Beyond Machine Interpretation: Learning from Expert Over-Reads Improves ECG Diagnosis

Sunwoo Kwak[1] SK3355@CORNELL.EDU
Fengbei Liu[1] FL453@CORNELL.EDU
Nusrat Binta Nizam[1] NN284@CORNELL.EDU
Ilan Richter[2] IR2498@CUMC.COLUMBIA.EDU
Nir Uriel[2] NU2126@CUMC.COLUMBIA.EDU
Peter M. Okin[3] POKIN@MED.CORNELL.EDU
Mert R. Sabuncu[1,4,5] MSABUNCU@CORNELL.EDU
[1] *Cornell Tech, New York, NY, USA*
[2] *Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA*
[3] *Department of Medicine, Weill Cornell Medicine, New York, NY, USA*
[4] *Department of Radiology, Weill Cornell Medicine, New York, NY, USA*
[5] *School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA*

Editors: Under Review for MIDL 2026

Abstract

Automated machine-read ECG interpretations are widely used in clinical practice but often unreliable, leading to systematic diagnostic errors. This work investigates how training with cardiologist over-reads impacts model accuracy and clinical reliability. Using a large paired corpus of over two million ECGs containing both machine and expert interpretations, we evaluate three learning paradigms: (i) supervised learning on expert over-read labels, (ii) Self-training that extends expert supervision to public ECGs, and (iii) multimodal contrastive learning with CLIP and NegCLIP. Across all settings, models trained with expert over-read data consistently outperform those trained on machine-read labels, especially for rare but clinically important conditions. Self-training and NegCLIP further demonstrate scalable strategies to propagate expert knowledge beyond labeled datasets. These findings highlight the essential role of expert over-reads in developing trustworthy and clinically aligned ECG AI systems.

Keywords: ECG, Expert Over-read, Self-training, Contrastive learning, Clinical AI

1. Introduction

Electrocardiography (ECG) is a cornerstone of early cardiovascular diagnosis, yet expert interpretation requires substantial specialty training, and the clinical workforce capable of performing high-quality ECG over-reads is shrinking—nearly half of U.S. counties have no cardiologist specializing in ECG interpretation (Kim et al., 2024). Consequently, many hospitals and clinics now rely heavily on automated machine-read interpretations produced by device-embedded algorithms (Smulyan, 2019; Brady et al., 2020). However, these systems are not reliably accurate: multiple studies report misclassification rates of 20–35% across arrhythmias, conduction abnormalities, and ischemic patterns (Schläpfer and Wellens, 2017; Kraik et al., 2025), prompting clinical guidelines to recommend that all automated ECG

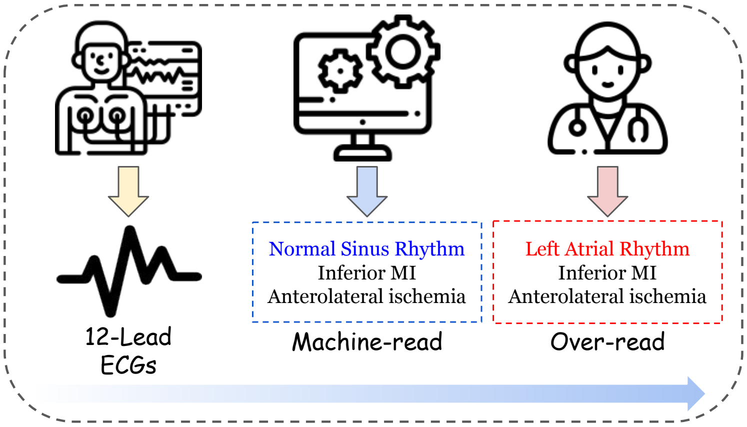

Figure 1: Comparison of machine-read and expert over-read ECG interpretations. Machine-read diagnoses are automatically produced by device algorithms, while expert over-reads are verified by cardiologists. This pairing enables direct analysis of diagnostic discrepancies for model training.

outputs be verified by cardiologist over-reads before informing patient care (McNamara et al., 2021). Despite this, both routine practice and much of ECG machine learning research continue to depend on machine-read labels, primarily because large-scale expert-annotated corpora remain difficult and costly to obtain (Hong et al., 2020). Influential public datasets—such as MIMIC-IV-ECG (Gow et al., 2023) and the PhysioNet 2021 Challenge (Moody et al., 2021)—have accelerated methodological progress but rely largely on automated or partially verified interpretations. Models trained on such supervision risk inheriting systematic errors embedded in machine-read diagnostics, especially for subtle or low-prevalence conditions, leaving a persistent gap between algorithmic performance and clinical reliability.

To address this challenge, we curated a large corpus of over two million 12-lead ECGs, each containing both the original machine-read interpretation and a cardiologist over-read provided by experts with more than 30 years of experience. Difference between two labels described in Figure 1. This paired design enables direct measurement of diagnostic discrepancies between machine and expert readings and allows models to learn from validated expert judgments rather than noisy automated labels. Our central hypothesis is that models trained on over-read data achieve higher diagnostic accuracy, robustness, and clinical alignment than those trained solely on machine-read interpretations.

In our study, the first two approaches—supervised and self-training—require label extraction from raw over-read text descriptions, while the contrastive approach can operate directly on the unstructured text itself. We evaluate three paradigms. **(i) Supervised learning:** a multi-label ResNet-50 classifier (He et al., 2016) trained on expert over-read labels, serving as a direct benchmark for expert-informed ECG interpretation. **(ii) Self-training:** a FixMatch-style framework (Sohn et al., 2020) that extends the supervised model by treating large public ECG datasets with machine-read labels (e.g., MIMIC-IV-ECG) as unlabeled samples, effectively leveraging expert supervision without requiring additional annotation. **(iii) CLIP-style contrastive learning:** a multimodal contrastive objective inspired by CLIP (Radford et al., 2021) and extended with NegCLIP ideas (Yuksekgonul et al., 2023), which jointly trains ECG signals with raw expert over-read text to align signal–text representations and explicitly separate machine biases from expert-level semantics.

Across these paradigms, each approach offers different strengths. Supervised learning provides a strong and interpretable baseline but is limited to the extracted label set. Self-training improves performance by combining expert-labeled and large public datasets, though it requires longer training and substantial unlabeled corpora. CLIP-style contrastive learning achieves comparable accuracy while remaining highly scalable, enabling new diagnostic phrases or expert text to be incorporated without full retraining. Crucially, in all settings, models trained on expert over-read labels consistently outperform those trained on machine-read data. This performance gap illustrates the clinical risk of relying solely on machine-read interpretations and highlights the necessity of large-scale expert over-read data for developing reliable ECG AI systems.

Key contributions: (1) We curate a large-scale paired ECG corpus—over two million 12-lead recordings with both machine-read interpretations and expert cardiologist over-reads—enabling direct quantification of diagnostic discrepancies at scale. (2) Using this corpus, we systematically evaluate supervised, self-training, and CLIP-style contrastive frameworks to assess how expert over-read supervision impacts model accuracy, robustness, and clinical alignment. (3) We show that expert-supervised models consistently outperform those trained on machine-read labels across major diagnostic categories.

2. Related Work

Machine-read ECGs are widely used in clinical workflows, yet clinically important inaccuracies persist; authoritative reviews consistently recommend that automated outputs be verified by clinician over-reads before informing care (Schläpfer and Wellens, 2017; Smulyan, 2019; McNamara et al., 2021). Despite this, large-scale expert annotations remain difficult to obtain, leading most public corpora to rely on partially verified or machine-generated interpretations.

Open ECG datasets such as MIMIC–IV–ECG and the PhysioNet 2021 Challenge have accelerated progress, but their label provenance is heterogeneous and often dominated by device-generated or mixed pipelines rather than comprehensive expert over-reads at scale (Gow et al., 2023; Moody et al., 2021). Models trained solely on such labels risk propagating systematic noise, especially for subtle, low-prevalence, or clinically consequential conditions (Hong et al., 2020). Only a few prior studies have examined machine–expert discrepancies, and none provide large paired corpora enabling systematic quantification across millions of ECGs.

In parallel, ECG foundation models emphasize scale and transfer. Recent efforts such as ECG-FM and ECGFounder pretrain on over one to ten million ECGs and report strong performance across diagnostic and prognostic tasks (McKeen et al., 2025; Li et al., 2025). However, these efforts also rely primarily on machine-read or heterogeneous labels, reinforcing the field's dependence on imperfect supervision.

Our work complements these directions by focusing on who provides the supervision. Using a uniquely paired corpus containing both machine-read interpretations and cardiologist over-reads, we quantify how expert-verified supervision alters downstream reliability relative to machine-read labels. We further explore scalable learning paradigms—supervised learning, FixMatch-style self-training (Sohn et al., 2020), and multimodal contrastive learning (Radford et al., 2021; Yuksekgonul et al., 2023)—to propagate expert signal and align model behavior with validated clinical judgment.

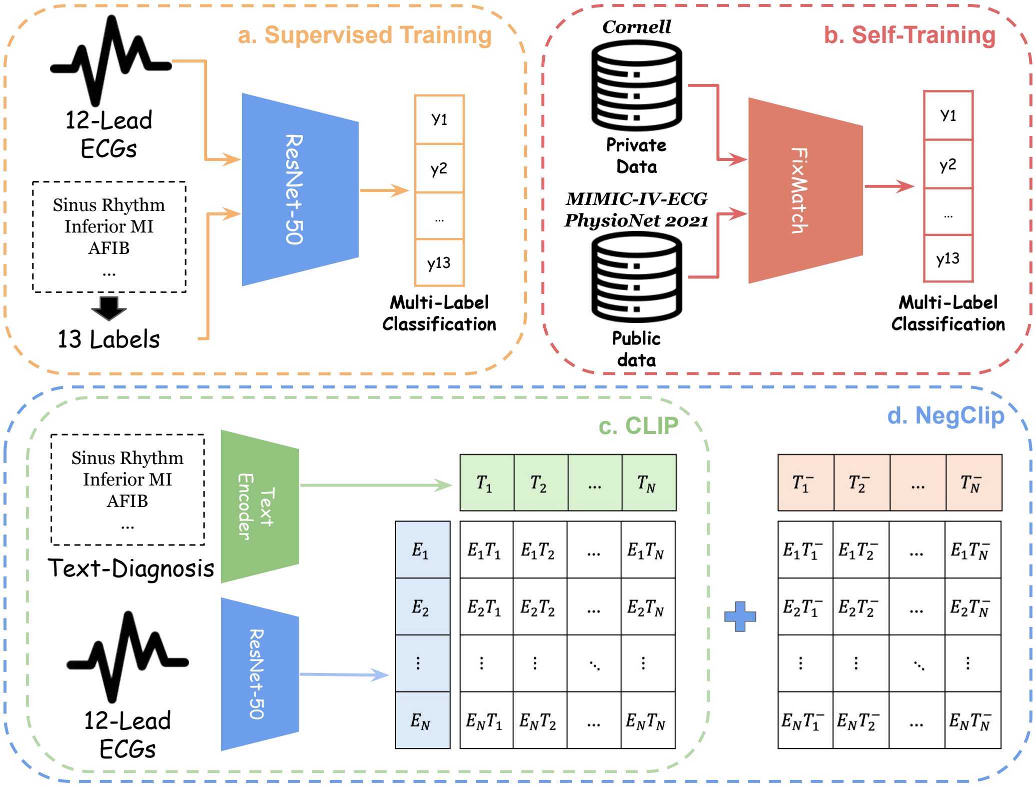
a. Supervised Training
12-Lead ECGs
Sinus Rhythm
Inferior MI
AFIB
...
13 Labels
ResNet-50
y1
y2
...
y13
Multi-Label Classification
b. Self-Training
Cornell
Private Data
MIMIC-IV-ECG PhysioNet 2021
Public data
FixMatch
y1
y2
...
y13
Multi-Label Classification
c. CLIP
Sinus Rhythm
Inferior MI
AFIB
...
Text-Diagnosis
Text Encoder
12-Lead ECGs
ResNet-50
T_1 T_2 ... T_N
E_1 E_2 ⋮ E_N
E_1T_1 E_1T_2 ... E_1T_N
E_2T_1 E_2T_2 ... E_2T_N
E_NT_1 E_NT_2 ... E_NT_N
d. NegClip
T_1^- T_2^- ... T_N^-
$E_1T_1^-$ $E_1T_2^-$ ... $E_1T_N^-$
$E_2T_1^-$ $E_2T_2^-$ ... $E_2T_N^-$
$E_NT_1^-$ $E_NT_2^-$ ... $E_NT_N^-$

Figure 2: Overview of the four learning paradigms evaluated in this study: (a) **Supervised learning** using expert over-read labels, (b) **Self-training** that leverages unlabeled public ECGs through pseudo-labeling, (c) **CLIP-style contrastive learning** aligning ECG signals with expert text, and (d) **NegCLIP** extending CLIP with hard-negative sampling for finer signal–text discrimination. Together, these frameworks represent complementary strategies for integrating expert knowledge into ECG AI models.

3. Methods

Preliminary We denote the over-read available training set as $\mathcal{D}_L = \{\mathbf{x}_i, \mathbf{t}_i, \hat{\mathbf{t}}_i\}_{i=1}^{|\mathcal{D}_L|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{L \times S}$ is the ECG signal with shape L leads and S length, and $\mathbf{t}_i$ and $\hat{\mathbf{t}}_i$ are the corresponding over-read and machine-read textual reports. We further derive over-read diagnostic labels $\mathbf{y}_i \in \{0,1\}^C$ from $\mathbf{t}_i$ and machine-read labels $\hat{\mathbf{y}}_i \in \{0,1\}^C$ from $\hat{\mathbf{t}}_i$ using the same label extraction process, C is the number of classes and both labels are multi-hot vectors. We also define public dataset with no over-read as $\mathcal{D}_U = \{\tilde{\mathbf{x}}\}_{i=1}^{|\mathcal{D}_U|}$.

An overview of the four learning paradigms evaluated in this study—supervised learning, self-training, CLIP-style contrastive learning, and NegCLIP—is provided in Figure 2. For neural network used for training, we denote the ECG encoder as $f_\theta :\to \mathbb{R}^d$ with linear classifier for supervised and self-training and MLP projection layer for CLIP based training (He et al., 2016). We also define text encoder as $h_\phi :\to \mathbb{R}^d$ for CLIP based training. θ and ϕ are learnable parameters. We now describe each training paradigm in detail.

Supervised training. The baseline model is trained end-to-end on expert over-read labels $\mathbf{y}_i$ using a binary cross-entropy (BCE) loss with class-balanced weight:

$$\begin{aligned}
\mathcal{L}_{sup} &= -\frac{1}{|\mathcal{D}_L| \times |C|} \sum_{i=1}^{\mathcal{D}_L} \sum_{c=1}^{C} [w^c \cdot \mathbf{y}_i^c \log \sigma(f_\theta(\mathbf{x}_i)) + (1 - \mathbf{y}_i^c) \log(1 - \sigma(f_\theta(\mathbf{x}_i)))], \\
w^c &= \min\left(\frac{\sum_{i=1}^{N} (1 - \mathbf{y}_i^c)}{\sum_{i=1}^{N} \mathbf{y}_i^c}, w_{\max} \right)
\end{aligned} \tag{1}$$

where σ is the sigmoid function, w^c is the positive class weight for class c, and $w_{\max}$ is a maximum cap to prevent extreme weights for highly imbalanced classes.

Self-training. To incorporate public ECGs without expert over-read labels, we employ a FixMatch-style framework (Sohn et al., 2020). The same architecture used in the supervised setting is trained jointly on labeled and unlabeled data. For labeled ECGs, we apply a weak augmentation and compute the supervised loss in the same manner as our supervised baseline. For unlabeled ECGs, the model first processes a weakly augmented view to obtain pseudo-labels $\mathbf{p}_i = \sigma(f_\theta(\tilde{\mathbf{x}}_i))$. We apply a single confidence threshold α shared across all diagnostic labels, masking out unlabeled samples whose maximum predicted probability does not exceed this threshold. The retained pseudo-labels are then used as regression targets for the corresponding strongly augmented views of the same unlabeled ECGs, following the FixMatch formulation based on masked binary cross-entropy. The overall self-training loss is defined as:

$$\begin{aligned}
&\mathcal{L}_{self} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{unsup}, \\
&\mathcal{L}_{unsup} = -\frac{1}{|\mathcal{D}_U| \times |C|} \sum_{i=1}^{\mathcal{D}_U} \sum_{c=1}^{C} \mathbb{1}(\sigma(f_\theta(\tilde{\mathbf{x}}_i)) \geq \alpha) \left[p_i \log \sigma(f_\theta(\tilde{\mathbf{x}}_i^{\text{str}})) + (1 - p_i) \log(1 - \sigma(f_\theta(\tilde{\mathbf{x}}_i^{\text{str}}))) \right],
\end{aligned} \tag{2}$$

where λ balances the supervised and unsupervised terms and $\tilde{\mathbf{x}}_i^{\text{str}}$ is strong augmented $\tilde{\mathbf{x}}_i$. In our implementation, pseudo-labels are taken from the supervised model's predictions, and the loss is calculated only on unlabeled samples that pass the confidence threshold.

CLIP-based training. To explore contrastive learning based approach on paired ECG and text data, we train a CLIP-style model that aligns ECG signals and expert over-read text (Radford et al., 2021). We split over-read reports into individual diagnosis statements, treating each statement as a positive pair for the corresponding ECG for fine-grained supervision. The CLIP loss is defined as:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2|\mathcal{D}_L|} \left[\sum_{i=1}^{\mathcal{D}_L} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{\mathcal{D}_L} \exp(s_{ij})} + \sum_{j=1}^{\mathcal{D}_L} \log \frac{\exp(s_{jj})}{\sum_{i=1}^{\mathcal{D}_L} \exp(s_{ji})} \right] \tag{3}$$

where $s_{ij} = f_\theta(\mathbf{x}_i)^\top h_\phi(\mathbf{t}_j)/\tau$ is the cosine similarity between ECG $\mathbf{x}_i$ and text embeddings $\mathbf{t}_j$ scaled by a learnable temperature τ.,

NegCLIP-based training. To fully exploit the paired nature of our dataset, we extend CLIP with diagnosis-level hard negatives derived from machine-read vs. over-read discrepancies (Yuksekgonul et al., 2023). We treat all over-read statements as ground-truth positives and any statement that appears in the machine-read report but not in the over-read as an explicit hard negative. The NegCLIP loss is defined as:

$$\mathcal{L}_{\text{neg-CLIP}} = -\frac{1}{2|\mathcal{D}_L|}\left[\sum_{i=1}^{\mathcal{D}_L}\log\frac{\exp\left(s_{ii}\right)}{\sum_{j=1}^{\mathcal{D}_L}\exp\left(s_{ij}\right)+\exp\left(\hat{s}_{ij}\right)}+\sum_{j=1}^{\mathcal{D}_L}\log\frac{\exp\left(s_{jj}\right)}{\sum_{j=1}^{\mathcal{D}_L}\exp\left(s_{ji}\right)+\exp\left(\hat{s}_{ji}\right)}\right] \tag{4}$$

where $\hat{s}_{ij} = f_\theta(\mathbf{x}_i)^\top h_\phi(\hat{\mathbf{t}}_j)/\tau$ is the similarity between ECG $\mathbf{x}_i$ and machine-read text $\hat{\mathbf{t}}_j$. This design pushes ECG embeddings away from machine-only textual hypotheses while pulling them toward expert-confirmed diagnoses, thereby leveraging the unique expert/machine pairing in our corpus.

4. Experiments

4.1. Dataset curation and label extraction

Data and Preprocessing We used a combination of private and public electrocardiogram (ECG) datasets to investigate the role of expert over-read data in automated ECG interpretation. Our primary corpus is a private dataset collected at Weill Cornell Medicine / NewYork–Presbyterian Hospital, containing approximately two million standard 12-lead ECGs, each with both the machine interpretation and a expert over-read provided by experts with more than 30 years of experience. All recordings were acquired at 250 Hz for 10 s and serve as the foundation for supervised and contrastive training. An overview of the full dataset flow is provided in Figure A.1.

To complement these data with additional diversity and unlabeled samples, we incorporated two public datasets: (i) the **PhysioNet 2021 Challenge** dataset (Moody et al., 2021), using six constituent sources (CPSC, CPSC-Extra, PTB-XL, Georgia, Ningbo, Chapman) and excluding PTB and St. Petersburg INCART due to short duration or inconsistent sampling rates; and (ii) the **MIMIC-IV-ECG v1.0** database (Gow et al., 2023), consisting of 10-s recordings sampled at 500 Hz. These public datasets were used primarily for self-training method.

All signals passed through a standardized preprocessing pipeline to ensure temporal and spectral consistency. ECGs with invalid values (e.g., `NaN`, constant-zero leads) were removed. Signals below 250 Hz were discarded, and those above 250 Hz were downsampled via linear interpolation. Longer recordings were split into non-overlapping 10-s windows. After temporal alignment, we applied: (1) baseline-wander removal with a 0.5 Hz high-pass filter, (2) per-lead z-score normalization, and (3) a 50–60 Hz notch filter to suppress power-line interference. All final waveforms were represented as 12-lead, 10-s signals at 250 Hz for downstream analyses.

Label extraction. Each ECG record in our private dataset contains two independent textual interpretations: the initial machine-generated report and the expert over-read diagnosis. To enable quantitative evaluation, we converted these free-text descriptions into a structured multi-label representation covering the most clinically relevant rhythm and morphological abnormalities. A detailed summary of label frequencies for both machine-read and over-read sources is provided in Table A.1.

In consultation with cardiologists, we defined a vocabulary of 13 diagnostic labels spanning rhythm abnormalities, conduction blocks, axis deviations, and ischemic findings. These categories reflect conditions routinely reported in clinical cardiology practice and capture the most prevalent and clinically actionable ECG patterns in our dataset.

To ensure consistent label assignment, we curated a synonym map that consolidated equivalent medical expressions, abbreviations, and spelling variants into a unified terminology. For example, phrases such as “normal sinus rhythm” and “NSR” were grouped under *Sinus Rhythm*, while terms like “left anterior fascicular block” and “left axis deviation” were standardized to *Left Axis Deviation*. This mapping was iteratively refined with cardiologist input to guarantee alignment between machine-read and expert over-read interpretations.

The final text-processing pipeline was applied identically to both machine-generated and expert reports, producing a harmonized label vocabulary across sources. Each ECG was ultimately represented as a binary vector of 13 diagnostic labels, enabling direct, systematic comparison of model performance on machine-read versus expert-validated interpretations.

4.2. Implementation Details and Evaluation

All experiments use a unified 1D ResNet–50 ECG encoder (He et al., 2016) and the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and cosine annealing over 100 epochs. Supervised and self-training models use a batch size of 512, while contrastive models use a batch size of 256 paired ECG–text samples. Early stopping is applied based on validation macro-AUPRC with a patience of 10 epochs. The dataset is divided into 80% development and 20% testing, with the development portion further split into 80% training and 20% validation. A fixed random seed ensures identical splits across all training paradigms.

Supervised. Supervised training uses BCEWithLogitsLoss with class-balanced positive weights (ratio of negatives to positives, capped at 50) and no data augmentation. Model selection is based on validation macro-AUPRC.

Self-training. For self-training, labeled samples use the same preprocessing as supervised training. For unlabeled ECGs, weak augmentations include Gaussian noise ($\sigma = 0.015$), amplitude scaling, and small temporal shifts. Strong augmentations include time warping, temporal masking, random resampling, lead dropout, and Gaussian noise with $\sigma = 0.02$ (Raghu et al., 2022; Sohn et al., 2020). A single confidence threshold ($\alpha = 0.95$) is used to filter pseudo-labels, ensuring that only highly reliable predictions contribute to the unsupervised loss. The relative weighting between labeled and unlabeled batches follows the standard FixMatch configuration.

CLIP and NegCLIP. For contrastive learning, the ECG encoder is paired with the original CLIP text transformer (Radford et al., 2021). ECG and text representations are pro-

Table 1: **Overall comparison across all paradigms.** Per-label performance (AUROC and AUPRC), reported in percentage (for supervised, self-training, CLIP, and NegCLIP models evaluated against expert over-read labels. Best performance per metric per row is shown in **bold**, and second best is underlined.

Diagnosis	**Supervised**		**Self-Training**		**CLIP**		**NegCLIP**	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Sinus Rhythm	99.7	99.9	**99.7**	**99.9**	89.5	98.6	98.9	99.8
RBBB	99.9	98.8	**99.9**	**99.0**	99.7	95.2	99.9	98.1
LBBB	99.9	98.0	**99.9**	**98.3**	99.5	79.4	99.9	97.5
Atrial Fibrillation	99.8	96.6	**99.8**	**97.3**	87.6	56.7	99.8	96.3
Atrial Flutter	98.7	80.0	**98.9**	**84.8**	62.5	8.3	98.9	75.9
Left Axis Deviation	99.3	94.0	**99.4**	**94.8**	97.6	73.3	98.8	88.6
Right Axis Deviation	99.7	87.6	**99.7**	**89.5**	99.4	80.9	99.6	86.3
Right Superior AD	99.8	74.4	**99.8**	**77.8**	99.2	46.3	99.8	70.6
LVH	94.0	77.4	**94.6**	**79.5**	90.5	66.7	93.4	73.3
RVH	99.3	77.9	**99.3**	**81.6**	98.5	66.1	99.1	77.8
Anterior MI	97.5	87.1	**97.7**	**88.2**	96.0	81.8	96.9	81.6
Inferior MI	98.7	90.3	**98.8**	**91.0**	98.4	88.3	98.4	87.8
Posterior MI	98.5	60.7	**98.7**	**65.0**	97.8	51.8	98.5	60.2

jected into a shared 512-dimensional space and ℓ_2-normalized. Over-read reports are decomposed into diagnosis-level statements, enabling multi-positive alignment between each ECG and the set of diagnoses appearing in the corresponding expert report. NegCLIP extends this setup by incorporating machine-read–only statements as explicit hard negatives (Yuksekgonul et al., 2023). Both contrastive variants use identical optimization settings for fair comparison.

5. Results

5.1. Evaluation Results

We evaluate all models using per-label AUROC and AUPRC, capturing both overall discrimination and precision under substantial class imbalance. For supervised and self-training models, these metrics are computed from sigmoid prediction scores, whereas for CLIP and NegCLIP, the ECG–text cosine similarity values (ranging from -1 to 1) serve directly as prediction scores for each diagnostic label. Table 1 provides an aggregated performance comparison across the four learning paradigms—supervised learning, self-training, CLIP, and NegCLIP—and serves as the primary reference for the cross-paradigm performance trends discussed below. For completeness, macro-averaged ROC and precision–recall curves for all paradigms are included in Figure A.2.

Supervised Learning Baseline The supervised results in Table 1 show that models trained on expert over-read labels and evaluated against the same ground truth achieve the strongest overall performance. These models form the baseline against which all other learning paradigms are compared. To contextualize the advantages of over-read supervision, we provide an ablation study in Section 5.2 that compares the same model architecture trained instead on machine-read labels. This experiment mimics common practice in prior ECG work and demonstrates that models trained on machine-read labels exhibit inflated performance when evaluated against machine-read test labels, but suffer notable drops when evaluated against expert over-read ground truth.

Self-Training with Unlabeled Public Data Self-training results in Table 1 demonstrate that incorporating unlabeled ECGs from public datasets through a FixMatch-style pseudo-labeling framework improves performance across nearly all diagnoses. By expanding the training signal beyond the expert-labeled private dataset, the model benefits from better representation of diverse ECG patterns and improves generalization, particularly for rare or underrepresented classes. The gains are most prominent in AUPRC, highlighting the benefit of leveraging large-scale unlabeled corpora under extreme class imbalance.

Contrastive Learning with CLIP and NegCLIP Table 1 also reports the contrastive learning results. Standard CLIP, which aligns ECG signals purely with expert over-read text, achieves competitive discrimination but struggles with precision for rare diagnoses. NegCLIP, which additionally leverages discrepancies between machine-read and over-read text as hard negatives, consistently improves both AUROC and AUPRC. Incorporating these disagreement-driven negatives sharpens the model's ability to differentiate subtle diagnostic cues, bringing performance closer to the supervised baseline while retaining the flexibility of a multimodal formulation. An additional experiment evaluating the impact of text granularity—independent diagnosis statements versus whole report encoding—is presented in Section 5.2.

5.2. Ablation Studies

Supervised: Machine-read vs. Over-read Labels. To further investigate how label source affects supervised training, Figure A.3 and Table A.2 evaluate the same model architecture trained on machine-read versus expert over-read labels. Models trained on machine-read labels appear highly accurate when evaluated against machine-read test labels, but their precision drops when evaluated against expert ground truth, revealing the noise and bias in automated interpretations. Because public machine-read datasets differ in sampling rate and signal format from our 250 Hz private ECGs, we replicated this setting by training on our own machine-read labels. Overall, this ablation shows that machine-read labels can give a misleading impression of performance, whereas expert over-reads provide far more reliable supervision.

Independent Diagnosis Text vs. Whole-text Formulation. In addition to the main NegCLIP formulation, we conducted an ablation to evaluate whether diagnosis-level text granularity is necessary for effective multimodal contrastive learning. Instead of splitting expert and machine-read reports into individual diagnosis statements, a whole-text variant encodes each report as one full sequence using the same BERT encoder (Devlin et al., 2019). As shown in Figure 3 and Figure A.4, this whole-text approach performs substantially worse across nearly all diagnoses. Because most diagnoses in an ECG report are unrelated to each other and machine–expert discrepancies typically occur in only one or two labels, collapsing the entire report into a single embedding dilutes these fine-grained disagreement signals. In contrast, our diagnosis-independent formulation preserves diagnosis-specific variation and yields markedly stronger supervision.

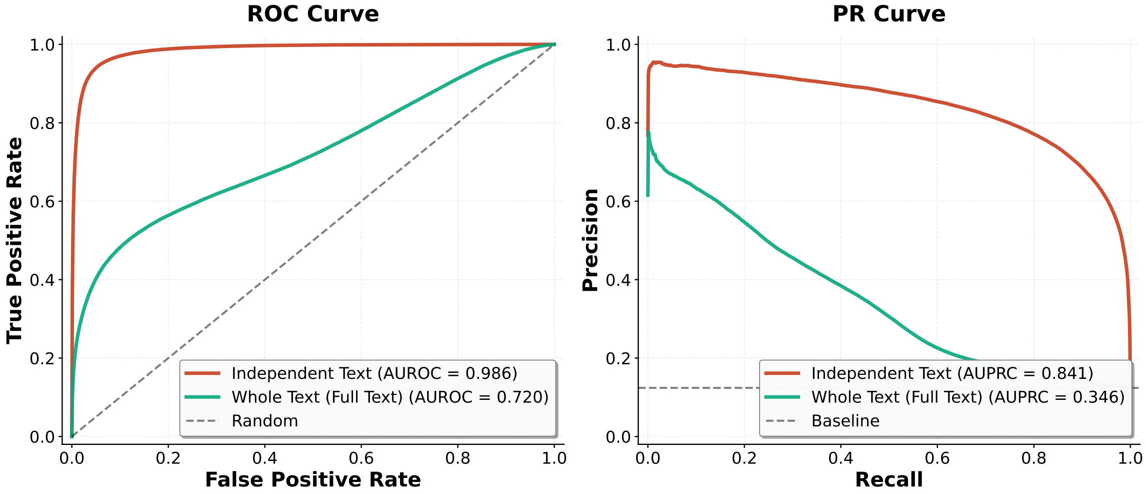

Figure 3: **NegCLIP: Independent diagnosis text vs. whole-text formulation.** Left panel shows the macro-averaged ROC curve, and the right panel shows the macro-averaged precision–recall curve comparing independent diagnosis text versus whole-text report embeddings.

6. Discussion

Our study underscores the critical importance of expert over-read ECG data for developing clinically reliable AI systems. We demonstrate that models trained on cardiologist-verified over-reads achieve consistently higher fidelity than those trained on machine-read interpretations, particularly for rare but clinically significant diagnoses where precision (AUPRC) is most critical. By leveraging a uniquely large paired dataset of over-read and machine-read ECGs, we systematically evaluate four methods—supervised, self-training, CLIP, and NegCLIP—and show that incorporating over-read supervision not only boosts diagnostic accuracy but also enables scalable frameworks. Self-training extends expert knowledge to unlabeled data, while NegCLIP introduces a multimodal pathway that exploits discrepancies between machine and expert interpretations to learn more robust representations. The key benefit of this work lies in demonstrating that expert-level supervision fundamentally shifts model behavior toward clinical reliability, while also offering practical strategies for data-limited environments.

Limitations include the use of data from a single health system, possible noise in text-to-label mappings, and the inability to externally validate on other public over-read datasets—since no comparable large-scale machine-read and over-read ECG pairs data currently publicly exist. In conclusion, our findings show that expert over-reads are not just helpful annotations but a critical foundation for building trustworthy, generalizable ECG AI systems. Moreover, self-training and NegCLIP offer practical pathways to extend their benefits far beyond the limited pool of labeled data.

Acknowledgments

This work was supported by funding from NewYork-Presbyterian for the NYP-Cornell Cardiovascular AI Collaboration.

References

William J. Brady, Amélie D. Perron, and Theodore C. Chan. Electrocardiographic interpretation: Clinical accuracy and challenges. *The American Journal of Emergency Medicine*, 38(5):995–1001, 2020. doi: 10.1016/j.ajem.2020.01.026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

Brian Gow, Alistair E. W. Johnson, Tom J. Pollard, Steven Horng, and Roger G. Mark. Mimic-iv-ecg database (version 1.0). PhysioNet, 2023. URL https://physionet.org/content/mimic-iv-ecg/1.0/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Seungwoo Hong, Yu Zhou, Xiyue Wang, Jing Shang, Chengliang Xiao, and Jimeng Sun. Ecg classification using deep neural networks: Current progress and future challenges. *IEEE Reviews in Biomedical Engineering*, 14:451–465, 2020. doi: 10.1109/RBME.2020.2981854.

Jeong Hwan Kim, Trinidad Cisneros, Amanda Nguyen, Jeroen van Meijgaard, and Haider J. Warraich. Geographic disparities in access to cardiologists in the united states. *Journal of the American College of Cardiology*, 84(3):315–316, 2024. doi: 10.1016/j.jacc.2024.04.054.

Krzysztof Kraik, Irena A. Dykiert, Joanna Niewiadomska, Marta Ziemer-Szymańska, Piotr Kukiełka, Adrian Martuszewski, and Małgorzata Poręba. The most common errors in automatic ecg interpretation. *Frontiers in Physiology*, 16:1590170, 2025. doi: 10.3389/fphys.2025.1590170.

Jun Li, Aaron Aguirre, Junior Moura, Che Liu, Lanhai Zhong, Chenxi Sun, Gari D. Clifford, Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains (ecgfounder). *NEJM AI*, 2025. doi: 10.1056/AIoa2401033.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122.

Robert L. McNamara, Nancy M. Albert, Anne B. Curtis, and et al. Clinical performance and quality measures for adults with atrial fibrillation or atrial flutter: 2021 update from the american heart association. *Circulation: Cardiovascular Quality and Outcomes*, 14 (12):e000102, 2021. doi: 10.1161/HCQ.0000000000000102.

George B. Moody, Ary L. Goldberger, Gari D. Clifford, Ikaro Silva, and et al. The physionet/computing in cardiology challenge 2021: Will two do? In *Computing in Cardiology (CinC)*, pages 1–4, 2021. doi: 10.22489/CinC.2021.438.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

Anirudh Raghu, Ming Zhang, Adam Rosenberg, Simon Kornblith, Ting Chen, Adrian Weller, and Harini Suresh. Data augmentation for electrocardiograms. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 351–371. PMLR, 2022.

Joerg Schläpfer and Hein J. J. Wellens. Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology*, 70(9):1183–1192, 2017. doi: 10.1016/j.jacc.2017.07.723.

Harold Smulyan. The computerized ecg: Friend and foe. *The American Journal of Medicine*, 132(2):153–160, 2019. doi: 10.1016/j.amjmed.2018.08.025.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it. In *International Conference on Learning Representations (ICLR)*, 2023.

Appendix A. Data Flow

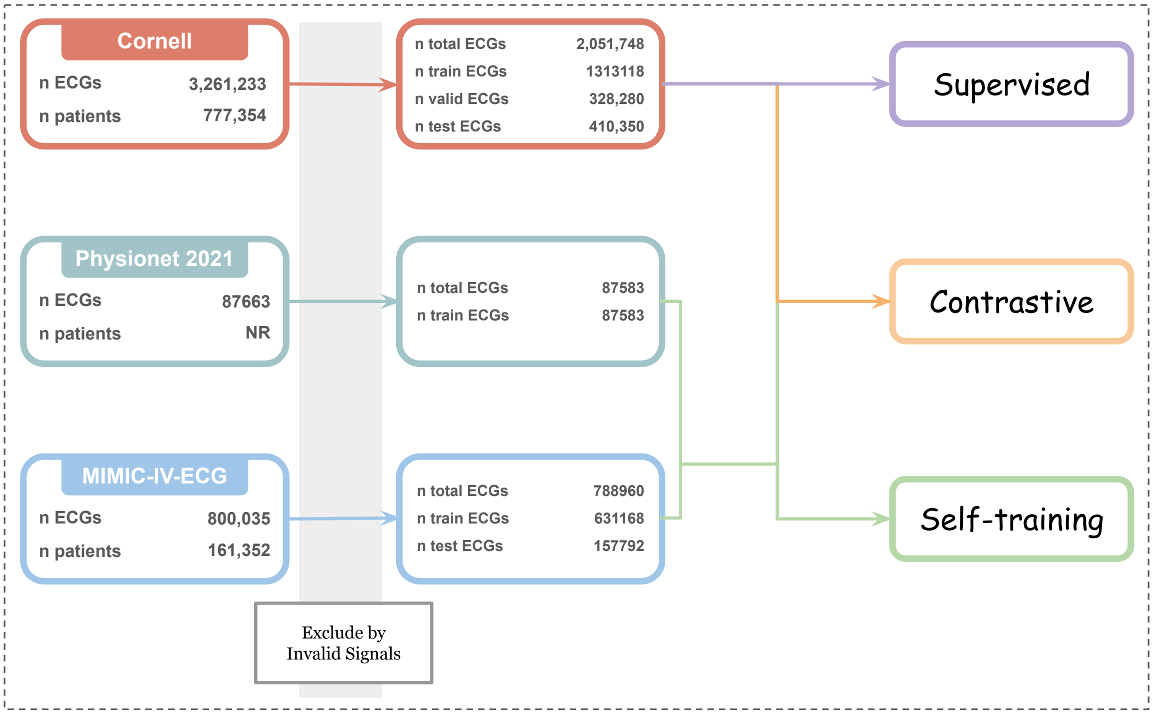

Figure A.1: **Data Flow of Private and Public Datasets.** Overview of dataset composition and flow used in our experiments, including the Cornell private ECG dataset, MIMIC-IV-ECG, and the PhysioNet 2021 Challenge dataset.

Appendix B. Cornell Private Data

Table A.1: **Cornell Private Dataset Summary.** Comparison of machine-read and expert over-read label counts. The Disagreement column reports the total number of studies where the two sources differ. *+Added%* column indicates the proportion of additional positive diagnoses introduced by experts—cases that were absent in the machine-read output but marked positive in the expert over-read. *-Removed%* column reflects positive labels removed by experts—cases that were labeled positive by the machine-read system but negated in the over-read. Together, these metrics show how expert review modifies automated ECG interpretations in both directions.

Diagnosis	**Machine-read**	**Over-read**	**Disagreement**	**+Added %**	**-Removed %**
Sinus Rhythm	1,826,249	1,848,631	43,848	+1.81%	-0.59%
RBBB	130,380	137,237	10,267	+6.57%	-1.31%
LBBB	52,687	59,449	11,678	+17.50%	-4.67%
Atrial Fibrillation	148,971	144,652	25,855	+7.23%	-10.13%
Atrial Flutter	25,566	43,071	26,189	+85.45%	-16.98%
Left Axis Deviation	198,282	205,834	12,554	+5.07%	-1.26%
Right Axis Deviation	37,367	50,708	21,383	+46.46%	-10.76%
Right Superior AD	5,090	9,019	5,491	+92.53%	-15.34%
LVH	270,853	315,212	64,415	+20.08%	-3.70%
RVH	13,157	42,192	31,579	+230.35%	-9.67%
Anterior MI	290,263	246,358	114,729	+12.20%	-27.33%
Inferior MI	179,658	179,836	54,542	+15.23%	-15.13%
Posterior MI	7,183	28,908	26,035	+332.45%	-30.00%

Appendix C. Overall Comparison AUROC and AUPRC Curves

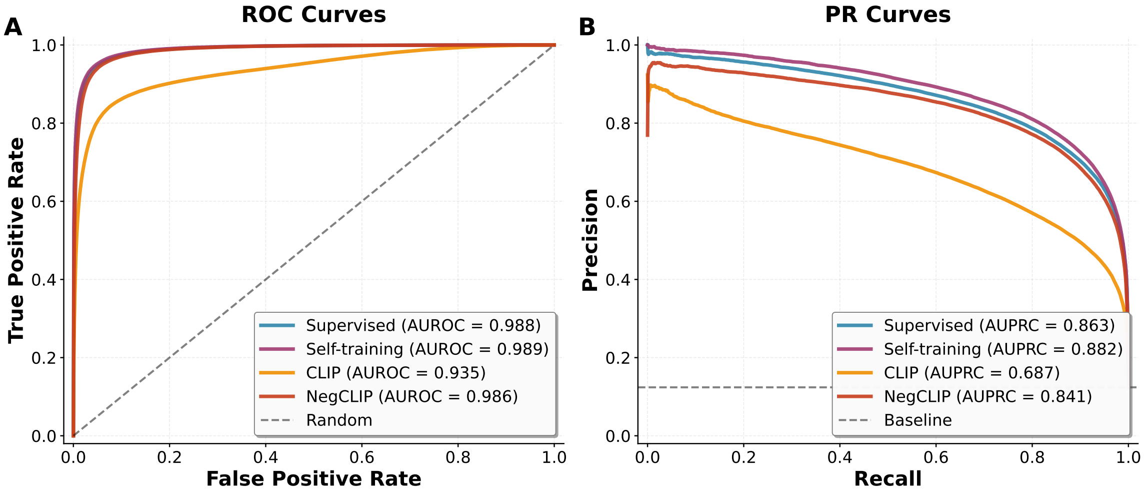

Figure A.2: **Overall comparison of AUROC and AUPRC curves across paradigms.** Panel (A) shows macro-averaged ROC curves (AUROC) for supervised, self-training, CLIP, and NegCLIP models. Panel (B) shows macro-averaged precision–recall curves (AUPRC) for the same paradigms.

Appendix D. Supervised Model Performance

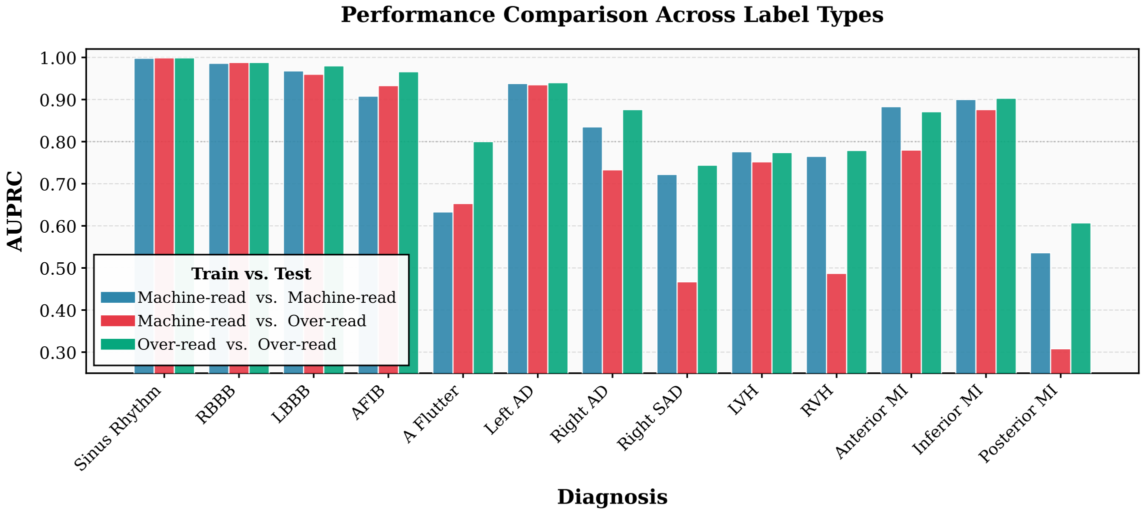

Figure A.3: **Comparison of AUPRC Supervised Models.** Bar plot comparing AUPRC across diagnostic categories for three supervised evaluation settings: (*i*) **Machine-read vs. Machine-read** — models trained on machine-read labels and evaluated against machine-read labels; (*ii*) **Machine-read vs. Over-read** — models trained on machine-read labels but evaluated against expert over-read labels, highlighting degradation when using noisier training labels; (*iii*) **Over-read vs. Over-read** — models trained and evaluated using expert over-read labels, representing the highest-quality supervision. Across nearly all diagnoses, training on expert over-reads yields superior performance, underscoring the impact of high-fidelity expert labels.

Table A.2: **Supervised learning.** Per-label evaluation comparing models trained on machine-read vs. expert over-read labels. “Machine” and “Over” indicate the evaluation label source. Bold numbers indicate the best value within each metric (AUROC or AUPRC) for each diagnosis.

Diagnosis	**Trained on Machine-read**				**Trained on Over-read**	
	AUROC–Machine	AUPRC–Machine	AUROC–Over	AUPRC–Over	AUROC	AUPRC
Sinus Rhythm	0.988	0.998	0.995	0.999	**0.997**	**0.999**
RBBB	0.999	0.986	0.998	0.988	**0.999**	**0.988**
LBBB	0.999	0.968	0.997	0.960	**0.999**	**0.980**
Atrial Fibrillation	0.993	0.908	0.996	0.933	**0.998**	**0.966**
Atrial Flutter	0.973	0.633	0.960	0.653	**0.987**	**0.800**
Left Axis Deviation	0.993	0.938	0.992	0.935	**0.993**	**0.940**
Right Axis Deviation	0.997	0.835	0.987	0.733	**0.997**	**0.876**
Right Superior AD	**0.999**	0.722	0.986	0.467	0.998	**0.744**
LVH	**0.948**	**0.776**	0.928	0.752	0.940	0.774
RVH	**0.994**	0.765	0.849	0.487	0.993	**0.779**
Anterior MI	0.975	**0.883**	0.962	0.780	**0.975**	0.871
Inferior MI	0.987	0.900	0.982	0.876	**0.987**	**0.903**
Posterior MI	**0.991**	0.536	0.933	0.308	0.985	**0.607**

Appendix E. Independent Diagnosis Text vs. Whole-text Comparison

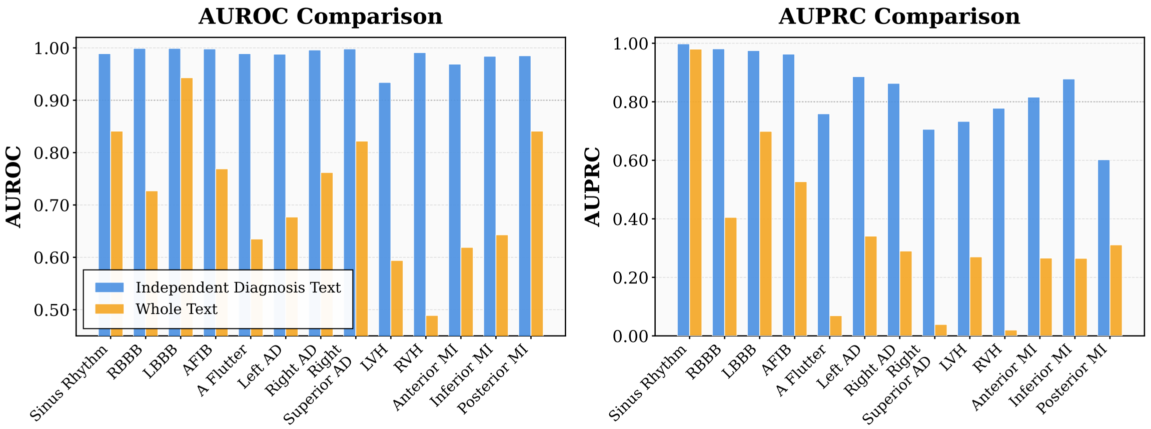

Figure A.4: **Impact of Diagnosis-level Text Granularity.** Comparison of per-label ARUOC AUPRC when using independent diagnosis statements versus whole-text report embeddings in multimodal contrastive learning. Independent diagnosis text provides substantially stronger supervision across most diagnostic categories.

Appendix F. Dataset Scaling Experiment

To quantify how unlabeled public ECGs contribute to self-training, we conduct a dataset scaling experiment in which increasing fractions of the public corpus are incorporated into the self-training pipeline. For each subset size, we evaluate AUPRC on the held-out test set. As shown in Figure A.5, model performance improves consistently as more unlabeled data is included, reflecting improved representation of low-prevalence ECG phenotypes. These results demonstrate that large-scale unlabeled corpora play a critical role in amplifying the benefits of expert over-read supervision when combined with pseudo-label–based training.

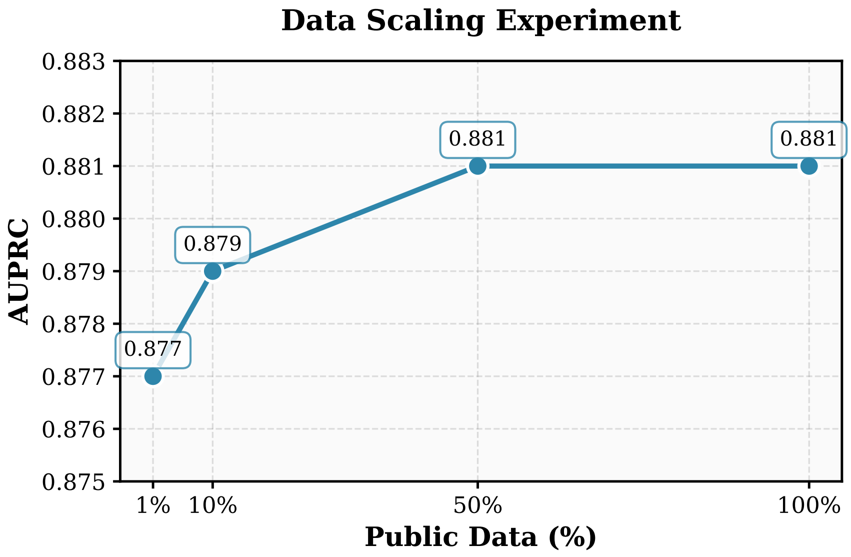

Figure A.5: **Public Dataset Scale on Self-Training Performance.** AUPRC improves steadily as larger portions of public ECG data are incorporated into the self-training pipeline.

Appendix G. Label Acronyms

Table A.3: **Explanation of Label Acronyms.** Full terminology corresponding to diagnostic abbreviations used throughout this paper.

Acronym / Term	**Full Description**
RBBB	Right Bundle Branch Block
LBBB	Left Bundle Branch Block
AD	Axis Deviation
LVH	Left Ventricular Hypertrophy
RVH	Right Ventricular Hypertrophy
MI	Myocardial Infarction
Anterior MI	Anterior Myocardial Infarction
Inferior MI	Inferior Myocardial Infarction
Posterior MI	Posterior Myocardial Infarction