# Cross-Modal Concept Transfer: From ECG Signals to Images for Explainable Disease Prediction

Chan Lee[1], Dasom Noh[1], Minjy Suh[1], Hyewon Lee[2,3], and Sunyoung Kwon[1,4]*

[1] Department of Information Convergence Engineering, Pusan National University, Busan, Republic of Korea
[2] Department of Cardiology and Medical Research Institute, Pusan National University Hospital, Busan, Republic of Korea
[3] College of Medicine, Pusan National University, Yangsan, Republic of Korea
[4] Center for Artificial Intelligence Research, Pusan National University, Busan, Republic of Korea
mldlcl2022@gmail.com, sy.kwon@pusan.ac.kr

**Abstract.** Electrocardiography (ECG) is widely used for cardiac diagnosis but is often stored as standardized ECG report images in clinical practice, with raw signals not consistently preserved. This discrepancy constrains the deployment of signal-dependent models and motivates image-based ECG analysis. However, image-based approaches rely on pixel-level representations, making it difficult to capture clinically meaningful quantitative measurements derived from signals. Moreover, existing explainability methods are largely limited to qualitative visualizations, lacking precise and quantitative interpretations. To address these challenges, we propose **Cross-Modal Concept Transfer (CMCT)**, which integrates image representations with signal-derived quantitative information. CMCT extends the Concept Bottleneck Model (CBM) to a cross-modal setting by using signal-derived measurements as intermediate concepts, guiding image representations toward clinically meaningful physiological features. The model is trained using both image and signal data, while predictions are made entirely in the concept space, enabling inference regardless of the input modality. Experimental results show that CMCT outperforms existing image-based methods and matches state-of-the-art signal-based models, showing that improved interpretability does not compromise performance. The predicted concepts are directly interpretable with clinical reference ranges, enabling quantitative explanations and facilitating concept-level interventions for validating the model's reasoning. These results highlight a practical path toward reliable and interpretable ECG image analysis. Our codes are available at: https://github.com/mldlcl2022/CMCT

**Keywords:** Electrocardiogram · Explainable AI · Multimodal AI.

* Corresponding author.

## 1 Introduction

Electrocardiography (ECG) is a fundamental tool widely used for cardiac screening and diagnosis, and the growing global burden of cardiovascular diseases has recently intensified research efforts in artificial intelligence for ECG analysis [1]. Recent studies have proposed Self-Supervised Learning (SSL) [2] approaches based on raw ECG signals, such as ST-MEM [3] and HeartLang [4], as well as Self-Supervised Multimodal Learning (SSML) [2] methods that combine ECG signals with diagnostic reports, including MERL [5] and MELP [6]. These advances predominantly focus on signal-centered modeling, reflecting the intrinsic nature of ECG as an electrical waveform. However, clinical practice does not always align with typical research settings. Although ECG is acquired as a signal, it is often stored as standardized ECG report images, with raw waveform data not consistently preserved [7,8,9]. Instead, only a limited set of signal-derived quantitative measurements, such as heart rate and R-R interval, are retained. Consequently, models that rely on direct access to raw signals may face practical constraints in real-world deployment.

Motivated by these practical limitations, recent studies have explored ECG analysis under image-based settings. Image-based approaches [7,10,12] can leverage pretrained computer vision models to obtain rich representations and directly benefit from well-established post-hoc explanation techniques, such as Grad-CAM [11]. However, images are inherently composed of pixel-level representations, which fundamentally limit their ability to accurately capture or directly extract quantitative signal-derived measurements, such as waveform durations and inter-beat intervals [9]. Moreover, existing explainability methods primarily highlight regions of model attention, which provide only qualitative insights and fail to offer precise and quantitative explanations of the model's decision-making process [7].

To address these challenges, we propose a method that simultaneously leverages the practical advantages of image-based representations and clinically meaningful quantitative measurements derived from ECG signals. To this end, we adopt the Concept Bottleneck Model (CBM) [13], which incorporates an intermediate layer of human-interpretable concepts, where predictions are made based solely on the predicted concept values. This design enforces the model to ground its decision-making in interpretable concepts rather than latent representations, thereby enhancing interpretability and verifiability. Originally defined in a single-modality setting, we extend CBM to a cross-modal framework by using physiologically meaningful information derived from ECG signals as concept labels. In this framework, the concept space acts not only as an interpretable bottleneck, but also as a signal-to-image transfer bridge that transfers signal-derived physiological supervision into the image representation. This supervision guides image representations to reflect clinically meaningful physiological concepts beyond pixel-level features. As a result, the proposed method overcomes the limitations of purely qualitative interpretation in existing image-based approaches and provides quantitatively grounded, clinically meaningful explanations.

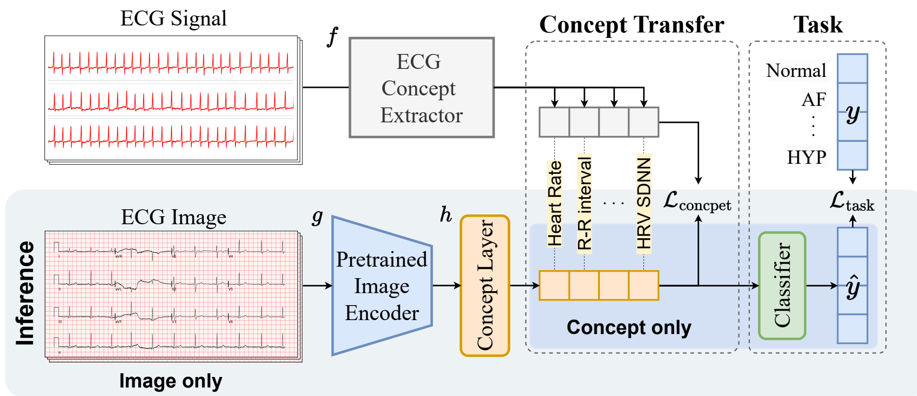


**Fig. 1.** Overview of the proposed framework.

Furthermore, as illustrated in Fig. 1, the final prediction is performed entirely within the concept space, making the decision-making process independent of any specific input modality. Since concepts can be derived from images or raw signals, the proposed framework supports diverse input modalities and provides consistent, concept-based explanations across real-world deployment scenarios. Finally, the main contributions of this paper are as follows.

1. We propose a **cross-modal learning framework** that transfers rich signal-derived information into image representations, enabling practical and deployable ECG analysis.
2. CMCT enables **quantitative interpretability** by incorporating signal-derived measurements into the concept bottleneck without sacrificing predictive performance.
3. **Modality-agnostic inference** is supported by concept-based prediction. The model can perform inference using any available modality, including images, signals, or concept representations, making it suitable for diverse real-world deployment scenarios.

## 2 Method

In this section, we present **Cross-Modal Concept Transfer (CMCT)**, a framework designed to ground image-based ECG modeling in physiologically meaningful signal-derived concepts. CMCT is inspired by Concept Bottleneck Models (CBM), which use human-interpretable concepts as an intermediate layer for prediction. We extend this paradigm to a cross-modal setting, where signal-derived quantitative measurements supervise the intermediate concept layer during training to guide image representation learning, as illustrated in Fig. 1.

### 2.1 Signal-to-Concept Extraction

A core component of concept-based learning is the construction of meaningful concepts. Prior CBM studies typically rely on manually defined concepts with human annotations, while recent approaches attempt to derive concepts automatically using LLMs [13,14,15]. However, manual annotation is costly and difficult to scale, and LLM-generated concepts may suffer from reliability issues. In contrast, CMCT derives concepts automatically from raw ECG signals using standard rule-based measurements, without extra annotations or LLMs.

Let $\mathbf{s}^i \in \mathbb{R}^{L \times T}$ denote the $i$-th raw ECG signal sample, where $L$ is the number of leads and $T$ is the number of time points. We extract concepts using NeuroKit2 [16] by detecting characteristic waveform peaks (e.g., P, QRS, and T complexes) and computing quantitative ECG variables based on clinically defined measurement rules. This yields a set of signal-derived concepts $\mathbf{c}^i = c_1^i, \ldots, c_j^i$, where $j$ indexes the concept dimensions.

These concepts are grouped into two categories: (1) peak- and interval-related features capturing waveform morphology (e.g., RR interval, QTc) [17], and (2) heart rate variability (HRV)-related features reflecting rhythm variability (e.g., SDNN, RMSSD) [18]. Since all concepts are continuous and vary in scale, we standardize them using training-set statistics. The resulting normalized concepts are used for cross-modal concept transfer (Section 2.2).

### 2.2 Cross-Modal Concept Transfer

In this section, we describe how image representations are mapped into the signal-derived concept space. Given paired samples $(\mathbf{s}^i, \mathbf{x}^i)$ of a raw ECG signal and its corresponding rendered image, we first extract a concept vector $\mathbf{c}^i$ from $\mathbf{s}^i$ as described in Section 2.1. The ECG image $\mathbf{x}^i$ is then processed by a pretrained image encoder to produce a latent representation. This representation is then passed through an intermediate concept layer to generate predicted concept values $\hat{\mathbf{c}}^i$. During training, the predicted concepts $\hat{\mathbf{c}}^i$ are optimized to approximate the signal-derived concepts $\mathbf{c}^i$ using the following SmoothL1 (Huber) loss:

$$\mathbf{c}^i = f(\mathbf{s}^i), \quad \hat{\mathbf{c}}^i = h(g(\mathbf{x}^i)), \quad \mathcal{L}_{\text{concept}} = \text{SmoothL1}(\hat{\mathbf{c}}^i, \mathbf{c}^i) \tag{1}$$

where $f$ denotes the concept extractor from ECG signals, $g$ is the image encoder, and $h$ is the concept prediction layer. Consequently, signal-derived physiological concepts guide the learning of image representations, allowing the model to perform inference using images alone without requiring raw signals. Importantly, the predicted concept vector $\hat{\mathbf{c}}^i$ serves as the sole input to the disease classifier, and final predictions are computed entirely within the concept space. This design ensures that decision-making is grounded in clinically meaningful quantitative variables while enabling consistent inference across different input modalities.

### 2.3 Objective Function with Dynamic Weighting

Directly combining the concept loss and the task loss with fixed weights can bias training toward a single objective, because the two terms may differ in scale and converge at different speeds. To address this issue, we adopt an Exponential Moving Average (EMA)-based dynamic loss-weighting strategy that balances the two objectives throughout training. The overall training objective is defined as a weighted sum of the concept loss and the task loss:

$$\mathcal{L}_{\text{total}}^{(t)} = \lambda_{\text{concept}}^{(t)} \mathcal{L}_{\text{concept}}^{(t)} + \lambda_{\text{task}}^{(t)} \mathcal{L}_{\text{task}}^{(t)} \tag{2}$$

At training step $t$, the dynamic weights are computed from EMA of each loss:

$$\text{EMA}_{\text{concept}}^{(t)} = \alpha \, \text{EMA}_{\text{concept}}^{(t-1)} + (1-\alpha) \, \mathcal{L}_{\text{concept}}^{(t)} \tag{3}$$

$$\text{EMA}_{\text{task}}^{(t)} = \alpha \, \text{EMA}_{\text{task}}^{(t-1)} + (1-\alpha) \, \mathcal{L}_{\text{task}}^{(t)} \tag{4}$$

where $\alpha \in [0, 1]$ is a smoothing factor. Larger $\alpha$ yields smoother, slower-updating EMAs, whereas smaller $\alpha$ reacts faster to recent loss changes but can make the weights noisier. Based on these EMAs, unnormalized weighting factors are defined as

$$u_{\text{concept}}^{(t)} = \left(\text{EMA}_{\text{concept}}^{(t)} + \epsilon\right)^p, \quad u_{\text{task}}^{(t)} = \left(\text{EMA}_{\text{task}}^{(t)} + \epsilon\right)^p \tag{5}$$

and the final weights are obtained by normalizing the unnormalized weighting factors:

$$\lambda_{\text{concept}}^{(t)} = \frac{u_{\text{concept}}^{(t)}}{u_{\text{concept}}^{(t)} + u_{\text{task}}^{(t)}} \in [\lambda_{\min}, \lambda_{\max}], \quad \lambda_{\text{task}}^{(t)} = 1 - \lambda_{\text{concept}}^{(t)} \tag{6}$$

Here, $\epsilon > 0$ prevents numerical issues; larger $\epsilon$ makes the weighting less sensitive to very small losses, while smaller $\epsilon$ increases sensitivity but may amplify numerical instability when a loss is near zero. $p$ controls the strength of reweighting; larger $p$ accentuates the contrast between losses and yields more aggressive weight shifts, whereas smaller $p$ makes the weights closer to uniform. The weights are constrained to lie within $[\lambda_{\min}, \lambda_{\max}]$.

## 3 Experiments and Results

### 3.1 Experimental Setup

**Datasets.** We evaluate our method on two public 12-lead ECG benchmarks, PTB-XL [19] and CPSC 2018 [20]. PTB-XL contains 21,837 recordings (10 seconds, 500 Hz) annotated into 5 diagnostic categories, while CPSC 2018 includes 6,877 recordings (10–20 seconds, 500 Hz) with 9 arrhythmia classes. Following clinical practice, we render each waveform into a standard ECG report image using ECG-Image-Kit [8], which serves as input to all image-based encoders.

**Baselines.** We compare CMCT with both image-based and signal-based baselines. For image-based baselines, we train standard 2D encoders on the rendered ECG images, including Inception [21], ResNet [22], EfficientNet [23], and

**Table 1.** Performance of image-based and signal-based methods.

**(a) Image-based methods**

| Downstream Modality | Method | CPSC 2018 | | | | PTB-XL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **F1** | **Pre** | **Re** | **Acc** | **F1** | **Pre** | **Re** | **Acc** |
| Image | Inception | 0.7362 | 0.7664 | 0.7160 | 0.9507 | 0.6728 | 0.6949 | 0.6581 | 0.8558 |
| | ResNet | 0.7555 | 0.8069 | 0.7241 | 0.9551 | 0.7015 | 0.7110 | 0.6953 | 0.8659 |
| | EfficientNet | 0.7029 | 0.7659 | 0.6683 | 0.9462 | 0.6794 | 0.7179 | 0.6537 | 0.8619 |
| | ViT | 0.6926 | 0.7155 | 0.6837 | 0.9390 | 0.6797 | 0.7018 | 0.6690 | 0.8552 |
| | Our | **0.7953** | **0.8308** | **0.7752** | **0.9624** | **0.7208** | **0.7534** | **0.6983** | **0.8776** |

**(b) Signal-based methods**

| Downstream Modality | Method | CPSC 2018 | | | | PTB-XL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **F1** | **Pre** | **Re** | **Acc** | **F1** | **Pre** | **Re** | **Acc** |
| Signal | ST-MEM | <u>0.7815</u> | <u>0.8082</u> | <u>0.7658</u> | <u>0.9605</u> | **0.7289** | 0.7434 | **0.7207** | **0.8798** |
| | MERL | 0.6567 | 0.6895 | 0.6388 | 0.9323 | 0.6887 | 0.7277 | 0.6626 | 0.8695 |
| | HeartLang | 0.6140 | 0.6382 | 0.6107 | 0.9230 | 0.7053 | **0.7785** | 0.6599 | 0.8747 |
| | MELP | 0.6948 | 0.7652 | 0.6697 | 0.9471 | 0.7183 | 0.7461 | <u>0.7025</u> | <u>0.8797</u> |
| Image | Our | **0.7953** | **0.8308** | **0.7752** | **0.9624** | <u>0.7208</u> | <u>0.7534</u> | 0.6983 | 0.8776 |

ViT [24]. Unless otherwise specified, CMCT uses ConvNeXt [25] as the default backbone. For signal-based baselines, we include recent state-of-the-art ECG models that operate directly on raw waveforms, namely ST-MEM [3], MERL [5], HeartLang [4], MELP [6], and fine-tune their publicly released pretrained checkpoints on our splits.

**Implementation Details.** Each dataset is split into train/valid/test sets with an 8:1:1 ratio. For image encoders, we follow the backbone-specific preprocessing pipeline, including the recommended input resizing and normalization. For signal-based models, we adhere to the original preprocessing and experimental configurations released by the authors. We optimize all trainable models with AdamW and a cosine annealing learning-rate schedule. For the EMA-based dynamic loss weighting, we set the smoothing factor $\alpha$ to 0.9, the reweighting exponent $p$ to 1.5, and $\epsilon$ to $10^{-8}$. All experiments are conducted on an NVIDIA A100 GPU with CUDA 12.2.

### 3.2 Performance Comparison

We first evaluate the proposed method under both the image-based and signal-based settings. As shown in Table 1(a), CMCT consistently outperforms existing image-based approaches by a considerable margin across all evaluation metrics. We attribute this improvement to aligning signal-derived clinical concepts with image representations, enabling image embeddings to capture physiologically meaningful information beyond shallow visual features. As a result, our method achieves performance comparable to state-of-the-art signal-based models that directly operate on ECG signals (Table 1(b)). This result is particularly meaningful, as our approach focuses on interpretability through a concept-based

**Table 2.** Predicted concept values relative to normal for AF samples.

| **Norm Avg** | **P wave duration** **101.0** $\pm$ **18.3** ms | **Heart Rate** **75.1** $\pm$ **10.2** bpm | **HRV_SDSD** **32.2** $\pm$ **30.3** ms | **HRV_SDNN** **28.7** $\pm$ **23.4** ms |
|---|---|---|---|---|
| AF sample 1 | 56.51 | 143.10 | 89.27 | 97.83 |
| AF sample 2 | 80.64 | 65.72 | 264.26 | 207.55 |
| AF sample 3 | 46.49 | 123.66 | 142.33 | 118.99 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **AF Avg** | **68.2** $\pm$ **16.6** ms | **105.5** $\pm$ **28.6** bpm | **200.2** $\pm$ **90.0** ms | **153.3** $\pm$ **54.6** ms |

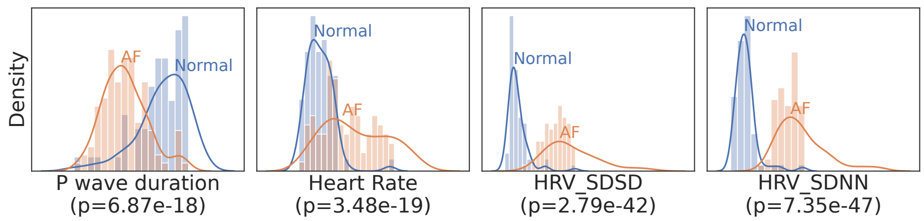


**Fig. 2.** Distribution of predicted concept values.

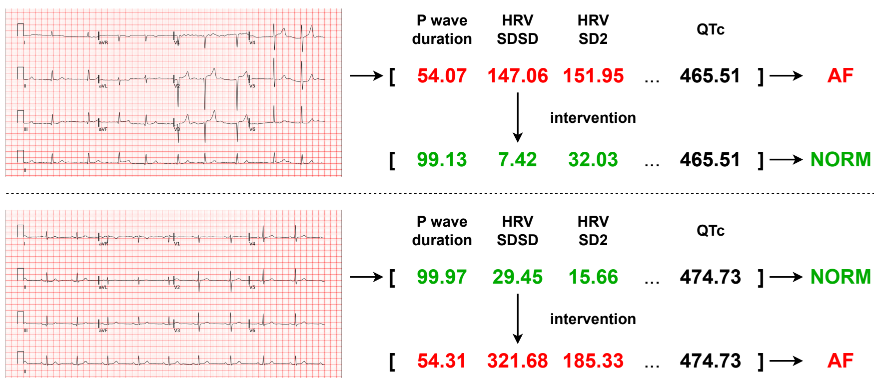


**Fig. 3.** Concept-level intervention on AF-related concepts.

framework without sacrificing predictive performance. These results demonstrate that our method effectively integrates the two modalities, achieving performance comparable to both image-based and signal-based models.

### 3.3 Concept-Based Clinical Interpretation

Prior concept-based and explainability methods [7,9,12,26,27] primarily rely on relative concept importance or qualitative visualizations. In contrast, CMCT provides quantitative interpretation by outputting concept values that can be directly compared with clinical criteria. As shown in Table 2, atrial fibrillation (AF) samples exhibit shorter predicted P-wave duration than the normal range,

**Table 3.** Results with different backbone encoders.

| Dataset | Setting | Inception | ResNet | EfficientNet | ViT | ConvNeXt |
|---|---|---|---|---|---|---|
| CPSC 2018 | w/o CMCT | 95.07% | 95.51% | 94.62% | 93.90% | 96.03% |
| | w CMCT | 95.28% | 95.58% | 95.06% | 94.78% | 96.24% |
| | Gap | **+0.21%** | **+0.07%** | **+0.44%** | **+0.88%** | **+0.21%** |
| PTB-XL | w/o CMCT | 85.58% | 86.59% | 86.19% | 85.52% | 87.12% |
| | w CMCT | 86.37% | 87.01% | 86.34% | 85.36% | 87.76% |
| | Gap | **+0.79%** | **+0.42%** | **+0.15%** | **-0.16%** | **+0.64%** |

consistent with the diminished or absent P wave in AF [28]. They also show elevated heart rate and HRV-related measures, reflecting increased variability due to irregular rhythm [18]. At the population level, Fig. 2 shows clear differences in concept distributions between normal and AF groups, with statistically significant separation. These results indicate that CMCT captures clinically meaningful concept patterns that support the plausibility of its predictions.

### 3.4 Concept-Level Intervention Analysis

To further validate the effectiveness of the learned concepts, we perform concept intervention experiments on variables related to atrial fibrillation (AF). As shown in Fig. 3, manipulating a small set of AF-relevant concepts induces meaningful changes in model predictions. Specifically, in the top row, an AF sample is transformed into a normal prediction after intervention, while in the bottom row, a normal sample is shifted toward AF. These results demonstrate that the learned concepts are not only predictive but also causally influence the model's decision-making process, confirming that the model relies on clinically meaningful concept representations.

### 3.5 Robustness Across Backbone Encoders

We further evaluate robustness across different image backbone encoders. As shown in Table 3, adding the proposed concept-transfer module consistently improves performance over the corresponding baselines across most backbones on both CPSC 2018 and PTB-XL. While the magnitude of improvement varies depending on the encoder and dataset, the overall trend indicates that the benefits of CMCT are not tied to a specific backbone architecture. This makes it easy to pair CMCT with newly developed backbones in the image or ECG-image domain. As a result, one can obtain performance improvements together with concept-based interpretability without redesigning the framework.

## 4 Conclusion

In this paper, we proposed **Cross-Modal Concept Transfer (CMCT)**, a framework that improves explainability in ECG modeling by transferring quantitative signal information to ECG images through continuous concept bottlenecks. By leveraging clinically established signal-derived measurements as intermediate concepts, CMCT trains image representations with explicit physiological supervision. Experiments demonstrate that CMCT consistently outperforms existing image-based methods while achieving performance comparable to state-of-the-art signal-based models, indicating that improved interpretability does not compromise predictive accuracy. Furthermore, CMCT provides quantitative interpretability through concept predictions aligned with clinical reference ranges, facilitating concept-level interventions for validating the model's reasoning. It also maintains strong performance across diverse backbone encoders, highlighting its backbone-agnostic and extensible nature.

**Acknowledgments.** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00254177); in part by the Leading Generative AI Human Resources Development (IITP-2026-RS-2024-00360227) grant funded by the Korea government(MSIT).

**Disclosure of Interests.** A Korean patent application related to the method described in this paper is currently pending. The authors declare no other competing interests relevant to the content of this paper.

## References

1. Vaduganathan, M., Mensah, G.A., Turco, J.V., Fuster, V., Roth, G.A.: The global burden of cardiovascular diseases and risk: a compass for future health. Journal of the American College of Cardiology **80**(25), 2361–2371 (2022)
2. Zong, Y., Mac Aodha, O., Hospedales, T.M.: Self-supervised multimodal learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 5299–5318 (2024)
3. Na, Y., Park, M., Tae, Y., Joo, S.: Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In: International Conference on Learning Representations (ICLR) (2024)
4. Jin, J., Wang, H., Li, H., Li, J., Pan, J., Hong, S.: Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In: International Conference on Learning Representations (ICLR) (2025)
5. Liu, C., Wan, Z., Ouyang, C., Shah, A., Bai, W., Arcucci, R.: Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In: International Conference on Machine Learning (ICML) (2024)
6. Wang, F., Xu, J., Yu, L.: From Token to Rhythm: A Multi-Scale Approach for ECG-Language Pretraining. In: International Conference on Machine Learning (ICML) (2025)
7. Ao, R., He, G.: Image based deep learning in 12-lead ECG diagnosis. Frontiers in Artificial Intelligence **5**, 1087370 (2023)
8. Shivashankara, K.K., Deepanshi, Mehri Shervedani, A., Clifford, G.D., Reyna, M.A., Sameni, R.: ECG-Image-Kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. Physiological measurement **45**(5), 055019 (2024)

9. Stenhede, E., Bjørnstad, A.M., Ranjbar, A.: Digitizing paper ECGs at scale: an open-source algorithm for clinical research. npj Digital Medicine (2026)
10. Nam, J.H., Park, S.H., Kim, S.J., Lee, S.C.: Vizecgnet: visual ECG image network for cardiovascular diseases classification with multi-modal training and knowledge distillation. In: IEEE International Conference on Image Processing (ICIP), pp. 3219–3223. IEEE (2024)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 618–626. IEEE (2017)
12. Perez Alday, E.A., Gu, A., J Shah, A., Robichaux, C., Ian Wong, A.K., et al.: Classification of 12-lead ECGs: the physionet/computing in cardiology challenge 2020. Physiological Measurement **41**(12), 124003 (2020)
13. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., et al.: Concept bottleneck models. In: International Conference on Machine Learning (ICML), pp. 5338–5348. PMLR (2020)
14. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: International Conference on Learning Representations (ICLR) (2023)
15. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: International Conference on Learning Representations (ICLR) (2023)
16. Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., et al.: NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods **53**(4), 1689–1696 (2021)
17. Kligfield, P., Gettes, L.S., Bailey, J.J., Childers, R., Deal, B.J., et al.: Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. Circulation **115**(10), 1306–1324 (2007)
18. Shaffer, F., Ginsberg, J. P.: An overview of heart rate variability metrics and norms. Frontiers in public health **5**, 290215 (2017)
19. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., et al.: PTB-XL, a large publicly available electrocardiography dataset. Scientific Data **7**(1), 154 (2020)
20. Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., et al.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics **8**(7), 1368–1373 (2018)
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., et al.: Going deeper with convolutions. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE (2015)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. (2016)
23. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), pp. 6105–6114. PMLR (2019)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

25. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986. IEEE (2022)
26. Rao, S., Mahajan, S., Böhle, M., Schiele, B.: Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In: European Conference on Computer Vision (ECCV), pp. 444–461. (2024)
27. Pang, W., Ke, X., Tsutsui, S., Wen, B.: Integrating clinical knowledge into concept bottleneck models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 243–253. (2024)
28. January, C.T., Wann, L.S., Calkins, H., Chen, L.Y., Cigarroa, J.E., et al.: 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology **74**(1), 104–132 (2019)