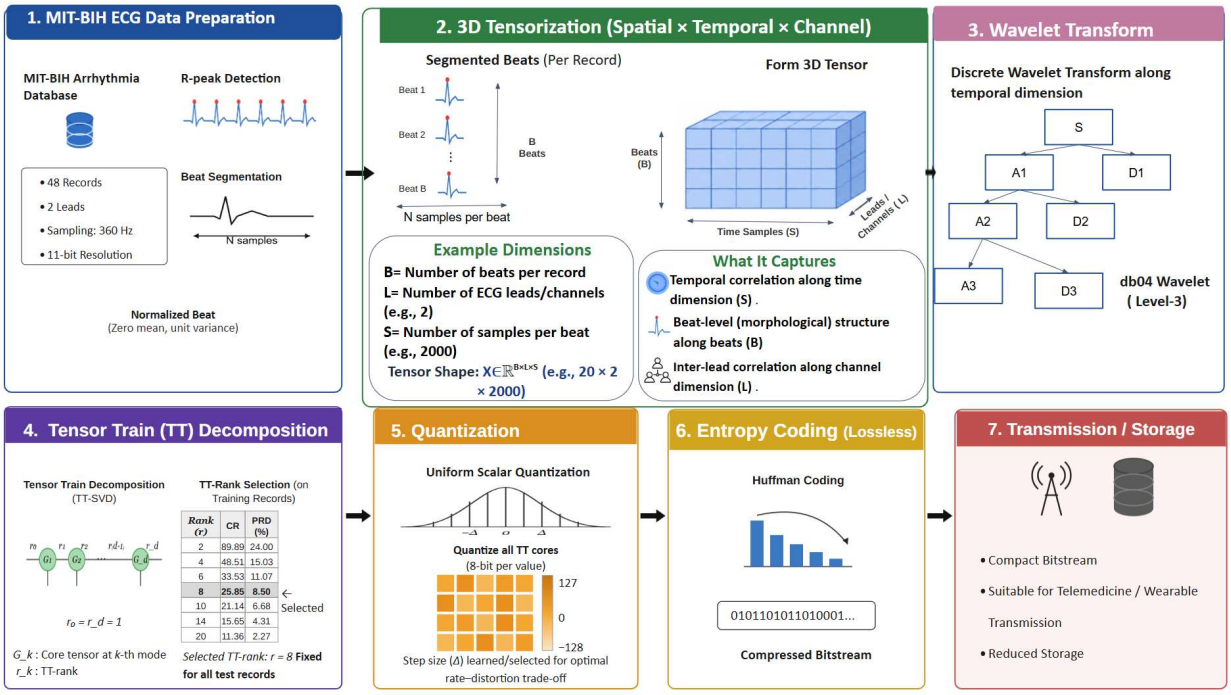


Fig. 1 Overview of the proposed TensorECG compression pipeline. The raw multi-lead ECG is band-pass filtered, segmented into fixed-length beats, and grouped into blocks of B consecutive beats. Each beat is transformed independently using a wavelet transform to construct the coefficient tensor $\mathbf{C} \in \mathbb{R}^{B\times L\times S}$. The resulting wavelet-domain tensor, rather than the raw beat tensor, is compressed using Tensor-Train (TT) followed by quantization and entropy coding.

4.2 Tensorization

Before the wavelet front end is applied, the segmented beats within a block are assembled into a single three-way tensor. For a block of B consecutive beats, each spanning L leads and S time samples per beat, the raw beat tensor is defined as

$$\mathcal{X} \in \mathbb{R}^{B\times L\times S}, \qquad \mathcal{X}(b,\ell,s) = x_b^{(\ell)}(s), \tag{4}$$

where $x_b^{(\ell)}(s)$ is the s-th sample of the ℓ-th lead of the b-th beat in the block. This is the same notation introduced; the tensorization step here is simply the construction of $\mathcal{X}$ from the per-beat, per-lead signals produced by R-peak detection and beat segmentation.

Each of the three axes is chosen to correspond to a distinct, physiologically meaningful source of redundancy in multi-lead ECG:

- **Beat axis (B).** Successive heartbeats from the same patient share highly similar morphology and timing, so consecutive slices $\mathcal{X}(b,:,:)$ are strongly correlated. This is the redundancy exploited by beat-to-beat (temporal) compression.

- **Lead axis (L).** The L leads are distinct linear projections of the same underlying cardiac dipole, so slices $\mathcal{X}(:, \ell, :)$ across leads are correlated rather than independent. This is the inter-lead (spatial) redundancy that per-lead coders discard.
- **Time-sample axis (S).** Within a single beat, the waveform has a well-defined morphological structure (P wave, QRS complex, T wave) that is itself compressible; this is the axis the wavelet front end subsequently sparsifies before TT decomposition acts on it.

Representing the block as a single tensor $\mathcal{X}$, rather than as $B \cdot L$ independent one-dimensional signals, is what allows the beat and lead axes to be jointly compressed by TT-SVD rather than requiring each redundancy to be exploited by a separate, hand-designed stage, as in the per-lead or 2-D block transform baselines

4.3 TT compression of the coefficient tensor

Given $\mathcal{C} \in \mathbb{R}^{B \times L \times S}$ and a target rank $\mathbf{r} = (1, r, r, 1)$, TT-SVD (Section 3) produces core tensors $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ and a reconstruction $\hat{\mathcal{C}} = \mathrm{TT}^{-1}(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)$ via Eq. 2. Each beat/lead's coefficient vector $\hat{\mathcal{C}}(b, \ell, :)$ is then inverse-wavelet-transformed (Eq. 1) to obtain the reconstructed beat $\hat{\mathcal{X}}(b, \ell, :)$. The single scalar r is the only knob exposed to the operator: increasing r monotonically shrinks reconstruction error at the cost of larger core tensors, giving a single-parameter rate–distortion trade-off in place of the multiple independent thresholds (wavelet level, coefficient-keep fraction, per-subband quantization step) that classical wavelet coders require. Because $\mathcal{G}_1$ and $\mathcal{G}_3$ carry boundary rank 1 by construction, only $\mathcal{G}_2$'s storage grows with r, which keeps the parameter count of DWT+TT linear in r rather than quadratic, as it would be for a bare (non-train-structured) low-rank factorization of the same tensor.

4.4 Quantization and Entropy Coding

The concatenated TT core values $\mathbf{g} = \left[\mathrm{vec}(\mathcal{G}_1); \mathrm{vec}(\mathcal{G}_2); \mathrm{vec}(\mathcal{G}_3)\right]$ are uniformly quantized to M levels with step size Δ,

$$q = \left\lfloor \frac{g - g_{\min}}{\Delta} \right\rceil, \qquad \hat{g} = q\,\Delta + g_{\min}, \tag{5}$$

where $M = 256$ (8-bit resolution, matching the $[-128, 127]$ quantizer illustrated in Figure 1) and $\Delta = (g_{\max} - g_{\min})/(M - 1)$ is set from the empirical range of $\mathbf{g}$ for each encoded block. This matches the fixed, 8-bit scalar quantization used throughout the pipeline and requires no additional tuning beyond the TT-rank selection of Section 4.7.

The empirical Shannon entropy of the quantized values is used as an (achievable, idealized) bit-rate estimate,

$$\hat{H}(\mathbf{g}) = -\sum_{m=1}^{M} p_m \log_2 p_m, \qquad \mathrm{bits}(\mathbf{g}) = |\mathbf{g}| \cdot \hat{H}(\mathbf{g}), \tag{6}$$

following the classical minimum-redundancy coding result of Huffman [25]. This estimate assumes the quantized core values are coded independently of their positions; in practice this is subject to integer codeword lengths and residual statistical dependencies between coefficients. To quantify this gap, we additionally implement an explicit Huffman coder over the same quantization levels and report the resulting integer-codeword-length bit count alongside $\hat{H}(\mathbf{g})$ throughout Section 6. The Huffman-achieved bit rate is, on average, 12.2% higher than the Shannon entropy estimate. Accordingly, Shannon-based CR values reported in this paper should be interpreted as *upper bounds* on the achievable compression ratio, whereas the Huffman-achieved CR is treated as the operative, implementation-backed value wherever both are available.

The term $\mathrm{bits}_{\mathrm{residual}}$ separately accounts for the quantization residual retained to bound peak reconstruction error (Section 3), coded using the same Huffman scheme described above; omitting it would understate the true bit budget whenever residual coding is enabled.

$$\mathrm{CR} = \frac{N \cdot b_{\mathrm{orig}}}{\mathrm{bits}(\mathbf{g}) + \mathrm{bits}_{\mathrm{residual}}}, \tag{7}$$

where $N = B{\cdot}L{\cdot}S$ is the number of original samples in the block and $b_{\mathrm{orig}} = 11$ bits/sample (a representative ADC resolution for clinical/Holter ECG hardware).

4.5 Rank-adaptive operation

TensorECG can select the smallest rank that satisfies target distortion by searching a rank grid at encode time. Because PRD is monotonically non-increasing in r (a larger TT-rank can only reduce the approximation error of Eq. 2), a simple ascending sweep with early stopping is guaranteed to return the lowest-rank (hence highest-CR) point in the searched grid that satisfies the target, without an exhaustive search. This monotonicity

Algorithm 1 TensorECG: DWT+TT encoding and decoding with low bit residual refinement

Require: Raw beat block $\mathcal{X} \in \mathbb{R}^{B\times L\times S}$; TT-rank r; quantization levels M; optional residual bit depth n_b
Ensure: Reconstructed block $\hat{\mathcal{X}}_{\text{final}}$; bit cost bits

Encoding
$\mathcal{C} \leftarrow \mathbf{0}$
for $b = 1, \ldots, B$ **do**
  for $\ell = 1, \ldots, L$ **do**
    $\mathcal{C}(b, \ell, :) \leftarrow \mathrm{DWT}\big(\mathcal{X}(b, \ell, :)\big)$ ▷ Eq. 1
  end for
end for
▷ $\mathcal{C} \in \mathbb{R}^{B\times L\times S}$
$\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\} \leftarrow \text{TT-SVD}\big(\mathcal{C}, \mathbf{r} = (1, r, r, 1)\big)$
$\mathbf{g} \leftarrow [\mathrm{vec}(\mathcal{G}_1); \mathrm{vec}(\mathcal{G}_2); \mathrm{vec}(\mathcal{G}_3)]$
Quantize $\mathbf{g}$ to M levels and estimate $\hat{H}(\mathbf{g})$ ▷ Eq. 6
bits $\leftarrow |\mathbf{g}|\hat{H}(\mathbf{g})$
$\hat{\mathcal{C}} \leftarrow \mathrm{TT}^{-1}(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)$ ▷ Eq. 2

if $n_b > 0$ **then**
  $\mathcal{E} \leftarrow \mathcal{X} - \hat{\mathcal{X}}$
  $e_{\min} \leftarrow \min(\mathcal{E})$, $e_{\max} \leftarrow \max(\mathcal{E})$
  $q \leftarrow \left\lfloor \dfrac{\mathcal{E} - e_{\min}}{e_{\max} - e_{\min}}(2^{n_b} - 1) \right\rceil$
  $\hat{\mathcal{E}} \leftarrow \dfrac{q}{2^{n_b} - 1}(e_{\max} - e_{\min}) + e_{\min}$
  Entropy-code q and obtain $\text{bits}_{\text{residual}}$
  bits $\leftarrow$ bits $+ \text{bits}_{\text{residual}}$
end if

Decoding
$\hat{\mathcal{C}} \leftarrow \mathrm{TT}^{-1}(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)$
for $b = 1, \ldots, B$ **do**
  for $\ell = 1, \ldots, L$ **do**
    $\hat{\mathcal{X}}(b, \ell, :) \leftarrow \mathrm{IDWT}\big(\hat{\mathcal{C}}(b, \ell, :)\big)$
  end for
end for
if $n_b > 0$ **then**
  $\hat{\mathcal{E}} \leftarrow \dfrac{q}{2^{n_b} - 1}(e_{\max} - e_{\min}) + e_{\min}$
  $\hat{\mathcal{X}}_{\text{final}} \leftarrow \hat{\mathcal{X}} + \hat{\mathcal{E}}$
else
  $\hat{\mathcal{X}}_{\text{final}} \leftarrow \hat{\mathcal{X}}$
end if
return $\hat{\mathcal{X}}_{\text{final}}$, bits

also permits replacing the linear sweep with a bisection search over $\mathcal{R}$, reducing the number of TT-SVD evaluations from $O(|\mathcal{R}|)$ to $O(\log|\mathcal{R}|)$ at encode time, which matters when the grid is fine-grained or the encoder runs on resource-constrained wearable hardware. If no rank in $\mathcal{R}$ meets τ (e.g., an unusually noisy record), TensorECG falls back to the largest available rank $r^{(\max)}$ and flags the block as below-target rather than silently violating the distortion requirement. Because the search is performed independently per block, the selected rank – and therefore the achieved compression ratio – can vary across leads and across time within the same recording, adapting compression aggressiveness to local signal complexity rather than fixing a single global operating point.

4.6 Residual refinement

We propose an optional second coding step for applications that require more accuracy than the base DWT+TT operating point without switching to a significantly larger TT-rank (which costs bits roughly quadratically in r per Eq. 3). Given the DWT+TT reconstruction $\hat{\mathcal{X}}$, the residual $\mathcal{E} = \mathcal{X} - \hat{\mathcal{X}}$ is uniformly

Algorithm 2 TensorECG-RankControl: leakage-free rank selection and adaptive-rank compression

Require: Training set $\mathcal{D}_{\text{train}}$; candidate ranks $\mathcal{R} = \{2, 3, 4, 6, 8, 10, 14, 20, 28\}$; PRD threshold $\theta = 9\%$; optional target PRD τ

Ensure: Frozen rank $r^\star$ or adaptive rank r; reconstruction $\hat{\mathcal{X}}$; bit cost bits

Phase I: Offline leakage-free rank selection
for each $r \in \mathcal{R}$ **do**
  Encode every record in $\mathcal{D}_{\text{train}}$ using TensorECG$(\cdot, r)$
  Compute mean PRD(r) and mean CR(r) over $\mathcal{D}_{\text{train}}$
end for
$\mathcal{R}_{\text{valid}} \leftarrow \{r \in \mathcal{R} : \text{mean PRD}(r) \le \theta\}$
if $\mathcal{R}_{\text{valid}} \neq \emptyset$ **then**
  $r^\star \leftarrow \arg\max_{r \in \mathcal{R}_{\text{valid}}} \text{mean CR}(r)$
else
  $r^\star \leftarrow \arg\min_{r \in \mathcal{R}} \text{mean PRD}(r)$
end if
Freeze $r^\star$ and apply it unchanged to MIT-BIH test and transfer databases

Phase II: Optional encode-time adaptive rank
if target PRD τ is specified **then**
  for $r \in \mathcal{R}$ in ascending order **do**
    $\hat{\mathcal{X}}, \text{bits} \leftarrow \text{TensorECG}(\mathcal{X}, r)$
    $p \leftarrow \text{PRD}(\mathcal{X}, \hat{\mathcal{X}})$ ▷ Eq. 9
    if $p \le \tau$ **then**
      break
    end if
  end for
  $r \leftarrow$ selected rank
else
  $r \leftarrow r^\star$
  $\hat{\mathcal{X}}, \text{bits} \leftarrow \text{TensorECG}(\mathcal{X}, r)$
end if
return $\hat{\mathcal{X}}, \text{bits}, r$

quantized to n_b bits per element,

$$q = \left\lfloor \frac{\mathcal{E} - e_{\min}}{e_{\max} - e_{\min}} (2^{n_b} - 1) \right\rceil, \qquad \hat{\mathcal{E}} = \frac{q}{2^{n_b} - 1}(e_{\max} - e_{\min}) + e_{\min}, \tag{8}$$

The residual bits are coded and transmitted separately from the TT core bits. Residual refinement lets an application move along a finer-grained rate–distortion curve than the discrete rank grid $\mathcal{R}$ alone permits, without re-running TT-SVD or changing the selected rank r. The two knobs are therefore complementary rather than substitutes: r sets the coarse operating point and n_b trims residual error on top of it, at a bit cost that scales linearly in n_b rather than with the rank-dependent cost of Eq. 3. This also makes residual refinement well suited to progressive or scalable transmission, since the base DWT+TT stream remains fully decodable on its own if the residual layer is dropped or arrives late.

4.7 Rank selection protocol (train/test disjoint)

The only sensitive hyperparameter in the fixed-rank configuration is r. Following Zigel's PRD quality taxonomy, we select r by grid search over $r \in \{2, 3, 4, 6, 8, 10, 14, 20, 28\}$ *exclusively* on the 35-record MIT-BIH TRAIN split (Section 5), choosing the rank with the best mean compression ratio among ranks whose mean PRD on TRAIN is $\le 9\%$ (the boundary of Zigel's "very good" class); if no rank meets that bound, the rank with the lowest mean PRD is used as a fallback. The selected rank is then *frozen* and applied, unmodified, to every record in the 13-record MIT-BIH TEST split and to all three transfer databases; no metric reported in Section 6 for a test record was used, directly or indirectly, to choose r. This protocol mirrors standard *leakage-free* hyperparameter selection practice from supervised learning — where a hyperparameter is tuned only on a held-out validation set and then fixed before evaluation on test data — adapted here to a signal-processing pipeline rather than a trained model. The train split plays the role of a validation set for a single scalar hyperparameter, and the test and transfer splits play the role of held-out and out-of-distribution evaluation, respectively. We select r using the MIT-BIH training set only, rather than re-tuning it for each transfer

database, because re-optimizing the rank for each target distribution could overestimate the generalization performance of a single deployed codec configuration under unseen recording conditions. The discrete grid $\{2, 3, 4, 6, 8, 10, 14, 20, 28\}$ is coarser at high ranks by design, reflecting the diminishing compression-ratio returns observed as r increases and reducing the need for fine-grained search in this regime.

5 Experimental Setup

5.1 Datasets

We use four PhysioNet databases, chosen to span a range of lead counts and sampling rates (summarized in Table 1): the MIT-BIH Arrhythmia Database (2-lead, 360 Hz, per-beat AAMI-style rhythm annotations), the St. Petersburg INCART 12-lead Arrhythmia Database [26] (12-lead, 257 Hz, per-beat annotations), the PTB Diagnostic ECG Database [27] (15-lead, 1000 Hz, per-*record* diagnostic class rather than per-beat rhythm labels), and the European ST-T Database [28] (2-lead, 250 Hz, per-beat annotations, curated for ST/T ischemic-morphology evaluation). Beats are windowed from 0.25 s before to 0.5 s after each

Table 1 PhysioNet databases used, and the fixed record-level train/test split

Database	Leads	f_s (Hz)	Annotation granularity	Train	Test
MIT-BIH (mitdb)	2	360	per-beat	35	13
INCART (incartdb)	12	257	per-beat	0	30
PTB (ptbdb)	15	1000	per-record	0	30
European ST-T (edb)	2	250	per-beat	0	30

detected/annotated fiducial point (MIT- BIH/INCART/European ST-T: annotated R-peak; PTB: R-peak from QRS detection, since PTB does not provide a per-beat annotator), band-pass filtered at 0.5 to 40 Hz, and grouped into blocks of B = 20 consecutive beats. Model selection uses 35 MIT-BIH records for training, while evaluation is carried out on 13 held-out MIT-BIH records together with 30 records each from INCART, PTB, and European ST-T, none of which are seen during training. In total, the model is evaluated on 13 + 30 + 30 + 30 = 103 unseen records spanning four databases, with INCART, PTB, and European ST-T serving as fully independent, out-of-distribution transfer sets.

5.2 Metrics

Distortion is measured using percentage root-mean-square difference (PRD) and signal-to-noise ratio (SNR),

$$\mathrm{PRD}(x, \hat{x}) = 100\sqrt{\frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i x_i^2}}\ (\%), \tag{9}$$

$$\mathrm{SNR}(x, \hat{x}) = 10 \log_{10} \frac{\sum_i x_i^2}{\sum_i (x_i - \hat{x}_i)^2}\ (\mathrm{dB}), \tag{10}$$

where x and $\hat{x}$ denote the original and reconstructed signals, respectively. Both metrics (Eqs. 9–10) are computed identically at either the beat or block level, depending on context. Rate is quantified using the compression ratio CR (Eq. 7).

We also present six clinically motivated fidelity metrics, calculated per beat on lead 0 and all anchored to an amplitude-threshold QRS-onset/offset detector applied independently to the original and reconstructed beat: the detector locates the R-peak as the sample of maximum absolute amplitude, then finds the onset/offset as the first/last sample within a ±60 ms search window whose absolute amplitude exceeds 25% of the R-peak amplitude.

- **R-peak amplitude error**: $||x_r| - |\hat{x}_r||$, the absolute difference in R-peak amplitude between original and reconstructed beat at the shared R-peak index.
- **QRS-duration error**: the absolute difference, in ms, between the QRS onset–offset interval measured on the original beat and on the reconstructed beat using the detector.
- **ST-segment deviation**: mean absolute error between original and reconstructed samples in the window 60–120 ms after the (original-beat) QRS offset.
- **T-wave amplitude error**: the absolute difference in peak amplitude between the original and reconstructed T-wave, each located independently as the sample of maximum absolute amplitude within the 120–300 ms window after QRS offset.

- **T-wave location error**: the absolute difference, in ms, between the original and reconstructed T-wave peak locations found.

These clinical-alignment metrics are *research proxies*: they are computed using simple, single-lead, threshold-based rules rather than validated clinical QT-measurement algorithms, and are intended only to detect relative timing distortions introduced by compression, not to serve as diagnostic-grade measurements. Because amplitude-threshold detectors are used rather than clinical-grade delineation algorithms, these metrics are intended for relative original-vs-reconstructed comparison within a method rather than as standalone diagnostic measurements.

According to the standard procedure defined in ANSI/AAMI EC57 [29], beat rhythm labels are categorized into the AAMI N/S/V/F/Q superclasses for the stratified evaluation.

Statistical significance between the proposed method and each baseline is assessed per record (per matched block, for the ablation study) with a paired t-test and a Wilcoxon signed-rank test on the paired metric values; p-values from the Wilcoxon test are corrected for multiple comparisons across all baselines compared in the same run using the Holm step-down procedure, and a comparison is reported as significant only if it survives.

5.3 Baselines

TensorECG is compared to three baselines that are representative of the literature reviewed in Section 2: **DCT+Huffman** and **DWT+Huffman**, per-beat, per-lead coefficient-thresholding coders entropy-coded and **Block-2-D-DCT**, a 2-D DCT applied over the (beat × sample) plane of a block, which represents the traditional method used to exploit inter-beat redundancy without a tensor factorization. Each method's native rate-control parameter is selected (keep-ratio for DCT/DWT/2-D-DCT coders; TT-rank for TT-only/DWT+TT), and each record's unique PRD-vs-CR curve is interpolated at five common target compression ratios.

Interpolation is performed *per record first, then averaged*: points are never pooled across records before interpolating, since doing so could bracket a target CR using curves from two different records and yield a physically inconsistent estimate. We additionally compare against **TT-only**, a fourth baseline that isolates the contribution of the wavelet front end. TT-only applies the same TT-SVD decomposition directly to the raw beat tensor $\mathcal{X}$ rather than to the wavelet coefficient tensor, using the same rank grid and the same entropy-coding stage. TT-only therefore differs from the full DWT+TT pipeline by exactly one stage – the absence of the wavelet transform – which is what makes it a valid ablation of the wavelet front end's contribution.

6 Results

6.1 TT-rank selection on the training split

The training-only grid search of Section 4.7 is displayed in Table 2. In line with the TT-SVD optimality property covered in Section 3, truncation error decreases as r grows, which in turn reduces reconstruction distortion but at the cost of requiring more bits to store the additional retained components. The grid search was evaluated with two additional constraints – mean R-peak amplitude error ≤ 0.05 mV and mean QRS duration error ≤ 5 ms – alongside the original PRD $\leq 9\%$ criterion. Rank $r = 8$ is the smallest rank meeting *all three* thresholds simultaneously, i.e. the same rank the original PRD-only criterion selected (CR= 25.85 at PRD= 8.50% on TRAIN, with mean R-peak error 0.026 mV and mean QRS error 1.79 ms, both comfortably inside the added clinical bounds); this rank is therefore confirmed by the additional clinical constraints and remains fixed at $\mathbf{r} = (1, 8, 8, 1)$ for all following results in this section. The next coarser rank in the grid, $r = 6$, already exceeds the 9% PRD threshold (PRD= 11.07%), so no finer rank between 6 and 8 was evaluated; a denser grid in this interval might yield a small additional CR improvement at the same quality bound, which we leave to future refinements of the rank-selection procedure. All test and transfer-database results reported below therefore use ranks selected exclusively from the training data, rather than ranks re-optimized per record or per dataset.

6.2 Rate–distortion evaluations

Every method achieves complete coverage of the five target CRs across the 13 test records: for each method, all five target CRs fall within the empirically achievable operating range for nearly every test record, as detailed in the Supplementary Material. The reported ranges represent the achievable CR range under the specified quality constraint; we summarize using the median across the 13 test records, and report the mean and standard deviation across the same 13 records for the central rate-distortion result. At every matched CR, DWT+TT achieves the lowest PRD of all five methods, and the gap widens with CR: at CR= 20, DWT+TT's PRD (0.95%) is 24× lower than DWT+Huffman's (23.10%) and 46× lower than DCT+Huffman's (44.19%); at CR= 80, DWT+TT (5.31%) is still 8.8× lower than DWT+Huffman (46.70%). TT-only, the same TT

Table 2 TT-rank grid search with clinical metrics, 35-record MIT-BIH TRAIN split only (block size 20). Rank $r = 8$ is selected (bold) as the smallest rank meeting all three thresholds: mean PRD $\leq 9\%$, mean R-peak amplitude error ≤ 0.05 mV, and mean QRS duration error ≤ 5 ms.

Rank r	CR	PRD (%)	R-peak err. (mV)	QRS err. (ms)
2	89.89	24.00	0.130	5.36
3	62.67	18.31	0.078	4.10
4	48.51	15.03	0.059	3.49
6	33.53	11.07	0.037	2.42
8	**25.85**	**8.50**	**0.026**	**1.79**
10	21.14	6.68	0.019	1.39
14	15.65	4.31	0.010	0.81
20	11.36	2.27	0.004	0.32

step without the wavelet front end, stays close behind DWT+TT at low CR but degrades faster at high CR (12.94% at CR= 80 vs. DWT+TT's 5.31%), showing that the wavelet front end's contribution grows with the aggressiveness of the rank truncation.

6.2.1 Statistical significance

The paired significance tests of DWT+TT against every baseline, across all five target CRs, are presented in Table 3. Every comparison is significant after Holm correction ($p_{\text{Holm}} \leq 0.001$ throughout), and the effect sizes are large: $|d_z| > 1.6$ for every baseline at every CR.

Table 3 Paired statistics, DWT+TT vs. baselines, PRD at CR= 20 ($n = 13$ test records). All comparisons significant after Holm correction.

Comparison	Mean diff.	d_z	p_{Holm}
vs. DCT+Huffman	−42.67	−2.73	< 0.001
vs. DWT+Huffman	−22.31	−5.08	< 0.001
vs. (2D-DCT)	−9.87	−3.61	< 0.001
vs. TT-only	−1.84	−1.68	< 0.001

6.3 Ablation study

The full DWT+TT+Entropy pipeline is broken down into its constituent stages, each ablation arm differing from the full pipeline by exactly one stage is shown in Table 4, all evaluated at each stage's *native* default operating point on the same 20-beat test blocks.

Table 4 Ablation study (mean ± std over 78 MIT-BIH test blocks). Each row differs from "Full" by exactly one stage.

Method	CR	PRD (%)	SNR (dB)
Proposed (DWT+TT+Entropy)	27.42±4.11	6.73±2.61	25.07±3.36
No-Entropy-Coding	1.43±0.00	6.23±2.61	24.07±3.36
No-Wavelet (TT-only)	11.77±1.59	6.61±2.55	24.21±3.35
(2D-DCT)	6.70±0.24	3.84±1.50	28.98±3.51

First, removing entropy coding while keeping the same TT-truncated core ("No-Entropy-Coding", raw 32-bit float storage of the TT cores) does not change PRD/SNR (since it does not touch the reconstruction) but collapses CR from 27.42 to 1.43 – a 19.2× compression-ratio penalty for identical distortion, confirming that most of the achieved compression ratio comes from entropy-coding the already-small set of TT core values, not from the TT truncation alone. Second, the wavelet front end is worth a further 2.3×. No-Wavelet (TT-only) reaches CR= 11.77 at essentially the same PRD (6.61% vs. 6.73%) as the full pipeline's CR= 27.42, so at this operating point the DWT stage alone roughly doubles the compression ratio. Third, TT truncation is the only lossy stage. The statistical analysis confirms significant differences for the relevant comparisons, except for Full versus No-Entropy-Coding in PRD, for which the two configurations produce identical reconstructions

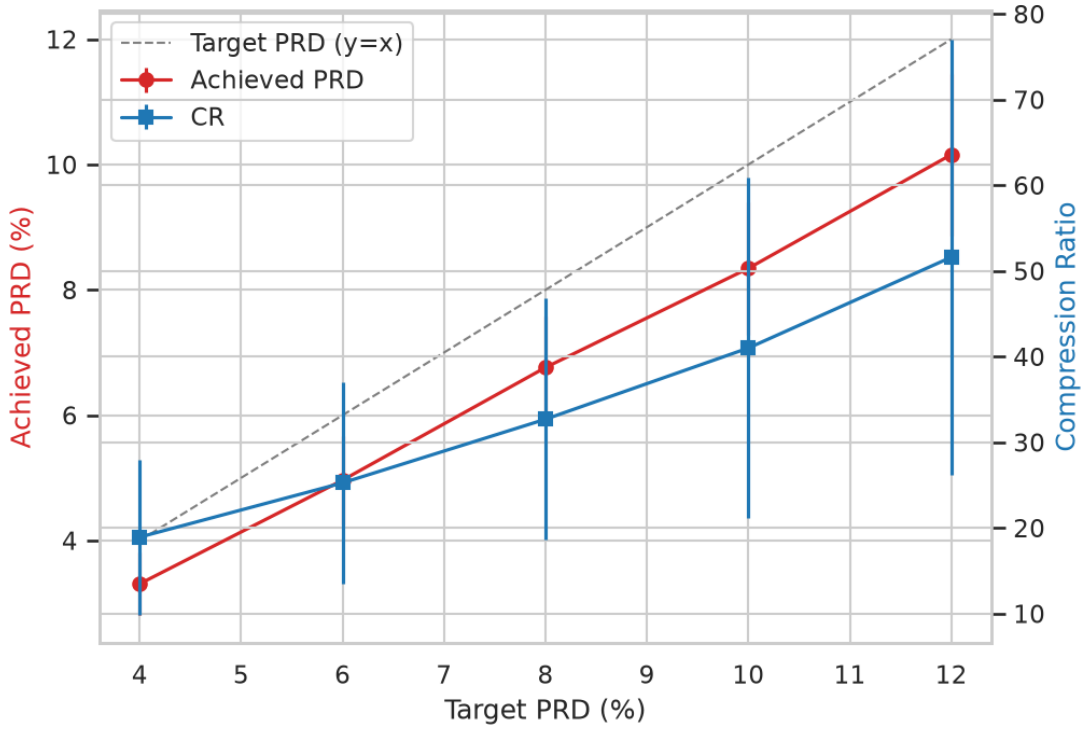


Fig. 2 Adaptive-rank DWT+TT: achieved PRD vs. target PRD, and the resulting compression ratio.

and therefore exactly the same PRD; their difference is instead reflected entirely in CR, since entropy coding affects the bit count without altering the reconstructed signal. Proposed method has a *higher* PRD than (2D-DCT) at these native operating points (6.73% vs. 3.84%), while also achieving a 4.1× higher CR (27.42 vs. 6.70). This is an artifact of comparing each method's default rank/keep-ratio rather than a matched-CR point, and is resolved by the matched-CR study in Section 6.2, where DWT+TT beats (2D-DCT) at every common CR. We report both views because they answer different questions: the matched-CR study is the fair rate–distortion comparison, while the native-operating-point ablation isolates each stage's individual contribution to the full pipeline's own default setting.

6.4 Rank-adaptive TT

Experiments are conducted on the MIT-BIH test set at five PRD target levels, as shown in Table 5. Two properties make this a practically useful operating mode. First, the achieved PRD closely tracks the requested target, within 0.2 pp at the tightest target of 4%, growing to approximately 1.8 pp of overshoot at the loosest target of 12% (reflecting the coarseness of the eight-point rank grid used here), so a caller can request a quality level directly, rather than having to reverse-engineer an appropriate rank. Second, the resulting compression ratio is far from flat: CR rises from 18.94 at the strictest target (PRD $\leq$ 4%) to 51.61 at the loosest (PRD $\leq$ 12%), a 2.7× range, driven by the chosen rank falling from 13.9 to 4.8 – i.e. the adaptive mode automatically trades detail for rate along the rate–distortion curve (Figure 2).

Table 5 Adaptive-rank DWT+TT, MIT-BIH test set, mean over test blocks.

Target PRD (%)	Achieved PRD (%)	CR	Chosen rank
4	3.31	18.94	13.92
6	4.96	25.27	10.15
8	6.76	32.69	7.59
10	8.34	41.01	6.10
12	10.16	51.61	4.82

6.5 Residual correction

The residual-refinement bit depth n_b (Eq. 8) is varied over the frozen-rank ($r = 8$) DWT+TT base in Table 6. Two distinct regimes are observed. At very low bit depth ($n_b = 2$), residual coding is counter-productive:

PRD slightly *increases* (6.73% → 7.97%) while CR drops sharply (27.42 → 8.90), because 2-bit uniform quantization of the residual introduces additional quantization error while still consuming entropy bits – a regime in which the distortion introduced by quantizing the residual outweighs the reduction in reconstruction error, so $n_b = 2$ is not beneficial in this setting. From $n_b \geq 3$ onward, residual coding maintains a smoothly tunable trade-off: at $n_b = 4$, PRD drops to 1.46% (a 4.6× reduction from the n_b=0 base) at CR= 3.85, while at $n_b = 6$, PRD reaches 0.35% (SNR= 49.9 dB) with CR still at 2.25. This traces a tunable path from the base operating point down to near-transparent reconstruction, at a bit-rate cost that remains far below the raw-signal rate (CR> 1 throughout the swept range).

Table 6 Residual-coding sweep (base: DWT+TT, $r = 8$), mean over MIT-BIH test blocks. $n_b = 0$ is the base pipeline with no residual stage.

n_b	CR	PRD (%)	SNR (dB)	n_b	CR	PRD (%)	SNR (dB)
0	27.42	6.73	24.07	**4**	**3.85**	**1.46**	**37.46**
2	8.90	7.97	22.76	5	2.84	0.71	43.76
3	5.82	3.15	30.82	6	2.25	0.35	49.92

6.6 Clinical fidelity evaluation

We compare two configurations, DWT+TT and DWT+TT with 4-bit residual refinement, on six beat-timing metrics computed per beat over the entire MIT-BIH test set (Table 7). Unlike Table 4 and Table 6, which report metrics over 20-beat blocks, this evaluation is performed on whole held-out records; the two evaluations therefore measure the same pipeline at different granularities. Residual refinement reduces every fiducial-point error substantially – R-peak amplitude error by 65%, QRS-duration error by 59%, ST-segment deviation by 39%, T-wave amplitude error by 42%, T-wave location error by 23%, and QT-interval error by 31% – for a CR that is only 3–4× lower than the unrefined base (Table 6, n_b=4: CR= 3.85 vs. the block-level base CR= 27.42; the whole-record clinical evaluation follows the same ratio). These fiducial-point metrics are computed with a simple amplitude-threshold detector rather than a validated clinical algorithm; because the same detector is applied to both the original and reconstructed beats, the *relative* improvements reflect genuine robustness rather than detector artifacts. Given how large and consistent this improvement is, we recommend the 4-bit residual-refinement configuration as the default TensorECG operating point whenever bandwidth headroom allows CR ≈ 4, reserving the higher-CR, no-residual configuration for the most bandwidth-constrained links. Paired significance tests (DWT+TT vs. the other two arms) are reported for every metric. Two patterns are notable. First, every DWT+TT-vs.-residual-4-bit comparison is significant with a small-to-moderate effect size, showing consistent improvement. Second, DWT+TT vs. TT-only is significant for PRD, SNR, R-peak amplitude, and ST-deviation, but *not* significant after correction for QRS-duration error, T-wave amplitude error, or T-wave location error – i.e. at this frozen rank, the wavelet front end mainly improves aggregate energy-based fidelity (PRD/SNR) and the R-peak/ST window, while having a statistically inconclusive effect on T-wave timing and QRS duration specifically, a nuance that a PRD-only evaluation would have missed.

6.7 Cross-dataset transferability

We trained on 35 MIT-BIH records and tested generalization on four datasets – a 13-record MIT-BIH held-out split plus 30-record splits from INCART, PTB, and European ST-T – using a single frozen rank $\mathbf{r} = (1, 8, 8, 1)$ selected solely from the training data, with no test or transfer record ever consulted and no further tuning applied. Figure 3 shows the corresponding per-record distributions of achieved PRD across the four databases for both methods, with the 8% target marked as a reference line.

At matched rank, the 2-lead datasets (European ST-T, MIT-BIH) achieved CR of 20.9 and 28.8 respectively, while the 12/15-lead datasets (INCART, PTB) achieved substantially higher CR of 117.2–192.2, without any degradation in reconstruction quality – PRD across all transfer databases (3.68–9.87%) remained within the same "good-to-very-good" band as the MIT-BIH test split (6.54%), despite zero per-database tuning. The wavelet front end's benefit was preserved across every database, with DWT+TT's CR exceeding TT-only's by 1.9–2.4× at closely matched PRD (e.g., INCART: 117.2 vs. 55.1 at PRD 9.87% vs. 9.55%), confirming the wavelet-then-TT mechanism is not MIT-BIH-specific. Once reconstruction quality is matched across datasets (PRD around 6.4–6.8%), the differences in compression ratio become more pronounced, not less. This confirms that the CR gap reflects each database's genuine redundancy structure – driven by inter-lead correlation that the tensor-train decomposition exploits more efficiently as channel count increases – rather than a mismatch between a fixed rank and specific databases, consistent with the joint beat-lead-sample factorization mechanism discussed in Section 7.

Table 7 Clinical validation metrics, mean ± std, MIT-BIH test set($n = 1560$ beats)

Method	R_{amp} err (mV)	QRS dur. err (ms)	ST dev. (mV)	T_{amp} err (mV)	T_{loc} err (ms)	QT err (ms)
DWT+TT	0.040±0.053	2.01±4.26	0.023±0.027	0.024±0.031	13.63±30.89	10.09±23.17
DWT+TT (res. 4-bit)	**0.014**±0.016	**0.82**±2.84	**0.014**±0.013	**0.014**±0.015	**10.52**±27.38	**6.93**±19.38

Table 8 Cross-database transfer at the *frozen* MIT-BIH-selected rank $\mathbf{r} = (1, 8, 8, 1)$ (matched-rank view), mean ± std over each database's test split. CR figures are Shannon-bound upper estimates (Eq. 6).

Database	Method	CR	PRD (%)	SNR (dB)
MIT-BIH (2-lead, $n = 13$)	DWT+TT	28.80±4.30	6.54±2.37	24.26±3.24
	TT-only	11.78±1.57	6.42±2.30	24.41±3.20
INCART (12-lead, $n = 30$)	DWT+TT	117.24±23.10	9.87±8.02	22.57±6.52
	TT-only	55.14±9.15	9.55±7.74	22.83±6.48
PTB (15-lead, $n = 30$)	DWT+TT	192.24±37.00	3.68±0.38	28.72±0.91
	TT-only	99.41±7.64	3.66±0.38	28.78±0.91
European ST-T (2-lead, $n = 30$)	DWT+TT	20.90±2.52	5.80±2.90	25.72±4.15
	TT-only	11.21±0.99	5.66±2.82	25.93±4.15

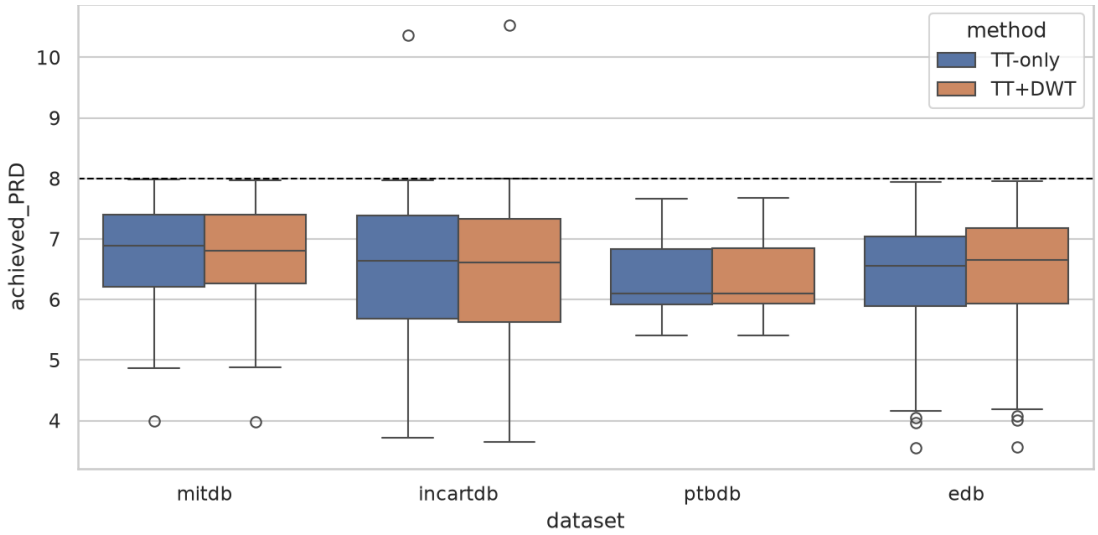


Fig. 3 Distribution of achieved PRD per record for TT-only and DWT+TT across the four evaluation databases (MIT-BIH, INCART, PTB, European ST-T) at the matched-quality operating point. Boxes show the interquartile range with the median; whiskers extend to 1.5×IQR and circles mark outlying records.

6.8 Arrhythmia and diagnostic-class-stratified evaluation

This section reports PRD separately by AAMI beat class (N: normal, S: supraventricular ectopic, V: ventricular ectopic, Q: unknown/paced) for the three per-beat-annotated databases, and by diagnostic class (myocardial infarction, the only class with sufficient PTB label coverage in our keyword-matched taxonomy) for PTB. This lets us inspect performance on less frequent beat classes separately from the dominant Normal-class contribution that characterizes every whole-record PRD figure reported thus far. Two patterns stand out as clinically relevant. First, ventricular-ectopic (V) beats are reconstructed with consistently higher PRD than Normal (N) beats across every database with enough V-labelled beats to compare, since ectopic morphology deviates from the P-QRS-T template that dominates the training statistics against which the TT-rank was chosen – and it is precisely the kind of subgroup gap that a PRD-only, class-blind evaluation would hide. Second, the gap between DWT+TT and TT-only is small and inconsistent in sign within every class (differences of at most 0.5 pp, sometimes favouring TT-only). This indicates that the wavelet front end's aggregate benefit (Section 6.2) is broad-based rather than concentrated in the majority-Normal population, and does not systematically favour or disadvantage any particular beat class. Finally, the very small sample sizes for MIT-BIH's S class (n=4) and European ST-T's V class (n=7) mean these two rows should be read as indicative rather than statistically reliable; we report them for completeness rather than to draw strong conclusions from them.

6.9 Noise robustness

Gaussian noise is injected at four controlled input SNR levels (40, 30, 20, 10 dB) before compression. Table 9 reports PRD with respect to both the noisy input, quantifying the distortion introduced by the compression process itself, and the clean reference, quantifying the overall end-to-end distortion after compression. This dual evaluation separates compression-induced degradation from any unintended denoising effect arising from low-rank truncation.

Table 9 Noise robustness, MIT-BIH test set, mean ± std. PRD is reported against the noisy input (coder-only distortion) and the clean reference (end-to-end distortion).

Input SNR (dB)	Method	PRD vs. noisy (%)	PRD vs. clean (%)	CR
10	DWT+TT	27.46±0.82	17.75±1.01	16.21
	TT-only	27.28±0.80	17.67±0.92	10.68
20	DWT+TT	11.25±1.67	8.41±2.13	22.34
	TT-only	11.16±1.63	8.28±2.08	11.54
30	DWT+TT	7.36±2.41	6.92±2.54	25.86
	TT-only	7.25±2.35	6.80±2.49	11.74
40	DWT+TT	6.79±2.58	6.75±2.60	27.12
	TT-only	6.68±2.53	6.63±2.55	11.76

At every input SNR, the PRD computed against the clean reference signal is consistently *lower* than the PRD computed against the noisy input (e.g. at 20 dB: 8.41% vs. 11.25% for DWT+TT). This shows that the low-rank TT truncation and wavelet coefficient thresholding together act as a mild, incidental denoiser: the compressed reconstruction is, on average, closer to the clean underlying signal than to the noisy input the coder actually received. This gap narrows as input SNR improves, shrinking from 9.7 pp at 10 dB to just 0.04 pp at 40 dB, where the injected noise contribution is already small. The trend is consistent with a fixed-rank low-rank approximation whose denoising effect is naturally more pronounced when the input noise level is higher. The compression ratio itself is essentially unaffected by noise level for TT-only (10.68 to 11.76 across the full 30 dB sweep) but rises noticeably for DWT+TT as input SNR improves (16.21 to 27.12), because injected noise raises the wavelet coefficients' own entropy and therefore their bit cost, an effect the wavelet front end is more sensitive to than the raw-sample TT-only path.

6.10 Computational and storage overhead

To evaluate the computational cost behind the accuracy-compression trade-offs reported throughout, Table 10 compares mean wall-clock time and bytes/beat across all four rate-controlled methods at their default operating point on the MIT-BIH test set. TensorECG's storage advantage is large – DWT+TT needs 26.97 bytes/beat, 8.6× fewer than DWT+Huffman's 232.49 and 4.1× fewer than the block-2-D-DCT baseline's 110.45 – but this comes at a real computational cost: DWT+TT's per-beat processing time (10.1 ms) is roughly 12× DWT+Huffman's (0.83 ms) and 6.5× Block-2-D-DCT's (1.56 ms), because the TT-SVD step requires factorizing an entire block's coefficient tensor rather than processing one beat independently. TT-only sits between the two, confirming that the wavelet front end itself is cheap and that the TT-SVD is the dominant cost driver. For a streaming wearable deployment this favours a blockwise pipeline (encode a block of B=20 beats every few seconds, well within real-time budgets on a modern microcontroller) rather than a strictly beat-synchronous one, and we flag encoder latency as an explicit design constraint rather than a free byproduct of the compression-ratio gains reported above.

Table 10 Computational/storage overhead, MIT-BIH test set, mean ± std per processed beat.

Method	Time (s/beat)	Bytes/beat
DWT+Huffman (per-beat)	0.00083 ± 0.00002	232.49 ± 8.32
(Block-2-D-DCT)	0.00156 ± 0.00028	110.45 ± 3.01
TT-only	0.00205 ± 0.00023	63.49 ± 8.73
DWT+TT	$\mathbf{0.01013 \pm 0.00046}$	$\mathbf{26.97 \pm 3.16}$

6.11 Real-time deployment constraints

The per-beat encoding time in Table 10 is only part of the end-to-end latency an application experiences: DWT+TT operates on a whole block of B beats at once, so a block cannot be encoded until all B beats have been acquired. At 60 bpm and $B = 20$, this algorithmic wait alone is ≈20 s, a substantial contributor to end-to-end latency that must be weighed against TensorECG's suitability for real-time use. We computed the full algorithmic + computational latency at 60 bpm across $B \in \{5, 10, 20, 40\}$ on the MIT-BIH test split (Table 11).

Table 11 Block-size sensitivity: quality vs. end-to-end latency at 60 bpm, MIT-BIH test split.

Block size B	CR	PRD (%)	Total latency (s)
5	10.96	5.63	5.00
10	18.27	6.92	10.00
20	**35.43**	**8.75**	**20.01**
40	67.71	10.24	40.02

At $B = 20$, total latency is ≈20 s (20.00 s algorithmic + 14 ms computational), which is incompatible with real-time arrhythmia detection on the encoded stream itself. We therefore recommend $B = 5$ (5 s latency, PRD 5.63%, still within acceptable quality range) as the default operating point for latency-sensitive deployments, and reserve $B = 20$ for post-hoc compression of stored Holter/ICU recordings, or for wearables that run a separate, low-latency arrhythmia detector directly on the raw signal in parallel with TensorECG's higher-latency, higher-CR compression path.

7 Discussion

The wavelet front end sparsifies each individual beat/lead signal (concentrating energy into a few large coefficients), which increases the amount of *cross-beat, cross-lead* redundancy that the wavelet coefficient tensor $\mathcal{C}$ exposes: for the same family of P-QRS-T morphology repeated with small jitter across dozens of beats and highly correlated across leads, the large-coefficient positions and magnitudes are themselves strongly correlated across the beat and lead axes, which is precisely the structure a low TT-rank can capture. This explains both why TT-only (TT applied directly to raw time samples, without this sparsification) needs a higher rank to hit the same distortion, and why the two stages' compression-ratio contributions compound multiplicatively rather than additively (Section 6.3): the wavelet transform improves the tensor's effective rank structure, the TT decomposition then exploits that structure, and entropy coding finally exploits the resulting core values' low entropy – each stage amplifies what the previous stage exposes. Three configurations emerge as practically distinct recommendations from our results: (i) the *bandwidth-constrained* configuration (frozen rank 8, no residual), CR≈27 at PRD≈6.7% (Table 4), appropriate for severely bandwidth-limited wireless links; (ii) the *quality-constrained* configuration using the adaptive-rank mode of Section 6.4, which lets a downstream application (e.g. an arrhythmia classifier with a known PRD tolerance) request a target distortion directly; and (iii) the *diagnostic-fidelity* configuration, base DWT+TT plus 4-bit residual refinement, which delivers the best fidelity-per-bit trade-off for the clinically motivated metrics (Section 6.6) and is our recommended default when link budget allows CR ≈4. All three configurations above use the default block size $B = 20$, whose ≈20 s algorithmic latency (Section 6.11) is unsuitable for real-time arrhythmia detection on the encoded stream itself; for latency-sensitive deployments we instead recommend $B = 5$ (Table 11), which keeps PRD in the "very good" band at a 4× shorter block-acquisition delay.

The (block-2-D-DCT) baseline is the closest prior-art analogue to TensorECG in that it also exploits inter-beat redundancy via a joint transform; DWT+TT beats it at every matched CR by a wide and statistically significant margin (Table 3), which we attribute to the TT decomposition's ability to also exploit the *lead* axis (which the per-lead 2-D DCT cannot) and to its data-adaptive (SVD-based) rather than fixed (DCT-basis) factorization. The four-database transfer study (Section 6.7) turns the single-database result above into a genuine generalization claim: a rank chosen once on 35 two-lead MIT-BIH training records, without ever seeing a test or transfer record, keeps PRD within the "good-to-very-good" band on three databases it has never been tuned against, spanning 2- to 15-lead recordings and 250–1000 Hz sampling rates (Table 8). The stratified breakdown (Section 6.8) further shows that this transferability holds broadly across AAMI beat classes rather than being an artifact of the majority-Normal population, and the noise-robustness sweep (Section 6.9) shows the same pipeline incidentally denoises rather than amplifies distortion under realistic acquisition noise. Taken together with the overhead benchmark (Section 6.10), these results let a practitioner

reason about TensorECG's expected behaviour on a new device or population without re-running the rank search, at the price of a roughly order-of-magnitude higher per-beat encoding time than the classical baselines.

8 Limitations and Future Work

The explicit Huffman coder implemented alongside the Shannon-entropy estimate (Section 6) achieves a bit rate 12.2% higher than the Shannon estimate on the MIT-BIH test split; we therefore report Shannon-based CR values as an upper bound throughout, with the Huffman-achieved figure given alongside it wherever measured. Three of our four transfer databases (INCART, PTB, European ST-T) contribute no records to the rank search itself, so the reported gains characterize transfer from MIT-BIH rather than a jointly optimized rank across sources. Our clinical fidelity metrics rely on a simple amplitude-threshold QRS/T detector rather than a validated fiducial-point algorithm, making their absolute magnitudes research-grade proxies, though the identical detector applied to original and reconstructed beats makes the relative comparisons between methods more robust. We quantified this further by comparing QRS duration estimates from two independently-parameterized amplitude-threshold detectors (25% threshold/120 ms window vs. 10% threshold/160 ms window) on the same 13 MIT-BIH test records: the two parameterizations disagreed by a mean of 41.8 ms per beat, large relative to typical QRS durations (80–120 ms), confirming that the absolute magnitude of our fiducial-point error metrics is materially sensitive to detector design choices and not solely to compression-induced distortion. Neither detector has been validated against manually annotated ground truth (e.g. PTB's annotated fiducial points); we therefore restrict all clinical-fidelity claims in this paper to *relative* (original-vs.-reconstructed, same detector) comparisons. PTB's diagnostic labels are derived from free-text header keyword matching and may not perfectly match PTB's official taxonomy. Finally, rank selection uses a single scalar mean-PRD threshold rather than one that directly incorporates clinical fidelity, motivating future work on per-class or per-lead rank selection, joint wavelet-level/rank optimization, and a learned entropy coder in place of the static estimate used here.

9 Conclusion

We presented TensorECG, a compression framework that factorizes a block of multi-lead heartbeats' per-beat, per-lead wavelet coefficients with a Tensor-Train decomposition, jointly exploiting beat-to-beat and lead-to-lead redundancy that per-lead and block-2D transform coders leave on the table. Using a strictly disjoint, record-level train/test protocol, we showed that a single TT-rank selected on 35 MIT-BIH training records generalizes to a held-out 13-record test split with large, statistically significant gains over DCT, DWT, and block-2-D-DCT baselines at every matched compression ratio (e.g. 24× lower PRD than DWT+Huffman at CR= 20, Cohen's d_z up to -6.9). The ablation study showed that the three stages of the pipeline combine multiplicatively rather than additively, the adaptive-rank variant let a caller hit a target distortion directly across a 2.7× range of achievable compression ratios, and a lightweight 4-bit residual-refinement stage nearly halved PRD and significantly improved every clinically motivated fidelity metric we measured for a moderate reduction in compression ratio. Extending the evaluation to three further PhysioNet databases (INCART, PTB, and European ST-T, spanning 12, 15, and 2 lead recordings) confirmed that the MIT-BIH-selected rank transfers without collapsing – PRD stays within the same good-to-very-good band on every database with zero per-database tuning – that this transferability is broad-based across AAMI beat classes rather than concentrated in the majority-Normal population, that the pipeline incidentally denoises rather than degrades under realistic acquisition noise, and that its storage savings (up to 8.6× fewer bytes/beat than a per-beat DWT+Huffman baseline) come at a quantifiable, roughly order-of-magnitude increase in per-beat encoding time relative to the classical baselines. Taken together, these results support tensor-train factorization – applied jointly across the beat and lead axes on top of a standard wavelet front end – as a practical and rate-distortion-efficient approach to multi-lead ECG compression, with adaptive-rank and residual-refinement extensions that make the operating point directly controllable by a downstream application's quality requirement.

Acknowledgements

The authors acknowledge the National Supercomputing Mission (NSM) for providing computing resources through PARAM UTKARSH at C-DAC Knowledge Park, Bengaluru, implemented by C-DAC and supported by MeitY and DST, Government of India. The authors also gratefully acknowledge Dr. Ruchi Juyal for providing valuable guidance for this work.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest/Competing interests:** The authors declare no conflict of interest.

- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** All four databases used (MIT-BIH, INCART, PTB, European ST-T) are publicly available via PhysioNet [2].
- **Author contribution:** Ashish Pratap: Writing – Original Draft, Review and Editing, Methodology, Hariharan Muthusamy: Investigation, Conceptualization and Critical Revision of the manuscript,Ramesh Naidu Laveti: Writing – Review,Visualization, Software, Data curation. Sithu D Sudarsan: Review, Validation, Conceptualization,

References

[1] G. B. Moody and R. G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[2] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[3] Z. Lu, D. Y. Kim, and W. A. Pearlman. Wavelet compression of ECG signals by the set partitioning in hierarchical trees algorithm. *IEEE Transactions on Biomedical Engineering*, 47(7): 849–856, 2000.

[4] S. O. Rajankar and S. N. Talbar. An electrocardiogram signal compression techniques: A comprehensive review. *Analog Integrated Circuits and Signal Processing*, 98(1):59–74, 2019.

[5] B. S. Kim, S. K. Yoo, and M. H. Lee. Wavelet-based low-delay ECG compression algorithm for continuous ECG transmission. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):77–83, 2013.

[6] N. Ahmed and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing*. Springer, Berlin, 2000.

[7] W. Philips and G. De Jonghe. Data compression of ECG's by high-degree polynomial approximation. *IEEE Transactions on Biomedical Engineering*, 39(4): 330–337, 1993.

[8] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[9] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[10] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.

[11] J. R. Cox, F. M. Nolle, H. A. Fozzard, and G. C. Oliver. AZTEC, a preprocessing program for real-time ECG rhythm analysis. *IEEE Transactions on Biomedical Engineering*, 15(2): 128–129, 1968.

[12] S. M. S. Jalaleddine, C. G. Hutchens, R. D. Strattan, and W. A. Coberly. ECG data compression techniques – a unified approach. *IEEE Transactions on Biomedical Engineering*, 37(4): 329–343, 1990.

[13] L. V. Batista, E. U. K. Melcher, and L. C. Carvalho. Compression of ECG signals by optimized quantization of discrete cosine transform coefficients. *Medical Engineering & Physics*, 23(2):127–134, 2001.

[14] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[15] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[16] C. T. Ku, K. C. Hung, T. C. Wu, and H. S. Wang. Wavelet-based ECG data compression system with linear quality control scheme. *IEEE Transactions on Biomedical Engineering*, 57(6): 1487–1494, 2010.

[17] S. G. Miaou and S. N. Chao. Wavelet-based lossy-to-lossless ECG compression in a unified vector quantization framework. *IEEE Transactions on Biomedical Engineering*, 52(3): 539–543, 2005.

[18] J. Chen and S. Itoh. A wavelet transform-based ECG compression method guaranteeing desired signal quality. *IEEE Transactions on Biomedical Engineering*, 45(12): 1414–1419, 1998.

[19] M. Blanco-Velasco, F. Cruz-Roldán, J. I. Godino-Llorente, and K. E. Barner. ECG compression with retrieved quality guaranteed. *Electronics Letters*, 44(3):206–207, 2008.

[20] Y. Zigel, A. Cohen, and A. Katz. The weighted diagnostic distortion (WDD) measure for ECG signal compression. *IEEE Transactions on Biomedical Engineering*, 47(11): 1422–1430, 2000.

[21] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[22] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann, Cambridge, MA, 5th edition, 2017.

[23] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[24] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4): 1253–1278, 2000.

[25] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[26] V. Tihonenko, A. Khaustov, S. Ivanov, A. Rivin, and E. Yakushenko. St. Petersburg INCART 12-lead arrhythmia database. *PhysioNet*, 2008.

[27] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik*, 40(s1):317–318, 1995.

[28] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. B. Moody, C. Zeelenberg, and C. Marchesi. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *European Heart Journal*, 13(9):1164–1172, 1992.

[29] Association for the Advancement of Medical Instrumentation. *ANSI/AAMI EC57:2012 – Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*. AAMI, Arlington, VA, 2012.