Quantifying the impact of Signal Simplification, Data Quantity, and Task Difficulty on Vision Transformer Performance for Electrocardiogram Rhythm Classification

Jarod P. Hartley P.J.HARTLEY@SWANSEA.AC.UK and **W. Joseph MacInnes** WILLIAM.MACINNES@SWANSEA.AC.UK
Swansea University

Editors: Under Review for MIDL 2026

Abstract

This paper investigates the interplay between signal simplification, data quantity and task difficulty in the context of electrocardiography rhythm classification using vision transformer models. Recognising the complexity and variability inherent in ECG signals, we examine how applying a simple pre-processing technique aimed at enhancing human readability, through noise reduction, affects the accuracy and robustness of machine learning models. We also consider the broader implications for model development, particularly in relation to the challenges posed by real-world ECG data. Given the inherent complexity and diversity of ECG diagnoses, it is often impractical to gather substantial amounts of data for every potential diagnosis. By assessing how different dataset sizes affect performance, we seek to understand the extent to which vision transformers rely on data quantity. This is particularly important given the limitations of real-world datasets and its potential impact on automated diagnostic systems. This paper further examines the scalability of ECG classification by employing a dataset encompassing ten distinct conditions. While previous research has demonstrated success in scenarios involving a limited number of conditions, such controlled environments are rarely representative of real-world practice. Therefore, it is crucial to understand how vision transformer models perform when faced with more complex and varied classification tasks and to evaluate their capacity to manage increased diagnostic diversity. Our findings provide insight into optimising ECG classification pipelines with regards to balancing the need for data clarity, quantity, and diagnostic breadth to enable reliable and scalable AI-driven cardiac assessment.

Keywords: Vision Transformer, Signal Simplification, Data Quantity, Task Difficulty, Electrocardiogram, ECG, Machine Learning

1. Introduction

Electrocardiogram signals present a unique challenge in both clinical and computational contexts due to their complexity and the wealth of information they contain. While there have been extensive studies looking at the classification of ECG signals, these often lack data consistency and are rarely reflective of the unique challenges contained within this complex classification task. While achieving high accuracy is crucial, it is equally important to recognise and account for the limitations inherent in real-world data. Understanding how these constraints affect model performance is essential to developing a robust and reliable diagnostic system.

1.1. Electrocardiograms

Electrocardiograms (ECGs) are graphical representations of the heart's electrical activity and are essential tools in both clinical settings and medical research. However, the signals captured by ECG devices are often complex and intricate, requiring considerable expertise to interpret accurately. For human readability, simplification techniques such as noise reduction and artifact removal are employed to enhance clarity (Blinowska and Zygierewicz, 2011).

The ECG recording process begins with the placement of electrodes on the skin at specific points on the chest and limbs to capture the heart's electrical signals (Strauss and Schocken, 2021). These electrodes detect the tiny electrical changes on the skin that arise from the heart during each heartbeat. The signals are then transmitted to an ECG machine, where they are amplified and filtered to remove any external electrical interference.

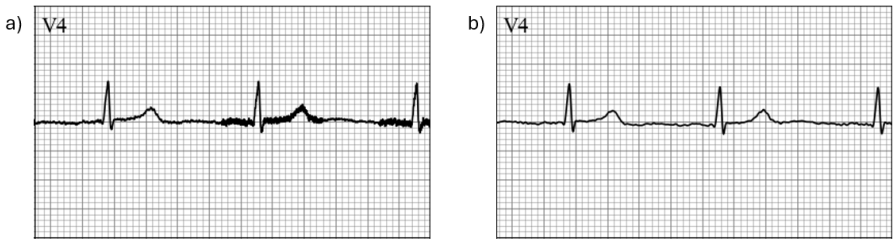

Figure 1: Image showing a raw plotted signal with filtering method applied (a) and the same signal plotted with a Butterworth bandpass filter to supress signal noise (b).

Once the raw signals are captured, they undergo digitisation, converting the analogue signals into a digital format that can be processed by computers (Gacek and Pedrycz, 2011). The digital signals are then subjected to further processing, such as baseline wander removal to correct any drift in the signal, and the elimination of power-line interference, see Figure 1. Advanced algorithms are used to enhance the signal quality and to extract meaningful features, such as the P wave, QRS complex and T wave, which are critical for diagnosing various cardiac conditions.

1.2. Proposal

In this paper we plan to explore the effects of data simplification for human readability on machine learning models. We intend to investigate how a simple signal enhancement technique, such as noise reduction, influences the accuracy and reliability of these models. By comparing models trained on raw versus increasingly simplified ECG signals, we aim to determine the extent to which simplification impacts the classification performance of the models.

The amount of information extractable from an ECG is extensive; however, the broad range of diagnosable conditions and the various settings and in which an ECG can be

recorded results in the classification task being extremely complex. Due to these challenges, acquiring large quantities of data is not always possible. Therefore, we will explore the importance of data quantity by using multiple models trained on progressively smaller datasets and review each model's performance.

Most ECG classification models are trained to classify a small number of conditions (Yildirim et al., 2019; Vo, 2025; Kachuee et al., 2018). While this helps to reduce variance by keeping the classification task simple, it is not reflective of reality as the variety of conditions diagnosable from an ECG is extensive. Therefore, to improve real-world fidelity we will utilise a 10-class dataset to train a model. While this is still far from being truly reflective, it will allow us an insight into how the model could handle a broader scope and an increased task difficulty.

2. Method

2.1. Vision Transformer

The SwinV2 vision transformer (Liu et al., 2021) was chosen due to its previous success classifying the MIMIC-IV-ECG dataset (Hartley et al., 2025). The model utilizes a shifted window approach to limit self-attention computation to non-overlapping local windows. This approach optimises efficiency and performance, especially when handling high resolution images. The model neatly divides the image into patches and windows. We chose to divide the images into 16x16 pixel patches and utilised a window size of 24 due to the divisions high overlap with the structure of a standard ECG, see Figure 2. With this configuration, each window corresponds to a single lead.

2.2. Dataset

The MIMIC-IV-ECG dataset is comprised of approximately 800,000 diagnostic ECGs that were collected from nearly 160,000 distinct patients (Gow et al., 2023). Each ECG is recorded using 12 leads, spans 10 seconds, and is sampled at a frequency of 500 Hz. This extensive dataset encompasses all ECGs for patients present in the wider MIMIC-IV Clinical Database (Johnson et al., 2023). The dataset was chosen due to its extensive quantity of data that covers a substantial number of cardiac conditions and its high level of consistency.

We selected the four most prominent diagnoses in the MIMIC-IV-ECG dataset — Sinus Rhythm, Sinus Bradycardia, Sinus Tachycardia, and Atrial Fibrillation — and extracted a total of 120,000 ECG recordings corresponding to these conditions. Each condition contributed 25,000 images to the training set and a further 5,000 images to the validation set. These ECG recordings were then converted into image format for training on the SwinV2 vision transformer, as illustrated in Figure 2. This curated dataset was designated as c04-s120, indicating four conditions and a total of 120,000 samples.

2.3. Signal Simplification

The Simple Moving Average (SMA) algorithm is a simple smoothing algorithm used to remove noise and minor artifacts from signals (Oppenheim and Schafer, 2010). It is calculated by computing the unweighted average of the most recent observations within a

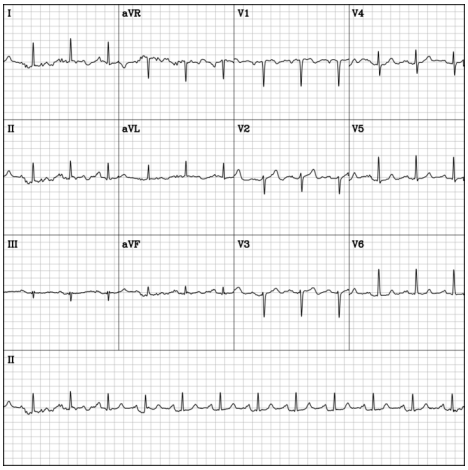

Figure 2: A standard ECG in grayscale, displayed across a grid background at 1536x1536 resolution. This image was part of the dataset used to train the models in the study and serves as an example of the images used during training.

specified window size. To simulate signal simplification, we used varying window sizes. As seen in Figure 3, the greater the window size (w), the simpler and more distorted the lead.

This method allowed us to incrementally mask features and, therefore, would allow us to observe model performance at observable thresholds. While the SMA algorithm is not used directly in normal ECG processing pipelines, it is a well established signal smoothing algorithm and acts as a computationally cheap and stable low pass filter.

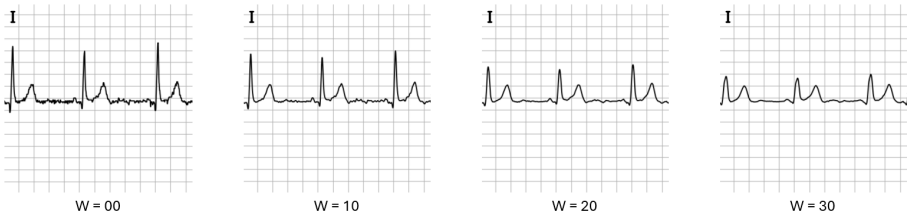

Figure 3: A standard ECG in grayscale, displayed across a grid background at 1536x1536 resolution. This image was part of the dataset used to train the models in the study and serves as an example of the images used during training.

We created four variations of the c04-s120 dataset, the variation w00 represented no signal simplification and the variations w10, w20 and w30 represented the number of neighbouring nodes used to determine the average, see Table 1 and Figure 3. We decided to limit the window size to 30 as extreme levels of masking of the P wave, QRS complex and T wave was present in w30.

Table 1: Table showing a breakdown of all the datasets used for training models

Dataset	Window Size	No. of Classes	Training Data Class
c04-s120-w00	0	4	25,000
c04-s120-w10	10	4	25,000
c04-s120-w20	20	4	25,000
c04-s120-w30	30	4	25,000
c04-s60-w10	10	4	12,500
c04-s30-w10	10	4	6,250
c04-s15-w10	10	4	3,125
c10-s84-w10	10	10	8,200

Each of these models was trained for 80 epochs. Although training for more epochs might have yielded higher accuracies, computation time limited our ability to train further, and we believe 80 epochs produced sufficient data to establish trends.

2.4. Data Count

To examine how data volume affects accuracy, three subsets were derived from the c04-s120-w10 dataset: c04-s60-w10, c04-s30-w10 and c04-s15-w10. The c04-s60-w10 subset contains exactly half the data count of c04-s120-w10, with the training and validation sets reduced to half their original size. Similarly, c04-s30-w10 and c04-s15-w10 comprise a quarter and an eighth, respectively, of the original data, and their training and validation sets were likewise reduced to a quarter and an eighth of the initial count. For more details, see Table 1.

Unlike the c04-s120-w10 model, which was trained for 80 epochs, the c04-s60-w10, c04-s30-w10 and c04-s15-w10 subsets were trained for 120, 160 and 200 epochs respectively. This adjustment was made to investigate whether increasing the number of training epochs could compensate for the reduction in available data. By extending the training duration, we aimed to assess if the models could achieve comparable performance despite having access to fewer samples, thereby providing insight into the interplay between data volume and training strategy.

The selected subset sizes were chosen to provide a clear, logical reduction in data volume, while also reflecting practical and achievable sample counts. To keep the dataset balanced, it is often necessary to make a substantial reduction in the amount of data per class. For example, when increasing the number of conditions from 4 to 10, the data per condition had to be reduced from 25,000 to 8,400. This highlights the challenge inherent in ECG classification: while each condition holds equal significance, sufficient data for training models on every condition may not always be accessible.

2.5. Task Difficulty

To introduce a higher level of task complexity, the c10-s84-w10 dataset was created. Unlike the datasets described in the signal simplification and data count sections, which each featured 4 conditions, c10-s84-w10 included 10 distinct conditions. Since there is an inverse relationship between the number of conditions and the amount of data available for each,

the training data per condition was limited to 8,200 samples, with 200 additional samples per condition being allocated for validation. The c10-s84-w10 model was trained for 80 epochs, in line with the training regimen applied to the other models.

3. Analysis of Results

3.1. Signal Simplification

The highest performing model was c04-s120-w10 which achieved a top accuracy of 96.23%, see Table 2. This was closely matched by model c04-s120-w00, which recorded an accuracy of 96.14%. Model c04-s120-w20 reached an accuracy of 95.69% and, notably, even with substantial smoothing, c04-s120-w30 attained a respectable accuracy of 94.30%. Across all models the precision, recall and F1 score are tightly coupled, indicating a high level of consistent class-wise performance.

Despite model c04-s120-w10 performing slightly better than model c04-s120-w00 in accuracy, model c04-s120-w00 performed slightly better in precision, recall and F1 score, which is an indication of a slightly more balanced classification. Ultimately these differences are minor, and the performance of both models indicates that, for this set of rhythms, there is no significant difference to the model between a raw plotted signal and a slightly smoothed signal.

A more pronounced difference is found between model c04-s120-w10, which achieved the highest accuracy at 96.23%, and model c04-s120-w30, which had the lowest accuracy at 94.30%. While a 2% drop in accuracy may seem small, in practical terms it means that around 1 in 50 patients could receive an incorrect diagnosis, highlighting its possible clinical impact.

This context can also be applied to the top performing model as, even with an accuracy of 96.23%, approximately 1 in 25 patients would be misdiagnosed. This becomes even more notable given that the rhythms under consideration — Sinus Rhythm, Sinus Bradycardia, Sinus Tachycardia, and Atrial Fibrillation — are quite distinct and, even for less experienced clinicians, identifying them is typically straightforward. While the loss in accuracy is likely due to borderline cases, meaning the clinical consequences aren't critical for this set of rhythms, it is still important to contextualise them.

Overall, the models are showing a trend that broadly matches with already established ideas; that a small amount of simplification can improve readability, but excessive simplification can mask import features and harm accuracy. Although the high degree of signal simplification did not drastically affect performance, it still had a noticeable impact. Therefore, while these results do demonstrate the ViTs ability to effectively manage high levels of masking, strict signal processing standards should still be enforced. This is especially important considering the simplicity of the classification task.

3.2. Data Count

Diagnosing an ECG is a complex task that relies on interpreting both local and global features. Clinicians must draw on substantial experience and remain adaptable as they assess and label an ECG. Given the wide range of possible diagnoses and the layered complexity of ECG reports, it is almost impossible to gather large quantities of data for every diagnosis.

Table 2: Table showing the accuracy, precision, recall and F1 score of the best epoch for the models c04-s120-w00, c04-s120-w10 and c04-s120-w20.

Model	Epoch	Accuracy	Precision	Recall	F1
c04-s120-w00	76	96.14	0.965	0.965	0.965
c04-s120-w10	80	96.23	0.964	0.964	0.964
c04-s120-w20	79	95.69	0.959	0.959	0.959
c04-s120-w30	78	94.30	0.943	0.943	0.942

As a result, models trained for ECG classification are often limited to a smaller set of conditions or must utilise imbalanced datasets for training. This makes it crucial to understand the impact of data quantity on a model's ability to accurately interpret and classify ECGs.

To fully understand the impact of data quantity, we took a relatively simple diagnostic problem that looked at the classification of four distinct rhythms — Sinus Rhythm, Sinus Bradycardia, Sinus Tachycardia, and Atrial Fibrillation — and trained four models with different data quantities, ranging from 25,000 units of training data per class to 3,125 units of training data per class, see Table 1 and Table 3.

We initially trained all four models for a total of 80 epochs and saw an immediate trend, see Table 3. The model with the most training data per class, c04-s120-w10, significantly outperformed the other three models and obtained an accuracy of 96.23%. Model c04-s60-w10 followed with an accuracy of 88.43% and subsequently models c04-s30-w10 and c04-s15-w10 were the worst performing models with accuracies of 73.36% and 62.92% respectively. This shows the direct impact of data quantity, the more data, the higher the accuracy.

Table 3: Table showing the accuracy, precision, recall and F1 score of the best epoch, before or at epoch 80, for the models c04-s120 -w10, c04-s60-w10, c04-s30-w10 and c04-s15-w10.

Model	Epoch	Accuracy	Precision	Recall	F1
c04-s120-w10	80	96.23	0.964	0.964	0.964
c04-s60-w10	80	88.43	0.887	0.887	0.886
c04-s30-w10	79	73.36	0.728	0.735	0.729
c04-s15-w10	78	62.92	0.596	0.631	0.583

Although a clear trend did present itself, we wanted to see if the loss in accuracy could be overcome by training the model for additional epochs. Therefore, we trained c04-s60-w10 for an additional 40 epochs, c04-s30-w10 for an additional 80 epochs and c04-s15-w10 for an additional 120 epochs, see Table 4.

On initial observation of Table 4, it becomes apparent that although the accuracy of each model did improve, the improvement was superficial and ultimately had a minor impact on the trend. This can be seen with the model c04-s60-w10 as, although the additional 40 epochs did help c04-s60-w10 draw closer, overall c04-s120-w10 still outperformed its

accuracy by more than 3%. The models c04-s30-w10 and c04-s15-w10 did not perform any better as even with the additional epochs, neither broke the 90% mark, with c04-s15-w10 failing to get over the 80% mark.

Table 4: Table showing the accuracy, precision, recall and F1 score of the best epoch for the models c04-s120-w10, c04-s60-w10, c04-s30-w10 and c04-s15-w10.

Model	**Epoch**	**Accuracy**	**Precision**	**Recall**	**F1**
c04-s120-w10	80	96.23	0.964	0.964	0.964
c04-s60-w10	118	93.17	0.935	0.935	0.935
c04-s30-w10	160	85.80	0.863	0.862	0.859
c04-s15-w10	194	77.96	0.780	0.748	0.730

Although the overall trend remained consistent, each model still demonstrated some improvement. To provide a comprehensive evaluation of each model's performance, we examined the training and testing loss at the best performing epoch for each model, see Table 5.

Table 5: Table showing the loss difference at the best epoch for the models c04-s120-w10, c04-s60-w10, c04-s30-w10 and c04-s15-w10.

Model	**Epoch**	**Test Loss**	**Train Loss**	**Difference**
c04-s120-w10	80	0.2315	0.0487	0.1828
c04-s60-w10	118	0.3200	0.0537	0.2663
c04-s30-w10	160	0.5231	0.0592	0.4639
c04-s15-w10	194	0.7034	0.0615	0.6419

Looking at the loss figures for the models not only helps us contextualise the improvements but also helps to highlight the models current training stage. By examining Table 5, we can determine that all four models are in the mid-training phase, meaning that they have yet to settle and still have the potential for improvement. However, where model c04-s120-w10, which has a test-train loss difference of 0.1828, is nearing the end-training phase, c04-s15-w10 has barely reached the mid-training phase with a test-loss difference of 0.6419. This tells us that the difference in accuracy previously observed may be due to the models being in different stages of training, and the number of epochs required for a model with less data to perform well is significantly higher.

3.3. Task Difficulty

The classification of ECG rhythms is far from a simple task. This is largely due to the high number of classes and the small details that determine each diagnosis. It is for this reason that ECG classification models are often trained with a small number of conditions. However, this is not reflective of reality where a high quantity of labels and diagnosis exist.

Therefore, in this section we look at how increasing the task difficulty, by increasing the number of conditions the model has to consider from 4 to 10, affects model performance.

This step-up in difficulty from a small number of conditions in a controlled setting is reflective of a medical students' transition from the classroom to hospital. However, while a medical student may struggle to adapt, the ViT has handled the increased difficulty well and obtained an accuracy of 85.60%, highlighting its flexibility and ability to handle the increased difficulty, see Table 6.

Table 6: Table showing the loss difference at the best epoch for the models c04-s120-w10, c04-s60-w10, c04-s30-w10 and c04-s15-w10.

Model	**Epoch**	**Accuracy**	**Precision**	**Recall**	**F1**
c04-s120-w10	80	96.23	0.964	0.964	0.964
c10-s84-w10	80	85.60	0.844	0.845	0.844

Although the model fails to obtain the same level of accuracy as model c04-s120-w10, which obtained an accuracy of 96.23%, the accuracy relative to chance is still impressive. As model c04-120-w10 only has four conditions, the chance of randomly selecting the correct class was 25%, this means that an accuracy of 96.23% is roughly 3.85 times greater than chance. Model c10-s84-w10, on the other hand, managed to obtain 85.60% with a chance of 10%. This means that it managed to obtain a result 8.56 times greater than chance. While this does not make the loss in accuracy justifiable, it does give it some important context that we must consider when evaluating the model's performance.

It is also important to remember that due to the increased condition count, a lower data per class had to be used for training. From the previous section we have already established that decreasing the training data quantity per class directly affects the model's ability to learn. While this can be overcome by increasing the number of epochs the model is trained for, it does not fully compensate for the data loss, see Table 3 and Table 4.

Model c10-s84-w10 had 8,200 units of data, see Table 1, this falls roughly in the middle of models c04-s60-w10 and c04-s30-w10 which had 12,500 and 6,250 units of training data per class respectively. By examining Table 3, we can see that model c04-s60-w10 obtained an accuracy of 88.43% and model c04-s30-w10 obtained an accuracy of 73.36%, meaning that model c10-s84-w10, which falls in the middle of these two models in terms of training data per class, outperformed its expected accuracy. This should be due to the fact that even though c10-s84-w10 had a lower training data count per class compared to c04-s60-w10, the overall data count was higher due to the increased class count. This shows that while increasing the class count does affect the accuracy, it can also result in a slight uptick in performance due to the overall increase in data.

Although model c10-s84-w10's precision, recall, and F1 are closely aligned, they diverge slightly from the overall accuracy, suggesting potential class-wise variability, see Table 6. This may just be statistical noise given the modest validation size of only 2,000 images; however, further investigation into per-class performance is required, see Table 7.

Table 7 immediately highlights the inconsistent performance across classes. Although most classes perform relatively well and achieve an accuracy of over 90%, the overall accu-

Table 7: Table showing the loss difference at the best epoch for the models c04-s120-w10, c04-s60-w10, c04-s30-w10 and c04-s15-w10.

Class	**Rhythm**	**Accuracy**
0	Atrial Fibrillation	78.0
1	Atrial Fibrillation with Rapid Ventricular Response	95.0
2	Sinus Arrhythmia	78.0
3	Sinus Bradycardia	93.0
4	Sinus Rhythm	82.0
5	Sinus Rhythm with 1st Degree A-V Block	69.0
6	Sinus Rhythm with borderline 1st Degree A-V Block	72.0
7	Sinus Rhythm with PAC(s)	90.0
8	Sinus Tachycardia	95.0
9	Ventricular Pacing	93.0

racy is affected a handful of poorly performing classes. The most notable of which is class 5 and 6, which achieve an accuracy of only 69% and 72% respectively. Although this is notably lower compared to the other rhythms, it does make sense when mapping the classes to the rhythms. Class 5 maps to the arrythmia "Sinus Rhythm with 1st Degree AV-block" and class 6 maps the arrythmia "Sinus Rhythm with borderline 1st Degree AV-block", which are two very similar rhythms. Due to the high overlap of the two conditions, it is understandable that misclassification can occur and it is possible that if the model were allowed to train for longer, stricter classification requirements between the two would arise.

The other poorly performing classes were 0, 2, and 4, corresponding to Atrial Fibrillation, Sinus Arrhythmia, and Sinus Rhythm respectively. While the lower accuracy of class 2 can be explained by the subtle nature of Sinus Arrhythmia, it is surprising to see classes 0 and 4 under perform, given the transformer's previous success in classifying both Atrial Fibrillation and Sinus Rhythm.

Several factors could contribute to this unexpected result, but the most plausible explanation is the presence of noise within the dataset and the limited training time. Furthermore, Sinus Rhythm has significant overlap with other conditions in the dataset, which can complicate classification. It is also important to recognise that the class labels were assigned using machine measurements, meaning the model's accuracy is ultimately dependent on the machine's ability to correctly categorise the ECGs. Consequently, any limitations or errors in the initial machine labelling will inevitably impact the overall performance of the classification model. Unfortunately, this is a well-documented issue within the field of medical imaging classification and while steps can be taken to limit data poisoning, it is near impossible to fully eliminate it.

Overall, the challenges presented by the classification of ECGs are extremely difficult to navigate, even with a thorough understanding of the problem, designing a reliable and robust diagnostic system is no simple task. However, by understanding the problems and employing effective training strategies along with strict data requirements, high accuracy classification is possible.

Acknowledgments

We acknowledge the support of the Supercomputing Wales (SCW) project, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government.

References

K.J. Blinowska and J. Zygierewicz. *Practical Biomedical Signal Analysis Using MATLAB®*. Series in Medical Physics and Biomedical Engineering. Taylor & Francis, 2011. ISBN 9781439812020. URL https://books.google.co.ke/books?id=uizRO7qiPxIC.

Adam Gacek and Witold Pedrycz. *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 0857298674.

Benjamin Gow, Tom Pollard, Laura A. Nathanson, Alistair Johnson, Benjamin Moody, Carla Fernandes, Nathan Greenbaum, Jonathan W. Waks, Pooyan Eslami, Thomas Carbonati, Akshay Chaudhari, Eric Herbst, Daniel Moukheiber, Scott Berkowitz, Roger Mark, and Steven Horng. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0), 2023. URL https://doi.org/10.13026/4nqg-sb35. RRID:SCR_007345.

P. J. Hartley, J. Edwards, E. Akinola, and W. J. MacInnes. Vision transformers for interpreting ecg diagrams. In *Artificial Intelligence in Healthcare: Second International Conference, AIiH 2025, Cambridge, UK, September 8–10, 2025, Proceedings, Part II*, page 396–405, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-00655-4. doi: 10.1007/978-3-032-00656-1_29. URL https://doi.org/10.1007/978-3-032-00656-1_29.

Alistair Johnson, Leo Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 2.2), 2023. URL https://doi.org/10.13026/6mm1-ek67. RRID:SCR_007345.

Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, page 443–444. IEEE, June 2018. doi: 10.1109/ichi.2018.00092. URL http://dx.doi.org/10.1109/ICHI.2018.00092.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. *CoRR*, abs/2111.09883, 2021. URL https://arxiv.org/abs/2111.09883.

Alan V. Oppenheim and Ronald W. Schafer. *Discrete-time Signal Processing*. Pearson, The University of Michigan, 2010. ISBN 0131988425, 9780131988422.

David G. Strauss and Douglas D. Schocken. *Marriott's Practical Electrocardiography, 13e*. Lippincott Williams & Wilkins, a Wolters Kluwer business, 01 2021. ISBN 978-1-496397-45-4.

Thien Nhan Vo. Heart rate classification in ecg signals using machine learning and deep learning, 2025. URL https://arxiv.org/abs/2506.06349.

Ozal Yildirim, Ulas Baran Baloglu, Ru-San Tan, Edward J. Ciaccio, and U. Rajendra Acharya. A new approach for arrhythmia classification using deep coded features and lstm networks. *Computer Methods and Programs in Biomedicine*, 176:121–133, 2019. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2019.05.004. URL https://www.sciencedirect.com/science/article/pii/S0169260718314329.

Appendix A. Performance Metrics for all Models

Table 8: Table showing the performance metrics for all the models trained. These metrics include the Accuracy, Precision, Recall, F1 score, Test Loss and Train Loss at each models best performing epoch.

Model	Epoch	Accuracy	Precision	Recall	F1	Test Loss	Train Loss
c04-s120-w00	76	96.14	0.965	0.965	0.965	0.2396	0.0491
c04-s120-w10	80	96.23	0.964	0.964	0.964	0.2315	0.0487
c04-s120-w20	79	95.69	0.959	0.959	0.959	0.2804	0.0494
c04-s120-w30	78	94.30	0.943	0.943	0.942	0.2882	0.0514
c04-s60-w10	118	93.17	0.935	0.935	0.935	0.3200	0.0537
c04-s30-w10	160	85.80	0.863	0.862	0.859	0.5231	0.0592
c04-s15-w10	194	77.96	0.780	0.748	0.730	0.7034	0.0615
c10-s84-w10	80	85.60	0.844	0.845	0.844	0.5538	0.0677