Enhanced Cardio Care: Explainable Vision Transformer Multimodal Pipeline For Cardiac Abnormalities Detection Using Electrocardiogram Image Reports

Ngoc M. To[1,2], Vu Q. Vo[1,2], Quoc Cuong Ngo[1,2], Dinesh Kumar[1,2], Minh N. Dinh[1,2], Dang V. Nguyen[3], Dan V. B. Do[3]
[1]School of Engineering, RMIT University, Australia
[2]School of Science, Engineering & Technology, RMIT University, Vietnam
[3]Tam Duc Cardiology Hospital, Ho Chi Minh, Vietnam
E-mail: ngoc.tmn19@gmail.com

Keywords: ECG classifier, multimodal pipeline, vision transformer, ECG image, cardiac diagnostic

Received: July 16, 2025

Electrocardiogram (ECG) based Artificial Intelligence (AI) analysis has evolved. Its performance in diagnosing arrhythmias is now comparable to that of human experts, and it has the potential to assist societies with limited healthcare resources. However, these settings often have paper-based ECG image archives only, while the current AI-ECG analysis requires digitised ECG signals. To address this, we previously introduced Cardio Care, a mobile-friendly diagnostic pipeline capable of analysing both ECG signals and scanned ECG images. In this extended study, we enhance the pipeline's explainability and expand its model benchmarking by comparing the Vision Transformer (ViT) with two of its data-efficient variants: DeiT and BEiT. These models were evaluated on two image-based ECG datasets—one public dataset (Mendeley) and one private dataset (Tam Duc Cardiometabolic). Our results show that ViT achieves the strongest classification performance among all three variants, with macro F1-scores of up to 0.99 on Mendeley and 0.81 on Tam Duc. Additionally, we integrate a Grad-CAM-based explainability feature to visualise model attention, improving interpretability for clinical use. The enhanced Cardio Care pipeline now has an explainable function using Grad-Cam, demonstrating significant potential for scalable, low-cost cardiac screening in underserved healthcare settings.

Povzetek: Študija predstavlja razložljiv multimodalni okvir Cardio Care za analizo slik ECG z ViT/DeiT/BEiT. ViT dosega najboljše rezultate, Grad-CAM izboljša interpretabilnost, sistem je uporaben v okoljih z omejenimi viri.

1 Introduction

Cardiovascular disease (CVD) has remained the leading cause of global mortality for over 100 years [21] [16] and is responsible for approximately 20 million deaths annually [3]. While various medical devices can assist cardiologists in identifying cardiac abnormalities, the electrocardiogram (ECG) plays a central role, offering a non-invasive, convenient, and economical tool in modern medicine for evaluating the electrical activity associated with the cardiac activities [18].

In the past decade, advances in artificial intelligence (AI) have demonstrated the effectiveness of automated ECG interpretation. Deep learning networks, particularly convolutional neural networks (CNNs), have achieved expert-level accuracy and shown promising results in detecting arrhythmias and other heart-related abnormalities from digital ECG signal, reducing the reliance on trained healthcare professionals [8]. This has the potential of supporting the under-resourced healthcare systems with few specialist cardiologists. However, these models are not practical in low-income and rural real-world settings that only have paper based ECG and digital ECG devices are unavailable and clinicians rely on paper-based ECG printouts. This makes the AI-based ECG analysis unsuitable in such settings, where expert-level readers are scarce [12]. Thus, by excluding image-based ECGs from AI development pipelines results in excluding those who need this the most, and will lead to a sharp divide between people who will benefit from AI in health and those who will not. Hence, to promote equality in the benefits of AI in healthcare, the AI model should be developed to support both, digital ECG, and ECG images that can be used by front-end health care providers without latest ECG equipment.

To bridge this gap, we have developed and validated Cardio Care, a smartphone-friendly deep learning pipeline capable of analysing a standard 10-second resting ECG test, suitable for receiving both digitised ECG and imaging ECG from scanned or printed ECG reports [26]. Built on the Vision Transformer (ViT) architecture [7], Cardio Care employs self-attention mechanisms to effectively recognise patterns in ECG image data, providing a flexible and deployable solution, which is suitable for resource-limited settings. Our innovative pipeline has the capability to predict multiple cardiac abnormalities, both multi-label and single-label. Unlike other semi-supervised zero-shot mod-

els for general image classification [9, 10, 17, 27], our ViT models, trained on supervised datasets with cardiologist-level labels, are fine-tuned specifically for ECG reports. Cardio Care takes a different approach from traditional methods at the clinics, as can be seen in **Figure 1**, in which patients can easily photograph their ECG reports and upload them via a mobile app, and our AI model can provide highly accurate predictions to assist both patients and healthcare providers.

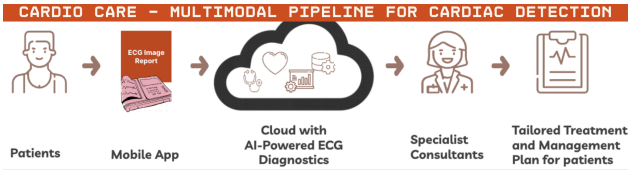

Figure 1: Simplified flowchart of Cardio Care application

In this extended study, we aim to improve both the architectural comparison and the explainability of the Cardio Care pipeline by introducing additional Vision Transformer variants. Specifically, aside from ViT, we evaluate two prominent extensions: the Data-efficient image Transformer (DeiT) [24], and the Microsoft Bidirectional Encoder representation from Image Transformers (BeiT) [2]. These models are designed for improved learning in environments with limited annotated data, making them well-suited for real-world clinical datasets. Since DeiT and BeiT are known for their performance on small-scale datasets, they will be trained on our two image-based datasets: the Mendeley (public) and Tam Duc (private) datasets. Furthermore, we enhance the transparency of Cardio Care by integrating a Grad-CAM-based explainability module, enabling visual interpretation of the model's attention on ECG waveform regions.

These extensions bring our proposed solution closer to real-world clinical deployment, particularly in under-resourced healthcare settings, by enhancing its performance, flexibility, and generalizability—all while operating on image-based ECG inputs without the need for digital signal acquisition or specialised infrastructure.

2 Methodology

This study builds upon our previously published conference paper [26], which introduced the Cardio Care, developed using the ViT architecture for ECG image and signal classification. In this extended version, we introduce two new Transformer variants (DeiT, BEiT), add an explainability module (Grad-CAM), and evaluate the performance across multiple datasets. We structured our methodology into three main components.

First, Section 2.1 - Datasets describes the three ECG datasets used for model development and evaluation. These include both signal- and image-based ECGs, covering a variety of dataset sizes and characteristics to represent real-world clinical variability. Second, Section 2.2 - Preprocessing outlines the preprocessing procedures applied to both signal and image ECG inputs. This involves preprocessing steps to transform ECG signals into usable waveform graphs, as well as cropping, augmentation, and normalisation of images to ensure consistency across modalities. Third, Section 2.3 - Training Pipeline presents the model architectures and training pipeline. We implement and compare three variants of Transformer-based algorithms: The Google's ViT [7], the Facebook's DeiT [24], and the Microsoft's BeiT [2] — for ECG classification. This section also details the training setup, evaluation metrics, explainable technique and cross-validation approach used to assess model performance across datasets.

For completeness, we retain the ViT model trained on the signal-derived ECG plots from our original study (using the CPSC dataset) in Section 3.2, as a baseline demonstrating Cardio Care's compatibility with signal inputs. However, no additional experiments were performed on this dataset in this extended work.

2.1 Datasets

To evaluate network performance across sample sizes and input types, we used three 12-lead ECG datasets, the characteristics of which are listed in Table 1.

Table 1: Distribution of abnormalities per datasets

Dataset	CPSC	Mendeley	Tam Duc
Input	signal	image	image
Sample	6877	929	170
Small-scale	No	Yes	Yes
Class	9	4	2
Balance	No	Yes	No
Access	Public	Public	Private

The China Physiological Signal Challenge (CPSC) [19] was released in 2018 and is publicly available at `http://2018.icbeb.org/Challenge.html`. This dataset comprises 6877 records in raw signal at 500 Hz with multi arrhythmias classes: normal sinus rhythm (SNR), atrial fibrillation (AF), first-degree atrioventricular block (IAVB), left bundle branch block (LBBB), right bundle branch block (RBBB), premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment depression (STD), and ST-segment elevation (STE). For comparison, the study utilised a 10-second ECG printout [20].

The 12-lead Mendeley "ECG Images dataset of Cardiac Patients" [11] is publicly available at `https://data.mendeley.com/datasets/gwbz3fsgp8/2`, consists of 929 ECG images in four classes: normal, myocardial infarction (MI), abnormal heartbeat, and previous history of myocardial infarction (MI his).

The third and new dataset is a private clinical dataset collected at Tam Duc Cardiology Hospital (Ho Chi Minh City, Vietnam), comprising 170 de-identified ECG images from

patients who visited between 2021 and 2023. The dataset is categorised into two classes: cardiometabolic (n = 71) and control (n = 99). All ECGs were standard 10-second, 12-lead printed reports scanned into high-resolution image format. The use of this dataset was approved by the hospital's ethics committee (Ref. No. 18.23/GCN-BVTD).

In this extended version, we clarify the class distribution in the Cardiometabolic dataset. The dataset contains 71 records labeled as disease and 99 as healthy, which corrects the reversed figures reported in our earlier conference paper [26]. That version mistakenly listed 71 as healthy and 99 as disease. All model training and evaluation use the corrected labels.

2.2 Preprocessing

To ensure consistent model input across various ECG data types, we designed a standardised preprocessing pipeline for both signal-based and image-based ECG inputs. The goal was to generate high-quality, normalised images from all modalities, suitable for Vision Transformer-based classification.

2.2.1 ECG signal preprocessing

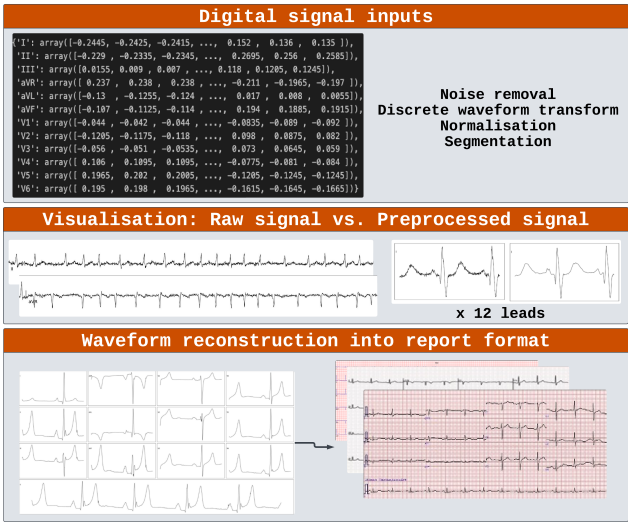

Figure 2: Preprocessing - Cardio Care framework for digital signal-based ECG inputs

Raw 12-lead ECG signals from the CPSC dataset were preprocessed in three stages [4, 6] before being converted into waveform images (**Figure 2**):

– Denoising: Signal noise was reduced using discrete wavelet transform with Daubechies-4 wavelet at level 4 decomposition [14, 25]. For noise thresholding, we applied the Median Absolute Deviation (MAD) method [15], a robust statistical estimator less affected by outliers, to identify and suppress high-frequency noise components while preserving clinically relevant waveform features.

– Normalisation: Signal was rescaled to a standardised amplitude range to reduce inter-record variability. This normalisation step improves signal consistency, enhances comparability across samples, and facilitates more reliable pattern recognition during model training.

– Segmentation: ECG records were segmented into a 10-second window, corresponding to 5000 samples at a 500 Hz sampling rate. Preprocessed signals were then converted into waveform plots, with 12 leads arranged in a 3x4 layout as a grayscale image to match model input requirements.

2.2.2 ECG image preprocessing

Image-based ECG reports, such as those from Mendeley and Cardiometabolic datasets, followed the standard format in clinical practice [13], underwent the following preprocessing steps: Non-relevant textual regions (e.g., patient information or hospital identifiers) were cropped, retaining only the 12-lead waveform area in a standardised layout. All ECG report images were then enhanced to 600 DPI and resized to 224 × 224 pixels to match the input resolution of the Transformer models. To simulate real-world variations such as misaligned scanning or handheld capture, we applied random rotations of ±10° as a form of data augmentation, inspired by prior work on image-based ECG interpretation [23]. This process enhances generalisability to real-world image acquisition settings. An illustration of this process is provided in the modelling overview **Figure 3**.

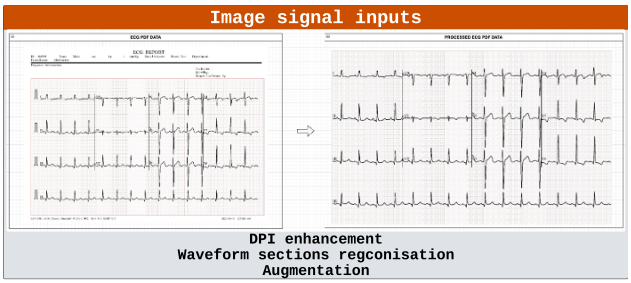

Figure 3: Preprocessing - Cardio Care framework for image-based ECG inputs

Finally, each 224 × 224 image was divided into non-overlapping 16 × 16 pixel patches, resulting in 196 patches per image. These patches were then flattened and embedded as input tokens to the Vision Transformer encoder. The patch size was chosen to capture local waveform features across multiple rows while maintaining spatial resolution consistent with standard ViT configurations.

2.3 Training pipeline

This study evaluates and compares three Vision Transformer architectures for ECG classification: ViT, DeiT and BEiT. All models were trained independently on each dataset using only preprocessed image-based ECG inputs. All experiments were conducted on Google Colab using NVIDIA A100 GPU (48GB VRAM). The pipeline is shown in **Figure 4**.

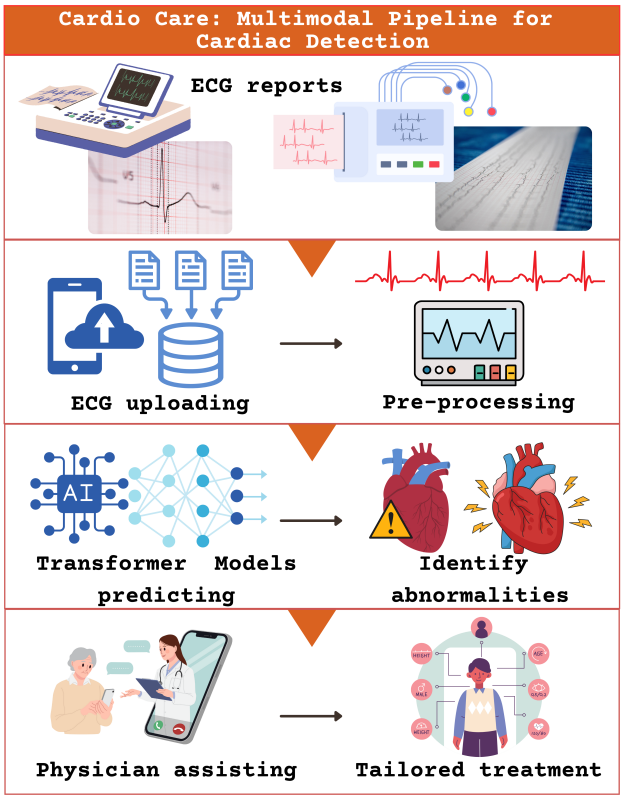

Figure 4: Multimodal pipeline of Cardio Care application

2.3.1 Model architectures

– The Google Vision Transformer (ViT) was introduced by Dosovitskiy et al. [7] is an advanced deep-learning architecture designed explicitly for visual recognition tasks. Unlike traditional deep-learning convolutional neural networks (CNN), ViT breaks down images into smaller patches and analyses the global relationships between them. This model's self-attention mechanism efficiently accesses the entire image and captures complex patterns, subtleties and anomalies in images. [Variant used: VIT-BASE-PATCH16-224-IN21K]

– The Facebook Data-efficient image Transformer (DeiT) was introduced by Touvron et al. [24] is particularly designed under constraints of limited data availability. Unlike the standard ViT, which requires large-scale datasets for optimal performance, DeiT incorporates knowledge distillation during training facilitated by a teacher-student paradigm. This strategy involves a distillation token that learns to mimic the output of a powerful, pre-trained teacher model (typically a CNN), effectively transferring the teacher's knowledge to the DeiT model. This enhances DeiT's ability to perform competitively with much smaller datasets than those required by traditional ViTs. [Variant used: DEIT-BASE-DISTILLED-PATCH16-224]

– The Microsoft Bidirectional Encoder representation from Image Transformers (BeiT) was introduced by Bao et al. [2] who proposed a masked image modelling (MIM) task to use two views for each image, image patches and visual tokens. Their study indicated that BeiT can improve the generalisation ability of fine-tuned models, particularly on small-scale datasets. [Variant used: BEIT-PATCH16-224]

2.3.2 Training and evaluation

Table 2: Training configuration per dataset

Datasets	CPSC	Mendeley	Tam Duc
Epoch	50	25	50
Batch	256	32	16
Learning rate	2e-4	2e-4	2e-5
Optimizer	AdamW	AdamW	AdamW
Train/Test	80/20	85/15	80/20
Multi-label	Y	N	N

To address class imbalance, we applied a stratified split and class-weighted cross-entropy loss, with weights inversely proportional to class frequencies in the training set. This approach ensures balanced accuracy, which is crucial in medical diagnostics, where missing rare cases can have serious consequences.

We employed 10-fold cross-validation on the full dataset to guarantee consistent performance estimates. For each fold, the performance metrics recorded include both Precision and Recall and F1-Score, which combines both metrics for a balanced assessment. Finally, the macro F1-score across n classes addresses class imbalance and reflects overall performance. Additionally, confusion matrices are visualised to aid interpretation.

3 Results

This section presents the performance of three Vision Transformer-based models (ViT, DeiT, and BeiT) trained and evaluated on three ECG datasets of different types and sizes. All models were trained solely using preprocessed image inputs. For consistency, CPSC signals were converted into 12-lead waveform plots.

3.1 Dataset summary

We utilised 12-lead ECGs from three datasets to demonstrate the performance of three model variants. In Section Method - Table 1 already summarises the characteristics of the datasets used in this study, including sample sizes, class labels, and modality. The categorical distribution of the abnormality can be seen below in **Table 3**. Due to the low prevalence of a few classes (236 cases or 3.43% LBBB; 220 cases or 3.20% STE), data were stratified based on clinical labels to ensure consistent distribution across both training and testing sets. A new cardiometabolic dataset (Tam Duc) comprising 170 samples has been introduced for evaluation utilising real-world clinical data.

Table 3: Distribution of abnormalities per dataset

	CPSC	Mendeley	Tam Duc
1	918 SNR	284 Normal	99 Control
2	1221 AF	240 MI	71 Disease
3	722 IAVB	172 MI his	
4	236 LBBB	233 Abnormal	
5	1857 RBBB		
6	616 PAC		
7	700 PVC		
8	869 STD		
9	220 STE		

3.2 Classification performances - ECG signal dataset

Although studies have been utilising the CPSC 2018 [28] [5], none of them have attempted shorter segments of the original signal or converted those to an imaging-based model. To evaluate our model, we compare it with machine learning classifiers as baselines for comparison. Our baseline classifiers used extracted statistical features as input for training, algorithms including Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), and Gradient Boosting Tree (GBT).

Table 4: Our ViT vs. baseline classifiers on CPSC dataset: Overall 10-fold CV performance

	LR	RF	MLP	GBT	Ours
Accuracy	0.40	0.36	0.45	0.50	**0.93**
Precision	0.58	**0.88**	0.54	0.84	0.71
Recall	0.41	0.34	0.48	0.49	**0.61**
F1-score	0.47	0.44	0.50	0.58	**0.65**

From **Table 5**, macro F1-scores are shown for all classes. With the exception of IAVB and STE, the harmonised F1-scores for the other classes ranged from 0.54 to 0.88. Our model also delivers the highest average performance across all classes, with a 7% improvement over the second model, GBT at 0.58.

To demonstrate the best fold's performance, confusion matrix is shown in **Figure 5**.

Table 5: Our ViT vs. baseline classifier on CPSC dataset: F1-score per class performance

	LR	RF	MLP	GBT	Ours
SNR	0.50	0.46	0.47	0.62	**0.60**
AF	0.58	0.58	0.62	0.74	**0.85**
IAVB	0.30	0.04	0.29	0.28	**0.36**
LBBB	0.77	0.84	0.70	**0.88**	0.79
RBBB	0.80	0.84	0.78	**0.85**	0.78
PAC	0.03	0.02	0.24	0.20	**0.54**
PVC	0.59	0.59	0.65	0.70	**0.74**
STD	0.41	0.38	0.50	**0.61**	0.59
STE	0.24	0.17	0.28	0.30	**0.48**
Average	0.47	0.44	0.50	0.58	**0.65**

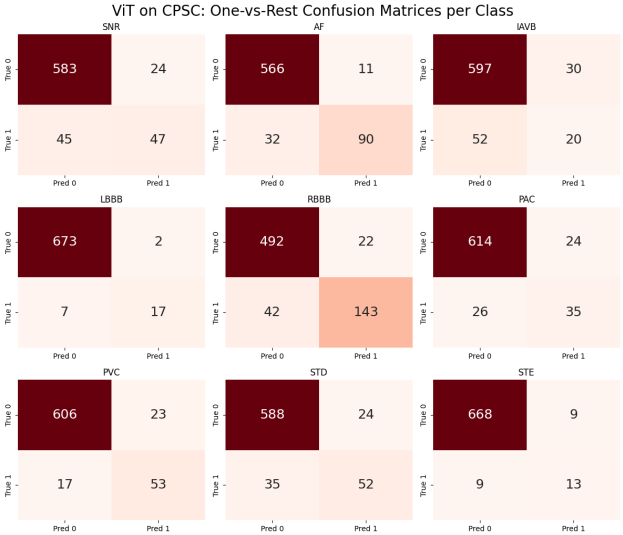

Figure 5: Best fold performed on CPSC dataset: Confusion matrix from fold no.5

Among all arrhythmia categories, the predictions for AF, LBBB, RBBB, and PVC were the most accurate in signal-based models, with F1 scores of 0.85, 0.79, 0.78, and 0.74, respectively. However, the model struggles to correctly identify positive cases of IAVB, resulting in a high number of false negatives — 52 out of 72 cases. This difficulty likely stems from the challenge of diagnosing IAVB in clinical practice due to its subtle features and overlap with other conditions. A similar pattern is observed in the STE class, as diagnosis can be challenged for clinical interpretation; [5] the performance of this class could reduce the overall macro metrics, as nearly half of the cases are being incorrectly identified (9 false negatives over 22 STE.). Therefore, interpretation should take this into account.

3.3 Classification performances - ECG image datasets

In this extended research, we enhance Cardio Care capabilities on small-scale datasets by utilising two variants of the ViT: Deit and Beit. On the Mendeley dataset (N=929), ViT

and DeiT models outperformed BeiT and previous studies, achieving a precision, recall and overall F1 score of 0.99. Meanwhile, Sadaq et al. achieved an overall F1 score of 0.98 with a lightweight 4-layer CNN [22], whereas Abubaker et al. obtained the same F1 score but with a slightly better recall of 0.99 compared to 0.98 using 2D CNN network [1].

Table 6: Our Transformer variants vs. CNNs on Mendeley dataset: Overall 10-fold CV performance

	2D CNN	**Light CNN**	**ViT**	**DeiT**	**BeiT**
Accuracy	0.98	0.98	**0.99**	**0.99**	0.86
Precision	0.98	0.98	**0.99**	**0.99**	0.85
Recall	**0.99**	0.98	**0.99**	**0.99**	0.85
F1-score	0.98	0.98	**0.99**	**0.99**	0.85

Table 7: Our best model ViT vs. CNNs on Mendeley dataset: per class performance

	Metric	**2D CNN**	**Light CNN**	**Ours**
Normal	Precision	0.97	-	1
	Recall	-	-	1
	F1	-	-	1
Abnorm beat	Precision	1	-	1
	Recall	-	-	0.98
	F1	-	-	0.99
MI	Precision	0.98	-	1
	Recall	-	-	1
	F1	-	-	1
MI his	Precision	0.98	-	0.96
	Recall	-	-	1
	F1	-	-	0.98
Average	Precision	0.98	0.98	**0.99**
	Recall	**0.99**	0.98	**0.99**
	F1	0.98	0.98	**0.99**

Overall, with the Mendeley sample size and balanced class distributions, ViT continues to deliver the strongest results among all models. DeiT serves as a comparable alternative to ViT, exhibiting similar performance (0.992 and 0.993, respectively).

In each individual class, ViT's performance per class can be found in **Table 7**, achieving 100% F1-scores for healthy and myocardial infarction subjects, 99% for abnormal heartbeat conditions, and 98% for individuals with a history of myocardial infarction.

In this study, we explore all three Transformer variants on another private image-based dataset (**Table 8**). Our ViT model achieves the highest F1-score of 0.81 compared to the other two variants.

Table 8: Our Transformer variants on the private Tam Duc dataset: Overall 10-fold CV performance

	ViT	**DeiT**	**BeiT**
Accuracy	0.82	0.82	0.82
Precision	0.85	**0.87**	**0.87**
Recall	**0.80**	0.78	0.78
F1-score	**0.81**	0.79	0.79

3.4 Explainability with Grad-Cam

To enhance model interpretability, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to visualise the attention distribution of the ViT on ECG images. Grad-CAM heatmaps were overlaid on the original input images to highlight regions that contributed most significantly to the model's predictions. Our Grad-Cam function successfully explains abnormal heartbeat and myocardial conditions in a balanced data set (Mendeley).

Representative examples are shown in Figure 6 and 7, drawn from the Mendeley dataset. In correctly classified abnormal ECGs with myocardial infarction, the attention maps consistently focused on waveform segments with clinical relevance—such as ST-segment deviations and irregular QRS morphology. Notably, Grad-CAM confirmed that the model does not depend on irrelevant regions (e.g., gridlines or metadata text), thereby further validating the efficacy of the preprocessing pipeline.

4 Discussion

This study offers notable contributions, including the following key points:

- We extend the Cardio Care pipeline by evaluating three Vision Transformer architectures—ViT, DeiT, and BEiT—for the classification of cardiac abnormalities from ECG images.
- We benchmark model performance on real-world ECG report images, using three datasets of varying size and modality.
- We demonstrate the feasibility of deploying Vision Transformer-based models in low-resource clinical settings where only image-based ECG inputs are available.
- We integrate a Grad-CAM-based explainability feature into the pipeline, enabling visual interpretation of the model's attention on ECG waveform regions to support clinical decision-making.

4.1 Model performance and generalisability

Despite the relatively small sample sizes of the image-based datasets, the Vision Transformer models demonstrated competitive performance in ECG classification. On

Figure 6: We evaluate our GradCam function (features extracted from the second and third-to-last layers: highlight in yellow and red for elevated ST segments) to predict myocardial infarction and a history of myocardial infarction cases [Random subjects - ID No. 10 in each class]

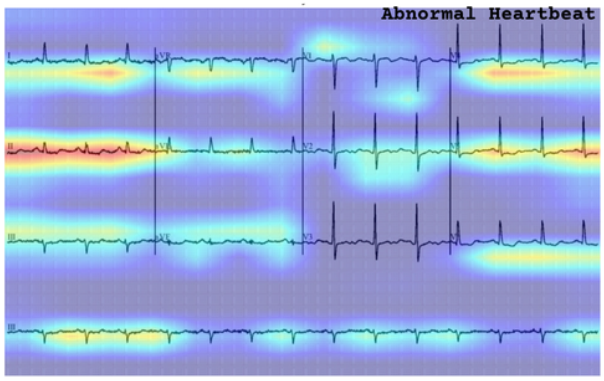

Figure 7: We evaluate our GradCam function (features extracted from the second and third-to-last layers: highlight in yellow and red for flutter or fibrillation segments) to predict abnormal heartbeat [Random subject]

the Tam Duc Cardiometabolic dataset, the best-performing model achieved a macro F1-score of 0.81, while the Mendeley dataset yielded extremely well performance (F1-score of 0.99), with balanced precision and recall across classes. These results indicate that Vision Transformers are capable of effectively capturing clinically relevant waveform features from real-world ECG images.

Among the evaluated architectures, ViT consistently outperformed or matched DeiT and BeiT across all datasets, reinforcing its suitability for ECG image interpretation tasks. Compared to prior CNN-based approaches [1, 22], ViT-based models achieved superior results, particularly in generalisation and consistency across input variations. This endorses the ongoing incorporation of transformer-based methodologies into image-oriented diagnostic procedures within resource-limited clinical settings.

A key addition in this study was the implementation of a Grad-CAM-based explainability module to visualise where the model concentrates on the ECG waveform. This feature is crucial for enhancing transparency and building clinical trust in AI systems. Grad-CAM heatmaps revealed that the models mainly focus on leads and segments that are pathologically relevant, which enhances the interpretability of the predictions and supports the decision-making process.

4.2 Limitations and further research

This study has several limitations. First, the CPSC and Tam Duc datasets exhibit class imbalance, which may introduce bias or skew evaluation metrics despite the use of class-weighted loss functions. Future work should explore advanced strategies for handling imbalance, such as focal loss or synthetic data augmentation. Although Transformer models such as ViT are capable of learning from small datasets due to pretraining, recent studies suggest that training at large-scale (from 100000 samples) may be necessary to fully exploit their capacity and improve generalisability in clinical applications.

Secondly, the Grad-CAM visualisations, while effective on the larger Mendeley dataset, showed limited interpretive clarity on the smaller Tam Duc dataset. This limitation is likely due to the restricted sample size, which may constrain the model's ability to form robust attention patterns. In low-data scenarios, the model may lack sufficient examples to produce consistent or clinically meaningful explanations. This highlights the need for either larger annotated datasets or the adoption of interpretability-focused architectures optimised for small-sample learning.

Lastly, although the model was evaluated across three datasets with different label structures, its diagnostic coverage remains limited to major rhythm classes and binary disease classification. Future work should aim to expand the model's label space to include more granular and rare ECG abnormalities, and explore multi-task learning to capture broader cardiovascular and metabolic risk profiles.

5 Conclusion

While modern ECG analysis techniques have demonstrated high diagnostic accuracy, their dependence on digital signal data presents limitations in low-resource and image-only clinical environments [12]. This study demonstrates that integrating Vision Transformer architectures into ECG image classification pipelines offers a viable and effective alternative.

By benchmarking ViT, DeiT, and BEiT models across datasets—including a real-world clinical ECG image dataset—we show that these models can achieve strong classification performance, even with limited data. The inclusion of a Grad-CAM-based explainability module further enhances the transparency of the pipeline, making it more suitable for clinical decision support.

These findings support using image-based deep learning in cardiac screening, especially where access to digitised ECG data is limited.

Acknowledgement

Ngoc Minh To: Conceptualisation, Data curation, Formal analysis, Visualisation, Writing - Original Draft. **Vu Vo**: Conceptualisation, Formal analysis, Investigation, Writing - Review & Editing. **Dang V. Nguyen**, **Dan V. B. Do** and **Quoc Cuong Ngo**: Resources, Methodology, Validation, Writing - Review & Editing. **Dinesh Kumar** and **Minh Ngoc Dinh**: Project administration and Supervision.

References

[1] Abubaker, M.B., Babayiğit, B.: Detection of cardiovascular diseases in ecg images using machine learning and deep learning methods. IEEE Transactions on Artificial Intelligence **4**(2), 373–382 (2023), 10.1109/TAI.2022.3159505

[2] Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. ArXiv **abs/2106.08254** (2021), https://api.semanticscholar.org/CorpusID:235436185

[3] Cesare, M.D., Perel, P., Taylor, S., Kabudula, C.W., Bixby, H., Gaziano, T.A., McGhie, D.V., Mwangi, J., Pervan, B., Narula, J., Piñeiro, D.J., Pinto, F.J.: The heart of the world. Global Heart **19** (2024), https://api.semanticscholar.org/CorpusID:267254220

[4] Chiang, H.T., Hsieh, Y.Y., Fu, S.W., Hung, K.H., Tsao, Y., Chien, S.Y.: Noise reduction in ecg signals using fully convolutional denoising autoencoders. IEEE Access **7**, 60806–60813 (2019), 10.1109/ACCESS.2019.2912036

[5] Chukwu, E.C., Moreno-Sánchez, P.A.: Enhancing arrhythmia diagnosis with data-driven methods: A 12-lead ecg-based explainable ai model. In: Särestöniemi, M., Keikhosrokiani, P., Singh, D., Harjula, E., Tiulpin, A., Jansson, M., Isomursu, M., van Gils, M., Saarakkala, S., Reponen, J. (eds.) Digital Health and Wireless Solutions. pp. 242–259. Springer Nature Switzerland, Cham (2024), https://doi.org/10.1007/978-3-031-59091-7_16

[6] Darmawahyuni, A., Nurmaini, S., Rachmatullah, M.N., Tutuko, B., Sapitri, A.I., Firdaus, F., Fansyuri, A., Predyansyah, A.: Deep learning-based electrocardiogram rhythm and beat features for heart abnormality classification. PeerJ Computer Science **8** (2022), https://peerj.com/articles/cs-825/

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://doi.org/10.48550/arXiv.2010.11929

[8] Gour, A., Gupta, M., Wadhvani, R., Shukla, S.: A comprehensive review of heart disease classification techniques utilizing ecg signal analysis. 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM) pp. 1–6 (2023), https://ieeexplore.ieee.org/document/10370226

[9] He, F., Nie, F., Wang, R., Jia, W., Zhang, F., Li, X.: Semisupervised band selection with graph optimization for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **59**(12), 10298–10311 (2021), 10.1109/TGRS.2020.3037746

[10] Ji, Z., Wang, Q., Cui, B., Pang, Y., Cao, X., Li, X.: A semi-supervised zero-shot image classification method based on soft-target. Neural Networks **143**, 88–96 (2021), https://www.sciencedirect.com/science/article/pii/S089360802100215X

[11] Khan, A.H., Hussain, M.: Ecg images dataset of cardiac patients (2021). https://doi.org/10.31449/inf.v49i3.1018010.17632/gwbz3fsgp8.2, https://data.mendeley.com/datasets/gwbz3fsgp8/2

[12] Khunte, A., Sangha, V., Oikonomou, E.K., Dhingra, L.S., Aminorroaya, A., Coppi, A.C., Shankar, S.V., Mortazavi, B.J., Bhatt, D.L., Krumholz, H.M., Nadkarni, G.N., Vaid, A., Khera, R.: Automated diagnostic reports from images of electrocardiograms at the point-of-care. medRxiv (2024), 10.1101/2024.02.17.24302976

[13] Kligfield, P., Gettes, L.S., Bailey, J.J., Childers, R., Deal, B.J., Hancock, E.W., van Herpen, G., Kors, J.A., Macfarlane, P., Mirvis, D.M., Pahlm, O., Rautaharju, P., Wagner, G.S.: Recommendations for the standardization and interpretation of the electrocardiogram. Circulation **115**(10), 1306–1324 (2007), 10.1161/CIRCULATIONAHA.106.180200

[14] Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K., O'Leary, A.: Pywavelets: A python package for wavelet analysis. Journal of Open Source Software **4**(36), 1237 (2019), https://doi.org/10.21105/joss.01237

[15] Li, Y., Li, Z., Wei, K., Xiong, W., Yu, J., Qi, B.: Noise estimation for image sensor based on local entropy and median absolute deviation. Sensors **19**(2), 339 (2019), https://doi.org/10.3390/s19020339

[16] Martin, S.S., et al.: 2024 heart disease and stroke statistics: A report of us and global data from the american heart association. Circulation **149**(8), e347–e913 (2024), 10.1161/CIR.0000000000001209

[17] Miao, Y., Wang, Q., Chen, M., Li, X.: Spatial-spectral hyperspectral image classification via multiple random anchor graphs ensemble learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 3641–3644. IEEE (2021), 10.1109/IGARSS47720.2021.9553932

[18] MSD, M.: Ecg: Reading the waves (2023), https://www.msdmanuals.com/home/multimedia/image/ecg-reading-the-waves

[19] Ng, E.Y.K., et al.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics (2018), https://doi.org/10.1166/JMIHI.2018.2442

[20] NUSSINOVITCH, U., ELISHKEVITZ, K.P., KAMINER, K., NUSSINOVITCH, M., SEGEV, S., VOLOVITZ, B., NUSSINOVITCH, N.: The efficiency of 10-second resting heart rate for the evaluation of short-term heart rate variability indices. Pacing and Clinical Electrophysiology **34**(11), 1498–1502 (2011), 10.1111/j.1540-8159.2011.03178.x

[21] Roth, G.A., et al.: Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study. Journal of the American College of Cardiology **76**(25), 2982–3021 (2020), 10.1016/j.jacc.2020.11.010

[22] Sadad, T., Safran, M., Khan, I., Alfarhood, S., Khan, R., Ashraf, I.: Efficient classification of ecg images using a lightweight cnn with attention module and iot. Sensors **23**(18) (2023), 10.3390/s23187697

[23] Sangha, V., Mortazavi, B., Haimovich, A.D., Ribeiro, A.H., Brandt, C.A., Jacoby, D.L., Schulz, W.L., Krumholz, H.M., Ribeiro, A.L.P., Khera, R.: Automated multilabel diagnosis on electrocardiographic images and signals. Nature Communications **13** (2021), https://rdcu.be/dRfuy

[24] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J'egou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (2020), https://api.semanticscholar.org/CorpusID:229363322

[25] Vonesch, C., Blu, T., Unser, M.: Generalized daubechies wavelet families. IEEE transactions on signal processing **55**(9), 4415–4429 (2007), 10.1109/TSP.2007.896255

[26] Vu, V.Q., Minh To, N., Nguyen Duc, T., Phung, N., Ngo, Q., Kumar, D., Dinh, M.: Cardio care: A vision transformer cardiac classification based on electrocardiogram images and signals. In: Buntine, W., Fjeld, M., Tran, T., Tran, M.T., Huynh Thi Thanh, B., Miyoshi, T. (eds.) Information and Communication Technology. pp. 199–209. Springer Nature Singapore, Singapore (2025), https://doi.org/10.1007/978-981-96-4285-4_17

[27] Xie, G.S., Zhang, Z., Liu, L., Zhu, F., Zhang, X.Y., Shao, L., Li, X.: Srsc: selective, robust, and supervised constrained feature representation for image classification. IEEE transactions on neural networks and learning systems **31**(10), 4290–4302 (2019), 10.1109/TNNLS.2019.2953675

[28] Zhang, D., Yang, S., Yuan, X., Zhang, P.: Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. iScience **24**(4), 102373 (2021), 10.1016/j.isci.2021.102373