AgenticX: An LLM-driven Interpretable ECG Classification Framework

Anonymous ACL submission

Abstract

Electrocardiogram (ECG) analysis is crucial for diagnosing cardiovascular disease, but existing automated ECG classification methods suffer from insufficient interpretability and accuracy, limiting their clinical application in practice. To address this issue, this paper proposes AgenticX, a multi-agent collaborative, LLM-driven interpretable ECG classification framework. We first invoke a base model to obtain model judgments, then, review whether the model's regions of interest conform to medical rules, and ultimately generate decision weights and medical reports from the LLM. Experiment results demonstrate that AgenticX outperforms existing work on the PTB-XL dataset, improving the accuracy and transparency of ECG classification.

1 Introduction

Cardiovascular disease (CVD) is one of the most severe and challenging global health issues (Battula et al., 2021), and the electrocardiogram (ECG) is one of the most important non-invasive tool for diagnosing heart diseases (Kaplan Berkaya et al., 2018). However, interpreting ECG signals is often time-consuming and labor-intensive. Additionally, with the rapid development of real-time health monitoring scenarios, such as wearable devices, the demand for automated and real-time ECG analysis has significantly increased. Therefore, developing efficient and interpretable automated ECG analysis systems has become an urgent necessity.

Currently, mainstream automated ECG analysis methods include mathematical methods, machine learning, deep learning, and others. Although these methods have achieved reasonable performance in classification tasks, they still exhibit significant shortcomings in terms of interpretability in clinical practice. Al-Zaiti and Bond (2023) categorized the interpretability of medical AI into four levels: assisting clinical decision-making, enhancing model robustness, discovering new medical patterns, and identifying and correcting errors (Al-Zaiti and Bond, 2023). However, most existing methods fail to even achieve the first level, limiting their practical application in clinical settings (Al-Zaiti and Bond, 2023; Tonekaboni et al., 2019).

The emergence of LLMs has opened new possibilities for developing model explanations that are more aligned with human cognition. However, despite some studies exploring the use of LLMs like ChatGPT to assist in explaining the behaviors of other AI models, such methods still constitute a minimal proportion of overall interpretability research (Sebin et al., 2024), or the involvement of LLMs is insufficient, such as using LLMs merely as "interpretation enhancement tools" to translate or expand existing XAI outputs, rather than performing independent reasoning and decision analysis (Sebin et al., 2024; Kroeger et al., 2024; Bhattacharjee et al., 2024). Therefore, the use of LLMs in ECG interpretability methods still holds significant research potential.

To address the existing limitations of ECG diagnostic models regarding clinical interpretability and reasoning transparency, we propose AgenticX, an interpretable ECG classification framework based on multi-model and multi-dimensional explanation fusion. Centered around LLMs, this framework integrates the predictive capabilities of deep neural networks with the reasoning capabilities of language models, thereby enhancing clinical reliability and diagnostic transparency.

Our contributions include the following key innovations:

Credibility Review Mechanism: We designed an LLM-driven intelligent agent (Review Agent) that integrates model explanation information (such as lead attention regions) with medical knowledge to identify and exclude untrustworthy predictions, significantly improving diagnostic accuracy.

LLM Gating Fusion Mechanism: By address-

ing the opacity of traditional deep learning gating models, we innovatively employed an LLM as an "expert system" gating controller. This approach outputs structured logic chains when integrating multiple model predictions and multiple category preditions for the same sample, achieving traceable and interpretable model fusion.

Global Performance Advantage and Empirical Validation: On the PTB-XL dataset, AgenticX surpasses existing best-performing studies across all five ECG disease classification tasks, establishing new benchmarks, particularly in core metrics accuracy, and validating the effectiveness and scalability of our approach.

Overall, the AgenticX framework enhances the accuracy and trustworthiness of AI models in automated ECG diagnosis and provides a more practical and transparent solution to interpretability challenges in medical AI.

2 Related Work

2.1 Explainable AI in ECG Classification

Due to the complexity and ambiguity of electrocardiogram (ECG) signals, ECG analysis has become a crucial task in medical AI. To improve model transparency, various Explainable AI (XAI) techniques have been introduced into ECG classification systems. By analysis, current ECG-XAI methods can be roughly categorized into three groups:

Saliency-based Visualization: Grad-CAM and its variants are the most commonly used techniques. These methods generate heatmaps to highlight the regions of the input signal that most contribute to the model's decision. Although intuitive, such visualizations often lack clinical interpretability and are difficult for domain experts to verify (Raza et al., 2021; Jahmunah et al., 2021; Moe et al., 2024; Hata et al., 2022; Singh and Sharma, 2022; K and U, 2025; Xiao et al., 2024; Zeng et al., 2024; Kim et al., 2022b; Jahmunah et al., 2022).

Feature Attribution: SHAP and similar model-agnostic methods aim to quantify each input feature's contribution to the prediction. These techniques are often applied to black-box models such as CNNs and Transformers. However, they also suffer from limited clinical interpretability and remain difficult for physicians to comprehend (Kim et al., 2022a; Agrawal et al., 2022; Kumar et al., 2023; Moe et al., 2024; Anand et al., 2022; Kilic et al., 2025; Vinh et al., 2025; Niu et al., 2025; Fikri and Isa, 2023; Katsushika et al., 2023).

Post-hoc Rule-based Reasoning: A small number of studies attempt to identify specific features strongly associated with certain cardiac conditions (e.g., ST elevation, T wave inversion) to provide backward interpretability. These methods rely on deterministic thresholds or rule-based logic but are constrained to narrow predefined conditions and lack generalizability across diverse ECG morphologies (de Chazal et al., 2004; Christov et al., 2006; Willems and Lesaffre, 1987; Dokur et al., 1999).

While these methods represent early advances in visualizing model behavior, they face three critical limitations: **Accuracy** — many methods still leave room for performance improvement; **Lack of Reasoning Capability** — most frameworks cannot integrate or validate explanations across multiple information sources; **Limited Clinical Usability** — the explanations are mostly visual or statistical and fail to map onto the reasoning framework of cardiologists.

2.2 LLM-based Approaches in ECG Classification

In recent years, LLMs, such as GPT-4, have demonstrated significant potential in enhancing explainability in AI by generating fluent natural language explanations based on information flow. However, such attempts remain limited and face evident technical challenges in the medical domain of ECG classification.

Most current studies use LLMs only for shallow post-hoc explanations, where LLMs generate natural language descriptions based on model outputs. For instance, Wu et al. (Wu et al., 2025) proposed a method using LLMs to enhance the interpretability of deep learning-based ECG models, but this approach is limited to post-hoc textual descriptions and does not influence the model's decision-making process, accuracy, or trust calibration.

Some studies have attempted to integrate LLMs more deeply into the diagnostic decision-making process. Roberto et al. (2024) showed that ChatGPT-4o achieved only 29.2% diagnostic accuracy in ECG image interpretation tasks, far below the 58.7% accuracy of cardiologists. Yang et al. (2025) developed ECG-LM, a multimodal framework combining ECG signals and natural language understanding, but found that the inclusion of LLMs actually degraded classification performance, failing to meet the expected goals. This indicates that accuracy remains a critical issue that

needs to be urgently addressed when using LLMs to assist in decision-making.

Furthermore, existing studies lack a coordinated agent-based framework to deeply embed LLMs into the diagnostic decision process, allowing them to truly make decisions and output the reasoning behind those decisions. Therefore, developing a collaborative, agent-based framework that coordinates LLMs with medical logic could become a major breakthrough in AI-driven ECG diagnosis, holding significant research and application potential.

3 Methodology

3.1 AgenticX

We propose **AgenticX**, a three-stage framework comprising: (i) multi-model diagnosis with structured explainability extraction, (ii) a review-agent–based reliability assessment mechanism, and (iii) an LLM-driven multi-agent decision fusion system. The overall architecture is illustrated in Fig. 1, where an ensemble of complementary backbones produces both predictions and model-specific attribution signals that are subsequently standardized and audited before final decision fusion.

A large language model (LLM) assistant was used to assist with code development and writing polish.

Algorithm 1 ECG Analysis

Input: sample
Output: classification_result, report
(base_results, base_explanation_results) = base_models(sample)
model_explanation_info = statistics(base_explanation_results)
model_judgment_info = adjust_weights(base_results)
review_results = review_model(model_explanation_info, model_judgment_info)
LLM_adjustment_values = LLM(model_explanation_info, model_judgment_info, review_results)
classification_result = adjust(base_results, LLM_adjustment_values)
report = LLM(model_explanation_info, model_judgment_info, review_results, classification_result)

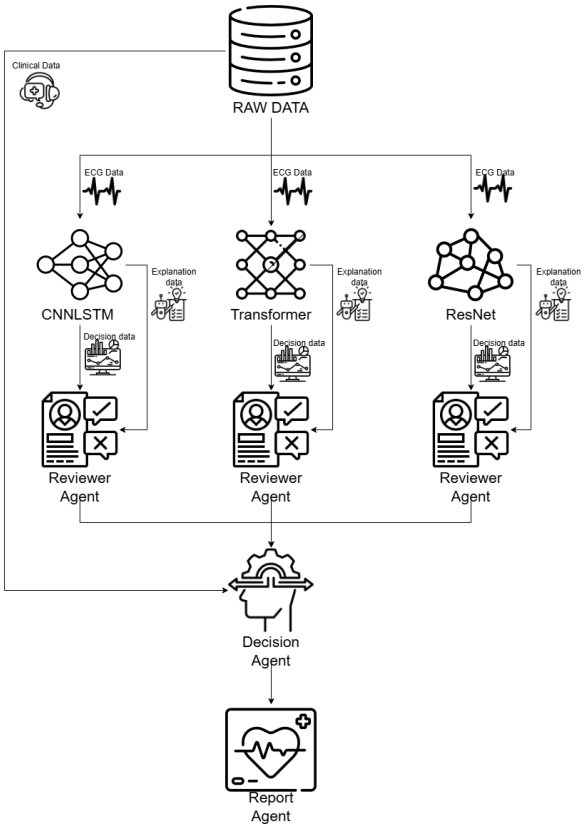


Figure 1: Overview of the **AgenticX** framework. The specific process is as follows: the data first enters the basic model to obtain decision data and explainable data. These two parts of data enter the reviewer agent to obtain the review results. All of the above information enters the decision agent to obtain the decision weight. After superimposing the decision weight on the decision data, the final judgment is obtained and a medical report is generated.

3.2 Stage I: Multi-Model Diagnosis

In the multilabel ECG classification setting, diverse deep architectures exhibit distinct strengths for modeling temporal, spatial, and cross-lead interactions. For example, for different classification tasks under the same dataset, different architectures may have their own performance advantages(Holguin-Garcia et al., 2024). To comprehensively capture heterogeneous diagnostic cues in 12-lead ECGs, we adopt three simple yet representative backbones: **CNN–LSTM**, **Transformer**, and **ResNet**.

CNN–LSTM. Convolutional layers extract local spatial features, while stacked LSTM layers model long-range temporal dependencies. This hybrid design is well-suited to arrhythmia patterns requiring rhythm analysis across multiple cardiac cycles. Compared with pure CNNs, CNN–LSTM yields stronger sequence modeling capacity.

Transformer. Multi-head self-attention enables learning long-range dependencies and cross-lead spatial interactions. For pathologies manifesting as coordinated multi-lead abnormalities (e.g., myocardial infarction), Transformers effectively capture inter-lead correlations and collaborative patterns.

ResNet. Residual connections alleviate vanishing gradients in deep networks and improve the learning of complex nonlinear representations. ResNets are adept at capturing higher-level spatio-temporal abstractions and, when paired with Grad-CAM, provide strong spatial interpretability.

Model choice rationale. Beyond classification accuracy, our framework explicitly requires *model-level interpretability*. Excessively complex models tend to have lower interpretability , so we prioritize simpler architectures with stronger attribution fidelity over highly complex but opaque models. Moreover, the three backbones are complementary across temporal and spatial dimensions, capturing features at different scales and from distinct perspectives. In this experiment, the hyperparameter settings for the three baseline models are as follows:

- **ResNet1D**: 4 residual stages with 64, 128, 256, and 512 channels; Adam optimizer (10^{-3}); 10 epochs; batch size 32.
- **CNN–LSTM**: Two 1×5 convolution blocks (64/128 filters) + bidirectional 2-layer LSTM (128 units per direction); Adam optimizer (10^{-3}); 15 epochs; batch size 32.
- **TimeSeries Transformer**: Model dimension 128, 8 attention heads, 2 encoder layers; Adam optimizer (10^{-3}); 5 epochs; batch size 32.

3.3 Stage II: Structured Explainability Extraction

For each backbone, we derive a standardized, structured explainability representation:

- **Peak detection:** Identify peaks in attribution heatmaps (from Grad-CAM or SHAP), followed by merging of proximal peaks.
- **Time–frequency mapping:** Using discrete wavelet transform (DWT)(Saritha et al., 2008), localize peak-aligned ECG segments (e.g., P, Q, R, S, T).
- **Segment-level indices:** For each cardiac cycle and segment (e.g., QRS, T wave), compute:
  - *Peak count*: total number of attribution peaks;
  - *Length ratio*: proportion of high-attribution frames within the segment;
  - *Total attribution*: sum of high-attribution values within the segment;
  - *Attention intensity*: total attribution divided by segment length (i.e., attention density).
- **Lead-wise attention statistics:** Aggregate attention distributions across leads per diagnostic class.

Example (Lead 11, Cycle 2). Lead-attention summary: (NORM): Lead 9 (13.7%), Lead 8 (10.8%), Lead 2 (10.0%).

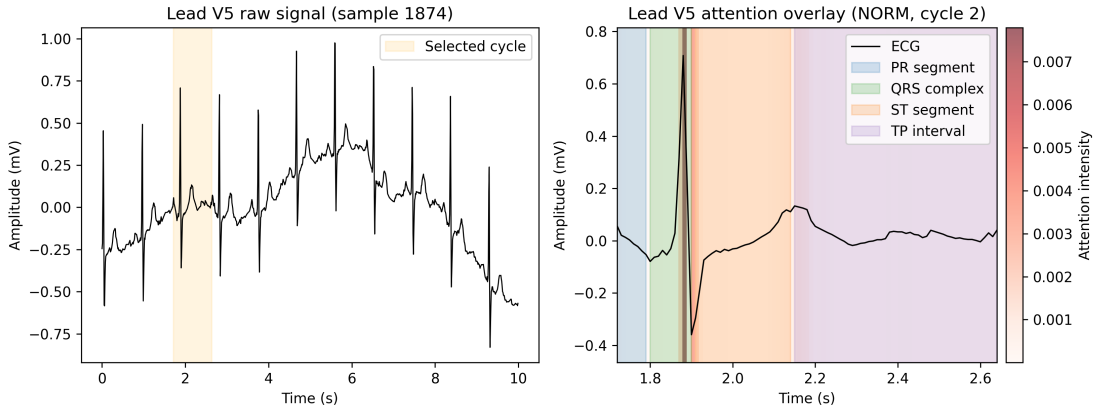
These structured indices constitute the input to the review agent and downstream decision agent, and are also consumed by the report generator.

3.4 Stage III: Multi-Agent Framework

In this stage, we invoke a *Review Agent* and a *Decision Agent* to produce the final prediction and generate a clinical-style report.

Reviewer agent based on model attention. Deep models are prone to *shortcut learning*, i.e., relying on spurious correlations rather than semantically relevant features. For instance, Beery *et al.* demonstrated large accuracy drops when animal-classification backgrounds changed (e.g., farm → beach) (Beery et al., 2018). In the medical context, if a model fixates on clinically irrelevant regions or leads, this may signal shortcut reliance. We therefore introduce a **Review Agent** that audits whether model attention concentrates on physiologically relevant waveforms/leads.

Inputs: predicted labels and probabilities, plus explainability metadata (lead, segment, weights). *Rule-driven prompts:* e.g., "STTC should focus on ST or T waves in V2–V6." If the attention deviates materially from medical priors, the Review Agent flags the output as unreliable and provides a rationale. MI (myocardial infarction) should focus on Q waves, ST-segment elevation, and T-wave inversion in leads V1-V4, while CD (heart failure) should focus on T-wave inversion in leads V3-V6.(Vogel et al., 2019) (Dressler and Roesler, 1947) (Zema, 1985) (Braga et al., 2004)




```
Sample 1874
Example (Lead 11, Cycle 2)
- Segment statistics:
  - Segment: PR segment | SHAP sum: 3.51e-04 | Peaks: 0 | Length ratio: 0.09 | Attention
intensity: 4.09e-03
  - Segment: QRS complex | SHAP sum: 1.39e-02 | Peaks: 2 | Length ratio: 0.12 | Attention
intensity: 1.18e-01
  - Segment: ST segment | SHAP sum: 4.11e-03 | Peaks: 1 | Length ratio: 0.26 | Attention
intensity: 1.59e-02
  - Segment: TP interval | SHAP sum: 4.81e-03 | Peaks: 8 | Length ratio: 0.54 | Attention
intensity: 8.94e-03
Lead-attention summary (NORM): Lead 10 (12.5%), Lead 2 (10.6%), Lead 12 (9.8%)
```

Figure 2: Example visualization of raw ECG signal (left) and corresponding attention overlay (right). Segment-level SHAP statistics and lead-attention summaries are shown at the bottom.

Decision Agent based on LLM-mediated adjustment. Empirical results indicate that using an LLM as a direct classifier yields suboptimal performance. To address this, we introduce a **Decision Agent** that leverages the LLM to return *adjustment weights* designed to modulate the fused model score. The Decision Agent takes the outputs from the deep models and incorporates the Review Agent's findings. Instead of evaluating each class independently, it aggregates *all classes for the same sample* to produce a final adjustment derived from the LLM. This collaborative framework allows the Decision Agent to refine the model's prediction by accounting for both the deep model outputs and the review of their attention mechanisms.

Score fusion across backbones

First, we compute a prior-weighted sum over models:

$$S(x, y) = \sum_{i=1}^{M} z_i(x \mid y)\, \alpha_i(x), \qquad (1)$$

where - x is the target class, - y is the input sample, - $z_i(x \mid y)$ denotes the logit of model i for class x on sample y, - $\alpha_i(x)$ is the class-conditional prior weight (set to prior accuracy in our experiments), - M is the total number of models.

LLM adjustment and final probability

We then obtain the final probability via a sigmoid mapping with LLM adjustment:

$$\hat{p}(x \mid y) = \sigma\big(S(x, y) + w_{\mathrm{LLM}}(x, y)\big), \qquad (2)$$

where $w_{\mathrm{LLM}}(x, y)$ is the LLM-derived adjustment and $\sigma(\cdot)$ is the sigmoid function. The probability $\hat{p}(x \mid y)$ is compared with a class-specific threshold to produce final decisions. On the validation set, we tune thresholds via k-fold grid search(Raschka, 2018) to optimize downstream criteria and then apply them to the test set.

Report generation. Given model predictions, explainability summaries, the Review Agent's verdicts, and the Decision Agent's outputs, a **Report Agent** consolidates *raw signals*, *explainability*, *review outcomes*, and *final decisions* into a structured, natural-language report designed for clinical interpretability.

4 Experiments and Results

4.1 Experimental Setup

We address three research questions (RQs):

- **RQ1 (§4.2):** Given the same data source, does **AgenticX** achieve higher accuracy than existing ECG classifiers?

- **RQ2 (§4.3):** Under the same data source, does **AgenticX** provide superior explainability compared with conventional LLM explanations?
- **RQ3 (§4.4):** Is the proposed framework effective overall, and how does each module contribute?

4.1.1 Dataset

We use PTB-XL (Wagner et al., 2022), one of the largest open ECG datasets, released under the CC BY 4.0 license and widely used as a common benchmark for arrhythmia research. PTB-XL contains 21,799 clinical 12-lead ECGs from 18,869 patients, each recording spanning 10 seconds. Recordings were annotated by up to two cardiologists, assigning potentially multiple ECG statements per record and grouped into five superclasses: NORM/MI/CD/STTC/HYP. We adopt these labels for classification.

We split the data into 80% training, 10% validation, and 10% testing. Concretely, we first hold out 80% for training and designate the remaining 20% as non-training data; the first half of the non-training split is used for hyperparameter tuning (validation), and the remaining half for testing, thereby mitigating leakage.

4.1.2 LLM Turbo

To fairly assess whether **AgenticX** enhances LLMs, we use general-purpose LLMs. Specifically, we test with `gpt4o-mini` and `gpt5-mini` to validate method effectiveness and generality.

4.1.3 Evaluation Metrics

We report standard classification metrics:

- **ACC (Accuracy):** proportion of samples with exactly matched multilabel predictions.
- **Precision:** among samples predicted as a class, the fraction that truly belong to that class.
- **Recall:** among all true instances of a class, the fraction correctly identified.
- **F1:** harmonic mean of Precision and Recall, capturing performance under class imbalance.

Since our final decision is adjusted by an LLM-derived weight, AUC—which primarily measures ranking ability of raw logits—is less informative in our setting, and thus we do not adopt it as a primary endpoint.

4.2 Performance Comparison (RQ1)

We compare **AgenticX** with competitive baselines on PTB-XL. Table 1 reports Accuracy, Precision, Recall, and F1.

AgenticX vs. other models. **AgenticX** achieves the highest **Accuracy** (0.900), surpassing all baselines and indicating strong overall discriminative ability. Its **Precision** (0.791) ranks third (below VGG16), while its **Recall** (0.709) is comparatively lower, which may be related to the fact that we did not use recall as the optimization metric.

Baselines vs. AgenticX. The three base backbones underperform most models in Accuracy. Simple voting improves accuracy but remains far behind top competitors. Using an LLM as a *standalone* gate (without explainability) increases Recall but substantially hurts Precision and overall F1, consistent with risk-averse over-diagnosis when unregularized. By contrast, **AgenticX** markedly improves Accuracy over the base models, evidencing the framework's effectiveness.

Overall analysis. **AgenticX** is best on Accuracy, implying more reliable overall judgments for clinical workflows. If comprehensive coverage of abnormalities is prioritized, ResGRU-Attention and 2D-CNN offer higher Recall. Future work could incorporate dynamic class rebalancing or recall-enhancement modules to improve Recall and F1 while maintaining high Accuracy.

4.3 Explainability Comparison (RQ2)

To address RQ2, we first selected a set of ECG samples and generated reports using a conventional LLM and the proposed **AgenticX** framework, respectively. To eliminate the confounding effect of accuracy differences, we then chose two cases on which both the LLM and AgenticX produced *correct* classifications, and focused solely on comparing explainability.

Participants and survey design. We collected 29 valid responses: 14 from medical professionals (e.g., cardiologists and medical students) and 15 from non-medical participants (e.g., patients and computer-science researchers). The survey presented two raw ECGs and asked respondents to rate the two corresponding reports on four dimensions using 5-point Likert scales (1–5):

- **Explainability**: clarity and coherence of the diagnostic reasoning;

Table 1: Main results on PTB-XL (multilabel, multiclass). Some prior works differ in labeling protocol; where applicable, we note such differences.

Model	**Acc.**	**Prec.**	**Recall**	**F1**
SincNet (2020) (Strodthoff et al., 2021)	0.765	0.714	0.662	0.680
VGG16 (2025) (Atwa et al., 2025)	0.796	**0.827**	0.766	0.676
ResGRU-Attention (2024) (Yang et al., 2024b)	0.884	0.726	0.800	0.759
2D-ECGNet (2024) (Tao et al., 2024)	0.892	0.752	0.790	0.770
STFAC-ECGNet (2024) (Yang et al., 2024a)	0.894	0.725	0.796	0.767
CAMV-RNN (2024) (Yang et al., 2024a)	0.889	0.748	0.783	0.765
2D-CNN (2024) (Elyamani et al., 2024)	0.899	0.811	0.783	**0.797**
basemodel_CNN	0.841	0.571	0.870	0.668
basemodel_RESNET	**0.936**	0.631	0.856	0.718
basemodel_TRANSFORMER	0.805	0.535	0.823	0.636
baseline (3-model sample vote)	0.852	0.586	0.883	0.685
baseline (3-model probability vote)	0.895	0.804	0.687	0.732
baseline (3-model + LLM decision)	0.772	0.435	**0.927**	0.559
AgenticX (Ours)	**0.900**	0.791	0.709	0.741

- **Clarity**: conciseness and lack of ambiguity in the language;
- **Accuracy**: concordance of the report with the ECG evidence;
- **Professionalism**: appropriateness of medical terminology.

The complete questionnaire is available online at this link.

Report One - ST-T Changes (STTC)

B. ST-T Changes (STTC) — Not supported

"Model outputs for STTC were low (well below thresholds) or, where near a threshold, lacked the right anatomical/segmental focus (i.e., insufficient emphasis on ST segments and T waves in canonical leads such as II, III, aVF, V1–V6)." "Review noted partial attention to anterior leads (e.g., V1–V3) but not the comprehensive, segment-specific pattern expected for clinically meaningful ST-T change."

Conclusion: "No reliable evidence of generalized ST-T abnormality."

Report Two - ST-T Changes (STTC)

ST segments: "Stable and isoelectric across all leads. T waves upright in most leads (inverted in aVR, which is normal)."

Report One provides a detailed model analysis of ST-T changes. It not only demonstrates that the model output is below the threshold but also details why the model fails to identify ST-T abnormalities: insufficient attention to specific anterior leads (e.g., V1-V3) and inadequate attention to ST segments and T waves in typical leads (e.g., II, III, aVF, V1–V6). Report Two is more concise, simply stating that ST segments remain stable and isoelectric in all leads (this is not necessarily true; morphological information derived solely from images can be contaminated by LLMs illusions, as discussed in the previous section), and noting that T waves in most leads are normal.

Results. *Accuracy.* Non-experts generally reported difficulty in judging diagnostic correctness, whereas medical professionals tended to view **AgenticX** reports as more accurate. For instance, Eric Zhou (medical student, University of Auckland) noted that the conventional LLM mentioned a "specific R-wave" pattern that was not present in the ECG, casting doubt on its correctness.
Professionalism. Non-experts tended to rate the conventional LLM as "more professional," likely due to denser, more arcane terminology that can create an impression of expertise despite hallucination risks; clinicians, however, perceived no substantial difference in professionalism between the two systems.
Explainability and clarity. Neither group observed a significant difference. A plausible reason is that our prompt design enforced a shared reporting template for both systems, yielding similar explanatory structure and presentation.

Summary. Overall, **AgenticX** maintains high classification accuracy while producing reports that clinicians deem more trustworthy; from the patient perspective, its reports are at least on par with those of a conventional LLM. Notably, AgenticX better avoids misleading descriptions, suggesting improved explainability and clinical usability, with

room for further enhancement.

4.4 Ablations and Analysis (RQ3)

We evaluate multiple baselines (see Table 1). **AgenticX (Ours)** denotes the full pipeline: multi-model diagnosis, structured explainability extraction, review-agent auditing, prior-weighted score fusion, LLM adjustment, and report generation. The *basemodel_{*}* rows are single-backbone results; *baseline (3-model sample vote)* aggregates by simple voting; *baseline (3-model probability vote)* uses prior-weighted fusion without LLM adjustment; and *baseline (3-model + LLM decision)* replaces fusion with an LLM-only gate.

Findings. Single backbones perform worse and show precision–recall imbalance, indicating insufficient capture of multi-dimensional ECG features. Simple voting improves accuracy (0.852), and prior-weighted fusion yields a further gain (0.895), demonstrating that weighted fusion surpasses naive voting. Using the LLM as a pure gate dramatically increases recall (0.927) but collapses precision (0.435), degrading F1 (0.559)—consistent with over-diagnosis when unregularized by explainability. The full **AgenticX** pipeline achieves the best accuracy (0.900) and improves overall reliability relative to all ablations. Future work may explore active learning and cost-sensitive thresholding to boost recall (and thus F1) while preserving high accuracy.

5 Discussion

The AgenticX framework proposed in this paper demonstrates certain advantages in accuracy and interpretability across multiple experiments, validating its feasibility and effectiveness in ECG disease classification and medical report generation tasks.

We found that the interpretable outputs of deep learning models themselves can serve as auxiliary decision-making information to improve accuracy. Post-hoc interpretation methods, such as heatmaps and SHAP values, are transformed into quantitative indicators at the "lead–heartbeat–waveform" level (e.g., peak count, coverage ratio, intensity density), which are then converted into standardized natural language descriptions. These structured explanations can be directly combined with medical rules and submitted to the reviewer agent. This not only serves a review function but also acts as direct input for report generation, providing traceability for decision-making reasons.

From the experimental results, AgenticX outperforms existing methods in classification accuracy and report fidelity, demonstrating its effectiveness and practical value. This suggests that LLMs can not only serve as an extension of interpretability outputs but also play a key role in the medical decision-making process, performing review and decision-making, and truly entering the decision chain to deliver clearer and more accurate medical reports.

6 Conclusion

This paper introduces AgenticX, an interpretable ECG diagnostic framework driven by multi-agent collaboration and LLMs, addressing the challenges of accuracy and interpretability in clinical practice. Experiments show that AgenticX outperforms existing methods in accuracy, and by introducing a review agent, it reduces false positives and hallucinations, improving the reliability of the reports. User studies reveal that both medical professionals and non-professionals find its reports accurate and easy to understand, aiding clinical decision-making and patient comprehension. AgenticX provides an effective pathway for the clinical adoption of medical AI.

7 Limitations

While AgenticX demonstrates strong performance and interpretability in ECG classification, several limitations remain that require further exploration:

Model Selection and Performance: Low accuracy can lead to missed or misdiagnosed patients, which can lead to serious consequences. The current model selection impacts system performance;Future work could explore more advanced architectures to improve accuracy and generalization.

Transferability and Validation: The framework has only been tested on the PTB-XL dataset. To ensure its adaptability across different datasets and domains, further transferability testing is needed.

Human Evaluation and Clinical Validation: The sample size for human evaluations is limited. Larger-scale clinical trials involving a broader range of professionals are needed to validate the framework's clinical value.

References

A. Agrawal, A. Chauhan, M.K. Shetty, M.D. Gupta, and A. Gupta. 2022. Ecg-icovidnet: Interpretable ai model to identify changes in the ecg signals of post-covid subjects. *Computers in Biology and Medicine*, 146:105540.

Salah S. Al-Zaiti and Raymond R. Bond. 2023. Explainable-by-design: Challenges, pitfalls, and opportunities for the clinical adoption of ai-enabled ecg. *Journal of Electrocardiology*, 81:292–294.

A. Anand, T. Kadian, M. K. Shetty, and A. Gupta. 2022. Explainable ai decision model for ecg data of cardiac disorders. *Biomedical Signal Processing and Control*, 75:103584.

A. E. M. Atwa, E.-S. Atlam, A. Ahmed, M. A. Atwa, E. M. Abdelrahim, and A. I. Siam. 2025. Interpretable deep learning models for arrhythmia classification based on ecg signals using ptb-x dataset. *Diagnostics*, 15(15):1950.

K. Battula, R. Durgadinesh, K. Suryapratap, and G. Vinaykumar. 2021. Use of machine learning techniques in the prediction of heart disease. In *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–5. IEEE.

Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–433, Munich, Germany. Springer.

A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu. 2024. Towards llm-guided causal explainability for black-box text classifiers. In *Proceedings of the AAAI ReLM 2024 Conference*, Vancouver, BC, Canada.

Simona Sarzi Braga, Raffaella Vaninetti, Antonio Laporta, Anna Picozzi, and Roberto F.E. Pedretti. 2004. T wave alternans is a predictor of death in patients with congestive heart failure. *International Journal of Cardiology*, 93(1):31–38.

Ivaylo Christov, Gèrman Gómez-Herrero, Vessela Krasteva, Irena Jekova, Atanas Gotchev, and Karen Egiazarian. 2006. Comparative study of morphological and time-frequency ecg descriptors for heartbeat classification. *Medical Engineering Physics*, 28(9):876–887.

Philip de Chazal, M. O'Dwyer, and R.B. Reilly. 2004. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51(7):1196–1206.

Z. Dokur, T. Ölmez, and E. Yazgan. 1999. Comparison of discrete wavelet and fourier transforms for ecg beat classification. *Electronics Letters*, 35(18).

William Dressler and Hugo Roesler. 1947. High t waves in the earliest stage of myocardial infarction. *American Heart Journal*, 34(5):627–645.

H. A. Elyamani, M. A. Salem, F. Melgani, and N. M. Yhiea. 2024. Deep residual 2d convolutional neural network for cardiovascular disease classification. *Scientific Reports*, 14:22040.

R. Fikri and S. M. Isa. 2023. Sleep stage classification on ecg signal and hrv using combination of random forest and shap value. *INNOVATIVE: Journal of Social Science Research*, 3(4):8680–8693.

Erika Hata, Chanjin Seo, Masafumi Nakayama, Kiyotaka Iwasaki, Takaaki Ohkawauchi, and Jun Ohya. 2022. Classification of aortic stenosis using ecg by deep learning and its analysis using grad-cam. *IEEE Transactions on Instrumentation and Measurement*, 71.

Sergio Alejandro Holguin-Garcia, Ernesto Guevara-Navarro, Alvaro Eduardo Daza-Chica, Maria Alejandra Patiño-Claro, Harold Brayan Arteaga-Arteaga, Gonzalo A. Ruz, Reinel Tabares-Soto, and Mario Alejandro Bravo-Ortiz. 2024. A comparative study of cnn-capsule-net, cnn-transformer encoder, and traditional machine learning algorithms to classify epileptic seizure. *BMC Medical Informatics and Decision Making*, 24:60. Open Access, Published: 01 March 2024.

V. Jahmunah, E. Y. K. Ng, Ru-San Tan, Shu Lih Oh, and U. Rajendra Acharya. 2022. Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals. *Computers in Biology and Medicine*, 146:105550.

V. Jahmunah, S.L. Oh, U.R. Acharya, T.R. San, E.J. Ciaccio, and R.S. Tan. 2021. Explainable deep learning-based approach for multilabel classification of electrocardiogram. *Computers in Biology and Medicine*, 134:104457.

Subhashini K and Prasanna E U. 2025. Interpretable deep learning framework for real-time ecg classification and diagnosis. In *Proceedings of the 2025 3rd International Conference on Signal Processing and Machine Learning (SPML)*. IEEE.

S. Kaplan Berkaya, A. K. Uysal, E. Sora Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu. 2018. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235.

S. Katsushika, S. Kodera, S. Sawano, H. Shinohara, N. Setoguchi, K. Tanabe, and I. Komuro. 2023. An explainable artificial intelligence-enabled electrocardiogram analysis model for the classification of reduced left ventricular function. *European Heart Journal – Digital Health*, 4(3):254–264.

M. E. Kilic, O. A. Tufekcioglu, Y. R. Yilancioglu, O. E. Turan, and E. E. Ozcan. 2025. Explainable ai for specific arrhythmia detection: Shap-based insights from multi-lead ecg data. *Journal of Medical and Biological Engineering*, pages 1–11.

J. Kim, S. Lee, H. Park, and Y. Choi. 2022a. Explainable ai decision model for ecg data of cardiac disorders. *Biomedical Signal Processing and Control*, 71:103079.

Jin-Kook Kim, Sunghoon Jung, Jinwon Park, and Sung Won Han. 2022b. Arrhythmia detection model using modified densenet for comprehensible grad-cam visualization. *Biomedical Signal Processing and Control*, 73:103408.

Nina Kroeger, Daniel Ley, Subhransu M. Krishna, Charu Agarwal, and Himabindu Lakkaraju. 2024. In-context explainers: Harnessing llms for explaining black box models. *arXiv preprint arXiv:2310.05797*.

R. Kumar, S. Singh, A. Sharma, and D. Gupta. 2023. Explainable ai-driven machine learning for heart disease detection using ecg signal. *Applied Soft Computing*, 136:110058.

A. Moe, L. Røstad, E.M. Støa, and M. Valla. 2024. Exploring interpretable ai methods for ecg data classification. *ACM Transactions on Computing for Healthcare*, 5(2):1–20.

Y. Niu, X. Chen, J. Fan, C. Liu, M. Fang, Z. Liu, and H. Fan. 2025. Explainable machine learning model based on eeg, ecg, and clinical features for predicting neurological outcomes in cardiac arrest patient. *Scientific Reports*, 15(1):11498.

Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv*. Accessed: 2025-10-05.

Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li. 2021. Designing ecg monitoring healthcare system with federated transfer learning and explainable ai. *Knowledge-Based Systems*, 232:107763.

Anna Maria De Roberto, Fabiola De Marco, Luigi Di Biasi, Domenico Rossi, and Genoveffa Tortora. 2024. Can chatgpt-4o enhance ECG interpretation accuracy compared to cardiologists? In *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6852–6858.

C. Saritha, V. Sukanya, and Y. Narasimha Murthy. 2008. Ecg signal analysis using wavelet transforms. *Bulg. J. Phys.*, 35:68–77. PACS numbers: 87.85.J; 02.30.Nw.

B. Sebin, N. Taskin, and N. Mehdiyev. 2024. Exploring the intersection of large language models (llms) and explainable ai (xai): A systematic literature review (research-in-progress). In *Proceedings of the International Conference on Information Resources Management (CONF-IRM)*. Association for Information Systems.

Prateek Singh and Ambalika Sharma. 2022. Interpretation and classification of arrhythmia using deep convolutional network. *IEEE Transactions on Instrumentation and Measurement*, 71.

N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek. 2021. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528.

Rui Tao, Lin Wang, Yingnan Xiong, and Yu-Rong Zeng. 2024. Im-ecg: An interpretable framework for arrhythmia detection using multi-lead ecg. *Expert Systems with Applications*, 237(Part A):121497.

Seyed Tonekaboni, Sahil Joshi, Matthew D. McCradden, and Ariel Goldenberg. 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the Machine Learning for Healthcare Conference*, pages 359–380. PMLR.

N. T. Vinh, B. T. Tung, and N. V. Ha. 2025. Multimodal deep learning for ecg heartbeat classification with shap-based interpretability. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Birgit Vogel, Bimmer E. Claessen, Suzanne V. Arnold, Danny Chan, David J. Cohen, Evangelos Giannitsis, C. Michael Gibson, Shinya Goto, Hugo A. Katus, Mathieu Kerneis, Takeshi Kimura, Vijay Kunadian, Duane S. Pinto, Hiroki Shiomi, John A. Spertus, P. Gabriel Steg, and Roxana Mehran. 2019. St-segment elevation myocardial infarction. *Nature Reviews Disease Primers*, 5.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. 2022. Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3). RRID:SCR_007345.

Jos L. Willems and Emmanuel Lesaffre. 1987. Comparison of multigroup logistic and linear discriminant ecg and vcg classification. *Journal of Electrocardiology*, 20(2):83–92.

Shi Wu, Jianlong Zhou, Yifei Dong, and Fang Chen. 2025. Enhancing explainability of deep learning-based ecg diagnosis using large language models. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence (ICAAI '24)*, ICAAI '24, pages 61–65, New York, NY, USA. Association for Computing Machinery.

Qiao Xiao, Khuan Lee, Siti Aisah Mokhtar, Iskasymar Ismail, Ahmad Luqman bin Md Pauzi, and Poh Ying Lim. 2024. Leveraging grad-cam for interpretability in lpkf-enhanced inceptiontime model for multilabel ecg classification. *International Journal of Public Health & Clinical Sciences (IJPHCS)*, 11(4):1–15.

Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Suyuan Zhao, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Ping Zhang, and 1 others. 2025. Ecg-lm: Understanding electrocardiogram with a large language model. *Health Data Science*, 5:0221.

Z. Yang, A. Jin, Y. Li, X. Yu, X. Xu, J. Wang, Q. Li, X. Guo, and Y. Liu. 2024a. A coordinated adaptive multiscale enhanced spatio-temporal fusion network

for multi-lead electrocardiogram arrhythmia detection. *Scientific Reports*, 14:20828.

Z. Yang, A. Jin, Y. Liu, W. Lv, and X. Zhu. 2024b. The fusion model of resnet and gru based on simplified self-attention for ecg classification on ptb-xl dataset. In R. Lee, editor, *Computer and Information Science and Engineering*, volume 1156 of *Studies in Computational Intelligence*. Springer, Cham.

Michael J. Zema. 1985. Q wave, s-t segment, and t wave myocardial infarction. useful clinical distinction. *American Heart Journal*, 78(3):391–398.

Wei Zeng, Liangmin Shan, Chengzhi Yuan, and Shaoyi Du. 2024. Advancing cardiac diagnostics: Exceptional accuracy in abnormal ecg signal classification with cascading deep learning and explainability analysis. *Applied Soft Computing*, 165:112056.