Faithfulness-Aware Visual-Evidence-Grounded Explanations for Cardiovascular Signal Image Classification

Muhammad Alfaiz Khalifah Alamsyah[1]★, Novanto Yudistira[1], and Wayan Firdaus Mahmudy[1]

Faculty of Computer Science, University of Brawijaya, Malang, Indonesia
faiz_doc@student.ub.ac.id

Abstract. Cardiovascular signals such as electrocardiogram and photoplethysmography are increasingly analyzed using deep neural networks after being transformed into time-frequency images. Although convolutional models can achieve strong classification performance, their predictions remain difficult to interpret, and natural language explanations generated by language models may sound plausible without being faithful to the visual evidence used by the classifier. This paper presents a faithfulness-aware explanation pipeline for cardiovascular signal image classification. The system first transforms cardiovascular signals into scalogram images and classifies them using a lightweight convolutional backbone. Grad-CAM is then used to extract structured visual evidence, including the dominant temporal-frequency region, activation area ratio, peak activation score, predicted class, and confidence. These evidence fields condition both template-based and lightweight language-model explanations. We evaluate the system using classification metrics and explanation reliability metrics, including evidence support rate, hallucination rate, label agreement, and evidence coverage. The proposed evidence-grounded explanation is expected to reduce unsupported clinical statements compared with unconstrained language-model explanations while preserving interpretability for non-expert users. This study is positioned as an explainable AI prototype, not a clinical diagnostic system. Our research shows that CNN can classify scalogram images with high accuracy, demonstrating good performance. Meanwhile, Grad-cam also provides good heatmaps. Future research should focus on using more advanced and complex LLMs.

Keywords: XAI · Faithfulness-Aware · Classification · Cardiovascular Signal Image.

1 Introduction

Cardiovascular signals such as the electrocardiogram (ECG) are widely used for detecting cardiac abnormalities and assessing cardiovascular status [1, 2]. In re-

★ Corresponding author.

© The Author(s) 2026
M. Murtadha et al. (eds.), *Proceedings of the 2026 11th International Conference on Sustainable Information Engineering and Technology (SIET 2026)*, Atlantis Highlights in Computer Sciences 26,
https://doi.org/10.2991/978-94-6239-828-3_9

cent years, deep learning has been increasingly applied to automated ECG analysis because it can achieve high classification performance from biosignals [3, 4]. These models can analyze ECG either as one-dimensional waveforms or as transformed image representations such as time-frequency scalograms, where continuous wavelet transform (CWT) preserves temporal and frequency information for downstream learning [5, 6]. This image-based formulation is attractive because convolutional neural networks (CNNs) can extract discriminative spatial patterns from scalogram images while reducing dependence on handcrafted features [7, 8].

Despite their predictive capability, deep neural networks remain difficult to interpret, and their black-box nature continues to limit clinical applicability [9, 10]. This limitation is especially important in healthcare, where trustworthy AI systems must provide transparent and accountable reasoning before they can be integrated into routine workflows [11, 12]. In ECG classification, explainable artificial intelligence (XAI) methods such as SHAP, LIME, saliency maps, and Grad-CAM are commonly used to expose the basis of model predictions [13, 14]. Among them, Grad-CAM is widely adopted because it generates class-discriminative localization maps by tracing gradients into the final convolutional layer and highlighting regions that contribute to a classifier decision [15]. For cardiovascular signal images, these highlighted regions can potentially correspond to informative temporal segments or frequency bands. However, existing ECG explainability methods still face important limitations, including inconsistent explanations, weak temporal dependency visualization, and difficulty capturing fine-grained patterns [16, 17].

Recent progress in language models has created new opportunities to convert model evidence into natural language explanations that are easier for end users to understand. Textual justifications are increasingly recognized as a useful form of explanation in medical AI, alongside visual and example-based explanations [18]. However, natural language generation also introduces a reliability challenge: a language model can produce a clinically plausible explanation that is not actually grounded in the classifier's decision process or visual evidence. For explainable cardiovascular AI, this creates an explanation faithfulness problem, in which the generated explanation should remain consistent with the predicted label, confidence, and supporting visual evidence rather than merely sounding medically reasonable [19].

To address this issue, this study proposes a faithfulness-aware, visual-evidence-grounded explanation pipeline for cardiovascular signal image classification. The pipeline transforms ECG signals into time-frequency images, trains a lightweight CNN classifier, extracts Grad-CAM-based visual evidence, and converts that evidence into a structured prompt for template-based and lightweight language-model explanation generation. The structured evidence includes the predicted class, confidence score, dominant region, activation area ratio, peak activation score, temporal focus, and frequency focus. By constraining the explanation generator to information explicitly derived from the classifier and its visual evi-

dence, the proposed design aims to reduce unsupported statements and improve explanation reliability.

The contribution of this work is threefold. First, it presents an end-to-end pipeline that links cardiovascular signal image classification, visual explanation, and natural language explanation generation. Second, it introduces a structured visual evidence prompt intended to reduce unsupported content in generated explanations. Third, it evaluates explanation reliability using faithfulness-oriented measures, including evidence support rate, hallucination rate, label agreement, and evidence coverage, in addition to standard classification metrics. The proposed system is intended as an explainable AI prototype and should not be interpreted as a clinical diagnostic tool.

2 Related Work

The first paper [20] focuses on cardiovascular disease (CVD) risk prediction through retinal fundus image analysis, based on the premise that the retina provides a non-invasive window into the body's vascular system. The proposed approach employs a hybrid deep learning architecture, specifically MobileNet combined with Long Short-Term Memory (LSTM), to capture both spatial features and sequential dependencies within retinal fundus images. The study utilized retinal fundus image datasets collected from the DRIVE, MESSIDOR, and an internal dataset. The experimental results demonstrated that the proposed MobileNet+LSTM hybrid model achieved the highest accuracy of 85%. Future work suggested further development and validation of the proposed system as a clinical diagnostic tool, particularly for deployment in remote and underserved healthcare settings.

The second paper [21] investigates post-hoc explanation methods for multimodal medical image analysis with the objective of enhancing clinical trust in deep learning-based decision-making. The researchers compared gradient-based explanation methods, including Gradient-weighted Class Activation Mapping (Grad-CAM) and Guided Backpropagation, with perturbation-based methods, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The evaluation was conducted using the BraTS 2020 (Brain Tumor Segmentation) dataset. The results demonstrated that the generated heatmap visualizations could identify clinically relevant regions contributing to brain tumor classification. These findings highlight the potential of post-hoc explainability techniques to improve the interpretability and transparency of deep learning models in medical image analysis.

The third paper [22] introduces a framework that integrates similarity learning and adversarial learning to generate sharper and more informative visual explanations for medical image classification. The study addresses the inconsistent performance of conventional saliency methods, which frequently produce noisy and poorly localized explanations. Using the RSNA Pneumonia Detection chest X-ray dataset, the proposed approach generates explanation maps based on the differences between images produced by a similarity-based generator and

an adversarial generator. The experimental results demonstrated superior localization performance compared with Grad-CAM, as measured by the Intersection over Union (IoU) metric. Future research directions include extending the proposed framework to multi-class classification tasks and three-dimensional (3D) medical imaging.

The fourth paper [23] proposes an automated ECG signal classification system to address the substantial manual workload associated with physician-based interpretation, which can be time-consuming and susceptible to diagnostic errors, particularly in settings with a shortage of cardiology specialists. The proposed methodology integrates Continuous Wavelet Transform (CWT) for converting one-dimensional (1D) ECG signals into two-dimensional (2D) scalogram images, which are subsequently processed using the AlexNet architecture through fine-tuning or transfer learning. The study employed data from PhysioNet, specifically the MIT-BIH Arrhythmia Database, covering arrhythmia, congestive heart failure, and normal sinus rhythm categories. The results demonstrated a high classification accuracy of 96%, outperforming conventional approaches such as Support Vector Machines (SVM) and standard convolutional neural networks (CNNs). Future work includes further refinement of the proposed methodology and evaluation of its scalability across broader patient populations and diverse healthcare settings.

The fifth paper [24] proposes medXGAN, a Generative Adversarial Network (GAN)-based framework for explaining medical image classification decisions through a generative latent space. The primary motivation is the limitation of conventional saliency-based methods, such as Grad-CAM, which may generate explanations with insufficient spatial detail and limited stability. Using COVID-19 chest X-ray and brain tumor MRI datasets, medXGAN decomposes anatomical structures and pathological features within the generative latent space, enabling their visualization through latent-space interpolation. The results demonstrated substantially finer-grained pathological localization compared with conventional Grad-CAM-based explanations. Future research directions include improving image resolution and extending the framework to multi-class classification tasks.

The subsequent paper [25] proposes a segmentation-to-explanation framework based on Large Language Models (LLMs) for generating textual reports from magnetic resonance imaging (MRI) data. The study is motivated by the black-box nature of deep learning models, which remains a major barrier to their clinical adoption. The proposed methodology employs SegResNet for image segmentation, followed by atlas-based anatomical mapping using the Julich Brain Atlas. The resulting anatomical information is structured in JavaScript Object Notation (JSON) format and subsequently provided as contextual input to LLMs, including Gemma, Llama, and Mistral, for automated medical report generation. Using the BraTS 2023 and Shifts 2022 datasets, the proposed framework demonstrated the ability to generate reports that were both factually accurate and contextually informative. Future work includes incorporating indi-

vidual patient medical histories to enable the generation of more personalized and clinically contextualized reports.

3 Methodology

This study uses a pipeline-based experimental design for cardiovascular signal image classification and explanation generation. The overall process consists of ECG signal preparation, transformation into time-frequency scalogram images, CNN-based classification, Grad-CAM visualization, structured evidence extraction, and natural-language explanation generation. The image-based formulation is used because CWT converts one-dimensional ECG signals into representations that preserve temporal and frequency information, while CNNs can learn discriminative image features from the resulting scalograms. Prior ECG studies have used this strategy to classify arrhythmia, congestive heart failure, and normal sinus rhythm, with reported accuracies of 98.74% in one CWT-CNN study and 96% in a transfer-learning scalogram study.

3.1 Beat Extraction

The beat extraction stage was performed to obtain individual heartbeat segments from each electrocardiogram (ECG) recording. Each ECG recording was processed together with its corresponding annotation file to identify the locations of annotated heartbeats within the signal. The ECG channel was selected by prioritizing the MLII channel when available, as this channel was used in the majority of the recordings. If the MLII channel was not available, another available ECG channel from the recording was used as an alternative.

Based on the beat locations obtained from the annotation files, the ECG signal was segmented into individual beat segments centered around each annotated beat. Each beat was represented using a signal window of 360 samples, consisting of 180 samples before and 180 samples after the annotation point. Segments whose index ranges extended beyond the boundaries of the recording were excluded from the extraction process to prevent the generation of incomplete segments. Consequently, all successfully extracted beats had a uniform length, allowing them to be processed consistently in subsequent stages.

Each ECG segment was then assigned a label based on the beat annotation type. The N annotation was categorized as a normal beat and assigned a label of 0, whereas beat annotations other than N were categorized as abnormal beats and assigned a label of 1. Annotations representing rhythm markers, such as the symbol, were excluded from the labeling process because they do not represent beat categories targeted by the classification task. All ECG segments and their corresponding labels were subsequently converted and compiled into a numerical dataset. The extraction process yielded a total of 111,298 ECG segments, each consisting of 360 samples, including 75,011 normal segments and 36,287 abnormal segments.

The extracted data were validated by calculating the distribution of segments across the respective classes to ensure that the labeling process produced the two predefined categories. Several ECG recordings were also examined to verify that the signal structure and channel information could be correctly read during the extraction process. In addition, representative examples of normal and abnormal beat segments were visualized to qualitatively inspect the morphological characteristics of the extracted signals. The output of this stage was a collection of ECG segments with uniform lengths and labels corresponding to the normal or abnormal class. The resulting dataset was subsequently used as the input for signal transformation into scalograms and classification using a Convolutional Neural Network (CNN).

3.2 Scalogram Generation

The ECG beat segments obtained during the extraction stage were subsequently transformed into scalograms. A scalogram is a two-dimensional representation of a signal that describes the distribution of signal characteristics across time and frequency. This representation was used to transform the one-dimensional ECG signals into images, allowing signal patterns to be learned using a Convolutional Neural Network (CNN)-based approach. Through this transformation, the temporal and frequency information contained in the ECG signals was represented as images that served as inputs to the classification model.

The signal transformation was performed using the Continuous Wavelet Transform (CWT) with a Morlet wavelet. CWT enables the analysis of signal characteristics across multiple scales and time positions, allowing changes in frequency components throughout the signal duration to be represented simultaneously. The transformation results were then represented as scalograms containing information on the intensity of the wavelet coefficients in the time-frequency domain. The resulting values were normalized to obtain a consistent value range, and the transformed images were subsequently resized to 128×128 pixels. Standardizing the image dimensions ensured that all images had the same size and could be consistently used as inputs to the CNN model.

Of the 111,298 ECG segments obtained during the extraction stage, 60,000 segments were selected for scalogram dataset generation. The dataset consisted of 30,000 normal-class segments and 30,000 abnormal-class segments. An equal number of samples from each class was used to achieve a balanced class distribution within the dataset. Samples from each class were selected randomly using a random seed of 42 to ensure that the sampling process was reproducible. The use of a subset of the available data also reduced storage requirements and computational load during the transformation process compared with converting all available ECG segments into scalograms.

Each transformed image was saved in PNG format and organized according to its class into two directories, namely normal and abnormal. This directory structure was used to preserve the class labels associated with each image while facilitating dataset loading during the model training and evaluation stages. After the transformation process was completed, the number of images in each

class was verified to ensure consistency between the generated dataset and the predetermined number of samples.

Visual inspection was performed by displaying representative scalograms from each class to qualitatively evaluate the transformation results. This inspection aimed to ensure that the ECG signals were successfully represented as images and produced visual patterns suitable for use as inputs in the classification process. The resulting scalogram dataset was subsequently used as input for classification using the ResNet-18 architecture. Following the classification process, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to identify regions of the scalograms that contributed to the model's classification decisions.

3.3 CNN Classification

The generated scalogram dataset was subsequently used as input for the Convolutional Neural Network (CNN)-based classification stage. The architecture employed in this study was ResNet-18, a relatively lightweight CNN architecture that utilizes residual learning to learn feature representations from images. In this study, ResNet-18 was used to perform binary classification of ECG scalograms into two classes, namely normal and abnormal. The trained model served as the primary classification model and provided the basis for the subsequent interpretation stage using Gradient-weighted Class Activation Mapping (Grad-CAM).

Before being used as model inputs, all scalogram images were standardized to a size of 128×128 pixels. Since the generated scalograms were grayscale images, they were converted into three channels to match the input format of ResNet-18, which uses RGB images. The pixel values were then normalized using normalization parameters consistent with the configuration of the pretrained ResNet-18 model. The dataset was loaded according to a directory structure consisting of normal and abnormal classes, allowing the label of each image to be determined based on the directory in which the image was stored.

The 60,000 scalogram images were divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The dataset was randomly split using a random seed of 42 to ensure that the partitioning process was reproducible. The training set was used to optimize the model parameters, while the validation set was used to monitor model performance during training. The test set was not involved in parameter updates and was used only after training was completed to evaluate the model's generalization ability on previously unseen data. The training data were processed using a mini-batch approach with a batch size of 32 images per iteration.

The ResNet-18 model was initialized with pretrained weights, and the final classification layer was modified to accommodate the binary classification task in this study. The original output layer was replaced with a classification layer producing two outputs, representing the normal and abnormal classes, respectively. Model optimization was performed using Cross-Entropy Loss as the loss function and the Adam optimizer with a learning rate of 1×10^{-4}. The model was

trained for 10 epochs. During each epoch, the model parameters were updated based on the training data, followed by performance evaluation on the validation set. Training loss and validation loss were recorded at each epoch to monitor the model learning process.

After training was completed, the model was evaluated using the test set, which had not been involved in parameter optimization. In addition to generating class predictions, the model produced class probabilities through the softmax function. The probability associated with the abnormal class was used as the prediction score for calculating the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Model performance was evaluated using several metrics, namely Accuracy, Macro-Recall, Macro-F1, Specificity, and AUC. In addition to numerical metrics, a confusion matrix was used to analyze the distribution of correct and incorrect predictions for each class. The use of multiple evaluation metrics enabled a more comprehensive assessment of model performance, particularly its ability to distinguish between normal and abnormal classes.

3.4 GRAD-CAM Interpretation

After the classification process using ResNet-18 was completed, the next stage involved interpreting the model's decisions using Gradient-weighted Class Activation Mapping (Grad-CAM). This method was used to identify regions of the scalogram that contributed to the model's classification decisions. Unlike the classification output, which only indicates the predicted class and the model's confidence, Grad-CAM provides a spatial representation in the form of a heatmap that illustrates the relative distribution of contributions across different regions of the image to the analyzed class. Thus, Grad-CAM was employed to obtain information regarding the locations of visual features that received attention from the model when generating its classification decisions.

Grad-CAM was implemented on the layer4 layer of ResNet-18. This layer was selected because it is located near the end of the convolutional network and produces feature representations with a higher level of abstraction before being passed to the classification layer. For each image in the test set, the model was executed to obtain the predicted class and confidence score. The predicted class was then compared with the ground-truth label to determine the prediction status, namely correct or incorrect. This information was retained together with the interpretation results, allowing Grad-CAM analysis to be performed on both correctly and incorrectly classified samples.

The Grad-CAM output was an activation map representing the relative contribution at each spatial location within the scalogram. Regions with higher activation values indicated areas of the image that had relatively greater contributions to the prediction of the analyzed class. The activation map was subsequently normalized and combined with the original scalogram to generate an overlay visualization. This visualization was used to directly illustrate the locations of activation on the scalogram. Each overlay result was saved as an image file so that it could be individually reviewed and used as part of the documentation of the model interpretation process.

In addition to generating a visual representation, the Grad-CAM results were converted into numerical information to form structured evidence. One of the extracted characteristics was the activation ratio, defined as the proportion of the activation map area with activation values above a predetermined threshold. This parameter was used to represent the spatial extent of the area considered relevant by the model. The maximum and mean activation values were also calculated to characterize the activation intensity of the activation map. The location with the highest activation value was identified as the hotspot, and its coordinates were subsequently used to determine the dominant region along the time and frequency dimensions.

To facilitate interpretation, the hotspot position was mapped into temporal and frequency region categories. The time dimension was divided into three regions: early, middle, and late, whereas the frequency dimension was divided into three categories: low, medium, and high. This mapping allowed the spatial information from the activation map to be represented as categorical descriptions that were more readily applicable to the subsequent stage of language-based explanation generation.

All classification and Grad-CAM information was subsequently compiled into an evidence record for each sample. Each record contained the sample identifier, predicted class, ground-truth class, prediction correctness status, confidence score, activation ratio, maximum and mean activation values, dominant temporal region, dominant frequency region, and hotspot position. The file path of the Grad-CAM overlay was also stored, allowing the numerical information to be traced back to its corresponding visualization. With this structure, the interpretation results were made available not only as images but also as structured data that could be further processed in the subsequent stage.

3.5 LLM Explanation

After the Grad-CAM interpretation evidence was obtained, the next stage involved generating language-based explanations using a Large Language Model (LLM). The model employed in this study was Qwen2.5-0.5B-Instruct, which has a relatively small number of parameters and is therefore suitable for experiments in computationally constrained environments. The LLM was not used to perform ECG classification directly. Classification was performed by ResNet-18, while the LLM was responsible for transforming information regarding the classification results and interpretability evidence into textual explanations.

The explanation experiments were conducted using two approaches. The first approach was the LLM without evidence, in which the LLM was provided with the predicted class and confidence score generated by ResNet-18 without any information regarding the regions receiving attention from the model based on Grad-CAM. This approach served as a baseline condition for examining the characteristics of explanations generated by the LLM when no structured evidence regarding the visual representation underlying the classification decision was provided.

The second approach was the evidence-grounded LLM, in which the LLM was provided with the prediction information along with the evidence obtained from Grad-CAM. The provided information included the predicted class, confidence score, activation ratio, maximum and mean activation values, dominant temporal region, dominant frequency region, and hotspot position. The prompt used in this approach was designed to guide the LLM to construct explanations based on the available evidence while avoiding the addition of information that was not contained in the evidence. Consequently, the generated explanations were expected to remain grounded in the information underlying the classification model's decision.

To ensure consistency in the comparison process, all samples were processed using a predefined prompt format. A total of 500 samples from the evidence dataset were selected as the LLM evaluation subset. The subset was selected based on computational requirements and the higher inference time of the LLM compared with the CNN classification process. Each sample was processed using both explanation approaches, resulting in two textual outputs: an explanation generated by the LLM without evidence and an explanation generated by the evidence-grounded LLM. All outputs were subsequently collected and stored in a csv file for further analysis.

The evaluation was conducted to assess the faithfulness of the generated explanations with respect to the information produced by the classification model and the Grad-CAM evidence. One of the metrics used was Label Agreement, which measures the consistency between the class mentioned in the explanation and the class predicted by ResNet-18. Evidence Coverage was then used to measure the extent to which the main evidence information was represented in the textual explanation. In this study, the primary information evaluated included the predicted label, dominant temporal region, and dominant frequency region. Meanwhile, Evidence Support Rate was used to measure the proportion of the key information stated in the explanation that was appropriately supported by the available evidence.

In addition to these three metrics, Hallucination Rate was calculated to identify the occurrence of information in the explanations that contradicted the available evidence. The assessment was performed across three main aspects: predicted class, temporal region, and frequency region. For example, if the evidence indicated that the dominant temporal region was categorized as middle, but the explanation stated that the dominant region was early, the statement was categorized as inconsistent with the evidence. Thus, Hallucination Rate served as an indicator of the extent to which the LLM generated information that was inconsistent with the evidence underlying the explanation.

The evaluation results for each sample were compiled and stored in the csv file. The mean values of Label Agreement, Evidence Coverage, Evidence Support Rate, and Hallucination Rate were subsequently calculated to obtain an overall performance summary for each explanation approach. These four metrics were used to compare the characteristics of explanations generated by the LLM without evidence and the evidence-grounded LLM, particularly in terms of the

effect of incorporating Grad-CAM evidence on the consistency of the generated explanations with the classification model's decisions.

Overall, this stage constituted the final component of the proposed explainability pipeline. ResNet-18 generated the classification predictions, Grad-CAM was used to obtain information regarding the regions contributing to the model's decisions, and this information was subsequently converted into structured evidence and provided to the LLM for textual explanation generation. This architecture maintained a clear separation between the classification and explanation processes, such that the LLM did not replace the primary classification model but instead functioned as an explanation-generation component grounded in information produced by the model. The developed system represents an explainability prototype for ECG classification models and is not intended to serve as a clinical diagnostic device. Therefore, the generated explanations are not intended to replace interpretation or diagnosis by qualified medical professionals.

4 Results and Discussion

A scalogram of the ECG dataset was successfully generated, and the final results are as follows. This study used 60,000 image datasets out of a total of 110,000. The datasets used were randomly selected, consisting of 30,000 normal ECG datasets and 30,000 abnormal ECG datasets.

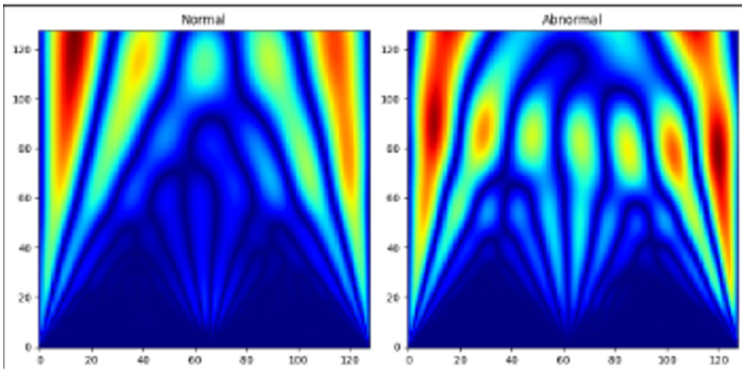


Fig. 1. scalogram image of normal and abnormal signal

The training results of the ResNet18 model obtained accuracy, recall, macro-f1, and specificity of 0.9913333333333333, 0.9913320397386288, 0.9913331069241266, 0.9915966386554622 with an AUC of 0.9994168749514805. For the confusion matrix results, the obtained true positives were 4484, false negatives were 38, false positives were 40, and true negatives were 4438. The Gradcam results are as shown in Fig. 1. During this stage, evidence from Gradcam is also created for each scalogram. This evidence serves as a dictionary for the future LLm.

Table 1. Result of RESNet18 Training

Matrix	Results
Accuracy	0.9913333333333333
Recall	0.9913320397386288
Macro-F1	0.9913331069241266
Specificity	0.9915966386554622
AUC	0.9994168749514805

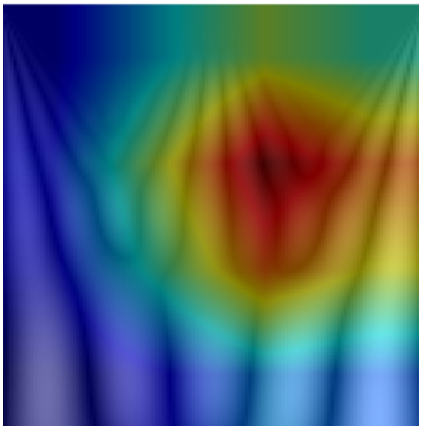


Fig. 2. heatmap of the scalogram from the grad-cam

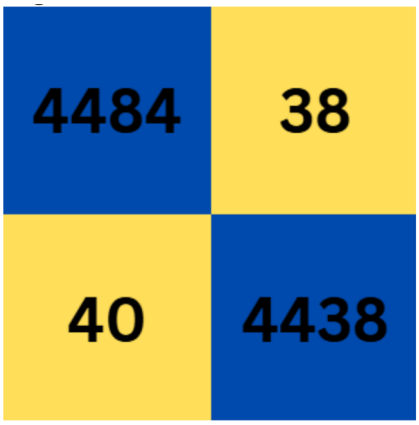


Fig. 3. confusion matrix of the Resnet-18

After obtaining the classification results from ResNet-18 and the corresponding evidence from Grad-CAM, the next stage involved generating textual explanations using a template-based explanation approach. This approach was employed as a baseline explanation method with a deterministic nature, in which the information contained in the generated statements was directly derived from the classification results and Grad-CAM evidence. Thus, the template-based explanation does not introduce new interpretations beyond the information provided by the model, but rather transforms the numerical and categorical information into a more readable textual format.

The information used to construct the template-based explanations included the predicted class, confidence score, dominant temporal region, dominant frequency region, activation ratio, and maximum Grad-CAM activation value. For example, when the model generated an abnormal prediction with a confidence score of 100%, with the dominant activation located in the middle temporal region and medium-frequency region, and an activation ratio of 16.55%, these values were directly incorporated into the explanation template. Through this mechanism, each component of the generated explanation can be traced back to the corresponding values contained in the Grad-CAM evidence.

For Sample 0, the ResNet-18 model generated an abnormal prediction with a confidence score of 100.00%, which was consistent with the ground-truth label. The Grad-CAM results showed an activation ratio of 16.55%, with the maximum activation located in the middle temporal region and medium-frequency region. Based on these results, the template generated an explanation stating that the model classified the ECG scalogram as abnormal with a confidence score of 100.00%, with the strongest Grad-CAM activation located in the middle temporal and medium-frequency regions. The activation ratio of 16.55% indicates that approximately 16.55% of the scalogram area had activation values above the predefined threshold.

For Sample 1, the model generated a normal prediction with a confidence score of approximately 100.00%, which was also consistent with the ground-truth label. The activation ratio for this sample was 23.46%, while the dominant activation region was located in the middle temporal and medium-frequency regions. Meanwhile, for Sample 2, the model generated an abnormal prediction with a confidence score of approximately 100.00%, again producing a prediction consistent with the ground-truth label. However, unlike the two previous samples, the dominant frequency region in this sample was categorized as high. This difference indicates that although several samples may have the same predicted class, the model attention patterns represented by Grad-CAM may vary across individual samples.

The template-based explanation used in this study has distinct characteristics compared with the explanations that are subsequently generated by the LLM. Since the template is constructed directly from the extracted evidence values, every piece of information presented in the explanation can be traced back to the model output and Grad-CAM results. This approach was therefore used as a baseline to characterize the form of explanation that can be obtained

without involving free-form language generation. Furthermore, its deterministic nature allows the template-based explanation to serve as a reference for subsequent evaluation of whether the use of an LLM can generate more informative explanations while maintaining consistency with the underlying model evidence.

It should be noted that the temporal and frequency regions reported in the Grad-CAM results represent positional regions within the scalogram image. The early, middle, and late temporal regions were determined based on the horizontal partitioning of the heatmap, whereas the low, medium, and high frequency regions were determined based on the vertical position of the heatmap. Therefore, these regions are used to describe the relative location of model activation within the scalogram and are not intended to represent specific physiological interpretations or clinical diagnoses. Similarly, the confidence score represents the output probability of the classification model and should not be interpreted as a measure of medical diagnostic certainty.

Thus, the template-based explanation serves as a bridge between visual explainability and the subsequent text explanation stage. The Grad-CAM results, which initially consist of a heatmap and numerical information, are transformed into consistent textual explanations that remain traceable to the underlying evidence. This approach subsequently serves as one of the baselines for comparison with two other approaches, namely LLM without evidence and evidence-grounded LLM. The comparison is intended to determine whether providing structured evidence to the LLM can produce explanations that are more consistent with the classification decisions and visual evidence generated by the CNN model.

The next stage of the study employed a Large Language Model (LLM) to transform the CNN classification results and Grad-CAM evidence into natural language explanations. The evaluation was conducted on 500 samples selected from the test set. Two approaches were considered in this experiment, namely LLM without evidence and evidence-grounded LLM. In the LLM without evidence approach, the model was provided only with the CNN prediction label and confidence score. In contrast, the evidence-grounded LLM approach provided additional information derived from Grad-CAM, including the activation ratio, maximum activation, mean activation, dominant temporal region, dominant frequency region, and hotspot position.

The generated outputs demonstrated distinct characteristics between the two approaches. In the LLM without evidence approach, the model tended to generate explanations based on general knowledge regarding ECG signals. For example, when provided with an abnormal prediction with a very high confidence score, the model could generate explanations concerning potential ECG characteristics, heart rate variations, rhythm, morphology, or recommendations for further examination. However, some of this information was not derived from the evidence available within the system. In several outputs, inconsistencies were also identified between the label provided as input and the interpretation generated by the LLM. These findings indicate that an LLM without evidence has

a tendency to generate explanations based on its general knowledge rather than solely on the rationale underlying the classification model's decision.

In contrast, in the evidence-grounded LLM approach, the information provided to the LLM was derived from the Grad-CAM analysis. For example, in one sample, the CNN generated an abnormal prediction with a confidence score of 1.0000, an activation ratio of 0.1655, a mean activation of 0.4172, and a dominant activation region located in the middle temporal and medium-frequency regions. These values were subsequently provided as the basis for generating the LLM explanation. Through this approach, the generated explanation was more closely associated with the information contained in the available evidence, thereby reducing the extent to which explanation generation depended solely on the general knowledge of the LLM.

Quantitative evaluation across the 500 samples resulted in a Label Agreement score of 1.0000. This result indicates that, for all evaluated samples, the prediction label stated in the LLM-generated explanation was consistent with the CNN prediction label. Thus, the LLM did not alter the primary classification decision when generating the explanation. This finding is important because one of the primary functions of an explanation in an explainable artificial intelligence (XAI) system is to explain the model's decision rather than generate a new classification decision that differs from the underlying model.

The Evidence Coverage score also reached 1.0000. In this study, Evidence Coverage measures whether the primary evidence used in the evaluation is represented in the LLM-generated explanation. The evaluated evidence consisted of the prediction label, dominant temporal region, and dominant frequency region. The obtained score indicates that all of these evidence components were present in the generated explanations across the evaluated samples. This result demonstrates that information derived from Grad-CAM can be propagated from the visual interpretation stage to the natural language explanation stage.

A similar result was obtained for the Evidence Support Rate, which also reached 1.0000. This score indicates that the primary evidence used in the study was either supported or explicitly mentioned in the explanations generated by the LLM. Therefore, the evidence-grounded approach successfully maintained the relationship between the information provided to the LLM and the information expressed in its explanations. This finding demonstrates that Grad-CAM can serve as an evidence source that connects CNN classification decisions with natural language explanations.

Nevertheless, the evaluation results also indicate that the incorporation of evidence does not completely eliminate the possibility of hallucination. The obtained Hallucination Rate was 0.4327, corresponding to approximately 43.27% based on the hallucination criteria defined in this study. This result indicates that although the LLM was able to preserve the provided evidence, it still exhibited a tendency to generate additional information that was not directly supported by the available evidence. For example, the LLM could provide interpretations regarding the presence of specific abnormalities, physiological conditions, or par-

ticular ECG characteristics, even though such information was not contained in the Grad-CAM evidence provided to the model.

These findings demonstrate a distinction between the LLM's ability to preserve evidence and its ability to generate explanations that are fully faithful to that evidence. The high Evidence Coverage and Evidence Support Rate indicate that the provided evidence was successfully incorporated into the generated explanations. However, the Hallucination Rate indicates that the LLM could still introduce information beyond the available evidence. Therefore, the presence of evidence in the prompt does not automatically guarantee that all content in the generated explanation is a direct consequence of the provided evidence.

Overall, the results demonstrate that the integration of CNN, Grad-CAM, and LLM can establish an explainable artificial intelligence pipeline capable of transforming classification results into more comprehensible textual explanations. The CNN produces the classification prediction, Grad-CAM provides information regarding the regions contributing to the model's decision, and the LLM translates this information into natural language. The evidence-grounded approach demonstrated a strong ability to preserve the primary information contained in the evidence, as indicated by the Label Agreement, Evidence Coverage, and Evidence Support Rate scores, all of which reached 1.0000.

However, these results also demonstrate that the system remains limited in its ability to generate explanations that are fully reliable and evidence-faithful. The Hallucination Rate of 0.4327 indicates that the LLM can still generate additional information that is not directly supported by the available evidence. Therefore, in the context of this study, the LLM is more appropriately positioned as a component for translating model evidence into natural language rather than as the primary source for determining the medical rationale underlying an ECG classification. The classification decision remains the responsibility of the CNN, while Grad-CAM provides the visual evidence and the LLM presents this evidence in a more accessible textual form.

5 Conclusion

Based on the findings of this study, ECG signals can be represented as scalograms using the Continuous Wavelet Transform (CWT), allowing signal information in both the time and frequency domains to be utilized as input for image-based classification models. In this study, a total of 60,000 scalograms consisting of normal and abnormal classes were used as the dataset for the classification process. This representation was subsequently provided as input to the ResNet-18 model to classify ECG signals into two classes, namely normal and abnormal.

The classification results obtained using ResNet-18 demonstrate that the scalogram-based approach can be effectively employed to distinguish between normal and abnormal ECG signals. The model was evaluated using Accuracy, Macro-F1, Recall, Specificity, and AUC, allowing its performance to be assessed from multiple perspectives rather than solely based on overall classification ac-

curacy. These classification results provided the basis for the subsequent stage, namely the interpretation of the model's decisions using Grad-CAM.

The application of Grad-CAM enabled the decisions of ResNet-18 to be represented not only through the predicted class and confidence score, but also through information regarding the regions of the scalogram exhibiting the strongest activation contributions to the model's decision. The resulting evidence was summarized in terms of activation ratio, maximum activation, mean activation, dominant temporal region, dominant frequency region, and hotspot coordinates. Thus, the CNN classification results could be complemented with information regarding the regions of the ECG representation that received the greatest attention from the model.

Subsequently, the evidence obtained from Grad-CAM was provided as input to an LLM to generate explanations in natural language. Evaluation on 500 samples demonstrated that the evidence-grounded LLM achieved a Label Agreement score of 1.0000, an Evidence Coverage score of 1.0000, and an Evidence Support Rate of 1.0000. These results indicate that the primary information derived from the CNN predictions and Grad-CAM evidence could be preserved during the LLM-based explanation generation process. Therefore, the integration of Grad-CAM and LLM can facilitate the transformation of numerical and visual model interpretation information into explanations that are more accessible and comprehensible.

Nevertheless, the study also identified limitations associated with the use of LLMs. The evidence-grounded LLM still produced a Hallucination Rate of 0.4327 based on the evaluation criteria employed in this study. This finding indicates that although the provided evidence could be preserved within the generated explanations, the LLM still exhibited a tendency to introduce information that was not directly supported by the available evidence. Therefore, within the context of this study, the LLM is more appropriately positioned as an evidence-based explanation generation component rather than as a source of clinical ground truth or a substitute for the underlying classification model.

Overall, this study developed an explainable artificial intelligence pipeline that integrates scalogram representation, CNN-based classification using ResNet-18, Grad-CAM-based interpretation, and LLM-based explanation generation. The proposed pipeline enables the ECG classification process to extend beyond the assignment of normal or abnormal labels by providing information regarding the regions receiving the greatest attention from the model and translating this information into natural language explanations. The findings also demonstrate that improving evidence grounding should be accompanied by effective hallucination control mechanisms to enable the resulting explainable AI system to provide more reliable and trustworthy explanations.

References

1. Haupt, M., Maurer, M.H., Thomas, R.P.: Explainable Artificial Intelligence in Radiological Cardiovascular Imaging-A Systematic Review. Diagnostics **15**(1), 100 (2025)

2. Manimaran, G., et al.: Explainable deep learning based techniques for ECG-Based heart disease classification: A systematic literature review and future direction. Computers in Biology and Medicine **180**, 108900 (2025)
3. Kumar, T.G.K., Ogare, M.K., Koolagudi, S.G.: ECG Signal Classification using Continuous Wavelet Transform Scalogram and Convolutional Neural Network. In: 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), pp. 1–6. IEEE (2024)
4. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision **128**(2), 336–359 (2016)
5. Ravi, V., Sowmya, V., Gopalakrishnan, E.A., Soman, K.P.: Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram. IEEE Transactions on Engineering Management **70**(2), 521–533 (2023). https://doi.org/10.1109/TEM.2021.3104751
6. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE (2018). https://doi.org/10.1109/WACV.2018.00097
7. Talukder, M., Talaat, A.S., Muna, N.J., Alazab, A., Kazi, M., Das, U.K.: An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals. Scientific Reports **15**(1), 1234 (2025)
8. S, A., Sanjay, M.: ECG Classification and Arrhythmia Detection Using Wavelet Transform and Convolutional Neural Network. In: 2021 International Conference on Communication, Control and Information Sciences (ICCISc), pp. 1–6. IEEE (2021)
9. Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C., Ou, C.: Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network. Entropy **23**(1), 119 (2021)
10. Goettling, M., Hammer, A., Malberg, H., Schmidt, M.: xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. Scientific Reports **14**(1), 8901 (2024)
11. Desai, S., Ramaswamy, H.G.: Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 972–980. IEEE (2020). https://doi.org/10.1109/WACV45572.2020.9093360
12. Wagner, P., Mehari, T., Haverkamp, W., Strodthoff, N.: Explaining Deep Learning for ECG Analysis: Building Blocks for Auditing and Knowledge Discovery. Computers in Biology and Medicine **155**, 106624 (2023)
13. Giriprasad, P., Sanjeeva Reddy, G.A., Sreehari, R.V., M.C.E, K.S.R.: Automatic Classification of Cardiac Arrhythmias based on ECG Signals Using Transferred Deep Learning Convolution Neural Network. Journal of Physics: Conference Series **1911**(1), 012015 (2021)
14. Fayyaz, A., Abdulkadir, S., Talpur, N., Al-Selwi, S.M., Hassan, S.U., Sumiea, E.H.H.: Grad-CAM (Gradient-weighted Class Activation Mapping): A systematic literature review. Computers in Biology and Medicine **198**(Pt B), 111200 (2025). https://doi.org/10.1016/j.compbiomed.2025.111200
15. Ennab, M., Mcheick, H.: Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models. Machine Learning and Knowledge Extraction **7**(1), 12 (2025). https://doi.org/10.3390/make7010012

16. Showrav, T.T., Lincoln, S.I., Hasan, M.K.: An explainable artificial intelligence framework for trustworthy arrhythmia detection from heart rhythm data. Healthcare Analytics **5**, 100455 (2026). https://doi.org/10.1016/j.health.2026.100455
17. Zhang, W.: 2D Transfer Learning for ECG Classification using Continuous Wavelet Transform. bioRxiv preprint (2024). https://doi.org/10.1101/2024.07.11.24310258
18. Soomro, S., Niaz, A., Choi, K.N.: Grad++ScoreCAM: Enhancing Visual Explanations of Deep Convolutional Networks Using Incremented Gradient and Score-Weighted Methods. IEEE Access **12**, 61104–61112 (2024). https://doi.org/10.1109/ACCESS.2024.3392853
19. Enitan, S.O.A., et al.: Enhancing Explainable AI in Medical Imaging: A Comprehensive Review of Techniques, Clinical Integration, and Future Directions. Journal of Medical Systems **49**(1), 20 (2025)
20. Seniaray, S., et al.: Cardiovascular Disease Prediction Through Retinal Image Analysis using Deep Learning. In: Proceedings of the 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), pp. 101–107 (2025)
21. Jin, W., et al.: Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. MethodsX **10**, 102100 (2023)
22. Charachon, M., et al.: Combining Similarity and Adversarial Learning to Generate Visual Explanation: Application to Medical Image Classification. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 45–49. IEEE (2021)
23. Kumar, T.G.K., et al.: ECG Signal Classification using Continuous Wavelet Transform Scalogram and Convolutional Neural Network. In: 2024 IEEE International Conference on Signal Processing (IConSCEPT), pp. 1–6. IEEE (2024)
24. Dravid, A., et al.: medXGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2938–2946 (2022)
25. Valerio, A.G., et al.: From segmentation to explanation: Generating textual reports from MRI with LLMs. Computer Methods and Programs in Biomedicine **270**, 108300 (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

