Evidence-Augmented Conformal Risk Control with Verifier-Gated LLM Reporting for Trustworthy ECG Decision Support

Khoa D. Pham[1,*], Dang Nguyen[2,*], Khanh Tran Quoc Le[1], Heath Rutledge-Jukes[3], Quan Le[1]
Hung N. Huynh[1], Trung Q. Le[4], Yanwen Xu[5], Om Prakash Yadav[6], Carl Yang[7]
Louise Y. Sun[8], Phat K. Huynh[1,†], Jacques Kpodonu[9,†]

[1]PASSIO Laboratory, North Carolina A&T State University, Greensboro, NC, USA
[2]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[3]Washington University in St. Louis School of Medicine, St. Louis, MO, USA
[4]Department of Industrial and Management Systems Engineering,
University of South Florida, Tampa, FL, USA
[5]Department of Mechanical Engineering, The University of Texas at Dallas, Richardson, TX, USA
[6]Department of Industrial and Systems Engineering,
North Carolina A&T State University, Greensboro, NC, USA
[7]Department of Computer Science, Emory University, Atlanta, GA, USA
[8]Division of Cardiothoracic Anesthesiology, Stanford University School of Medicine,
Palo Alto, CA, USA
[9]Division of Cardiac Surgery, Beth Israel Deaconess Medical Center,
Harvard Medical School, Boston, MA, USA

[*]Khoa D. Pham and Dang Nguyen contributed equally.
[†]Corresponding authors: Phat K. Huynh (pkhuynh@ncat.edu)
and Jacques Kpodonu (jkpodonu@bidmc.harvard.edu).

***Abstract*—Electrocardiography (ECG) is ubiquitous, yet clinical deployment of deep learning remains constrained by two reliability gaps: decision-aligned uncertainty (how likely true findings are missed) and verifiable communication, requiring AI reports not to invent diagnoses or measurements. We propose an ECG decision support pipeline that outputs risk-controlled multi-label diagnosis sets and produces clinician-facing text under deterministic verification. A ResNet-1D encoder produces per-label probabilities over $K = 23$ diagnostic subclasses. Conformal risk control (CRC) calibrates a single threshold on a held-out calibration fold, bounding the expected missed-finding loss at risk level α and yielding diagnosis sets rather than point decisions. To improve efficiency, we introduce evidence-augmented CRC (EA-CRC), combining model confidence with a retrieval-derived support score from a FAISS nearest-neighbor index built exclusively on training folds. On PTB-XL patient-wise folds, EA-CRC (TOPK = 10, β = 0.8) maintained near-target miss-risk and improved utility: at α = 0.10, micro-F1 increased from 0.568 (CRC) to 0.574 while mean set size decreased from 2.92 to 2.86. Risk-coverage curves supported selective deployment (AURC=0.0952 (CRC) vs AURC=0.0968 (EA-CRC)). Clinician-facing narratives are generated using a local instruction-tuned LLM to rewrite deterministic base report, with a deterministic verifier blocking out-of-set diagnoses. This framework provides an auditable pathway toward missed-finding, risk-controlled ECG review sets with verifier-gated reporting.**

***Index Terms*—Electrocardiography, conformal prediction, conformal risk control, uncertainty quantification, retrieval-based evidence, large language models, clinical decision support.**

I. Introduction

Electrocardiography (ECG) is a ubiquitous, low-cost, and noninvasive test that anchors clinical screening, triage, and longitudinal monitoring for cardiovascular disease. Deep learning has demonstrated strong performance across multiple ECG tasks, including screening for structural dysfunction from the 12-lead ECG [1], arrhythmia detection in ambulatory recordings [2], and large-scale 12-lead abnormality classification [3]. Despite these advances, deployment in high-stakes workflows is limited by two practical barriers. First, clinicians need decision-aligned uncertainty, meaning a reliable estimate of how likely the model is to miss true findings. Second, clinicians need verifiable communication so that any AI-generated report cannot invent diagnoses or measurements.

Recent representation learning and foundation-model efforts further improve ECG feature quality and label efficiency. Contrastive and self-supervised pretraining strategies (*e.g.,* CLOCS [4] and lead-agnostic objectives [5]) learn transferable ECG representations that reduce reliance on large labeled datasets. More recently, open ECG foundation models such as ECG-FM demonstrate strong generalization and label efficiency across downstream tasks [6]. The first research gap is that even highly capable encoders typically output probabilities or fixed label lists, leaving open the central question for clinical decision support: how to produce reliable, decision-aligned outputs that explicitly control the risk of missed findings.

Uncertainty estimation and calibration methods can provide partial answers but are not enough. Bayesian approximations and ensembles can improve uncertainty quantification, but they do not provide distribution-free guarantees and are computationally expensive [7], [8]. Selective prediction provides a risk–coverage trade-off by abstaining on uncertain cases

[9], but still requires a principled mechanism to translate model outputs into guarantees that align with clinical objectives. Conformal prediction offers distribution-free uncertainty quantification under exchangeability [10], and conformal risk control (CRC) generalizes conformal methods to control the expected value of any bounded monotone loss, enabling guarantees that directly target missed-finding risk in multi-label prediction [11]. A second research gap is how to leverage these guarantees while improving practical utility, since naive risk control can yield overly large prediction sets.

In parallel, there is rapid growth in ECG–text and language-enabled ECG systems. Multimodal ECG–report representation learning aligns ECG signals with clinical text to improve classification or enable text interfaces [12]–[14]. Large language model (LLM)-based report generation has shown promising fluency and perceived usefulness [15], and retrieval-augmented ECG reporting systems demonstrate that nearest-neighbor evidence can improve text generation quality [16]. However, these approaches face reliability risks: free-form generation can hallucinate unsupported diagnoses and fabricate measurements, and retrieval is often used heuristically rather than embedded into a formal, auditable reliability objective. This creates a third research gap: how to convert retrieval from a helpful add-on into principled evidence signal that supports reliability and safe clinical communication.

To address these gaps, we propose an evidence-aware, risk-controlled ECG decision support pipeline that replaces single-label decisions with *risk-controlled diagnosis sets* and to improve their utility using *retrieval evidence*, while enforcing verifiable report generation. Specifically, we (i) apply CRC to produce diagnosis sets calibrated to a target miss-risk α under a multi-label missed-finding loss; (ii) introduce *evidence-augmented CRC (EA-CRC)* that integrates model probability with a retrieval-derived support score computed from a similarity index (FAISS-style nearest-neighbor search [17]) to preferentially retain evidence-supported labels and prune low-evidence labels while preserving risk control; (iii) propose verified reporting via a code-locked LLM rewrite and deterministic verifier that prevents the generated report from introducing diagnoses outside the calibrated set and from fabricating numeric measurements.

II. Methods

A. PTB-XL Dataset Description

We used PTB-XL, a large open-access clinical 12-lead ECG resource comprising 21,799 10-second recordings from 18,869 patients [18]–[20]. Each ECG was annotated by up to two cardiologists using the SCP-ECG standard, yielding 71 statement codes spanning diagnostic, form, and rhythm categories. There are totally $K = 23$ diagnostic subclasses, including `NORM` (normal ECG), `AMI` (anterior/anteroseptal/anterolateral myocardial infarction or injury patterns), `IMI` (inferior/inferolateral/inferoposterior myocardial infarction or injury patterns), `LMI` (lateral myocardial infarction), `PMI` (posterior myocardial infarction), `STTC` (ST/T change statements, including T-wave abnormalities, long QT interval, digitalis/electrolyte-related patterns, and ventricular aneurysm-compatible ST/T changes), `NST` (non-specific ST changes), `ISC` (non-specific ischemic ST–T changes), `ISCA` (ischemic ST–T changes in anterior/lateral leads), `ISCI` (ischemic ST–T changes in inferior/inferolateral leads), `LVH` (left ventricular hypertrophy), `RVH` (right ventricular hypertrophy), `SEHYP` (septal hypertrophy), `LAO/LAE` (left atrial overload/enlargement), `RAO/RAE` (right atrial overload/enlargement), `LAFB/LPFB` (left anterior or posterior fascicular block), `IRBBB` (incomplete right bundle branch block), `CRBBB` (complete right bundle branch block), `CLBBB` (complete left bundle branch block), `ILBBB` (incomplete left bundle branch block), `IVCD` (non-specific intraventricular conduction disturbance), `AVB` (atrioventricular block, including first-, second-, and third-degree), and `WPW` (Wolff–Parkinson–White pattern).

The present study focuses on the 23 PTB-XL diagnostic subclasses and does not attempt to cover the full space of clinically actionable ECG interpretation, including detailed rhythm adjudication, interval measurements, axis measurements, waveform morphology descriptors, device-specific artifacts, or downstream management recommendations. Thus, the output set $C(x)$ should be interpreted as a calibrated review list over the available PTB-XL diagnostic taxonomy. This distinction is important because missing different findings may carry different clinical consequences. For example, missed acute ischemic patterns, conduction disease, or pre-excitation may not have the same clinical tolerance as lower-acuity nonspecific ST–T abnormalities. We therefore view the global miss-risk target α as a first-order safety objective, with diagnosis-specific or workflow-specific risk targets left for future clinical deployment studies.

B. Overview and Study Design

Fig. 1 summarizes the end-to-end pipeline and the evaluation protocol. Our method integrates three coupled components: (i) *risk-controlled multi-label set prediction* via conformal risk control (CRC), which targets a decision-aligned missed-finding loss [11]; (ii) *evidence-augmented risk control* (EA-CRC), which incorporates retrieval evidence by combining model probabilities with a nearest-neighbor support score within the same CRC framework; and (iii) *verified reporting*, in which an LLM is restricted to code-locked rewriting and is gated by a deterministic verifier. We followed the PTB-XL patient-wise folds with a strict separation of roles: folds 1–7 were used to train the encoder, fold 8 was used to select the epoch budget, fold 9 was reserved exclusively for CRC/EA-CRC calibration, and fold 10 was held out for final evaluation. The retrieval index and evidence summaries were constructed using folds 1–8 only, ensuring that calibration and test data did not influence evidence construction or model selection.

C. Problem Formulation

Let $x \in \mathbb{R}^{L\times T}$ denote a 12-lead ECG waveform with $L = 12$ leads and T samples, and let $y \in \{0, 1\}^K$ denote the associated multi-label target over K diagnostic subclasses. A

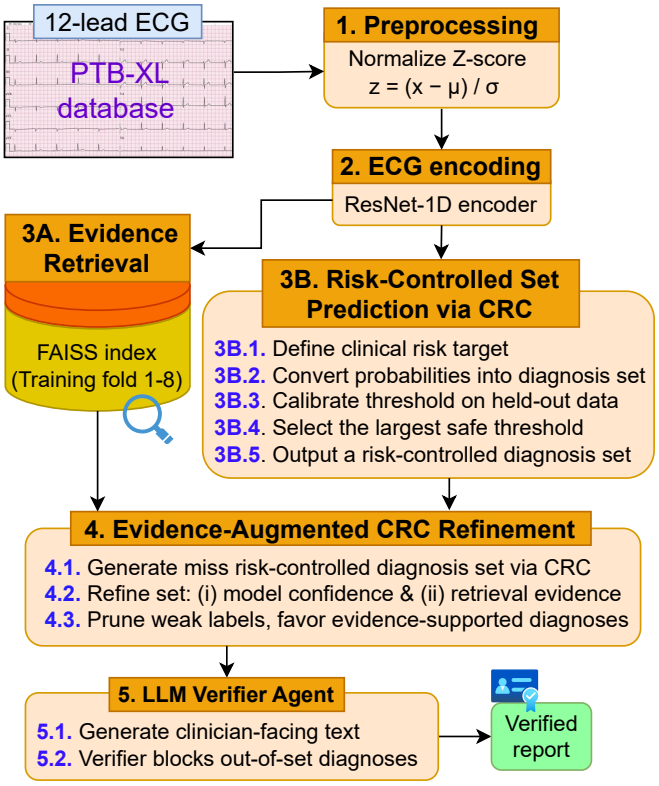

Fig. 1. Overview of the proposed framework and protocol.

trained encoder with parameters θ outputs logits $z_\theta(x) \in \mathbb{R}^K$ and per-label probabilities $p_\theta(x) = \sigma(z_\theta(x)) \in [0,1]^K$, where $\sigma(\cdot)$ is the element-wise sigmoid. Next, we output a *diagnosis set* $C(x) \subseteq \{1, 2, \ldots, K\}$, interpreted as the list of diagnoses that warrant clinical review. To quantify missed findings in the multi-label setting, we define a normalized missed-finding loss as the fraction of true labels omitted from $C(x)$:

$$\ell_{\text{miss}}(C, y) = \begin{cases} 1 - \dfrac{|C \cap Y^+|}{|Y^+|} & if\ |Y^+| > 0 \\ 0 & if\ |Y^+| = 0 \end{cases} \quad (1)$$

where $Y^+ = \{k \in \{1, \ldots, K\} : y_k = 1\}$ is the set of ground-truth positive labels. By construction, $\ell_{\text{miss}}(C, y) \in [0, 1]$ and is monotone in C: if $C \subseteq C'$, then $\ell_{\text{miss}}(C', y) \leq \ell_{\text{miss}}(C, y)$. Our reliability objective is to control the *expected miss-risk* $R(C) = \mathbb{E}\big[\ell_{\text{miss}}(C(X), Y)\big] \leq \alpha$, for a user-specified target level $\alpha \in (0, 1)$.

D. Preprocessing and Signal Quality

[I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6] is used as 12-lead ECG ordering, and we extract a single 10 s segment per record. We use the PTB-XL 100 Hz release so each lead contains $T = 1000$ samples. Let μ_ℓ and σ_ℓ denote the mean and standard deviation of lead ℓ estimated on training folds only. We apply per-lead z-score normalization to obtain the model input $\tilde{x}$. PTB-XL provides structured signal-quality annotations including baseline drift static noise burst noise and electrode problems as well as event indicators including extra beats and pacemaker [19]. We summarize these annotations into a binary quality flag used only for selective deployment analysis. Let $a_m(x)$ denote the mth quality field and M the number of quality fields then $q(x) = \sum_{m=1}^{M} \mathbb{I}(a_m(x) \neq \varnothing)$.

E. Credibility score for risk–coverage ordering

For selective deployment we sort test examples by a scalar credibility score computed from the predicted diagnosis set size, retrieval evidence, and a binary signal-quality flag. Let $C(x)$ denote the predicted diagnosis set and let $s_k(x)$ denote the retrieval support score for label k. We define the minimum in-set support as follows:

$$s_{\min}(x) = \begin{cases} \min_{k \in C(x)} s_k(x) & |C(x)| > 0 \\ 0 & |C(x)| = 0 \end{cases} \quad (2)$$

and define a binary quality flag from PTB-XL annotations $g(x) = \mathbb{I}(q(x) \geq 1)$. We then compute credibility as:

$$\text{cred}(x) = \frac{s_{\min}(x)}{1 + |C(x)|} - g(x) \quad (3)$$

Higher $\text{cred}(x)$ indicates a smaller set with stronger evidence and no quality flag. Risk–coverage curves are obtained by accepting examples in decreasing order of $\text{cred}(x)$ and computing selective miss-risk on the accepted subset.

F. Encoder and Training

We use a ResNet-1D encoder [21] that maps the normalized ECG $\tilde{x}$ to an embedding $h_\theta(\tilde{x}) \in \mathbb{R}^d$ with $d = 512$. A linear classification head produces logits $z_\theta(\tilde{x}) \in \mathbb{R}^K$ and probabilities $p_\theta(\tilde{x}) = \sigma(z_\theta(\tilde{x}))$, where $\sigma(\cdot)$ is the element-wise sigmoid. We train using weighted multi-label cross entropy:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \, \text{BCE}(y_{ik}, p_{\theta k}(\tilde{x}_i)) \quad (4)$$

$$\text{BCE}(y, p) = -y \log p - (1 - y) \log (1 - p) \quad (5)$$

Model selection is performed on fold 8 and the final encoder is retrained on folds 1–8 using the selected epoch budget.

G. Risk-Controlled Set Prediction via CRC

Given per-label probabilities $p_\theta(\tilde{x})$, we output a diagnosis set by thresholding with a single global parameter t:

$$C_t(x) = \{k \mid p_{\theta k}(\tilde{x}) \geq t\} \quad (6)$$

Larger t produces smaller sets and typically increases the missed-finding loss defined in Sec. II-C. We calibrate t on the held-out calibration fold using conformal risk control [11]. Let $\{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$ denote the calibration examples. The empirical miss-risk at threshold t is:

$$\widehat{R}_{\text{cal}}(t) = \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} \ell_{\text{miss}}(C_t(x_i), y_i) \quad (7)$$

Since $\ell_{\text{miss}} \in [0, 1]$, we set the loss upper bound $B = 1$ and choose the largest threshold $t^\star$ that satisfies the CRC finite-sample inequality:

$$\frac{n_{\text{cal}}}{n_{\text{cal}} + 1} \widehat{R}_{\text{cal}}(t^\star) + \frac{B}{n_{\text{cal}} + 1} \leq \alpha \quad (8)$$

This choice yields the smallest diagnosis sets among all thresholds that are certified to meet the target expected miss-risk level. We compute $t^\star$ by monotone bisection because $\widehat{R}_{\mathrm{cal}}(t)$ is nondecreasing in t.

H. Evidence and Support Retrieval

1) Nearest-neighbor search: To obtain an evidence signal for each query ECG, we perform nearest-neighbor retrieval in the encoder embedding space. Let $h_\theta(\tilde{x}) \in \mathbb{R}^d$ denote the penultimate-layer embedding produced by the trained encoder. We convert this embedding into a unit-norm vector: $e(x) = \frac{h_\theta(\tilde{x})}{\|h_\theta(\tilde{x})\|_2}$, where $\|\cdot\|_2$ denotes the Euclidean norm.

We build a FAISS index over $\{e(x_j)\}$ from folds 1–8 only and retrieve the top K_{NN} nearest neighbors [17]. Since all embeddings are ℓ_2 normalized, similarity between a query x and a retrieved neighbor x_j is computed by cosine similarity $\mathrm{sim}_j = e(x)^\top e(x_j)$, where $\mathrm{sim}_j \in [-1, 1]$. We evaluate $K_{\mathrm{NN}} \in \{1, 3, 5, 10\}$ and use the retrieved neighbor labels and similarities to compute per-label evidence support.

2) Per-label support score: Given a query ECG x, let $\mathcal{N}(x)$ denote the set of retrieved neighbors with $|\mathcal{N}(x)| = K_{\mathrm{NN}}$. For each neighbor $x_j \in \mathcal{N}(x)$ we denote its cosine similarity to the query by sim_j and its multi-label target by $y_j \in \{0,1\}^K$ with component y_{jk} for label k. We define two complementary evidence signals for label k. First, the count-based support is the fraction of retrieved neighbors that contain label k:

$$s_k^{\mathrm{cnt}}(x) = \frac{1}{K_{\mathrm{NN}}} \sum_{j \in \mathcal{N}(x)} \mathbb{I}(y_{jk} = 1) \tag{9}$$

Second, the similarity support upweights closer neighbors:

$$s_k^{\mathrm{sim}}(x) = \frac{1}{K_{\mathrm{NN}}} \sum_{j \in \mathcal{N}(x)} \mathrm{sim}_j \, \mathbb{I}(y_{jk} = 1) \tag{10}$$

We combine these into a single bounded evidence score:

$$s_k(x) = \frac{1}{2} \left(s_k^{\mathrm{cnt}}(x) + s_k^{\mathrm{sim}}(x) \right) \tag{11}$$

which increases when label k appears frequently among neighbors and when those neighbors are highly similar to the query.

I. Evidence-Augmented CRC (EA-CRC)

EA-CRC refines risk-controlled diagnosis sets by combining two complementary signals for each label k: $p_{\theta k}(\tilde{x})$ and $s_k(x)$ from Sec. II-H. We define a per-label score that penalizes low confidence and low evidence:

$$\mathrm{score}_k(x) = (1 - p_{\theta k}(\tilde{x})) + \beta \, (1 - s_k(x)) \tag{12}$$

where $\beta \geq 0$ controls the strength of the evidence term. Smaller scores correspond to labels that are both more probable under the encoder and better supported by retrieved neighbors. Given a threshold t, EA-CRC outputs the score-thresholded diagnosis set $C_{t\beta}^{\mathrm{EA}}(x) = \{k \mid \mathrm{score}_k(x) \leq t\}$.

For fixed β, increasing t can only add labels to the set and therefore cannot increase the missed-finding loss, implying a monotone risk–size trade-off. For each candidate β, we compute the empirical miss-risk on the calibration fold:

$$\widehat{R}_{\mathrm{cal}}(t, \beta) = \frac{1}{n_{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{cal}}} \ell_{\mathrm{miss}}\big(C_{t\beta}^{\mathrm{EA}}(x_i), y_i\big) \tag{13}$$

and select the smallest threshold $t_\beta^\star$ that satisfies the CRC inequality at target level α. We then choose $\beta^\star$ by grid search over a fixed candidate set, selecting the value that yields the smallest calibration-set mean set size among all risk-feasible pairs $\left(\beta, t_\beta^\star\right)$. Final performance is reported once on the held-out test fold using $\beta^\star$ and $t_{\beta^\star}^\star$.

J. Verified Report Generation (Verifier-Gated LLM Rewrite)

To produce clinician-facing text while preventing unsupported clinical claims, we use a verifier-gated code-locked rewrite procedure [22]. For each ECG x, the calibrated diagnosis set $C(x)$ is treated as the sole source of diagnostic content. We first generate a deterministic base report that enumerates only the diagnostic codes in $C(x)$ and intentionally contains no numeric measurements.

A local instruction-tuned LLM (Phi-3-mini-4k-instruct [23]) rewrites the base report into fluent clinical prose under an explicit constraint prompt that (i) forbids introducing any diagnostic codes beyond the calibrated set, (ii) disallows numeric measurements or interval values, and (iii) keep the output concise (at most eight lines), ensuring that the model functions purely as a constrained surface-realization module. We use deterministic decoding (sampling disabled) with a mild repetition penalty (1.05) to ensure reproducibility. In addition, to enforce this constraint, we apply a deterministic verifier to the rewritten output. The verifier checks two conditions: *code fidelity* and *measurement safety*. Code fidelity requires that every diagnosis mentioned in the rewritten report belongs to $C(x)$ and that no additional codes are introduced. Measurement safety blocks measurement-like text patterns to prevent fabricated quantitative statements. If either condition fails, the output is rejected and a single repair attempt is made using a stricter rewrite prompt. If the second attempt fails, the system falls back to the deterministic base report.

K. Evaluation and Statistical Analysis

We evaluate on the held-out PTB-XL test fold 10. Primary set-prediction outcomes are the achieved miss-risk $\widehat{R}_{\mathrm{te}} = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \ell_{\mathrm{miss}}(C(x_i), y_i)$ and the average diagnosis-set size $\widehat{S}_{\mathrm{te}} = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} |C(x_i)|$, where n_{te} is the number of test examples. For utility we also report micro-averaged precision recall and F1 computed by aggregating true positives false positives and false negatives across all labels and all test examples. Specifically let $\mathrm{TP} = \sum_{i=1}^{n_{\mathrm{te}}} \sum_{k=1}^{K} \mathbb{I}(k \in C(x_i)) \, \mathbb{I}(y_{ik} = 1)$, $\mathrm{FP} = \sum_{i=1}^{n_{\mathrm{te}}} \sum_{k=1}^{K} \mathbb{I}(k \in C(x_i)) \, \mathbb{I}(y_{ik} = 0)$, and $\mathrm{FN} = \sum_{i=1}^{n_{\mathrm{te}}} \sum_{k=1}^{K} \mathbb{I}(k \notin C(x_i)) \, \mathbb{I}(y_{ik} = 1)$. Then $\mathrm{Prec}_\mu = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$, $\mathrm{Rec}_\mu = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$, and $\mathrm{F1}_\mu = \frac{2 \, \mathrm{Prec}_\mu \, \mathrm{Rec}_\mu}{\mathrm{Prec}_\mu + \mathrm{Rec}_\mu}$.

III. Results

A. Point-Prediction Encoder-Based Classification Baseline

We first evaluated the ResNet-1D encoder to show whether it produced a strong probabilistic foundation before introducing risk-controlled set prediction. Metrics (Prec, Rec, Spec, and F1) were evaluated after selecting a per-subclass probability threshold that maximized F1 on fold 9. The tuned thresholds were then applied once to the test fold 10. The one-vs-rest ROC curves in Fig. 2 showed consistently strong separability for most diagnostic subclasses. Table I summarized tuned operating-point behavior across all $K = 23$ subclasses. Several rare subclasses exhibited unstable operating-point metrics due to extremely limited support.

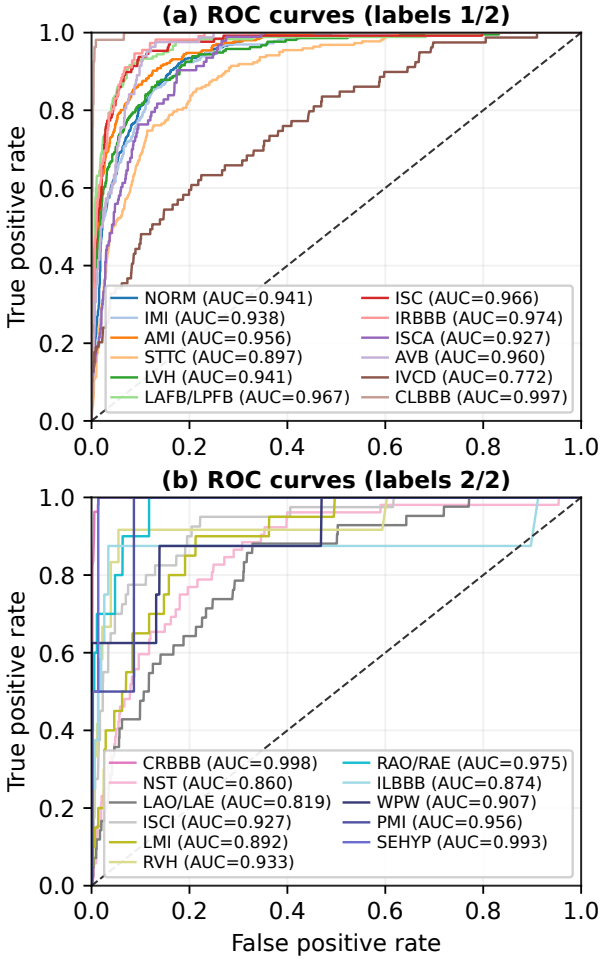

Fig. 2. One-vs-rest ROC curves for the 23 diagnostic subclasses on test fold 10. For readability, the subclasses are shown in two panels (a) and (b).

B. Risk-Controlled and Evidence-Augmented Set Prediction

After grid-search tuning on the held-out calibration fold, we selected TOPK$= 10$ and $\beta = 0.8$ as the default EA-CRC configuration for all subsequent experiments, since it achieved near-optimal utility while maintaining the target miss-risk constraint. CRC was first used to convert encoder probabilities into a clinically interpretable diagnosis set with explicit control of missed-finding risk. For each target level α, CRC calibrated a single threshold on the calibration fold (fold 9) under the multi-label miss-loss in Sec. II-C and was evaluated once on the held-out test fold (fold 10). EA-CRC followed the same protocol but incorporated retrieval-derived support into the decision rule, and was evaluated under identical splits.

TABLE I

Per-subclass operating-point metrics on PTB-XL test fold 10 for $K = 23$ diagnostic subclasses. Pos(%) denotes the number of positive test examples. Thr denotes per-label thresholds. Rare subclasses with prevalence $< 1\%$ are marked †.

Label	Pos (%)	Thr	Prec↑	Rec↑	Spec↑	F1↑
NORM	963 (43.8%)	0.48	0.774	0.936	0.787	0.847
IMI	327 (14.9%)	0.58	0.610	0.731	0.918	0.665
AMI	306 (13.9%)	0.72	0.707	0.775	0.948	0.739
STTC	222 (10.1%)	0.66	0.425	0.725	0.890	0.536
LVH	214 (9.7%)	0.63	0.579	0.738	0.942	0.649
LAFB/LPFB	179 (8.1%)	0.77	0.682	0.754	0.969	0.716
ISC	128 (5.8%)	0.90	0.650	0.609	0.980	0.629
IRBBB	112 (5.1%)	0.81	0.565	0.696	0.971	0.624
ISCA	93 (4.2%)	0.83	0.389	0.452	0.969	0.418
AVB	82 (3.7%)	0.59	0.423	0.573	0.970	0.487
IVCD	79 (3.6%)	0.82	0.233	0.253	0.969	0.242
CLBBB	54 (2.5%)	0.81	0.855	0.870	0.996	0.862
CRBBB	54 (2.5%)	0.74	0.743	0.963	0.992	0.839
NST	52 (2.4%)	0.69	0.157	0.346	0.955	0.216
LAO/LAE	42 (1.9%)	0.70	0.130	0.167	0.978	0.146
ISCI	40 (1.8%)	0.57	0.281	0.450	0.979	0.346
LMI†	20 (0.9%)	0.36	0.083	0.250	0.975	0.125
RVH†	12 (0.5%)	0.97	0.188	0.250	0.994	0.214
RAO/RAE†	10 (0.5%)	0.94	0.571	0.400	0.999	0.471
ILBBB†	8 (0.4%)	0.50	0.143	0.375	0.992	0.207
WPW†	8 (0.4%)	0.50	0.714	0.625	0.999	0.667
PMI†	2 (0.1%)	0.50	0.125	0.500	0.997	0.200
SEHYP†	2 (0.1%)	0.50	0.200	0.500	0.998	0.286
Macro	–	–	0.445	0.563	0.964	0.484

Fig. 3 summarizes the operating characteristics across risk targets. In Fig. 3a, the achieved miss-risk tracked the target α closely for both CRC and EA-CRC. As α increased from 0.05 to 0.20, the system behaved as an explicit "risk knob": higher tolerated miss-risk produced more selective outputs and smaller diagnosis sets. Fig. 3b quantified this efficiency trade-off. CRC reduced the average set size from 4.13 at $\alpha = 0.05$ to 1.95 at $\alpha = 0.20$, while EA-CRC produced comparable or smaller sets over most operating points. Fig. 3c reported corresponding utility using micro F1. EA-CRC improved micro F1 at the nominal operating point $\alpha = 0.1$ (from 0.568 to 0.574), consistent with evidence acting primarily to prune weakly supported labels and improve set composition.

C. Selective Deployment via Risk–Coverage Analysis

To quantify this operating mode, we evaluated *risk–coverage* behavior on the PTB-XL test fold by sorting ECGs using the credibility score $\mathrm{cred}(x)$ and then computing performance on the accepted fraction of cases. Coverage denotes the fraction of test cases accepted (highest-credibility first), and selective miss-risk is the mean missed-finding loss computed only over the accepted subset.

Fig. 4 shows that both CRC and EA-CRC isolate a low-risk subset at low coverage, indicating that credibility-based

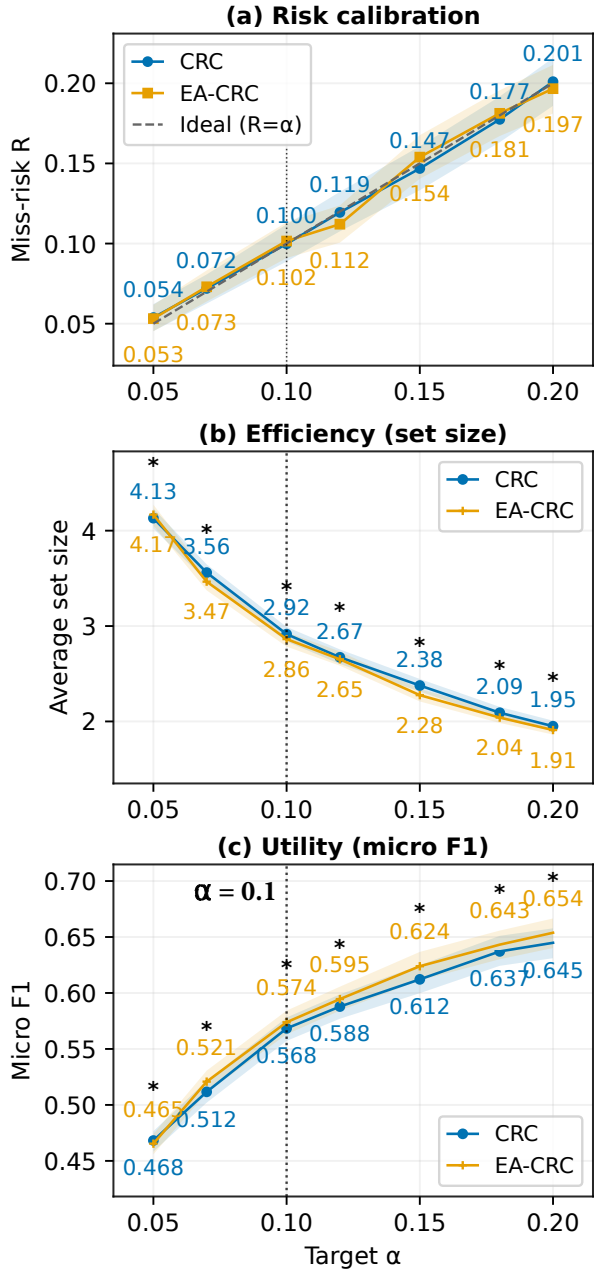

Fig. 3. Operating characteristics across risk targets α (TOPK=10, fixed $\beta = 0.8$). (a) Achieved miss-risk R on the test fold versus target α with the ideal reference $R = \alpha$. (b) Average diagnosis-set size versus α (efficiency). (c) Micro F1 versus α (utility). Shaded bands denote 95% bootstrap confidence intervals. Asterisks indicate operating points where the paired bootstrap confidence interval of Δ(EA-CRC$-$CRC) excluded zero.

ordering concentrates reliability in the earliest accepted cases. As coverage increases, selective miss-risk rises and approaches the overall operating regime near the target level $\alpha = 0.10$, consistent with the expected trade-off between automation coverage and error. We summarize ranking quality using the area under the risk–coverage curve (AURC), where lower values indicate better separation between low- and high-risk cases. CRC achieved AURC $= 0.0952$, while EA-CRC achieved AURC $= 0.0968$, suggesting that evidence augmentation primarily improves *set composition* under comparable miss-risk, while credibility-based ranking remains similar between methods.

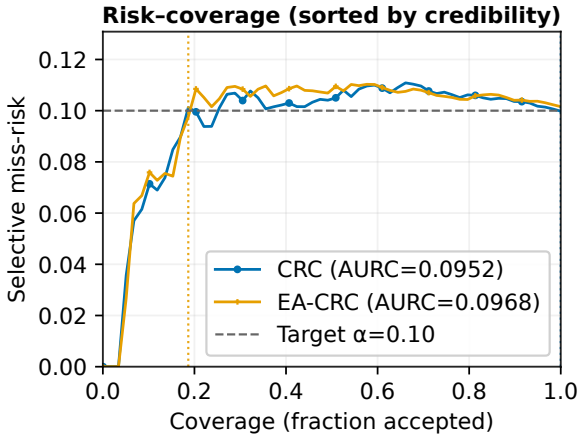

Fig. 4. Risk–coverage curves on PTB-XL test fold (10) with examples sorted by a credibility score computed from the predicted diagnosis set and retrieval evidence with an additional penalty for PTB-XL signal-quality flags. Coverage is the fraction accepted, and selective miss-risk is the mean missed-finding loss over accepted cases.

These results indicate that EA-CRC provides a modest improvement in review-set composition rather than a large overall performance gain. The primary value of EA-CRC is that retrieval evidence participates directly in the calibrated decision rule, allowing weakly supported labels to be pruned while maintaining near-target missed-finding risk.

D. Case Study: Certificates and Mechanistic Interpretation

To complement aggregate metrics, we present one case-study certificate for EA-CRC at TOPK=10, $\beta = 0.8$, and $\alpha = 0.10$. Each certificate reports the reference findings, the CRC review list, the EA-CRC review list, the labels pruned by EA-CRC, and retrieval-based evidence summaries. The goal of this analysis is to demonstrate that EA-CRC's pruning behavior aligns with case-based evidence while preserving missed-finding risk control.

Case 1 — *ecg_id=14534, GI=14497*

Setting	TOPK=10, $\beta = 0.8$, $\alpha = 0.10$
Thresholds	CRC $t^\star = 0.236$; EA $\tau^\star = 1.480$
Summary	$\lvert C_{\mathrm{CRC}}\rvert = 7 \rightarrow \lvert C_{\mathrm{EA}}\rvert = 5$; pruned 2 labels; no missed reference findings

Reference findings: `AMI, CRBBB, LAFB/LPFB, AVB`
CRC candidate list: `AMI, CRBBB, IRBBB, LAFB/LPFB, LMI, RVH, AVB`
EA-CRC candidate list: `AMI, CRBBB, LAFB/LPFB, LMI, AVB`
Pruned: `IRBBB, RVH`; **Added:** `none`
Evidence for EA-CRC findings: `AMI` 8/10, support=0.763; `CRBBB` 10/10, support=0.953; `LAFB/LPFB` 9/10, support=0.858; `LMI` 1/10, support=0.095; `AVB` 3/10, support=0.285.
Top similar ECGs: sim=0.921, labels=`AMI,CRBBB,LAFB/LPFB`; sim=0.916, labels=`AMI,CRBBB,LAFB/LPFB`; sim=0.914, labels=`AMI,CRBBB,LAFB/LPFB`.

IV. Discussion

Empirically, both CRC and EA-CRC track the target miss-risk across operating points while providing a practical accuracy–efficiency trade-off: higher tolerated miss-risk yields smaller diagnosis sets, and evidence augmentation improves set composition with comparable reliability. The key idea in EA-CRC is to treat retrieval as an evidence signal that participates directly in the decision rule, rather than as a post-hoc justification step. Using learned embeddings for nearest-neighbor support connects naturally to dense retrieval paradigms that have proven effective for grounding downstream generation

and decisions in retrieved examples [24]. Efficient approximate nearest-neighbor indexing further supports the feasibility of retrieval-backed evidence at scale [25]. In our case certificates, evidence support helps prune weak labels that are plausible under the encoder alone, improving review-list precision without sacrificing reference findings.

The risk–coverage analysis suggests a second deployment mode beyond fixed α, which is selective automation. Classic reject-option theory formalizes trading coverage for lower error, and modern selective classification frameworks analyze the same risk–coverage frontier that appears in our credibility-ordered curves. Operationally, this supports a sustainable workflow in which high-credibility ECGs can be auto-processed with bounded miss-risk while lower-credibility cases are deferred for clinician review, reducing alert fatigue and concentrating human attention where it matters most.

There are several limitations that motivate our future work. First, distribution-free guarantees rely on calibration data being representative of deployment; covariate shift can violate exchangeability and degrade calibration, suggesting the need for shift-aware conformal methods and external validation across sites and acquisition conditions. Second, retrieval evidence can inherit dataset biases and embedding blind spots; future work should incorporate quality-aware retrieval, uncertainty-aware similarity, and clinically meaningful subgroup audits. Finally, healthcare algorithms can encode structural inequities even without using sensitive attributes directly.

V. Conclusion

We introduced an ECG decision-support framework that couples decision-aligned uncertainty with verifiable communication. CRC provides diagnosis sets with explicit control of missed-finding risk, and EA-CRC injects nearest-neighbor evidence to improve efficiency by pruning weakly supported labels while preserving near-target risk. A risk–coverage analysis further supports selective deployment, enabling low-risk automation with principled deferral. To prevent unsafe free-form narratives, we restrict an instruction-tuned LLM to code-locked rewriting and enforce a deterministic verifier that blocks out-of-set diagnoses and measurement-like text. Future work will prioritize external validation, quality-aware and bias-audited retrieval, and integration with stronger ECG foundation encoders for improved generalization.

References

[1] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger, *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram," *Nature Medicine*, vol. 25, pp. 70–74, 2019.

[2] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, pp. 65–69, 2019.

[3] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixao, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, no. 1, p. 1760, 2020.

[4] D. Kiyasseh, T. Zhu, and D. A. Clifton, "CLOCS: Contrastive learning of cardiac signals across space, time, and patients," 2021.

[5] J. Oh, H. Chung, J.-m. Kwon, D.-g. Hong, and E. Choi, "Lead-agnostic self-supervised learning for local and global representations of electrocardiogram," 2022.

[6] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: an open electrocardiogram foundation model," *JAMIA Open*, vol. 8, no. 5, p. ooaf122, 2025.

[7] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, vol. 48 of *Proceedings of Machine Learning Research*, 2016.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," 2017.

[10] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.

[11] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *International Conference on Learning Representations (ICLR)*, 2024. Also available as arXiv:2208.02814.

[12] C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci, "MERL: Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement," 2024.

[13] J. Chen, X. Dong, W. Wang, S. Zhou, L. Yu, and X. Hu, "DERI: Cross-modal ECG representation learning with deep ECG-report interaction," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 4824–4832, 2025.

[14] M. Pham, A. Saeed, and D. Ma, "C-MELT: Contrastive enhanced masked auto-encoders for ECG-language pre-training," 2024.

[15] Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, J. Xiong, R. Arcucci, H. Yao, and M. Zhang, "MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation," in *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14510–14527, 2025.

[16] J. Tang, T. Xia, Y. Lu, C. Mascolo, and A. Saeed, "Electrocardiogram report generation and question answering via retrieval-augmented self-supervised modeling," 2024.

[17] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," 2017.

[18] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3)," *PhysioNet*, 2022. RRID:SCR_007345.

[19] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.

[20] 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*, vol. 101, no. 23, pp. e215–e220, 2000.

[21] N. Sakli, H. Ghabri, B. O. Soufiene, F. A. Almalki, H. Sakli, O. Ali, and M. Najjari, "ResNet-50 for 12-lead electrocardiogram automated diagnosis," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 7617551, 2022.

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] M. Ranjit, S. Srivastav, and T. Ganu, "RadPhi-3: Small language models for radiology," arXiv preprint arXiv:2411.13604, 2024.

[24] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Association for Computational Linguistics, 2020.

[25] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020.