Machine Learning for Healthcare

Avishag Nevo

October 2024 https://github.com/avishagnevo/VAE-ECG

The Research

The core question guiding this research is: *Can artificial intelligence recognize a patient's sex identity from ECG records by learning how to compress and reconstruct these signals?*

To explore this, I undertook a multi-faceted approach involving the extraction, processing, and analysis of ECG data, followed by building and training generative models aimed at minimizing reconstruction error and KL divergence. My objective was to determine whether an unsupervised model, trained using ECG data from one sex, could generalize to data from the other sex, thereby suggesting that there's latent representational differences between male and female ECG signals learned by the model.

This experimental project was full of challenges. Training several unsupervised machine learning models, on a 12-lead ECG dataset with limited size, necessitated iterative experimentation and examination, which required carefull decision making. Throughout the process, architectural components and hyperparameters were adjusted thoroughly, learning from each iteration, and refining the model design to better handle the data.

The hypotheses formulated for this research were:

$$H_0 : \text{Reconstruction Error}_{ID} = \text{Reconstruction Error}_{OOD}$$

$$H_1 : \text{Reconstruction Error}_{ID} \prec \text{Reconstruction Error}_{OOD}$$

In this context, I initially assumed that ECG records from males and females were representative of each other. Under the null hypothesis, a generative model trained on data from one group (In-Distribution, or ID) should generalize to the other group (Out-Of-Distribution, or OOD) during testing, implying that the distribution of reconstruction errors should be similar for both populations.

However, the alternative hypothesis suggests a disparity between the two populations, indicating that ECG records from the ID group are easier to reconstruct compared to those from the OOD group. In other words, the reconstruction error for the OOD population is expected to be stochastically larger than that for the ID population, reflecting an inherent difference in the complexity or information contained within the ECG signals of males versus females.

Background and related work

The study by Nishikimi et al. ((2024)) presents an efficient method for generating ECG data by conditioning on cardiac parameters, which connects to my approach of conditioning on age and

multi-hot encoded diagnoses. Their work enhances the model's ability to capture relevant features, aligning with my project's goal of incorporating supplementary information to improve ECG reconstruction.

The review by Hong et al. ((2020)) is crucial to my project, as it examines various deep learning methods applied to ECG data, with a focus on recurrent and convolutional neural networks, both of which I have explored and implemented. The review also discusses challenges in interpretability and scalability, which are issues I've faced. Understanding the open problems in this review helped me anticipate some of the difficulties and guided my architectural decisions.

Choi et al. ((2024)) contribute to my work by investigating how deep learning models can be improved through augmentation techniques, specifically for ECG data. While their focus is on diagnosis, not reconstruction, their use of augmentation resonates with my own preprocessing strategies, particularly dynamic windowing and focusing on critical ECG features. These methods were helpful in making my models more stable in reconstructing signals.

Finally, Jang et al. ((2021)) addresses unsupervised feature learning using a convolutional variational autoencoder for ECG data, which is similar to the architecture I implemented. Their findings support my hypothesis that a convolutional approach is more effective than LSTMs in reconstructing key ECG features, affirming my architectural shift, to be discussed thoroughly later in this report.

The data

Dataset

The dataset used for this study called "Chapman-Shaoxing Dataset", sourced from the *PhysioNet/Computing in Cardiology Challenge 2021*. It is composed of twelve-lead electrocardiogram recordings from 10,646 patients, sampled at a frequency of 500 Hz for 10-second intervals. Each patient record is labeled with a subset of 54 different conditions 2 (such as myocardial infarction and U wave abnormalities), as well as the patient's age 1 and sex.

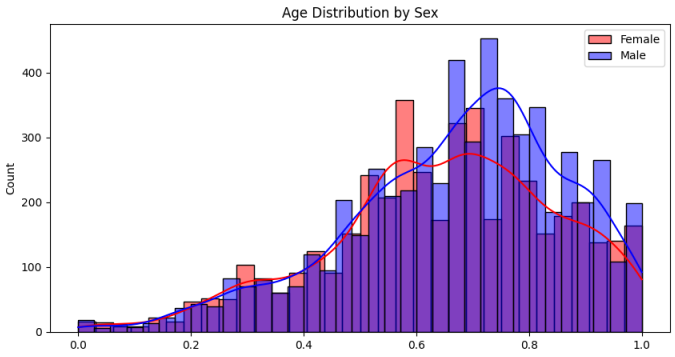

Figure 1: Age distribution by sex, we can observe that age distribution are similar

Of the patients included, 5,956 are males and 4,690 are females. Notably, 17% of the population presents a normal sinus rhythm, while 83% have at least one abnormal condition. The age distri-

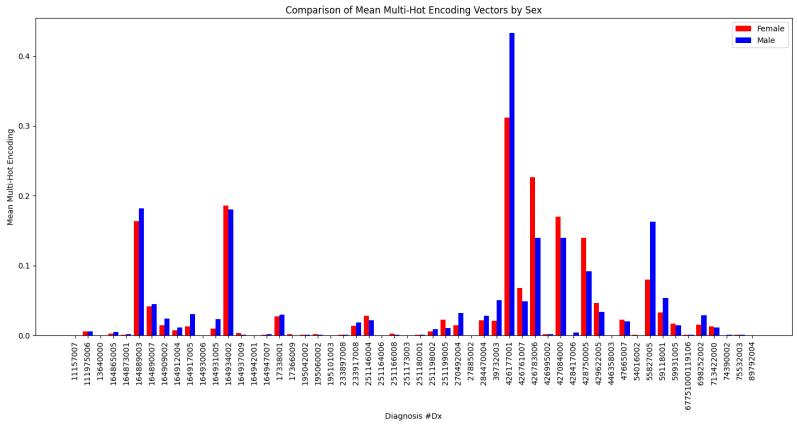

Figure 2: Diagnosis by sex, the largest difference is 0.12 for Dx = 426177001

bution reveals that most patients fall within the age groups of 51–60 years (19.82%), 61–70 years (24.38%), and 71–80 years (16.9%). The amplitude unit for the ECG signals is microvolt (µV).

ECG Recordings

An electrocardiogram is a diagnostic measurement that captures the electrical activity of the heart using electrodes placed on the patient's body. The size, shape, and positioning of different ECG components provide insights into the heart's function and potential pathologies. Below is an example of the ECG header content from the dataset:

```
JS00001 12 500 5000
JS00001.mat 16x1+24 1000.0(0)/mV 16 0 -254 21756 0 I
JS00001.mat 16x1+24 1000.0(0)/mV 16 0 264 -599 0 II
...
# Age: 74
# Sex: Male
# Dx: 426783006
# Rx: Unknown
# Hx: Unknown
# Sx: Unknown
```

How to interpret this information: - The first line indicates that the recording identifier is `JS00001`, containing 12 leads, each recorded at a sampling frequency of 500 Hz, and comprising 5,000 samples. The subsequent lines provide details for each lead, such as the signal resolution, units (`1000/mV`), and an offset applied during data acquisition. Finally, the patient information is displayed, which includes age (74 years), sex (male), and diagnosis (`426783006`), which corresponds to the SNOMED-CT code for sinus rhythm. Other medical history fields (`Rx`, `Hx`, `Sx`) are not available.

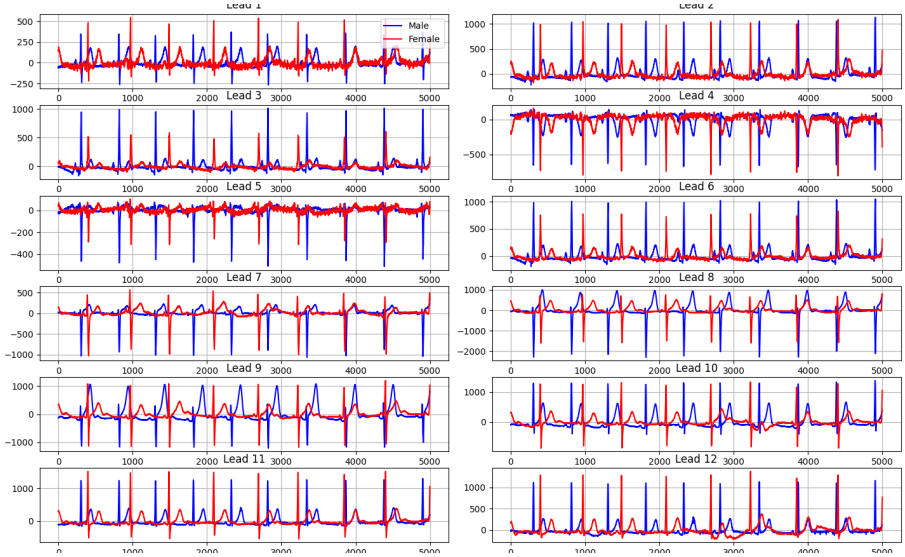

Figure 3: # Age: 31, 35 # Sex: Male, Female # Dx: 426177001, 426177001 raw 12-lead ECG records

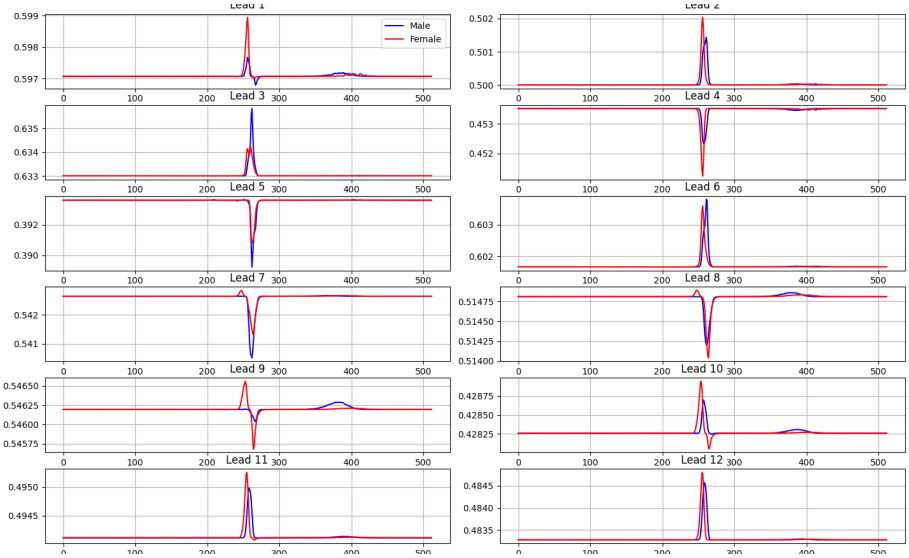

Figure 4: Same as in 3 # Age: 31, 35 # Sex: Male, Female # Dx: 426177001, 426177001 window size of 512 around a QRS complex, standardize by the personal mean and twice the personal standard deviation, raises to the power of 3, scaled by the global dataset min-max values 12-lead ECG records

Method

Preprocessing

In this section, I will chronologically outline the steps I took to preprocess the data and explain the triggers behind each decision.

#Dx Multi-Hot Encoding Since the project is unsupervised, I aimed to leverage not just the ECG signals themselves but also the diagnostic labels (#Dx), which provide context into each record. Each record can have multiple diagnostic labels, so to handle this, I employed a multi-hot encoding strategy to represent multiple diagnoses for each record simultaneously. This encoding was performed once and stored in an HDF5 file, to allow quick access during the iterative model training phase.

Global Min-Max Scaling To ensure the data were on a comparable scale, I applied Min-Max scaling. This was an easy choice for the age field due to its natural bounded range (e.g., 0–100 years), allowing proportionality with the multi-hot encoding component. The ECG signals were first normalized based on the lead-specific parameters provided in the header file (such as ADC gain and baseline). Then, global Min-Max scaling was experimentally applied across all leads to account for amplitude variations due to physiological differences or sensor placement, and also to mitigate the risk of some leads dominating others, or overshadowing the other components, during model training. This linear scaling was performed with respect to the full dataset, calculating global min-max values for both male and female patients for each lead independently, preserving the integrity of the later hypothesis, which compares stochastic differences between sexes. By scaling the ECG signals between 0 and 1, I encouraged smoother gradient descent and, hopefully, more stable training.

	Global Min	**Global Max**
Age	4.0	89.0
Lead 1	-10370.0	6998.0
Lead 2	-10263.0	10263.0
Lead 3	-13342.0	7735.0
Lead 4	-8472.0	10209.0
Lead 5	-5388.0	8301.0
Lead 6	-11712.0	7754.0
Lead 7	-8564.0	7218.0
Lead 8	-14928.0	14069.0
Lead 9	-17080.0	14191.0
Lead 10	-14742.0	19681.0
Lead 11	-16411.0	16802.0
Lead 12	-13469.0	14401.0

Table 1: Global Min and Max Values for Age and ECG Leads

Fixed-length Windowing Considering the periodic nature of ECG signals, I chose to feed only a portion of each record into the model. This was done to reduce computational overhead and improve training efficiency without sacrificing the quality of the input data. Given that the critical features of an ECG signal (the PQRST cycle) repeat with each heartbeat, a window of samples was considered sufficient to capture the essential characteristics of the ECG while reducing sequence length. This approach also allowed for experimentation with different window sizes.

Augmentation with Dynamic Windowing Rather than using a fixed portion of each ECG signal, I implemented a function that dynamically selects a random window of size N from the data each time the model retrieves a sample. This approach simulates natural variations by exposing the model to different segments of the ECG signal during each training iteration. Allowing the window location to vary dynamically within the `__getitem__(self, idx)` function acts as a form of natural data augmentation, enhancing the model's ability to generalize.

Critical Features Focus During the training process, it became clear that the model excelled at reconstructing the smooth, monotonic portions of the ECG signals but struggled with accurately predicting sharp transitions and peaks. This was expected, as the loss minimization tends to favor the larger, consistent sections of the signal, leading to an underrepresentation of the critical transitions, such as the QRS complex or other sharp changes. To address this issue, I experimented with new preprocessing transformations aimed at amplifying the key signal features while reducing the dominance of smoother parts:

- *Standardization by Twice the Standard Deviation:* Instead of relying solely on Min-Max scaling, the signal was standardized by dividing each value by twice its standard deviation. This method helps smooth out large amplitude differences between patients or recordings while preventing extreme values from dominating the training process.
- *Raising the Signal to a Power:* To further reduce the influence of smaller, less relevant peaks and enhance the prominence of extreme values post-standardization, I applied a non-linear transformation by raising the signal to a power (configurable as 3 or 4). This transformation squashes minor local extrema while emphasizing the differences between the smooth parts and the peaks.
- *Hamming Filter for Smoothing:* I experimented with applying a Hamming filter to further smooth the signal, reducing noise while preserving the primary peaks and transitions critical for diagnosing abnormalities.
- *Logarithmic Transformation:* Finally, I tested the logarithmic transformation $\log(1 + x)$ to restore the relative amplitudes of the peaks. This adjustment ensures that peaks maintain their significance after other scaling transformations, providing a more balanced representation of both high and low amplitude regions.
- *Global Min-Max Scaling:* As previously mentioned, Min-Max scaling was applied globally across the entire dataset to maintain proportionality between different leads and components.

Throughout the experiments, I tested various combinations of these transformations, applied in this sequence. This preprocessing pipeline not only addressed the model's shortcomings in reconstructing critical features but also served as a form of data augmentation, increasing the variety of signal representations the model encountered and ultimately improving its generalization abilities.

High-Importance Signal Segments Detection Alongside the dynamic windowing approach, I recognized that focusing the model on the most critical parts of the ECG signal—such as the P-waves, QRS complexes, and T-waves—could significantly improve its ability to reconstruct important features. To achieve this, I incorporated the Pan-Tompkins algorithm, widely used for QRS detection, was implemented alongside simpler peak detection methods like `scipy.signal.find_peaks`.

Once the key features were detected, dynamic windows were extracted around these points. By focusing the model on these high-importance signal segments, the models was able to better reconstruct the more complex parts of the ECG signal, such as the rapid transitions.

Models & Training

In this section, I will chronologically present the model architectures I developed to learn the data reconstruction process, alongside the corresponding training strategies, conclusions, and the adjustments that prompted me to refine the architecture. The generative models were based on a *Variational Autoencoder* framework. The training data, designated as In-Distribution (ID), consists of ECG records without sex labels. I conducted two types of experiments: one with ID as female and Out-of-Distribution (OOD) as male, and vice versa. In each experiment, the ID training data was split into an 80:20 train-test ratio to prevent data leakage. The following diagrams, created using the draw.io platform, visually illustrate the design of each model.

Contextual Recurrent Autoencoder 5 The network was trained to minimize the reconstruction error, experimenting with both the mean and the sum of squared errors as well as absolute errors. Training the network on the full dataset proved difficult due to time constraints and limited computational resources. This led to the introduction of the windowing techniques described earlier, specifically *"Fixed-length Windowing"* and *"Augmentation with Dynamic Windowing"*.

This architecture was tested with several hyper-parameters: $\{hidden\_dim : [2^5, 2^6, 2^7, 2^8], lr : [1e-3, 1e-5], window\_size = [2^6, 2^7, 2^8, 2^9, 2^{10}]\}$. Despite numerous attempts, I struggled to achieve stable training on the dataset. Consequently, I decided to improve the network structure in order to move forward.

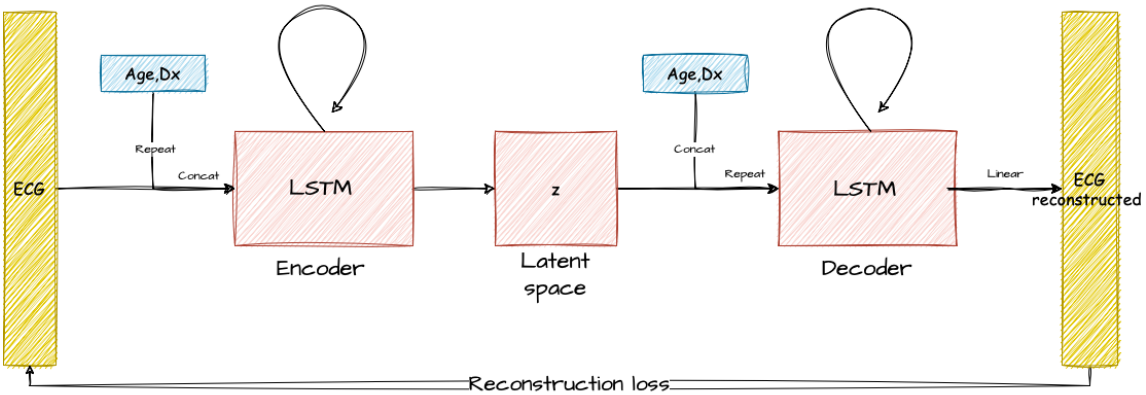

Figure 5: Contextual Reccurent Auto Encoder

Contextual Recurrent Variational Autoencoder 6 The network was trained to minimize the sum of the reconstruction error and the KL divergence term weighted. I experimented with both the mean and the sum of squared errors, as well as absolute errors. This architecture was

tested with various hyper-parameters: $\{hidden\_dim : [2^6, 2^7, 2^8], latent\_dim : [2^5, 2^6, 2^7, 2^8], lr : [1e-3, 1e-5], kl\_weight : [\frac{1}{2^0}, \frac{1}{2^1}, \frac{1}{2^2}], window\_size : [2^5, 2^6, 2^7, 2^8, 2^9]\}$.

During training, I encountered two major issues commonly discussed in the literature: the KL term diverging after a number of epochs, and the KL term dropping to zero—known as KL vanishing. I attempted both the monotonic annealing schedule and the cyclical annealing schedule suggested in Fu et al. ((2019)), but despite many efforts, I was unable to achieve training stability.

Another well-documented challenge in LSTM-based models is their limitation with longer input sequences. The longer the input, the more difficult it is for the model to learn effectively. In my case, only the first few timestamps of the ECG could be reconstructed with some accuracy, while the later parts of the signal were poorly restored.

Throughout the training process for this model and the previous one, it became clear that the model performed well at reconstructing the smooth, monotonic portions of the ECG signals but struggled with predicting sharp transitions and peaks. This is expected, as the loss minimization tends to focus on the larger, consistent sections of the signal, leading to an underrepresentation of critical transitions, such as the QRS complex or other sharp changes. This led me to incorporate the *"Critical Features Focus"* preprocessing workflow described earlier. After experimenting with several subsets of this routine, the best results were achieved with the combination of *Standardization by Twice the Standard Deviation, Raising the Signal to a Power of 3, and Global Min-Max Scaling*. However, even with these improvements, the ECG reconstruction remained suboptimal.

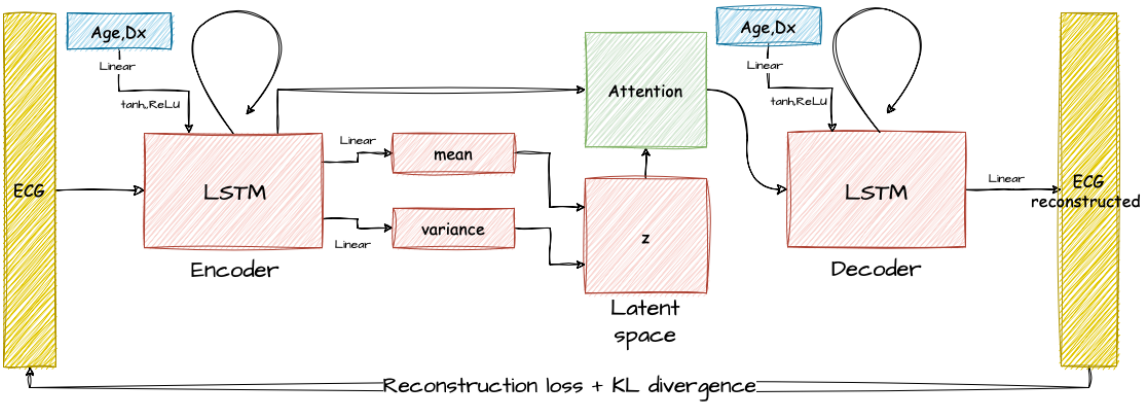

Figure 6: Contextual Reccurent Variational Auto Encoder

Convolutional Variational Autoencoder 7 In a significant shift, I decided to completely change the architecture of the encoder-decoder components, replacing the LSTMs with convolutional blocks. This allowed me to process the data in a 1D format using only the first lead of the ECG (Lead I), which helped with faster training and experimentation. The intuition behind this shift was that convolutional layers, with their ability to capture local features through sliding windows, would overcome the limitations I faced with LSTMs, especially in handling the sharp transitions in the ECG signals.

The network was trained to minimize the sum of the reconstruction error and the KL divergence

term, with the KL term weighted using an annealing schedule. After experimenting with various error metrics, I found that the sum of squared errors worked best in this setup. The architecture was tested with several hyper-parameters: $\{hidden\_dims : [[2^4, 2^5, 2^6], [2^5, 2^6, 2^7]], latent\_dim : [2^5, 2^6], lr : [1e-5], kl\_weight : [\frac{1}{2^1}, \frac{1}{2^2}], window\_size : [2^5, 2^6, 2^7, 2^8, 2^9]\}$.

I experimented with this architecture using several subsets of the *"Critical Features Focus"* preprocessing routine. Surprisingly, unlike the previous models, this architecture did not tend to favor the larger, smoother sections of the signal. The best results were achieved by using only *Min-Max Scaling*, which suggested that this convolutional architecture was better suited for the task of ECG reconstruction, preserving the ECG's natural structure. This finding indicated that the convolutional model is more capable of capturing both smooth and sharp transitions in the ECG signals compared to the LSTM-based models.

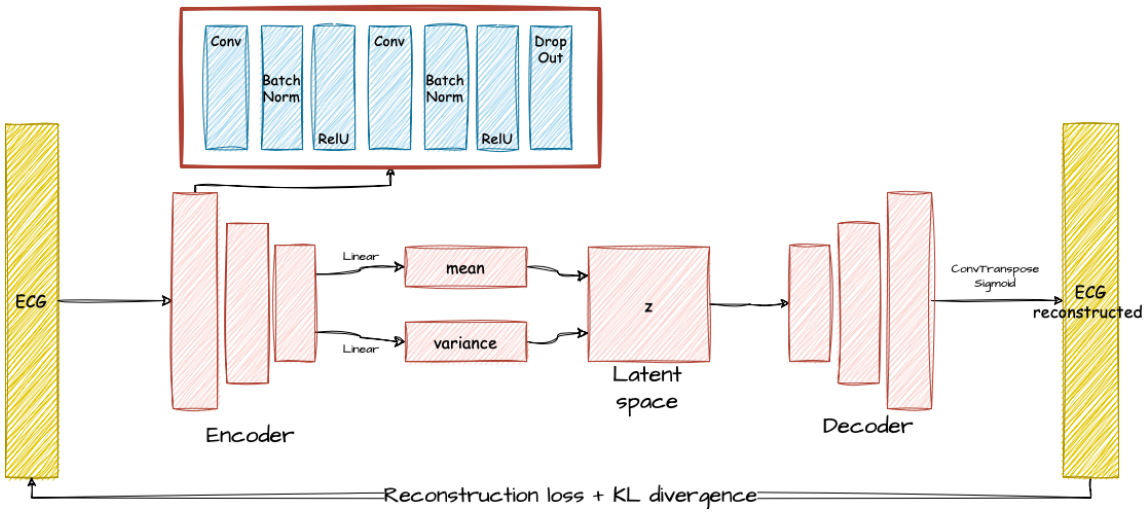

Figure 7: Convolutional Variational Auto Encoder

Hypothesis Testing

As a remider, The hypotheses formulated for this research were:

$$H_0 : \text{Reconstruction Error}_{ID} = \text{Reconstruction Error}_{OOD}$$

$$H_1 : \text{Reconstruction Error}_{ID} \prec \text{Reconstruction Error}_{OOD}$$

Hypothesis testing was conducted on the ID test data, compared against the test dataset from the OOD population. To assess whether the two independent samples (ID and OOD) come from different distributions, I employed the *Permutation Test*, using the *Wilcoxon Rank-Sum Test* and the *Mann-Whitney U-Test* that was first introduced in Mann and Whitney ((1947)), which are equivalent for this purpose. These are non-parametric methods that are well-suited for this analysis, as they rely on ordered metric-scale data—in this case, the reconstruction errors for each ECG

record, also because they do not assume a specific distribution for the data, making them ideal for comparing the reconstruction errors.

The null hypothesis states that the distribution underlying the ID sample is the same as the OOD distribution. The tests were chosen because they do not assume a specific distribution for the data, making them ideal for comparing the reconstruction errors.

To ensure statistical rigor, I estimated the p-value confidence interval using 1000 bootstrap samples, combined with a resampling method that involves 10^4 permutations for each bootstrap sample. Although ties in the data are possible, I represented the reconstruction errors as float32 to minimize this issue. Additionally, the Mann-Whitney algorithm is equipped to handle ties naturally, hence I present its results.

Main Results + Discussion

The results of the Contextual Recurrent Variational Autoencoder, as shown in the loss curves8 for both male and female datasets, reflect the challenges I observed during training. The reconstruction loss drops rapidly in the initial epochs, which is typical of VAE models as the network begins by finding an approximate solution. After this sharp decline, the loss stabilizes and decreases more gradually, indicating that the model finds some structure in the ECG data but struggles to improve further as training continues, to be specific, learning an expressive latent space.

Additionally, the KL loss remains consistently low throughout the process, confirming the issue of KL term vanishing that I encountered, despite experimenting with both annealing and cyclical annealing schedules. This imbalance between the reconstruction loss and KL divergence suggests that the latent space isn't contributing as much as expected, which likely explains the model's difficulty with reconstructing sharp transitions in the ECG signals.

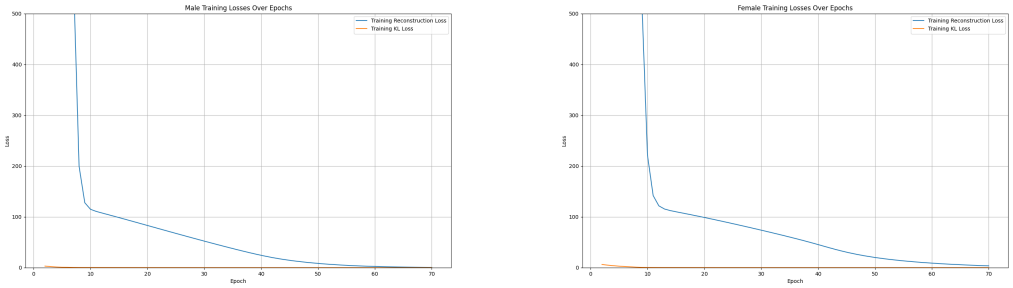

Figure 8: Training Loss of Contextual Recurrent Variational Autoencoder over 70 epochs for both datasets $hidden\_dim : 2^6$, $latent\_dim : 2^5$, $lr : 1e{-}5$, $kl\_weight : \frac{1}{2^2}$, $window\_size : 2^8$

As seen in the 12-lead reconstruction9 by CRVAE, the model struggles to accurately capture the main features of the ECG, particularly the sharp transitions. This outcome aligns with my observations during training, where the loss minimization process, as expected, favored the larger, smoother sections of the signal, leading to a weaker representation of this features and stronger representation of the "easier" parts.

The *Critical Features Focus* preprocessing workflow helped improve the overall structure of the reconstructions to some extent, the imporovment is that the early timestamps show some similarity to a shape of a QRS complex. However, despite these improvements, the model's ability to reconstruct the entire ECG sequence remains very limited.

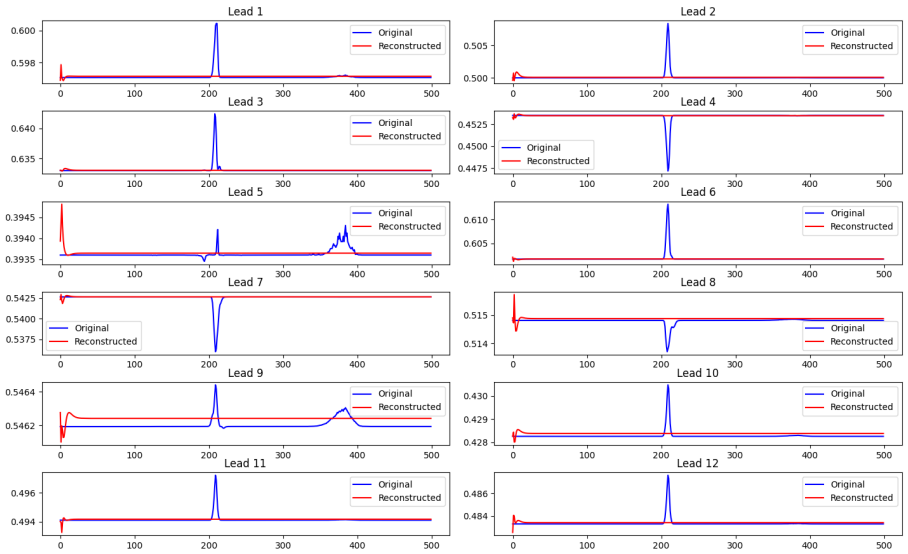

Figure 9: Representative 12-lead record reconstruction by Contextual Recurrent Variational Autoencoder, pre-processed with *"Critical Features Focus"*

The Convolutional Variational Autoencoder showed improved training results10 for both male and female datasets, with minimal preprocessing, with more stable reconstruction loss and better convergence. Its overall performance confirmed that convolutional simple architecture is better suited to capturing both smooth and sharp transitions in the ECG signal, on this setup.

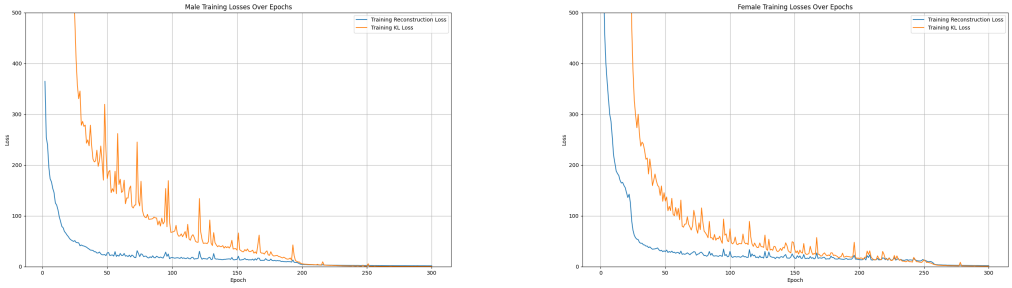

Figure 10: Training Loss of Convolutional Variational Autoencoder over 300 epochs for both datasets $hidden\_dims : [2^4, 2^5, 2^6]$, $latent\_dim : 2^6$, $lr : 1e{-}5$, $kl\_weight : \frac{1}{2^2}$, $window\_size : 2^9$

The improved results of the Convolutional Variational Autoencoder are evident in the recon-

structed 1-lead ECG signals11. In both male and female reconstructions, the model captures the general shape of the ECG waveform, although there are still noticeable discrepancies between the original and reconstructed signals, this matches my earlier observations. Nevertheless, the ConvVAE demonstrates a improvement over previous architectures in handling this data, suggesting that with more training and addition of complexity, parameters and tuning the model could lead to further performance gains.

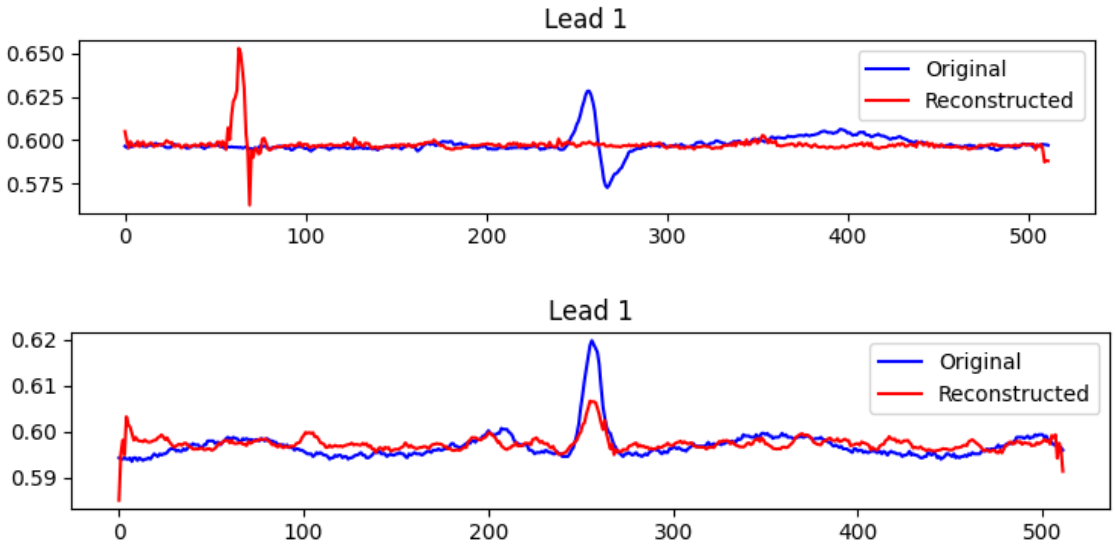

Figure 11: # Age: 31, 35 # Sex: Male, Female # Dx: 426177001, 426177001
Representative 1-lead ECG and reconstructed ECG records, the data is Min-Max scaled

Although the models did not achieve significant success, they still learned from a single sex at a time. If there were inherent differences between male12 and female13 ECG signals, we would likely have observed them even with the current results. The hypothesis testing, using the *Mann-Whitney U-test*, has a low standard deviation, which implies that for this setup, we cannot reject the null hypothesis. This means that under this specific models, there is no strong evidence of significant differences between the two groups based on the reconstruction error distributions. The p-values bootstrap distribution14 suggest that the variation is minimal, further reinforcing the idea that the model's inability to distinguish between sexes. Those results liability can be negatively impacted, as stated in Thiese et al. ((2016)), by the small sample size, magnitude of effect of the sex on the ECG shape, and random errors along the way like systematic error caused by choosing the wrong testing or evaluating methods and more.

Project's Limitations + Discussion

One significant limitation I haven't addressed is that even a slight misalignment in the peaks of the reconstructed signal compared to the original can cause a large MSE, even if the two signals look

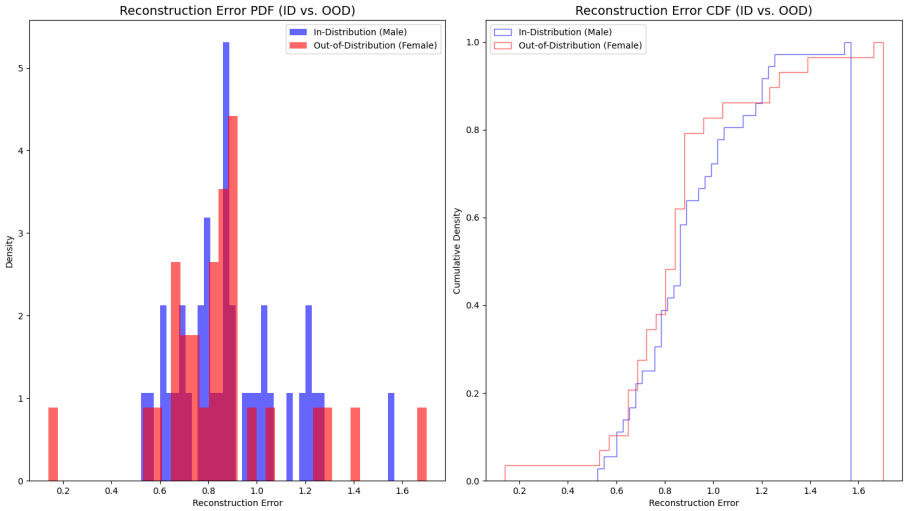

Figure 12: Male trained ConvVAE reconstruction error distribution

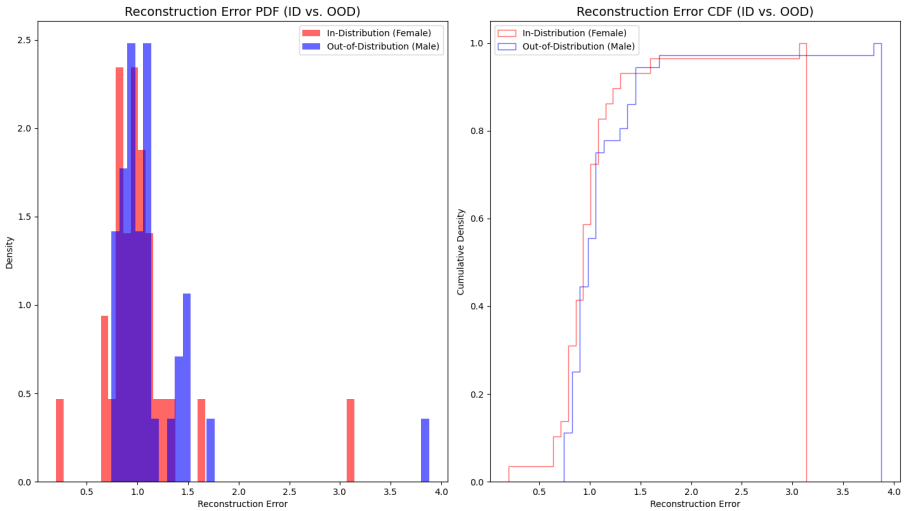

Figure 13: Female trained ConvVAE reconstruction error distribution

visually identical. Focusing more on the shape of the signal rather than just the MSE might be a more effective way to train the model for reconstruction accuracy.

Another shortcoming is that I did not explore modifying the architecture to generate periodic signals. Although this wasn't the core goal of the model, and the hypothesis testing could have revealed differences between the sexs ECG nature without it, the periodic nature of ECG signals

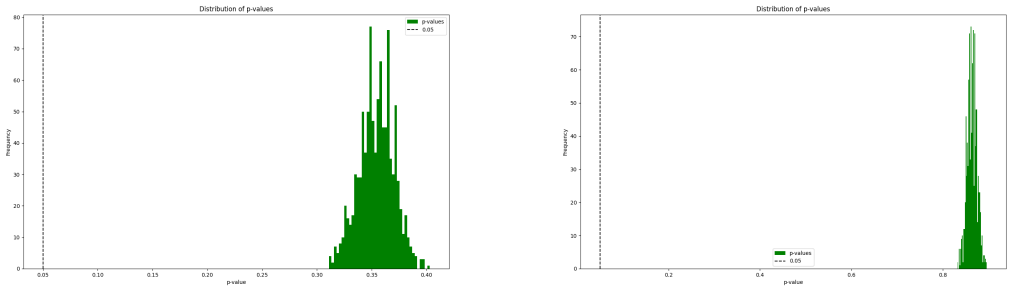

Figure 14: p-value 1000 bootstrap distribution for Male and Female,
CI for p-value: $[0.864 \pm 0.01]$, $[0.383 \pm 0.015]$ respectively

may still play a role. To create a high-quality test besed on synthetic ECG data, the model should better capture the structured, periodic nature of ECG signals. My results showed that simply sampling from a VAE trained on this data often leads to samples that miss periods, although the periods was present in the training dataset. To address this, I could have proposed an architecture designed to enforce periodicity, or just duplicate the decoder outputs, but this approach presents challenges as the period of the input signals isn't fixed. As a result, even small deviations in periodicity could lead to disproportionately large penalties in the loss function.

Another possible source of variation in the ECG data comes from differences in ECG lead placement between males and females, which might influence the underlying signal distribution. While I can't directly address this, scaling the data may help reduce this bias. Additionally, though it's unlikely, other types of distribution shifts (beyond the intended sex distribution) cannot be entirely ruled out. Despite assuming that the dataset is balanced in terms of time, age, location, sex and measurement methods, unanticipated variations might still occur.

Finally, although the dataset is limited to a single source and may not generalize well to other datasets, I chose to focus on this particular dataset due to very limited computational resources. The primary objective was to identify differences within this dataset between the sexes, which makes the issue of using a single dataset less problematic. In fact, it simplified the model training and allowed me to focus on the core task.

Code

Full project code is available on https://github.com/avishagnevo/VAE-ECG

References

J.-W. Choi, D.-Y. Hong, C. Jung, E. Hwang, S.-H. Park, and S.-Y. Roh. A multi-view learning approach to enhance automatic 12-lead ecg diagnosis performance. *Biomedical Signal Processing and Control*, 93:106214, 2024. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2024.106214. URL `https://www.sciencedirect.com/science/article/pii/S1746809424002726`.

H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1021. URL https://aclanthology.org/N19-1021.

S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. 2020. URL https://arxiv.org/abs/2001.01550.

J.-H. Jang, T. Y. Kim, H.-S. Lim, and D. Yoon. Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. *PLOS ONE*, 16(12):1–16, 12 2021. doi: 10.1371/journal.pone.0260612. URL https://doi.org/10.1371/journal.pone.0260612.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

R. Nishikimi, M. Nakano, K. Kashino, and S. Tsukada. Variational autoencoder–based neural electrocardiogram synthesis trained by fem-based heart simulator. *Cardiovascular Digital Health Journal*, 5(1):19–28, 2024. ISSN 2666-6936. doi: https://doi.org/10.1016/j.cvdhj.2023.12.002. URL https://www.sciencedirect.com/science/article/pii/S266669362300110X.

M. S. Thiese, B. Ronna, and U. Ott. P value interpretations and considerations. *Journal of Thoracic Disease*, 8(9), 2016. ISSN 2077-6624.