A novel approach to classification of ECG arrhythmia types with latent ODEs

Angelina Yan[1] **Matt L. Sampson**[2] **Peter Melchior**[2]
[1]Columbia University [2] Princeton University
ay2672@columbia.edu
{matt.sampson, peter.melchior}@princeton.edu

Abstract

12-lead ECGs with high sampling frequency are the clinical gold standard for arrhythmia detection, but their short-term, spot-check nature often misses intermittent events. Wearable ECGs enable long-term monitoring but suffer from irregular, lower sampling frequencies due to battery constraints, making morphology analysis challenging. We present an end-to-end classification pipeline to address these issues. We train a latent ODE to model continuous ECG waveforms and create robust feature vectors from high-frequency single-channel signals. We construct three latent vectors per waveform via downsampling the initial 360 Hz ECG to 90 Hz and 45 Hz. We then use a gradient boosted tree to classify these vectors and test robustness across frequencies. Performance shows minimal degradation, with macro-averaged AUC-ROC values of 0.984, 0.978, and 0.976 at 360 Hz, 90 Hz, and 45 Hz, respectively, suggesting a way to sidestep the trade-off between signal fidelity and battery life. This enables smaller wearables, promoting long-term monitoring of cardiac health.

1 Introduction

Cardiovascular diseases (CVDs) are one of the leading causes of death worldwide [Martin et al., 2025]. Many CVDs manifest in arrhythmias, which are routinely captured by electrocardiograms (ECGs), the standard tool in clinical settings. The morphological characteristics of the ECG waveform reflect normal versus abnormal heart beats that can be classified into distinct categories. Classifying heart beats is the basis for the diagnosis of arrhythmias and underlying heart problems [Acharya et al., 2017].

Traditionally, 12-lead ECGs are considered the gold standard for diagnosis and offer the most spatially comprehensive view of the electrical activities of the heart [Schlant et al., 1992]. However, the 12-lead ECG acquisition process requires trained professionals and provides only a short-term, spot-check measurement. In contrast, modern wearable ECGs (typically single-channel) require no additional setup and do not disrupt daily life [Bouzid et al., 2022]. The shift in public health goals from reactive treatment to proactive prevention has led to a rapid increase in the popularity of wearables. Their ease of use has made them ideal for long-term, continuous monitoring and early detection of paroxysmal or asymptomatic events [Babu et al., 2024]. For instance, wearable patches can be worn for weeks at a time and automatically acquire signals, producing continuous waveforms that can detect abnormalities and deviations from individual baselines not present in short-term 12-lead ECG waveforms [Sana et al., 2020]. This capability is crucial for closing the diagnostic gap in detecting infrequent or transient cardiac arrhythmias.

Despite their advantages, wearable ECGs often suffer from low sampling frequencies in the interest of battery conservation. This makes accurate morphology analysis from wearable data challenging

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from time series for health.

since low sampling frequencies compromise signal fidelity [Kwon et al., 2018]. In this work, we present an end-to-end classification pipeline robust to sampling frequency variations. This is achieved by training a latent ODE on high-frequency 360 Hz ECG signals, which handles both potential signal irregularity and noise. It models the underlying continuous signal as a differential equation in latent space [Rubanova et al., 2019]. As part of this generative modeling process, the latent ODE trains an encoder that maps each continuous time-series into a single latent vector. This encoding process has been shown to be effective for creating automated feature vectors to then be used for classification in scientific domains [Sampson and Melchior, 2025]. To perform arrhythmia classification, the latent feature vectors are then fed into a gradient boosted decision tree [Friedman, 2001]. Importantly, the trained encoder model of the latent ODE is able to create informative latent feature vectors from both low and high frequency data.

Background and related work

Given the immense volume of waveform data produced by wearables, it is infeasible to solely rely on the availability of cardiologists for diagnosis. In addition, manually interpreting subtle visual changes in ECG waveforms is challenging, dependent on physician experience, and subject to interpretative discrepancies [Ansari et al., 2023]. Fortunately, modern computational predictors have shown promise in assisting clinical decision-making. Over the last decade, the field of computational ECG analysis has undergone a paradigm shift from manual feature engineering to end-to-end deep learning models with fully automated feature extraction [Montenegro et al., 2022]. Recent work has demonstrated that modern deep learning neural network (DNN) models can not only assist but even rival or outperform human experts in the detection and classification of arrhythmias from 12-lead ECGs and single-lead ambulatory ECGs [Ribeiro et al., 2020, Hannun et al., 2019]. We present an alternative approach in end-to-end modeling, where we automate the creation of the feature vectors yet allow for a wide variety of downstream classification algorithms to be used.

2 Methodology

Here we describe the training dataset as well as our multi-step classification routine. We report further details on model architecture and training routines for all models in Appendix A.

2.1 Training data

We obtained the ECG heart beat signals from the widely used open access PhysioNet MIT-BIH Arrhythmia Database [Moody and Mark, 2001, Goldberger et al., 2000]. The database consists of 48 two-channel ECG recordings taken from 47 individuals at a sampling rate of 360 Hz. For this work, we chose the modified limb lead II (MLII) data for each recording because many wearables also acquire single-lead ECG data [Bouzid et al., 2022]. We performed a denoising and peak matching routine on the raw ECG beats as described in Liu et al. [2020]. We follow the Advancement of Medical Instrumentation (AAMI) standard for the classification of each beat into one of five classes: normal (N), ventricular (V), supraventricular (S), fusion of normal and ventricular (F), and unknown beats (Q). A total of 88887 ECG beats were extracted from this dataset. For the classification testing, we perform a naive downsampling of the ECG data by taking every nth entry to achieve the desired downsampled frequency. The downsampled data is then used for classification by our latent-ODE/decision tree pipeline shown in Algorithm 1.

2.2 Path-minimized latent ODEs

We train a latent ODE [Rubanova et al., 2019] on the ECG time-series data in the MIT-BIH dataset. Specifically, we use a path-minimized latent ODE as described in [Sampson and Melchior, 2025], which replaces the common variational component from the loss with an ℓ_2 regularizer that acts to minimize the point-to-point distance within a single latent trajectory. This has been shown to improve generative fidelity and inference performance of classifiers trained on the latent encodings. The latent ODE modeling has two goals: to provide a generative model of the full ECG dataset and to perform a robust encoding of each time-series into a low-dimensional latent vector $\mathbf{z}_0$, which we then use for classification.

2.3 Gradient boosted decision tree

To perform the arrhythmia classification, we use a gradient boosted decision tree (GBDT) [Friedman, 2001]. This classifier is trained on the encoded latent vectors, $\mathbf{z}_0$, of the original ECG segments and their class labels. A benefit of this multi-step approach is that by training the GBDT on the latent vectors instead of features directly from the ECG itself, we leverage the latent ODE model to produce a latent vector that closely matches a higher quality signal, even if the original ECG signal were of lower quality due to noise, sampling frequency, or other forms of corruption.

We perform a synthetic minority over-sampling technique (SMOTE) [Chawla et al., 2002] on the latent feature vectors such that we have the same number of samples in each of the 5 classes during the *training* of the GBDT classifier. Our latent feature vectors are conditionally sampled from the latent ODE encoder. During *testing*, we take n random latent vectors for each ECG and select the mode of the n random samples to be the final classification.

2.4 Full classification pipeline

We present pseudocode for the full classification routine in Algorithm 1.

Algorithm 1 Arrhythmia Classification Pipeline

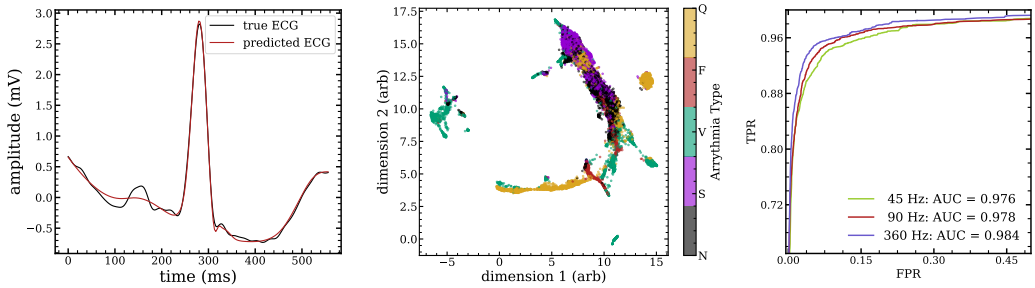
Input: ECG time-series vector (x, t); trained latent ODE; trained gradient boosted decision tree classifier (GBDT); ensemble size n, a random seed sampling
Return: Final class prediction $y^\star$

```
function PREDICT(x, t, n)
    for i = 1 to n do
        z_{0,i} ← LATENTODE(x, t, seed_i)          ▷ Encoding step
        ŷ.append(GBDT(z_{0,i}))                    ▷ GBDT predicts one label per z_{0,i}
    end for
    y* ← mode(ŷ)                                   ▷ Final class is the majority vote over {ŷ_i}_{i=1}^n
    return y*
end function
```

3 Experiments and Results

We perform a set of arrhythmia classification experiments on the MIT-BIH dataset using our latent ODE to GBDT classifier routine. We split our dataset into training/validation/test sets with a 70/15/15 split, respectively. All results reported come from the model with the best validation performance re-evaluated on the test set.

3.1 Latent ODE performance



a) b) c)

Figure 1: **a)** Reconstruction of a randomly sampled ECG timeseries with the latent ODE prediction in red and the true signal in black. **b)** UMAP of the latent feature vectors from the test set of the BIH-MIT data, with colors indicating the arrhythmia class. **c)** AUC-ROC curves from the GBDT classifier based on latent ODE encodings of ECG curves sampled at 360, 90, and 45 Hz.

We present a sample result from the trained latent ODE in panel **a)** of Figure 1. By using only 45 latent dimensions, we see excellent agreement between the true (black) and sampled (red) ECG curve, indicating the model's ability to learn the underlying structure of the ECG signals, while ignoring potential minor signal corruptions. We present additional samples in Appendix B.

In panel **b)** we show a UMAP [McInnes et al., 2018] of the test set of encoded ECG timeseries vectors, where each point is colored by the arrhythmia class. We see a clear structure in the projection with similar class labels clustering together in this reduced latent space. This give confidence that the latent ODE is providing meaningful encodings for the different types of ECG signals observed.

3.2 Classification results

We show the classification results in Table 1, displaying accuracy, precision, recall, and F1 score for all 5 classes. We report the classification results from the ECG series at sampling frequencies of 360 Hz, 90 Hz, and 45 Hz. We also show a set of one-vs-all AUC-ROC plots in panel **c)** of Figure 1 for all three frequencies. For the multiclass classification setting, predicted class labels were determined by selecting the class with the highest predicted probability from the gradient boosted decision tree. Performance metrics (accuracy, precision, recall, and F1) were computed from these predicted labels using `scikit-learn`'s default implementations with the prediction threshold set to 0.5.

Table 1: Classification results for ECG data sampled at 360 Hz, 90 Hz, and 45 Hz. We show the per-class accuracy, precision, recall and F1 scores. We also report the macro-averaged results for each sampling frequency in the bottom row.

		Accuracy (%)			Precision			Recall			F1		
class	count	360 Hz	90 Hz	45 Hz	360 Hz	90 Hz	45 Hz	360 Hz	90 Hz	45 Hz	360 Hz	90 Hz	45 Hz
N	10988	98.0	97.7	97.7	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
S	327	75.2	70.5	70.1	0.69	0.66	0.64	0.75	0.70	0.71	0.72	0.68	0.67
V	918	93.9	93.6	92.6	0.93	0.90	0.91	0.94	0.94	0.93	0.93	0.92	0.92
F	99	72.3	69.6	60.8	0.72	0.75	0.65	0.72	0.70	0.61	0.72	0.72	0.63
Q	1001	95.4	94.3	93.3	0.95	0.93	0.92	0.95	0.94	0.93	0.95	0.94	0.93
macro-avg	–	87.0	85.9	82.9	0.85	0.84	0.82	0.87	0.85	0.83	0.86	0.85	0.82

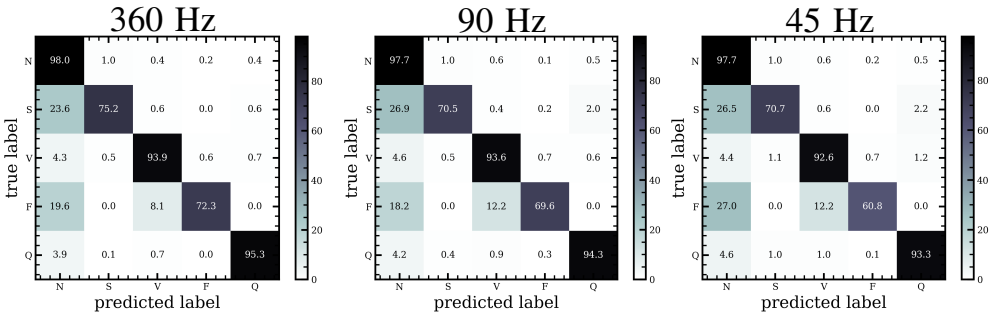


Figure 2: Normalized confusion matrices from the GBDT on the ECG data at samping frequencies of 360 Hz, 90 Hz, and 45 Hz from left to right, respectively.

In Figure 2 we show normalized confusion matrices for the GBDT performance on the test data for the 360 Hz, 90 Hz, and 45 Hz trials, respectively. From this we see the minority classes S and F are most commonly mislabelled as N, with this issue becoming particularly clear in the 45 Hz trial.

4 Discussion

The primary advantage of our classification pipeline is its inherent robustness to the challenges posed by real-world ECG data, specifically sparsity, incompleteness, and low sampling frequency. We empirically demonstrate this robustness in Table 1, which details our model's per-class classification metrics (accuracy, precision, recall, and F1 scores) across data sampled at 360 Hz, 90 Hz, and 45 Hz as well as the macro-averaged results for each metric. The results show a minimal degradation in performance as the sampling rate decreases, indicating our pipeline's ability to construct high-quality feature vectors even with a significantly reduced data stream.

We note that the difference (for all sampling frequencies) between the relatively high AUC-ROC scores of 0.984, 0.978, and 0.976 and the lower macro-averaged F1 scores in Table 1 of 0.86, 0.85, and 0.82 hints that our model still suffers from class imbalance issues, with the relatively poor performance on the minority classes reducing the macro-averaged F1 scores. While our pipeline shows promise as a novel method for treating irregular and low frequency ECG data, its current performance is limited by the relatively small size of the MIT-BIH dataset. In future work, we hope to improve our results by training on larger datasets with a greater representation of the undersampled S and F classes. The main quantitative difference we note here is: the accuracy for the S class decreases from 75.2% at 360 Hz to 70.1% at 45 Hz, and the accuracy for the F class decreases from 72.3% at 360 Hz to 60.8% at 45 Hz. Larger datasets as well as alternate oversampling techniques may help alleviate these issues in future efforts.

5 Limitations and future work

We point out this study represents preliminary work in this area. A clear limitation is the lack of practical testing of this new classification pipeline on edge devices, which would allow for accurate estimates of inference and memory costs. Another limitation is the relatively limited dataset used. We perform all testing on the publicly available MIT-BIH dataset, which contains significant class imbalance, a total of only 47 individuals, and is not a wearable dataset. In future work, we plan on exploring the performance of this pipeline over a larger wearable dataset containing raw ECG signals.

6 Conclusion

By combining a latent ODE model with a decision tree classifier, we demonstrate that high-quality classification is achievable at lower frequencies. Our work enables the use of smaller, less power-hungry sensors with longer battery life and the development of smaller, more comfortable wearable devices. Extending battery life also increases the length of continuous monitoring that can be provided by a wearable, which catches more infrequent or transient pathologies. Taken together, these improvements could encourage more patients to adopt long-term monitoring, aligning with the modern attitude shift toward a proactive approach to public health.

References

U Rajendra Acharya, Hamido Fujita, Muhammad Adam, Oh Shu Lih, Vidya K Sudarshan, Tan Jen Hong, Joel EW Koh, Yuki Hagiwara, Chua K Chua, Chua Kok Poo, et al. Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ecg signals: A comparative study. *Information Sciences*, 377:17–29, 2017.

Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, and Erchin Serpedin. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, 14:1246746, 2023.

Mohan Babu, Ziv Lautman, Xiangping Lin, Milan HB Sobota, and Michael P Snyder. Wearable devices: implications for precision medicine and the future of health care. *Annual review of medicine*, 75(1):401–415, 2024.

Zeineb Bouzid, Salah S Al-Zaiti, Raymond Bond, and Ervin Sejdić. Remote and wearable ecg devices with diagnostic abilities in adults: a state-of-the-science scoping review. *Heart Rhythm*, 19 (7):1192–1201, 2022.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Ohhwan Kwon, Jinwoo Jeong, Hyung Bin Kim, In Ho Kwon, Song Yi Park, Ji Eun Kim, and Yuri Choi. Electrocardiogram sampling frequency range acceptable for heart rate variability analysis. *Healthcare informatics research*, 24(3):198–206, 2018.

Shu Liu, Jie Shao, Tianjiao Kong, and Reza Malekian. Ecg arrhythmia classification using high order spectrum and 2d graph fourier transform. *Applied Sciences*, 10(14):4741, 2020.

Seth S Martin, Aaron W Aday, Norrina B Allen, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Nisha Bansal, Andrea Z Beaton, et al. 2025 heart disease and stroke statistics: a report of us and global data from the american heart association. *Circulation*, 151(8):e41–e660, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Larissa Montenegro, Mariana Abreu, Ana Fred, and Jose M Machado. Human-assisted vs. deep learning feature extraction: an evaluation of ecg features extraction methods for arrhythmia classification using machine learning. *Applied Sciences*, 12(15):7404, 2022.

G.B. Moody and R.G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Matt L Sampson and Peter Melchior. Path-minimizing latent odes for improved extrapolation and inference. *Machine Learning: Science and Technology*, 6(2):025047, 2025.

Furrukh Sana, Eric M Isselbacher, Jagmeet P Singh, E Kevin Heist, Bhupesh Pathik, and Antonis A Armoundas. Wearable devices for ambulatory cardiac monitoring: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 75(13):1582–1592, 2020.

Robert C Schlant, Robert J Adolph, JP DiMarco, Leonard S Dreifus, Marvin I Dunn, Charles Fisch, Arthur Garson Jr, LJ Haywood, Herbert J Levine, and John A Murray. Guidelines for electrocardiography. a report of the american college of cardiology/american heart association task force on assessment of diagnostic and therapeutic cardiovascular procedures (committee on electrocardiography). *Circulation*, 85(3):1221–1228, 1992.

A Model details

A.1 Latent ODE details

Our latent ODE-RNN architecture is implemented in `jax` and `equinox` [Bradbury et al., 2018, Kidger and Garcia, 2021]. We use a feed forward neural network to model our ODE function as in [Rubanova et al., 2019]. We use Tanh activation functions with 2 layers and a width of 50. We use ODE solvers from the `diffrax` package [Kidger, 2021], specifically the 5th order Tsit5() solver, Tsitouras 5/4 method (5th order Runge-Kutta), with adaptive steps and an initial dt = 0.001. We use 45 hidden and 45 latent dimensions and train for 50000 steps. We follow the implementation from Sampson and Melchior [2025] for the path-regularization. The training time is approximately 2 hours on a single NVIDIA A100.

A.2 Boosted classifier details

We use a gradient boosted classifier from the `scikit-learn` package [Pedregosa et al., 2011]. We train with 1000 trees and a max depth of 8. We find increasing the depth tends to bias the classifier towards the majority class during validation even with equal sized training classes during training (from SMOTE oversampling).

B Additional results

We show additional results here for both the classification performance and the latent ODE modeling.

We show some extra random samples of reconstructed ECG time-series from the trained latent ODE in Figure 3. We can see that the trained model accurately reconstructs a wide variety of ECG shapes.

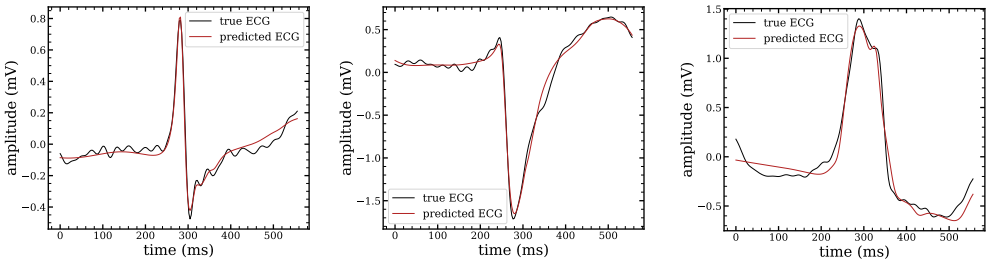


Figure 3: Additional reconstructions of randomly sampled ECG timeseries with the latent ODE prediction in red and the true signal in black.