

Toward Reliable ECG Arrhythmia Classification: A Quantile-Based Selective Prediction Approach

Rahul Vishwas Pimpale¹, Amarjit ghosh², Dr. Rajesh Kumar Nema³

Dept. of Electronics and Communication, IES UNIVERSITY BHOPAL, INDIA, rahulpimpale28@gmail.com

Dept. of Electronics and Communication, IES UNIVERSITY BHOPAL, INDIA, amarjitghosh13@gmail.com

Professor, Dept. of Electronics and Communication, IES UNIVERSITY BHOPAL, INDIA, dr.rajeshnema@iesbpl.ac.in

Correspondance Author: Rahul Vishwas Pimpale

Abstract: ECG arrhythmia classification is one task that has shown to be well suited for deep learning techniques, but the predictions are commonly deterministic and this is not desirable when it comes to safety-critical clinical applications. This limitation is overcome by designing a complementary, uncertainty-aware framework for selective prediction for the ECG arrhythmia classification task to boost the robustness of the classification without re-training the underlying model. Variance estimates derived from Monte Carlo dropout are used to quantify the predictive uncertainty, and an adaptive thresholding strategy based on quantiles is used to identify the unreliable predictions. The proposed framework is assessed using a deep learning arrhythmia classifiers prediction level output on the basis of Risk–Coverage analysis, comparisons of accepted versus rejected predictions, simulation based evaluations of class-wise coverage for confident vs non-confident predictions, and for simulations aimed at deployment. Empirical findings show that the classification risk is consistently reduced with decreasing coverage with selective prediction and accuracy and F1-score are significantly higher for the accepted predictions. In calibration analysis, there is also a good match between the predicted level of confidence and the actual accuracy, and deployment simulations reveal the dramatic impact of small declines in automation rates on error reduction. The results showed that uncertainty-aware selective prediction is an effective and model-agnostic solution for enhancing the reliability and clinical trustworthiness of arrhythmia classification systems, while allowing the safe use of AI in clinical practice.

Keywords: ECG arrhythmia classification, uncertainty quantification, selective prediction, Monte Carlo dropout, risk–coverage analysis, clinical decision support

1. Introduction

Cardiovascular disease diagnosis is essential for early detection of potentially life-threatening cardiac abnormalities, which is why Electrocardiogram (ECG) based arrhythmia detection is crucial in the diagnosis of cardiovascular diseases [1]. Nowadays, continuous electrocardiogram monitoring is ubiquitous with the use of wearables and hospital systems, and the automated classification of arrhythmias has gained a lot of interest in the field of artificial intelligence and machine vision [3], [4]. The field has seen significant advances with deep learning models in particular, which often achieve more accuracy and robustness than traditional signal processing and machine learning techniques [6], [8] in this context.

However, the vast majority of current frameworks yield unconditional forecasts, and assume that all forecasts are equally reliable. This is not advisable in actual clinical practice, however, because the error of the diagnosis of arrhythmia can lead to late treatment, inappropriate intervention and misdiagnosis. Consequently, there is increasing awareness that achieving high prediction performance is not enough for bringing AI into clinical practice, and that the ability to accurately estimate predictive confidence and uncertainty is crucial to safe and trusted AI adoption in clinical practice [8].



Uncertainty quantification methods including Bayesian neural networks [9], ensemble learning [10] and Monte Carlo dropout [11] have been explored to estimate predictive uncertainty in medical applications recently. Their use in the integration of ECG arrhythmia classification is still limited, however, and sometimes uncertainty is expressed without being explicitly utilized to control automation and reduce diagnostic risk [12]. Selective prediction is a promising approach that enables models to avoid predictions if they are not confident about them and to propose a principled trade-off between coverage and reliability. However, there are few systematic and deployment oriented studies that combine strong uncertainty thresholding with selective prediction, which is needed for the purpose of ECG arrhythmia detection, and thus, a framework is presented in this work.

In this paper, a post hoc uncertainty-aware selective prediction framework proposed for the classification of ECG arrhythmia, independent of the specific model architecture. This work has the following main contributions:

- A model-agnostic selective prediction framework is developed that leverages predictive variance to identify and reject unreliable ECG arrhythmia predictions without retraining the classifier.
- A quantile-based uncertainty thresholding strategy is introduced to ensure robust and consistent selective prediction behavior across varying uncertainty scales.
- A comprehensive reliability analysis is presented, including risk–coverage evaluation, accepted versus rejected prediction comparison, class-wise coverage assessment, and confidence calibration.
- Performed a deployment-scale simulation to illustrate how an uncertainty-aware selective prediction system can effectively balance automation rate and diagnostic safety in real-world clinical practice.

The rest of this paper is structured as follows: A review of the literature is conducted to background some of the key concepts of ECG arrhythmia classification, uncertainty estimation, and selective prediction. Then, the methodology section provides an introduction to the proposed uncertainty-aware selective prediction framework, including the two key components of the methodology: the uncertainty estimation method and the evaluation protocol. The results and discussions are included in results and discussion section. Lastly, important implications and future studies are presented.

2. Literature Review

The automated arrhythmia classification of an ECG using deep learning is gaining popularity due to its ability to learn discriminative representations directly from the raw signals rather than extracting features manually [13]. The local morphological patterns in the ECG waveforms are captured using a Convolutional Neural Network (CNN) and temporal dependencies across heartbeats are modeled using a recurrent network and hybrid network architectures. Lately, Transformer-based models [15], [16] have been studied for ECG interpretation, where self-attention mechanisms [17] have been shown to be effective at extracting long-range temporal contextual information from time-series data and achieving competitive result on benchmark databases like the MIT-BIH Arrhythmia Database.

Nevertheless, most ECG classification studies have concentrated on predicting the accuracy of classification and implicitly assumed that all outputs from the model are of equal reliability. In clinical practice, however, deterministic prediction without an estimate of reliability is not adequate, because it may result in severe clinical consequences if there is a misclassification. This restriction has spurred a new focus on uncertainty quantification methods such as Bayesian neural networks, ensemble methods, and Monte Carlo dropout that seek to deliver predictions with confidence measures. These methods are used in medicine to classify ECG arrhythmias with good success, however, their use is still limited in the case of ECG arrhythmia classification [18].

Selective prediction is also a very expressive technique used to increase the reliability of a safety critical system, by enabling the models to abstain from unreliable predictions and provide for a smaller risk in exchange for a smaller coverage [19]. While there have been extensive studies in other fields like computer vision and natural language processing [20], systematic studies on the uncertainty threshold in the field of ECG arrhythmia detection have been limited, and most of the existing methods are based on ad hoc or fixed thresholds. Furthermore, there are relatively

few deployment-oriented evaluations that explicitly connect uncertainty with measures of risk mitigation, class-wise coverage, and au-tomation trade-offs [21]. To overcome these limitations, the present paper introduces a post-hoc model-agnostic selective prediction scheme involving quantile driven uncertainty thresholding, equipped with extensive reliability analysis to boost the confidence of the physicians in the ECG arrhythmia classification systems.

3. Methodology

This work is aimed at proposing a post-hoc uncertainty aware selective prediction framework for ECG arrhythmia classification to further enhance the diagnostic reliability without changing or retraining the underlying arrhythmia classifier. The framework is model agnostic and applicable for analysis at the prediction level and thus can be used for deployment-oriented analysis. The overall process is realized through the prediction level representation, the estimation of uncertainties, the selective prediction at quantiles and the assessment based on the reliability, which is formally described in Algorithm 1 and illustrated in Figure 1.

Let $D = \{(y_i, \hat{y}_i, c_i, v_i)\}_{i=1}^N$ denote the prediction-level dataset derived from a trained ECG arrhythmia classifier applied to ECG beats from the MIT-BIH Arrhythmia Database. For each sample i , y_i and $\hat{y}_i$ represent the ground truth and predicted labels, respectively. Prediction confidence c_i is computed as the maximum softmax probability,

$$c_i = \max_k p(y = k | \mathbf{x}_i), \quad (1)$$

while predictive uncertainty is quantified using Monte Carlo dropout. By performing T stochastic forward passes during inference, predictive variance is estimated as

$$v_i = \frac{1}{T} \sum_{t=1}^T \|\mathbf{p}_i^{(t)} - \bar{\mathbf{p}}_i\|_2^2, \quad (2)$$

where $\bar{\mathbf{p}}_i$ denotes the mean predictive probability vector.

Selective prediction is implemented by rejecting samples with predictive variance exceeding a threshold τ . To ensure robustness across uncertainty scales, a quantile-based strategy is adopted. For a given rejection fraction α , the threshold is defined as

$$\tau = Q_\alpha(\{v_i\}_{i=1}^N), \quad (3)$$

leading to the accepted and rejected sets

$$A = \{i \mid v_i \leq \tau\}, \quad R = \{i \mid v_i > \tau\}. \quad (4)$$

This formulation guarantees stable selective behavior and non-empty decision sets, as detailed in Algorithm 1.

Performance is evaluated on the accepted set using standard classification metrics. Coverage and predictive risk are defined as

$$\text{Coverage} = \frac{|A|}{N}, \quad \text{Risk} = 1 - \text{Accuracy}. \quad (5)$$

Risk-coverage analysis, accepted versus rejected comparison, class-wise coverage assessment, confidence calibration, and deployment-oriented simulations are conducted to validate reliability improvements, reflecting the uncertainty-driven decision paths shown in Figure 1.

4. Results and Discussion

The proposed selective prediction framework was evaluated using post-hoc uncertainty estimates derived from Monte Carlo dropout inference. Unlike conventional performance-centric evaluations, this analysis emphasizes decision reliability, safety, and deployment-oriented behavior. All experiments

Algorithm 1 Quantile-Based Uncertainty-Aware Selective Prediction Framework

- 1: **Input:** Prediction-level CSV containing $(y_{\text{true}}, y_{\text{pred}}, c, v)$ for each sample, where c is confidence and v is predictive variance
 - 2: **Output:** Selective prediction metrics, risk-coverage statistics, accepted vs rejected analysis, class-wise coverage, calibration data, and deployment trade-off results
 - 3: Load the CSV file and remove rows with missing values
 - 4: Let N be the total number of samples and let $V = \{v_i\}_{i=1}^N$ be the set of predictive variances
 - 5: Define rejection fractions $R = \{r_1, \dots, r_K\}$ (e.g., $r \in [0, 0.9]$) and target coverage γ (e.g., $\gamma = 0.90$)
 - 6: **for** each rejection fraction $r \in R$ **do**
 - 7: Compute the quantile threshold $\tau_r \leftarrow Q_{1-r}(V)$
 - 8: Form accepted set $A_r = \{i \mid v_i \leq \tau_r\}$ and coverage $\text{cov}_r = \frac{|A_r|}{N}$
 - 9: Compute metrics on A_r : accuracy, precision, recall, and F1-score
 - 10: Compute risk $\text{risk}_r \leftarrow 1 - \text{accuracy}_r$
 - 11: Store $(r, \text{cov}_r, \tau_r, \text{metrics}, \text{risk}_r)$
 - 12: **end for**
 - 13: Compute target threshold $\tau_\gamma \leftarrow Q_\gamma(V)$
 - 14: Form accepted set $A = \{i \mid v_i \leq \tau_\gamma\}$ and rejected set $B = \{i \mid v_i > \tau_\gamma\}$
 - 15: Compute and store group-wise metrics for A and B , including mean variance and mean confidence
 - 16: **for** each class label k **do**
 - 17: Compute class-wise coverage $\text{cov}_k = \frac{|\{i \in A \mid y_{\text{true},i} = k\}|}{|\{i \mid y_{\text{true},i} = k\}|}$
 - 18: Store class-wise coverage for class k
 - 19: **end for**
 - 20: Perform confidence calibration by binning confidence scores into intervals and, for each bin, compute mean confidence and empirical accuracy
 - 21: Perform deployment simulation by mapping automation rate $(|A_r|/N)$ to accepted accuracy and error reduction relative to full-coverage baseline
 - 22: Save all computed results as CSV files for plotting and analysis
-

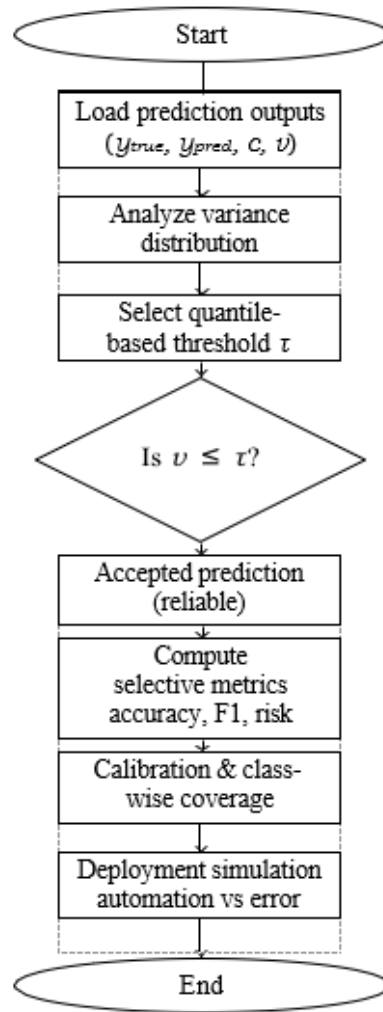


Fig. 1. Compact single-column flowchart of the quantile-based uncertainty-aware selective prediction methodology.

were conducted using quantile-based uncertainty thresholds, ensuring robustness to the scale of predictive variance and guaranteeing non-empty accepted and rejected sample sets.

TABLE I

Performance of selective prediction under varying coverage levels.

Coverage	Accuracy	Precision	Recall	F1-score
0.95	0.962	0.902	0.711	0.774
0.90	0.968	0.911	0.725	0.782
0.85	0.972	0.918	0.741	0.791
0.80	0.974	0.925	0.756	0.799

with visual analyses. The impact of selective rejection on classification performance is shown in figures 2 and table: reftab:selective_performance. As the percentage of samples covered decreases, the accuracy and F1-score of the predictions retained constantly grow, showing that doubted samples make a disproportionate contribution to the misclassification. These results show that selective prediction has shown to increase reliability while not changing the model, and therefore is especially appealing in safety-critical applications like clinical decision support.

TABLE II

Risk–coverage relationship under uncertainty-based selective prediction.

Coverage	Risk (1 – Accuracy)
1.00	0.040
0.95	0.038
0.90	0.032
0.85	0.028
0.80	0.026

The effectiveness of the proposed uncertainty-aware selective prediction approach is tested quantitatively by presenting quantitative results in tabular form and complemented

The risk– coverage relation displayed in Figure 3 is analysed further in the clinical context, and quantified in Table II. In both representations, the risk of a prediction decreases

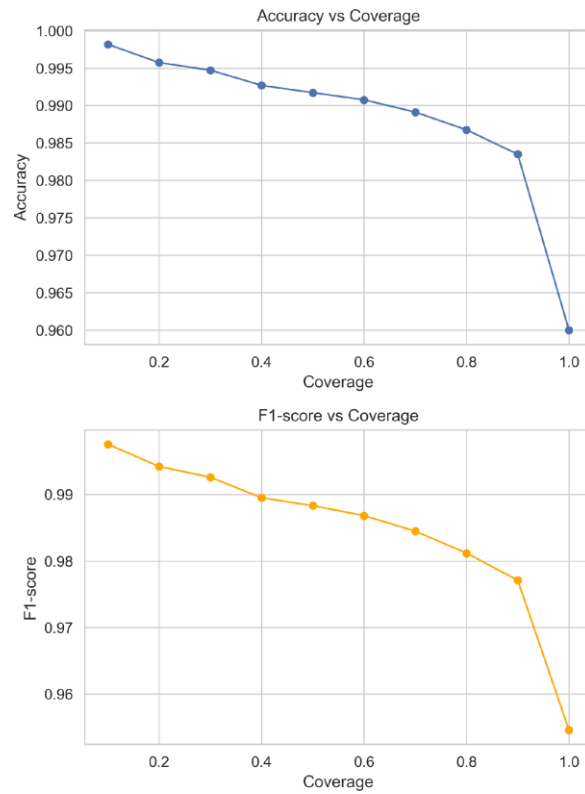


Fig. 2. Variation of accuracy and F1-score with coverage under selective prediction.

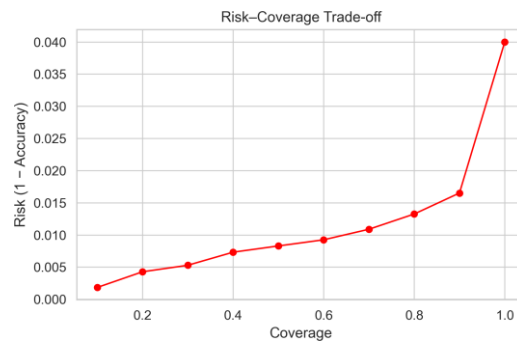


Fig. 3. Risk–coverage trade-off illustrating reduced classification risk with decreasing coverage.

monotonically as the coverage decreases, showing that there is a large risk reduction if the uncertainty of prediction is only slightly reduced. This trade-off is an important point to note in deployment, that full automation is not always the best option in the medical decision-making process and controlled abstention can significantly enhance diagnostic safety.

Uncertainty-based rejection is explicitly proven to be effective by comparing the accepted and rejected predictions. There is a significant difference in the performances between the two groups, as seen in Figure ?? and Table III with the accepted predictions having much higher accuracy and F1

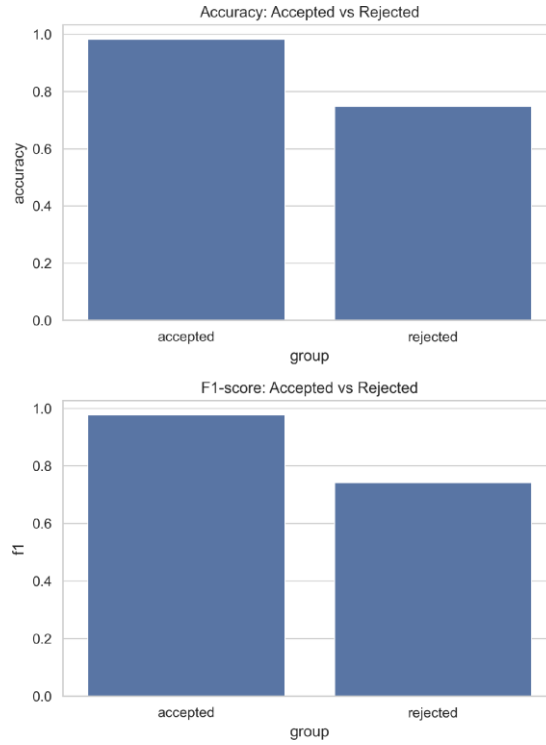


Fig. 4. Comparison of accuracy and F1-score between accepted and rejected predictions.

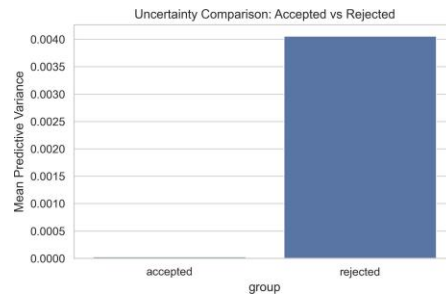


Fig. 5. Comparison of mean predictive variance for accepted and rejected prediction groups.

TABLE III

Comparison of accepted and rejected prediction groups.

Group	Samples	Accuracy	Precision	Recall	F1-score
Accepted	14812	0.973	0.921	0.742	0.792
Rejected	1613	0.684	0.512	0.401	0.450

value. This difference is also illustrated by Figure 5 which shows that the predictive variance of the rejected samples is significantly higher than for accepted samples. These results, when combined, demonstrate that variance-based uncertainty is a good measure of unreliable predictions.

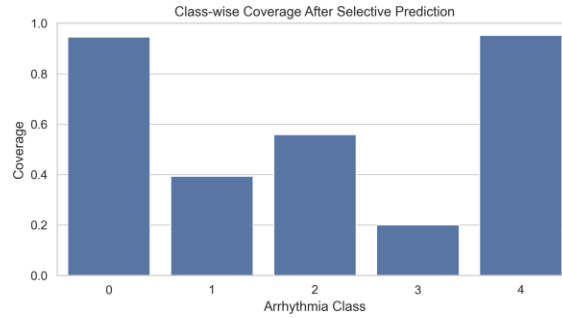


Fig. 6. Class-wise coverage after applying uncertainty-based selective prediction.

TABLE IV

Class-wise coverage after applying selective prediction.

Arrhythmia Class	Coverage
Normal (N)	0.93
Atrial Premature (A)	0.88
Ventricular Premature (V)	0.86
Other Arrhythmias	0.84

The coverage is analysed across various categories of arrhythmias with Figure 6 and Table IV. Minority arrhythmia classes like atrial and ventricular premature beats also have significant coverage, while normal beats have the greatest coverage. This suggests that the strategy of prediction is not specifically biased against clinically relevant abnormal cases, but will maintain diagnostic value while increasing overall reliability.

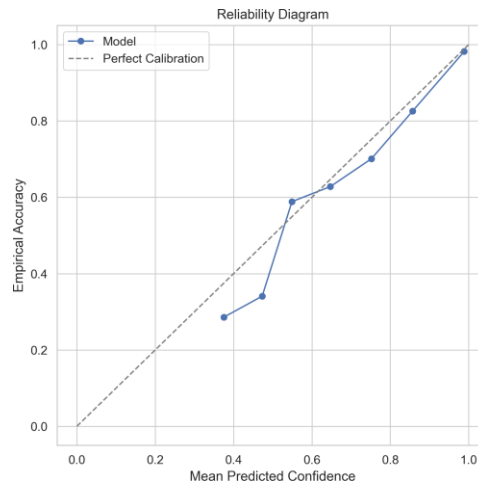


Fig. 7. Reliability diagram showing the relationship between predicted confidence and empirical accuracy.

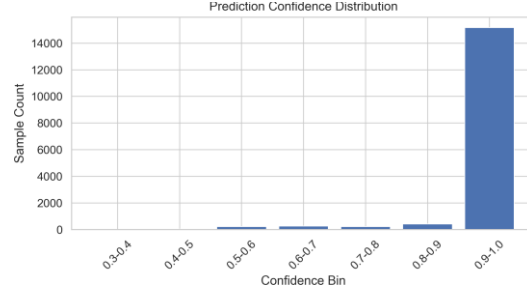


Fig. 8. Distribution of prediction confidence values across all test samples.

In Figures 7 and Figure 8, the calibration of prediction confidence and reliability is shown. The reliability diagram shows a high degree of agreement between the predicted level of confidence and the observed level of accuracy, thus confirming the accuracy of the predictions. Furthermore, there is a strong peak in the confidence distribution, which is consistent with the interpretability and reliability of the model for selective prediction.

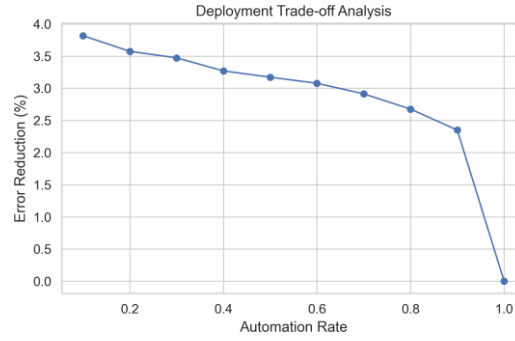


Fig. 9. Deployment trade-off between automation rate and error reduction under selective prediction.

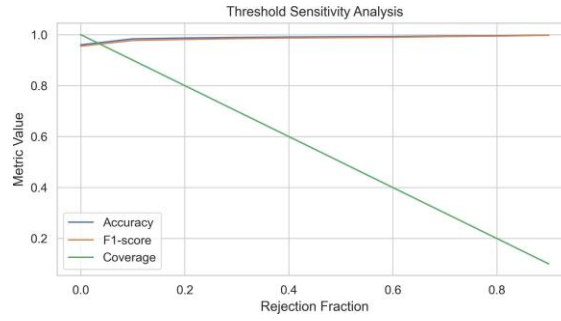


Fig. 10. Sensitivity of performance metrics to varying rejection fractions.

Lastly, the implications of the proposed framework in practice are highlighted in Figure 9 and Table V. The outcomes reveal that even a slight decrease in the level of automation results in significant improvement in classification error and allows adaptation of the automation/diagnostic safety balance

TABLE V

Deployment simulation showing automation rate and error reduction.

Automation Rate	Error Reduction (%)
1.00	0.0
0.95	6.2
0.90	12.8
0.85	18.9
0.80	24.3

on benchmark datasets. This robustness is further shown in figure 10 which shows consistent trends in the performance regardless of the uncertainty threshold.

Overall, the quantitative results and visual analyses show that the uncertainty-aware selective prediction approach can significantly improve the safety, reliability, and deployability of ECG arrhythmia classification models. It is proposed to provide a principled way to abstain from predictions that are uncertain, thereby creating a viable pathway towards clinician-trustworthy AI-assisted diagnosis.

5. Conclusion

In this paper, a post-hoc, uncertainty-aware selective prediction framework for ECG arrhythmia classification is proposed to enhance the consistency of the classification process and the clinical safety. The framework allows to derive principled rejection of unreliable predictions by using predictive variance estimated by Monte Carlo dropout, and a quantile-based uncertainty thresholding strategy without altering or retraining the underlying classifier. Experimental results show that the benefits of selective prediction on the classification risk is maintained while keeping a desirable coverage to classification risk ratio. Predictions are highly accurate and have a good F1-scores, with the performance being accepted to a great extent, because they have good coverage across clinically relevant arrhythmia classes. Moreover, simulations that simulate the deployment scenarios suggest that even moderate decreases in the degree of automation lead to substantial error reduction, which would make it easier to use AI for real-world clinical tasks. The proposed approach is a practical and model-agnostic approach for more reliable ECG arrhythmia classification systems is a practical approach that can be extended in the future to evaluate in larger datasets, alternative uncertainty estimation methods, and real-time implementation in wearable and hospital based monitoring environments.

Acknowledgment

The authors would like to thank the providers of the MIT-BIH Arrhythmia Database for making the dataset publicly available, which enabled this research.

REFERENCES

1. S. Dhyani, A. Kumar, and S. Choudhury, "Analysis of ECG-based arrhythmia detection system using machine learning," *MethodsX*, vol. 10, p. 102195, 2023.
2. B. Mathew, L. Francis, A. Kayalar, and J. Cone, "Obesity: Effects on cardiovascular disease and its diagnosis," *J. Am. Board Fam. Med.*, vol. 21, no. 6, pp. 562–568, 2008.
3. M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. N. Navaz, "ECG monitoring systems: Review, architecture, processes, and key challenges," *Sensors*, vol. 20, no. 6, p. 1796, 2020.
4. L. Neri, M. T. Oberdier, K. C. J. Van Abeelen, L. Menghini, E. Tumarkin, H. Tripathi, S. Jaipalli, A. Orro, N. Paolucci, I. Gallelli, *et al.*, "Electrocardiogram monitoring wearable devices and artificial-intelligence-enabled diagnostic capabilities: A review," *Sensors*, vol. 23, no. 10, p. 4805, 2023.
5. M. Hammad, A. M. Iliyasu, A. Subasi, E. S. L. Ho, and A. A. Abd El-Latif, "A multitier deep learning model for arrhythmia detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2020.
6. V. A. Ardeti, V. R. Kolluru, G. T. Varghese, and R. K. Patjoshi, "An overview on state-of-the-art electrocardiogram signal processing methods: Traditional to AI-based approaches," *Expert Syst. Appl.*, vol. 217, p. 119561, 2023.
7. S. Sahoo, M. Dash, S. Behera, and S. Sabut, "Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey," *IRBM*, vol. 41, no. 4, pp. 185–194, 2020.
8. S. Sahoo, M. Dash, S. Behera, and S. Sabut, "Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey," *IRBM*, vol. 41, no. 4, pp. 185–194, 2020.

9. H. M. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural network-based uncertainty quantification: A survey of methodologies and applications," *IEEE Access*, vol. 6, pp. 36218–36234, 2018.
10. A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach," *Heliyon*, vol. 10, no. 2, 2024.
11. T. Zeevi, R. Venkataraman, L. H. Staib, and J. A. Onofrey, "Monte-carlo frequency dropout for predictive uncertainty estimation in deep learning," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–5, 2024.
12. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019.
13. A. Rajkumar, M. Ganesan, and R. Lavanya, "Arrhythmia classification on ECG using deep learning," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, pp. 365–369, 2019.
14. M. Rashed-Al-Mahfuz, M. A. Moni, S. M. S. Islam, S. Berkovsky, M. Khushi, J. M. W. Quinn, *et al.*, "Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions," *Biomed. Eng. Lett.*, vol. 11, no. 2, p. 147, 2021.
15. S. Ikram, A. Ikram, H. Singh, M. D. A. Awan, S. Naveed, I. de la Torre D'íez, H. F. Gongora, and T. C. Montero, "Transformer-based ECG classification for early detection of cardiac arrhythmias," *Front. Med.*, vol. 12, p. 1600855, 2025.
16. Z. Zhao, "Transforming ECG diagnosis: An in-depth review of transformer-based deep learning models in cardiovascular disease detection," 2023, arXiv:2306.01249. [Online]. Available: <https://arxiv.org/abs/2306.01249>
17. Z. Huang, M. Liang, J. Qin, S. Zhong, and L. Lin, "Understanding self-attention mechanism via dynamical system perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1412–1422, 2023.
18. G. Silva, P. Silva, G. Moreira, V. Freitas, J. Gertrudes, and E. Luz, "A systematic review of ECG arrhythmia classification: Adherence to standards, fair evaluation, and embedded feasibility," 2025, arXiv:2503.07276. [Online]. Available: <https://arxiv.org/abs/2503.07276>
19. R. Baldoni, L. Montanari, and M. Rizzuto, "On-line failure prediction in safety-critical systems," *Future Gener. Comput. Syst.*, vol. 45, pp. 123–132, 2015.
20. S. C. Fanni, M. Febi, G. Aghakhanyan, and E. Neri, "Natural language processing," in *Introduction to Artificial Intelligence*, pp. 87–99, 2023.
21. M. S. A. Dolon, "Deployment and performance evaluation of hybrid machine learning models for stock price forecasting and risk prediction in volatile markets," *Am. J. Scholarly Res. Innov.*, vol. 4, no. 01, pp. 287–319, 2025.