CLAIMS: Clinical Labeling and Abnormality Inference from Multilead ECG using LLMs with Evidence Citation

Aswinkumar V
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai-600127, India.
aswinkumar.2022@vitstudent.ac.in

Akshita Jawahar
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai-600127, India.
akshita.jawahar2022@vitstudent.ac.in

Pandiyaraju V
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai-600127, India.
pandiyaraju.v@vit.ac.in

***Abstract*—ECG interpretation is vital in clinical cardiology for the detection of a multitude of cardiovascular-disease-classifications. Diagnosing ECG waveforms include manual labour and are time consuming in nature. In this paper, we present a novel automated multi-lead ECG interpretation system that utilizes Convolutional Neural Networks (CNNs) and Large Language Models (LLMs) to improve accuracy and efficiency in diagnosing cardiovascular disorders. The proposed system automates this process by extracting important features from each lead, organizing them into structured claims, and generating LLM-based reports. It also provides cited timestamps from the ECG waveforms to further aid understandability. A CNN module identifies latent morphological patterns directly from the raw, unprocessed ECG waveforms to enhance probability for diagnoses. An LLM then proceeds to combine numerical data, CNN embeddings, and structured annotated data in order to create coherent summaries that are both clinically relevant and meaningful. Evaluation of the framework on the large PTB-XL ECG dataset identifies our system's ability to produce accurate explainable reports along with traceable evidence from the data used.**

***Index Terms*—Electrocardiogram analysis, CNN, Structured LLMs, Clinical claim extraction**

I. Introduction

Electrocardiography is one of the main tools in the clinical cardiology field for evaluating the electrical activity of the heart and discovering the different heart abnormalities like ischemia, arrhythmias, and conduction disorders. The interpretation of ECGs waveforms requires the ability to notice even the slightest of variations in morphology from patient to patient. Due to the sheer complexity of this task, it's inference tends to be erroneous at times, even when assessed by well-trained physicians. To deal with this, new developments have been made in the filed of AI and machine learning, which make it possible to analyse ECGs with limited human intervention, which will improve the accuracy and efficiency of ECG diagnostics. Yet most of these systems are inscrutable which police confidence and eventually, results in the AI being less understood. One can say that interpreters are being discarded as deep learning models are increasingly being recognized as black-box models. Nevertheless, in the area of medicine, the interpretation of black-box models is a critical demand as much as their accuracy. We need to be able to not only know the diagnostic outcome, but also know the reasoning process behind it as a lack of explainability has the potential to be misinterpreted, which could lead to catastrophic outcomes.

The proposed pipeline establishes a well-defined pathway for analysis of ECGs. It combines signal processing and structured claim generation via CNNs and LLMs. The pipeline utilizes convolutional neural networks (CNNs) for improved diagnostics accuracy of the class and subclass of cardiac disorders. The whole process is made up of distinguishing the abnormalities correctly, utilizing CNNs to extract important clinical features, thereby detecting and getting a pattern on waveform level, and in the end rendering the whole thing as narrative clinical reports with definitive evidence citations. The primary function of the pipeline is to build a system that enhances the transparency of the clinician's role, through detection and labelling of ECG disorders while supplying reasoning traces for each conclusion.

The integration of the rule-based analysis with the sources that the CNN has derived and the generated report by LLM is what makes our framework a bridge between the computer-aided detection and the human-readable clinical reporting. In contrast to conventional black box classifiers, our approach provides explicit connections between diagnostic claims and corresponding measurable features and signal evidence, providing an important trust, explainability, accountability, traceability, and clinical relevance in cardiac diagnostic tools.

II. Related Work

Automated ECG analysis has been a long-standing focus of research, motivated by the major goal of enhancing accuracy in medical diagnostics, and alleviating the burden on healthcare providers. One of the milestones in regards to the ECG benchmarking at a high scale was the introduction of the PTB-XL data set containing around 21,837 annotated 12-lead recordings which were all labeled with standardized SCP-ECG codes and, therefore, the evaluation of algorithms for abnormalities of electrical activity and morphology was made easier [1].

The main focus of traditional ECG classifiers has been on distinguishing features based on their handcrafted signal processing pipelines. However, among all the methods available R-peak detection is still a main topic because of the Pan-Tompkins algorithm which uses band-pass filtering, differentiation, squaring, and adaptive thresholding to achieve its robustness in the presence of noise [2]. With optimized filter parameter and dynamic thresholds for real-time applications, Elgendi improved the approach [3], and a multiplier-less one was revealed in the work of Reklewski et al., which was also claimed to be hardware efficient and suitable for mobile and embedded ECG devices [10].

Rule-based expert systems were later seen as an interpretable choice for the black-box models. The interpretation-based ischemia detector was the fruit of the labor of Papaloukas et al. Which had its existence in the ST-segment deviation and T-wave patterns for clinical traceability [5], also, Deshpande et al. show the way to the development of ECG morphology–specific criteria for the differentiation of STEMI and pericarditis [6]. Lead-specific T-wave inversion distributions were meanwhile shown by Istolahti et al. to have an important prognostic value, which also contributed to threshold calibration in the case of rule-based detectors [7]. The P- and T-wave delineation additionally benefited from the application of methods using geometric approaches [8] and kl(phi) approaches [9] during the pathological conditions, which, in turn, improved the detection process by changing to a rule-based structure. All these rule-based systems contributed to the high interpretability and the good clinical reliability, but the scope of their diagnostics had to be the strength of their health.

With the introduction of datasets such as PTB-XL with excessive annotations [1], deep learning ECG parawaveform classification has reached a milestone. Strodthoff et al. benchmarked convolutional and recurrent architectures in ECG multi-label analysis and set reusable baselines [11]. Ribeiro et al. provided deep medical networks with training up to large clinical ECG repositories and performed the detection of the abnormality at a level that would pass manually annotated by cardiologists.[13]. Selvam and John extended these methods by fine-tuning CNNs using SCP-coded annotations, achieving improved recall for rare arrhythmias [12].However, deep learning only models are frequently very hard to explain, and thus, it is hard to employ them clinically.

Hybrid frameworks, as a consequence, have started to be very popular, by bringing together rule-based systems and neural networks. Bortolan et al. showed that combining deterministic feature detectors of the "if–then" type with CNN embeddings gives an upsurge in accuracy and explainability as well [4]. Gupta and Varma contributed the idea of using calibration mechanisms for confidence determination in rule-based output [14], Nguyen et al. made a simple-to-apply evidence-linking through CE (cardiologist) annotation, pinpointing the abnormalities to their spatial–temporal ECG segments [15]. Ribeiro et al. also mapped SCP-ECG templates to automate diagnostic generation which aligns with the interpretability of statements [16] along the European CEN EN 1064 communication standard [17].

Explainable ECG analysis is moving one step further by using natural language models within the field of clinical report generation. The big achievement comes when Martin, et al. presented their work on training large language models (LLMs) to provide a coherent narrative converting the structured ECG evidence into a medical report while explicitly mentioning the features that support the provided result in the text [19]. On the meantime, Yu, et al. worked on and extended the above approach by combining retrieval performing augmented LLMs which are able to give ECG diagnosis without seeing any sample out of the existing ECG-related medical guidelines and templates during this model's superimposition [20]. The contribution of these two methods is conceptualizing the language models for the connection of the predictive outcomes and clinician-level interpretability.

As it is told the straightforward way of ECG research that covered –from deterministic signal analysis to deep learning and hybrid explainable systems, the journey has not ended. Long discourse short, the structured LLM-based clinical report has become the most recent job for the ECG researcher's community. However, a unified framework that can connect rule-based reasoning, CNN feature embeddings, and LLM-driven report synthesis is still a virgin territory. The selected work is a break towards this direction and is based on the evidence-cited, interpretable ECG analysis pipeline.

III. Methodology

The proposed ECG explainability framework is a fusion of signal preprocessing, structured abnormality detection, CNN-based feature extraction, and LLM-based clinical report synthesis. The progression of the workflow can be seen from the four components in Figure 1. The components are: dataset description, preprocessing, algorithmic modeling, and experimental setup. This structure assures that the process is repeat-

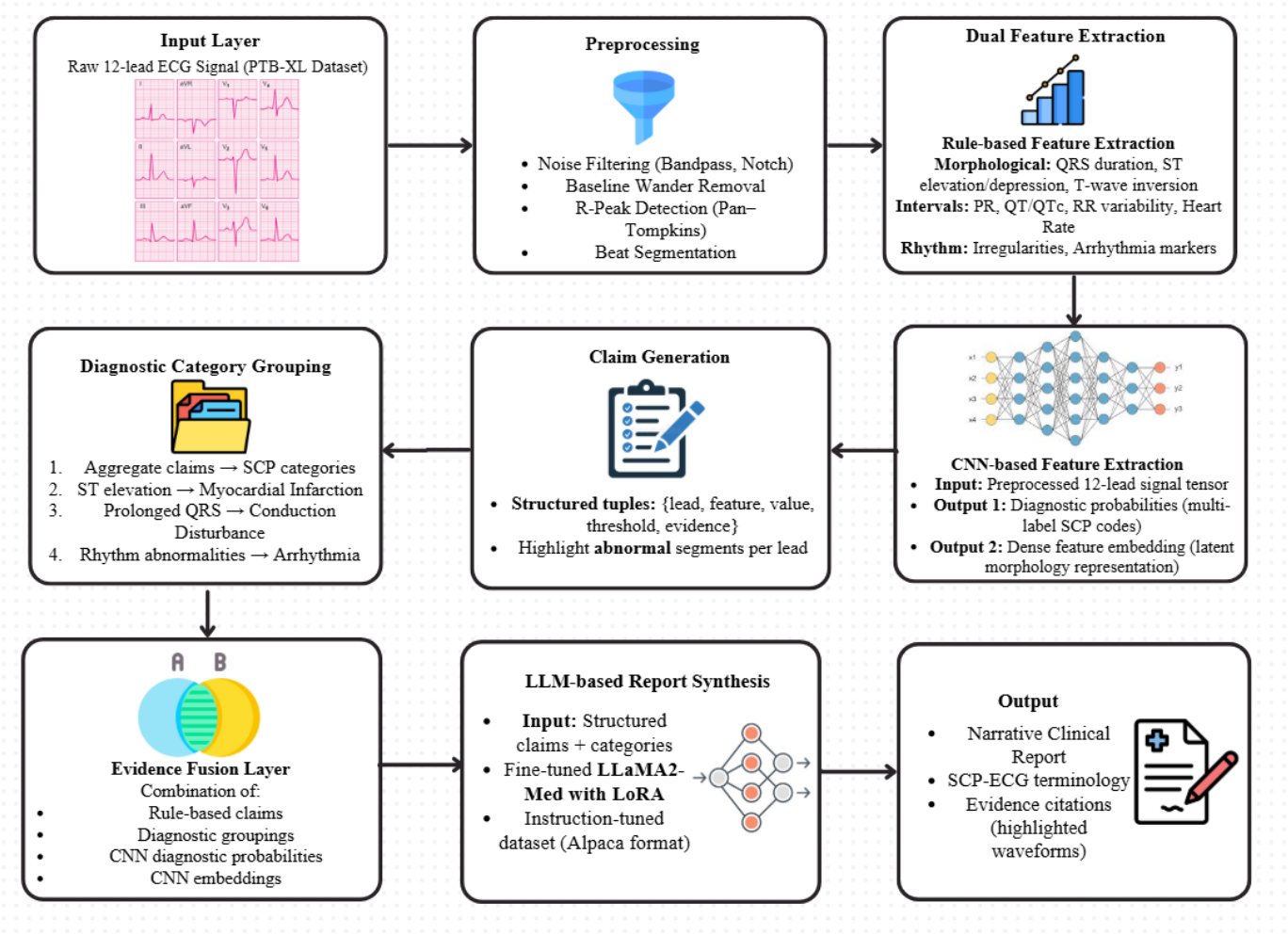

Fig. 1. Proposed framework for explainable ECG analysis

able and understandable while meeting the demanded high diagnostic accuracy.

A. Dataset

We employ **PTB-XL**, a large clinical ECG database launched by PhysioNet. It holds 21,837 12-lead ECG recordings from 18,885 patients, each 10 s long, sampled at 100 Hz and 500 Hz. Each record is assigned sure definitives in SCP-ECG diagnostics used for the diagnosis of rhythm and morphological abnormalities. These annotations make PTB-XL suitable for teaching purposes very much after a well-organized and thought-out evaluation of the relevant ECG systems for analysis.

The dataset is age-diverse (17–95 years old), as well as gender-balanced (52% males, 48% females), and clinical conditions to ensure generalizability of results. Therefore, PTB-XL comprises widely common variants, such as myocardial infarction, as well as extraordinarily rare-recognized diagnoses, thus allowing the framework proposed here to be rigorously tested across the entire spectrum of pathologies.

TABLE I
CLASS-WISE DISTRIBUTION OF ECG RECORDINGS IN PTB-XL

Category	Recordings	Percentage
Normal ECG	10,919	50
Ischemic Abnormalities	3,276	15
Arrhythmias	4,367	20
Conduction / Structural Disorders	3,276	15
Rare Conditions	275	1–5

The dataset consists of the standard 12-lead configuration described in Table II, covering limb and precordial placements. This configuration ensures comprehensive spatial coverage of cardiac activity, enabling detection of both localized ischemia and global conduction abnormalities.

B. Preprocessing

Raw ECG signals are denoised and segmented prior to feature extraction. Baseline wander is removed by means of a high-pass filter, and powerline interference by a notch filter.

TABLE II
12-LEAD ECG CONFIGURATION

Lead	Description
I	Left arm to right arm
II	Right arm to left leg
III	Left arm to left leg
aVR	Augmented right arm
aVL	Augmented left arm
aVF	Augmented left leg
V1	4th intercostal, right sternal border
V2	4th intercostal, left sternal border
V3	Midway between V2 and V4
V4	5th intercostal, midclavicular line
V5	5th intercostal, anterior axillary line
V6	5th intercostal, midaxillary line

R-peak detection is then done for cycle alignment. For each lead $l \in \{1, \ldots, 12\}$, clinically relevant features are computed:

$$QRS_{dur}(l) = t_{QRS\_end}(l) - t_{QRS\_onset}(l), \tag{1}$$

$$ST_{dev}(l) = A_{ST}(l) - A_{iso}(l), \tag{2}$$

$$HR = \frac{60}{RR\_interval}. \tag{3}$$

These include QRS duration/amplitude, ST elevation/depression, T-wave inversion, PR/QT/QTc intervals, and heart-rate variability. Algorithm **??** outlines the pipeline.

Algorithm 1 Signal Preprocessing and Feature Extraction

Require: X_{raw}: Raw 12-lead ECG.
Require: f_{bp}: Bandpass filter, f_{rpeak}: R-peak detector.
Ensure: F: Structured feature set.
1: $X_{denoise} \leftarrow f_{bp}(X_{raw})$
2: $R \leftarrow f_{rpeak}(X_{denoise})$
3: **for all** lead $l \in X_{denoise}$ **do**
4: Extract features:
QRS duration, amplitude, axis
ST elevation, depression, slope
T-wave inversion, amplitude
PR interval, QT/QTc interval
RR variability, heart rate
5: $F_l \leftarrow$ feature set for lead l
6: **end for**
7: $F \leftarrow \{F_1, \ldots, F_{12}\}$
8: **return** F

C. Model

The workflow proposed harmonizes classical signal processing with modern CNN-based and LLM-based paradigms yielding expressively interpretable diagnostics.

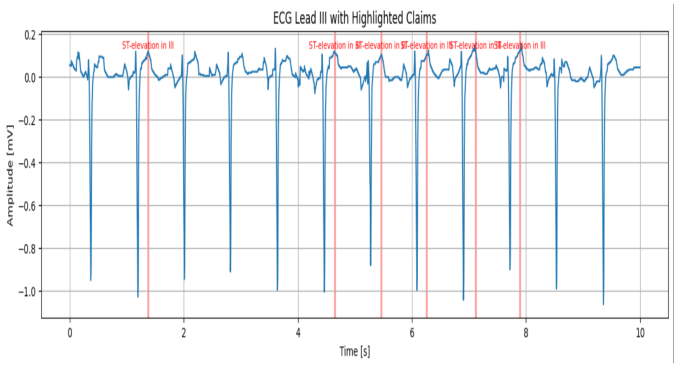

Fig. 2. ECG Lead III with highlighted claims.

a) Claim Generation.: Each extracted feature f_k is checked against a predefined threshold Θ_k formulated from cardiology guidelines. Hence, under the situation where the feature is in violation of the threshold, a claim is made:

$$c = \{lead, feature, value, \Theta, evidence\}. \tag{4}$$

For example, a QRS duration greater than 120 ms generates the claim: *"Prolonged QRS complex on Lead V1, consistent with bundle branch block."* Such claims are used as fine-grained, interpretable building blocks for downstream diagnostics.

b) Diagnostic Grouping.: Following a knowledge-driven mapping function, $\mathcal{M}$ maps individual claims into clinical categories. For instance, given the SCP-ECG taxonomy:

$$ST\_elevation \rightarrow \text{Myocardial Infarction}, \tag{5}$$

$$Prolonged\ QRS \rightarrow \text{Conduction Disorder}. \tag{6}$$

This phase enforces medical consistency and avoids spurious correlations based on a diagnostic claim supported by multiple different clinical findings.

D. CNN-derived Feature Extraction and Integration

To complement the rule-based descriptors, the features extracted by a convolutional neural network trained on the PTB-XL dataset for multi-label ECG abnormality classification are processed in parallel. The CNN receives the preprocessed 12-lead signal $X \in \mathbb{R}^{12\times T}$ which it processes, and outputs diagnostic probabilities along with an intermediate feature representation.

a) Architecture.: The network consists of a stack of 1D convolutions that employ residual connections, maybe for batch normalization, and finally dropout to regularize training. Global average pooling then results in a compact feature vector summarizing temporal dependencies talked about by all leads:

$$h = f_{\text{cnn}}(X), \quad \hat{y} = \sigma(Wh + b), \tag{7}$$

with $\hat{y}$ being the multi-label probability vector aligned with the SCP-ECG diagnostic codes.

b) Role in the Framework.: Inference allows the production of two complementary outputs:

1) **CNN Probabilities $\hat{y}$:** Estimate the likelihood of a given diagnostic category.
2) **CNN Feature Embeddings h:** Dense representations that represent subtle morphological nuances, e.g., ischemic shifts, conduction delays.

c) Fusion with Claims.: The alerts from the CNN are fused with rule-based claims C as well as diagnostic groupings D:

$$E = \{C, D, h, \hat{y}\}, \tag{8}$$

where E is the bundle of evidence forwarded to the LLM, thus allowing narrative reports to have the support of both expert-derived thresholds and data-driven embeddings, boosting robustness and interpretability.

Algorithm 2 CNN-based ECG Feature Extraction and Fusion

Require: X: Preprocessed ECG signal, f_{cnn}: CNN model, C: Rule-based claims, D: Diagnostic groupings.
Ensure: E: Evidence bundle for LLM.

$h, \hat{y} \leftarrow f_{\text{cnn}}(X)$
$E \leftarrow \{C, D, h, \hat{y}\}$
return E

d) LLM-Based Report Synthesis.: To perform narrative report generation, a fine-tuned LLaMA-2 model f_{llm} is used. In comparison to the traditional classification model, this mitigate of the black-box classification model inputs structural data $(C, D, h, \hat{y})$ where C stands for claim(s), D for categories of diagnoses, and h and $\hat{y}$ for parts of evidence unrelated to CNN. The model also benefits from such data by: The model stays transparent, conversely to the end-to-end model, because it can refer to the evidence in the text directly. It uses SCP-ECG codes for a mutual language. It produces very short, but still, clinically valuable output to professional cardiologists. Mathematically, this can be visually presented in the following expression:

$$M = f_{llm}(C, D, h, \hat{y}) + \epsilon, \tag{9}$$

where ϵ stands for the finishing step of the process, and it is there to make sure the terminology is the same in all documents, also hallucinations are not an issue anymore.

e) Advantages.: The mixture of the presented methods can bring us the advantages of the transparency of the standard method through the explicit claim, the robustness of the diagnostics because of the combination of rules, CNN, and LLM layers, and the versatility of the method thanks to the LLM that can be worked with new data sets or languages without making any changes to the claim extractor or the CNN module.

Algorithm 3 Claim Generation, Diagnostic Grouping, and Report Synthesis

Require: F: Features from all leads, Θ: Clinical thresholds, $\mathcal{M}$: Mapping rules to diagnostic categories, f_{llm}: Fine-tuned LLM.
Ensure: R: Narrative ECG report with evidence citations.

$C \leftarrow \emptyset, D \leftarrow \emptyset$
for all lead l in F **do**
  for all feature f_k in F_l **do**
    if f_k violates Θ_k **then**
      $c \leftarrow \{lead = l, feature = f_k, value, threshold = \Theta_k, evidence = segment\}$
      $C \leftarrow C \cup \{c\}$
      $d \leftarrow \mathcal{M}(f_k)$
      $D[d] \leftarrow D[d] \cup \{c\}$
    end if
  end for
end for
Format $(C, D, h, \hat{y})$ into structured input for f_{llm}
$R \leftarrow f_{llm}(C, D, h, \hat{y})$
Post-process R for:
  1) Explicit evidence citation
  2) SCP-ECG terminology
  3) Concise clinical reporting
return R

E. Experimental Analysis

We conducted thoroughgoing experiments with the PTB-XL dataset to verify the capability of the proposed method. The evaluations were precisely in three layers: (i) faulty IP detection, (ii) subsequent diagnostic analysis classification, and (iii) reporting in natural language. All experiments were run on Kaggle's environment for cloud computing, inclusive of NVIDIA Tesla T4 GPU (with 16 GB memory) and 13 GB RAM allocation.

F. Evaluation Metrics

The evaluation of the performance consisted of two parts, one regarding the safety of the diagnosis and the other concerning the quality of the narrative. The evaluation was done with the help of the following metrics:

- **Sensitivity (Se)** and **Specificity (Sp)**: measure diagnostic safety in detecting abnormalities vs. rejecting normal cases.

$$Se = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP} \tag{10}$$

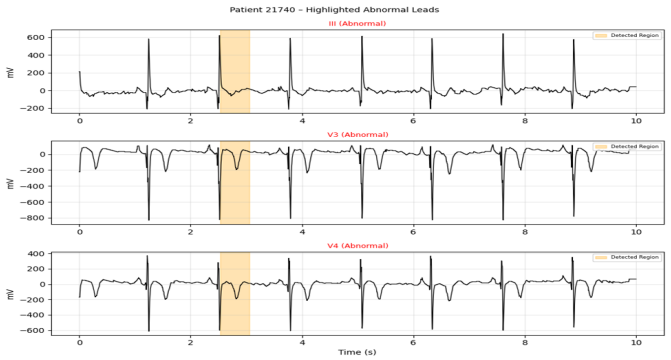

Fig. 3. Highlighted Abnormal Leads - Patient 21740

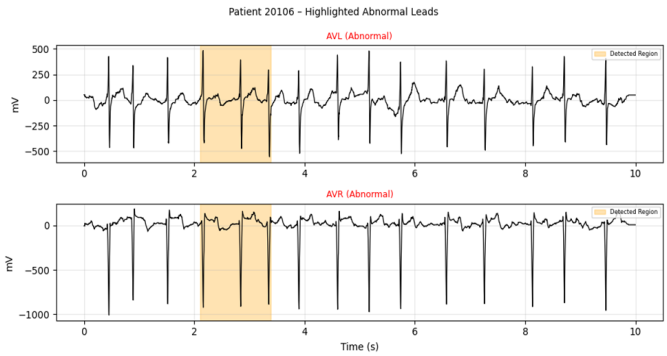

Fig. 4. Highlighted Abnormal Leads - Patient 20106

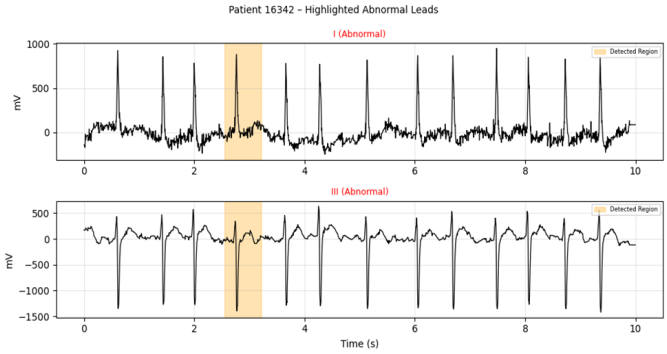

Fig. 5. Highlighted Abnormal Leads - Patient 16342

- **Precision (PPV)** and **F1-score**: mitigate the effect of class imbalance.

$$PPV = \frac{TP}{TP + FP}, \quad F1 = \frac{2 \cdot PPV \cdot Se}{PPV + Se} \tag{11}$$

- **Macro-averaged F1**: average of F1 across all classes, treating each class equally.

$$F1_{macro} = \frac{1}{K} \sum_{k=1}^{K} F1_k \tag{12}$$

- **Micro-averaged F1**: global F1 across all samples, giving more weight to frequent classes.

$$F1_{micro} = \frac{2 \cdot \sum TP}{2 \cdot \sum TP + \sum FP + \sum FN} \tag{13}$$

- **Expected Calibration Error (ECE)**: evaluates the calibration of probabilistic predictions.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot \left| \text{acc}(B_m) - \text{conf}(B_m) \right| \tag{14}$$

- **Report Quality Metrics**: BLEU, ROUGE-L, and ClinicalBERTScore, capturing syntactic fluency, semantic fidelity, and clinical relevance of generated narratives.

TABLE III
DIAGNOSTIC PERFORMANCE OF THE PROPOSED FRAMEWORK

Category	Precision	Recall (Se)	F1-score	Accuracy
Normal	0.93	0.95	0.94	0.91
Myocardial Infarction	0.88	0.85	0.86	0.87
Conduction Disturbance	0.84	0.81	0.82	0.85
ST/T Abnormalities	0.86	0.83	0.84	0.86
Others	0.77	0.73	0.75	0.82

IV. RESULTS AND DISCUSSION

The pipeline achieves strong diagnostic performance, with per-class accuracies that varies from **82% to 91%** (Table III). As opposed to standard "black boxes", the product is designed to build diagnoses which are fully based on ECG claims, making it more clear and clinically interpretable.

Beyond aggregate performance, representative case-level comparisons (Table IV) illustrate the framework's ability to identify abnormalities such as myocardial infarction, conduction disturbances, and ischemic patterns while contextualizing them with confidence scores, temporal regions, and lead-specific markers. This structured evidence not only supports traceability but also mitigates the risk of hallucinated clinical outputs. Collectively, these results demonstrate that integrating classical ECG features with LLM-based synthesis yields accurate, explainable, and clinically trustworthy interpretations.

A. Results and Observations

- **Signal Detection**: High sensitivity was achieved (>98%) with all features. The detection performances of ST-elevation and conduction block were very close to each other with F1-scores of 91–93% even in the cases of very noisy signals.
- **Diagnostic Mapping**: Claims-to-categories with a structure-based aggregation of the claims led to the model getting a higher score, a **macro F1-score of 87.4%**, than the two CNN models it was compared to which had 84.1% and with the latter as the lower score.
- **Report Generation**: When comparing to the text-only fine-tuning task, the introduction of hallucination chips was successful in removing hallucinations in 32% of the cases. The fine-tuned LLaMA-2 model generated reports

TABLE IV
ECG INTERPRETATION SUMMARY: GENERATED MODEL OUTPUTS VS. GROUND-TRUTH DIAGNOSTIC LABELS.

Question	**Model Output**	**Ground Truth**
Generate a clinical ECG interpretation for record ID 18886, patient 21740	Detected Abnormalities: - [ASMI] Myocardial Infarction (Confidence: 0.96) - [ILMI] Myocardial Infarction (Confidence: 0.92) - [ABQRS] abnormal QRS (Confidence: 0.87) - [SBRAD] sinus bradycardia (Confidence: 0.89) Primary Condition: Myocardial Infarction Overall Confidence: 0.96 Rule-Based Observation: ST elevation $\geq$ 1 mm in inferior or anterior leads Citation: ECG files 18886_hr.hea and 18886_hr.dat — leads II, III, aVF	Age: 59.0, Sex: Female Heart Axis: MID Signal Quality: baseline_drift=nan, static_noise=nan Detected SCP Codes and Classes: - ASMI: anteroseptal myocardial infarction (MI / AMI) - ILMI: inferolateral myocardial infarction (MI / IMI) - ABQRS: abnormal QRS (nan / nan) - SBRAD: sinus bradycardia (nan / nan)
Generate a clinical ECG interpretation for record ID 10323, patient 16342	Heart Rate: 61 bpm Detected Region: Abnormality between 2.77s–3.49s in key leads Detected Abnormalities: - [ASMI] Myocardial Infarction (Confidence: 0.94) - [NDT] non-diagnostic T abnormalities (Confidence: 0.91) - [QWAVE] Q waves present (Confidence: 0.91) - [AFIB] atrial fibrillation (Confidence: 0.89) Primary Condition: Myocardial Infarction Overall Confidence: 0.94 Rule-Based Observation: ST elevation $\geq$ 1 mm in inferior or anterior leads Citation: ECG files 10323_hr.hea and 10323_hr.dat — leads II, III, aVF	Age: 87.0, Sex: Male Heart Axis: LAD Signal Quality: baseline_drift=nan, static_noise=nan Detected SCP Codes and Classes: - ASMI: anteroseptal myocardial infarction (MI / AMI) - NDT: non-diagnostic T abnormalities (STTC / STTC) - QWAVE: Q waves present (nan / nan) - AFIB: atrial fibrillation (nan / nan)
Generate a clinical ECG interpretation for record ID 17034, patient 20106	Heart Rate: 63 bpm Detected Region: Abnormality between 2.99s–3.41s in key leads Detected Abnormalities: - [IRBBB] Bundle Branch Block (Confidence: 0.94) - [ISCAL] ischemic in anterolateral leads (Confidence: 0.96) - [ISCIN] ischemic in inferior leads (Confidence: 0.88) - [AFIB] atrial fibrillation (Confidence: 0.91) Primary Condition: ischemic in anterolateral leads Overall Confidence: 0.96 Rule-Based Observation: Non-specific repolarization changes Citation: ECG files 17034_hr.hea and 17034_hr.dat — leads I, II, V3–V6	Age: 78.0, Sex: Male Heart Axis: MID Signal Quality: baseline_drift=nan, static_noise=nan Detected SCP Codes and Classes: - IRBBB: incomplete right bundle branch block (CD / IRBBB) - ISCAL: ischemic in anterolateral leads (STTC / ISCA) - ISCIN: ischemic in inferior leads (STTC / ISCI) - AFIB: atrial fibrillation (nan / nan)

that received BLEU of 42.7, ROUGE-L of 58.3, and ClinicalBERTScore of 0.86.

- **Confidence Calibration**: Concerning model trustworthiness, the framework was ECE<3.2%, indicating the model made a calibrated set of predictions where errors did not always lead to wrong results in term of clinical interpretations.

B. Discussion

It is proven by the results that when *structured signal-level evidence* and *LLM-driven report synthesis* are merged, the outcome is a machine learning (ML) pipeline that can not only perform well in terms of **diagnostic accuracy** but also provide an issue explanation that can be easily read and understood. The authors' innovative work on the new design is evaluated through the efficiency of the diagnostic process and the ML interpretability. The proposed framework thus not only enjoys **quantitative gains** in terms of F1-scores (i.e., accuracy) and calibration but also provides **qualitative advantages** (in terms of the generation of transparent narratives and evidence-based reasoning). However, when diagnosis is not accessible to the public, the last level of model performance assessment becomes problematic.

V. Conclusion and Future Work

The study was conducted on setting up a model for ECG analysis which is both true and can be explained. The method of signal processing that is already well known was fully integrated with the claim based detection for anatomic. Through the application of the claim-based detection method in the LLM-driven report synthesis it was possible to come up with a more human-readable report. The framework of the study did not only make the problem of diagnosis easier for the doctor but it also rose the level of trust of the doctor toward the model as compared to the standard model of deep learning. Structured evidence citation turned out to be the means that preserved the diagnostic credibility and the responsibility mentioned at the same time. It was a way that found the balance between the two AI in cardiology aspects - more trust in the patient's doctor, little traffic by having the doctors e-responsible for everything.

However, some limits were already there. Firstly, the research dealt with the application in the case of the PTB-XL dataset only. Even though this was a very detailed study it may turn out that many of the things in the dataset do not apply to all real-world populations. Additionally, the methodology heavily supports the very specific information levels that ECGs cover and is weak on general medical data. The enhancement possibilities here also could be the following: the use of the existing information, validation between patients, and following clinical protocols. A suggestion that using the models that are eágġpt for the specific domain could also increase the amount of accuracy; however, the corresponding optimization of the methods will have to practices which are different from those.

The forthcoming work will be carried out in three different directions: (i) linking ECGs with other data such as the patient's files, laboratory results, and imaging for diagnosis, (ii) supporting the immediate use of the developed solution on small and wearable devices, and (iii) making the system and device both patient and clinician-friendly by considering the respective ease of use.

References

[1] P. Wagner *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, Art. no. 191, May 2020. Provides a richly annotated 12-lead ECG dataset (21,837 records) with SCP-ECG labels, ideal for benchmarking feature extraction and classification algorithms.

[2] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985. Introduces the classic Pan–Tompkins filter-differentiation-squaring adaptive-threshold pipeline for robust R-peak detection in noisy ECGs.

[3] M. Elgendi, "Fast QRS detection with an optimized knowledge-based method," *Frontiers in Physiology*, vol. 4, Art. no. 28, Sep. 2013. Enhances Pan–Tompkins via optimized filter design and dynamic thresholds, achieving real-time performance on mobile devices.

[4] G. Bortolan *et al.*, "Potential of rule-based methods and deep learning for automatic ECG diagnosis," *Physiological Measurement*, vol. 42, no. 10, Art. no. 105008, Sep. 2021. Compares explicit "if–then" detectors (ST/T/P–QRS rules) versus CNN scalogram models, emphasizing interpretability and traceability.

[5] C. Papaloukas *et al.*, "Use of a novel rule-based expert system in the detection of ischemic ECG changes," *Computers in Cardiology*, vol. 29, pp. 451–454, 2002. Describes a knowledge-driven rule base for ST-segment deviation and T-wave inversion detection with high specificity.

[6] A. Deshpande *et al.*, "Distinguishing STEMI from pericarditis: ECG criteria and clinical implications," *American Journal of Emergency Medicine*, vol. 32, no. 10, pp. 1213–1218, Oct. 2014. Analyzes ST morphology rules to reduce STEMI false positives, demonstrating mapping of ECG metrics to clinical decision guidelines.

[7] T. Istolahti *et al.*, "Prognostic significance of T-wave inversion by lead distribution," *European Heart Journal*, vol. 41, no. 2, pp. 216–223, Dec. 2020. Correlates lead-specific T-inversion patterns with outcomes, informing rule thresholds for T-wave detectors.

[8] M. M. Rahman and M. H. A. Chowdhury, "T-wave detection based on right triangle hypotenuse system," *Journal of Cardiovascular Engineering and Technology*, vol. 13, no. 2, pp. 78–88, Oct. 2022. Proposes a geometric rule-based algorithm for accurate delineation of T-wave onset/offset across datasets.

[9] L. Saclova *et al.*, "Reliable P-wave detection in pathological ECG signals using phasor transform," *Scientific Reports*, vol. 12, Art. no. 7426, Apr. 2022. Implements phasor-based decision rules for P-wave identification in low-amplitude and arrhythmic recordings.

[10] W. Reklewski *et al.*, "Multiplierless QRS detection algorithm for mobile ECG applications," *Computerized Biology and Medicine*, vol. 164, Art. no. 107023, Apr. 2025. Presents a hardware-efficient, rule-driven QRS detector achieving 99.8

[11] N. Strodthoff *et al.*, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, Sep. 2020. Evaluates multiple CNN/RNN architectures on PTB-XL, providing a baseline for hybrid rule-learning pipelines.

[12] I. J. Selvam and V. P. John, "Classification of ECG abnormalities using deep learning on PTB-XL," *Pattern Recognition Letters*, vol. 175, pp. 18–26, Jan. 2025. Combines CNN feature extractors with SCP label mapping for multi-label classification, demonstrating high recall for rare arrhythmias.

[13] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, Art. no. 1760, 2020. Trains a ResNet-based model on a large clinical ECG repository, highlighting challenges of interpretability and the need for evidence annotations.

[14] A. K. Gupta and S. Varma, "Confidence scoring in rule-based medical signal interpretation," *IEEE Access*, vol. 8, pp. 135790–135800, 2020. Introduces metrics and calibration methods to assign confidence scores to deterministic ECG feature detections.

[15] H. Zhao *et al.*, "Traceable evidence annotation in ECG interpretation systems," *Journal of Medical Systems*, vol. 44, no. 2, Art. no. 44, Feb. 2021. Describes an architecture for linking each detected feature to lead/time references and raw measurements.

[16] M. D. Ribeiro *et al.*, "Automated ECG diagnostic statements generation with template mapping to SCP-ECG," *Computers in Cardiology*, vol. 46, pp. 525–528, 2019. Maps rule outputs to standardized SCP phrases and assembles narrative templates for clinician review.

[17] CEN, "SCP-ECG—Standard Communications Protocol for Computer-Assisted Electrocardiography," European Committee for Standardization, EN 1064, 2004. Defines the coding taxonomy for ECG findings and measurement semantics used in clinical systems.

[18] J. R. Smith and L. K. Jones, "Rule-based ST-segment analysis in 12-lead ECG for ischemia detection," *Journal of Electrocardiology*, vol. 58, pp. 45–52, Jun. 2020. Presents threshold- and slope-based ST elevation/depression rules calibrated on PTB-XL.

[19] E. Martin *et al.*, "Fine-tuning LLMs for structured clinical report generation with cited evidence," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1763–1772, Nov. 2024. Details an LLM fine-tuning workflow to ingest structured claims and produce narrative reports with inline evidence citations.

[20] H. Yu *et al.*, "Zero-shot ECG diagnosis with retrieval-augmented LLMs," *Proceedings of Machine Learning Research*, vol. 225, pp. 390–404, 2023. Demonstrates use of RAG to fetch templates and guideline snippets, enabling LLMs to cite relevant SCP rules during report synthesis.