Methods draft

# Data

The data was collected from 285 patients, all with AF either persistent or longstanding , who have undergone the Catheter ablation pulmonary vein isolation procedure. Information from the procedure and after has been recorded like : RF or cryoablation , Was the patient in AF after the ablation and was there a recurrence in Arrhythmia after the procedure had been completed . The ECG and CS data came in the form of TXT file with additional info about the signals recorded like the label, the frequency of recording, samples per channel and start/end time of the recording. The patient data like pretreatment and ablation procedure details where all kept in csv files that can be open in excel. Each patient will have at least one sample taken for 60 seconds at 1000hz resulting in 60,000 timestamps for each sample. All participants will at least have one 60 second sample pre ablation at the point where the catheter has been placed within the heart to record a intracardiac reading using an EGM paired up with 4 surface ECG leads recorded with an ECG. Depending on the conditions during the ablation a EGM/ECG reading was also recorded straight after the ablation to the tissue which provided us with samples post ablation to also look it. Some participants have more than one post and pre ablation recording depending on how ideal it was to take them during the procedure so that it is not detrimental to the patient, here we use every sample we can obtain from every patient as to take advantage of the dataset in full.

**Lead focus**

In this study we focus on four leads as the procedure in which data has been collected resulted in the 4 leads being more well aligned with the intracardiac reading as oppose to the full 12 leads, the four leads chosen where: I , aVF , V1,V6. Both I and V6 represent the lateral surface of the heart while V1 gives signals related to the right atrium and cavity of the left ventricle and aVF is related to the inferior surface of the heart(Meek and Morris, 2002). Being able to analyse the signals coming from these areas can help identify any abnormalities in the signal and therefore inform cardiologists of what area of the heart is being affected to aid in a diagnosis and determine an area for exploration within the left atrium.

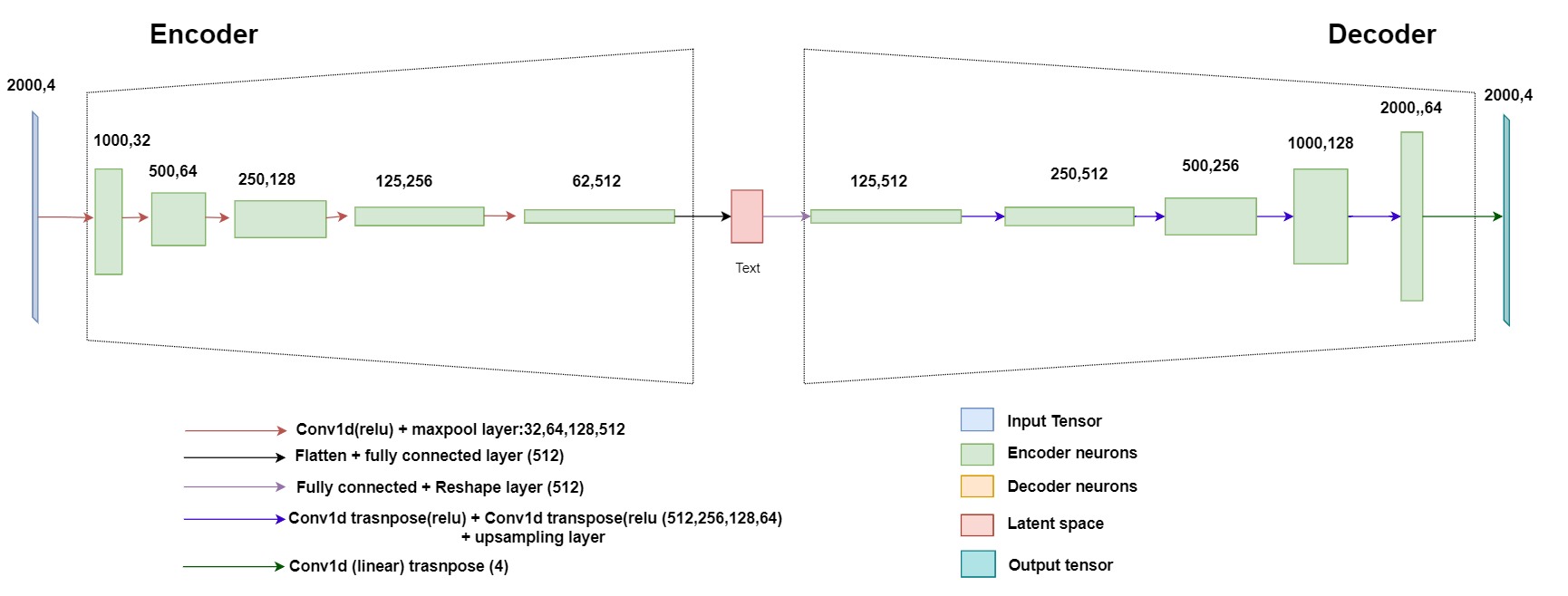
# Study pipeline

Figure 1: this figure details the data processing steps taken through this study. The study is split into a two-pronged approach the primary objective being the recreation of the intracardiac rhythm /predicting sample entropy. The secondary objective involved feature extraction from the CNN sample entropy prediction model clustering upon those features.

# Details of the models used for both approaches.

Autoencoder

For the autoencoder, a model was adapted from the work of Kuznetsov et al who used a variational autoencoder to generate a new cardiac cycle (QRS peak and small proportion surrounding signal including the p and t wave) that is similar to the original cycles the model was trained on(Kuznetsov *et al.*, 2021). However, as we only want to reconstruct the exact ECG signal or intracardiac rhythm that we want, the variational (sampling) component of the model was removed as we did not need to generate new signal just recreate them. Note that we are also attempting to recreate the complete signal for both ECG and intracardiac rhythm, so we are not constrained by only reconstructing the cardiac cycle. From this model we did take the convolutional then Maxpooling layer organisation to construct the encoder linking the encoder to the decoder using a flatten and reshape layer. Also due two the larger signal size of 2000 compared to 400 we used differing kernel sizes instead of going for 8 to 16 to 32 we went from 32,64,128,256,512 leaving us with a larger latent space. The architecture was developed in TensorFlow (v2.7.1) in collaboration with CUDA (V11.4.1) in order to utilise a Nvidia graphics card for a speedier process as oppose to using a GPU. To crate the architecture we used Oleszaks work who shows what a basic autoencoder should look like in which we modify this architecture to use more layers, to output a linear activation function and use a 1dimensional input as oppose two a 2 dimensional input(Oleszak, 2023). The details of the model will be depicted below.

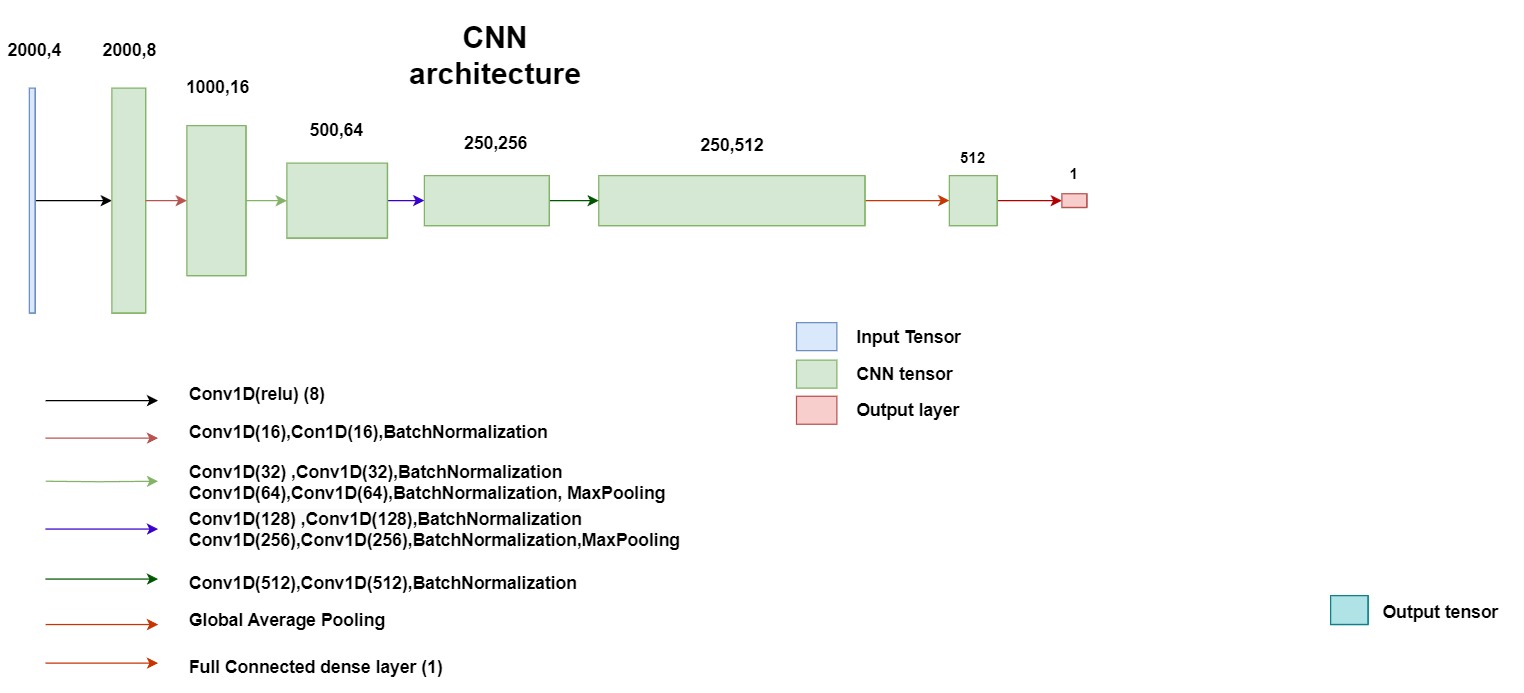


This figure shows the autoencoder architecture used. Within the encoder a convolutional 1d layer was used while I have detailed the back propagation in the introduction , it is useful to explain the forward propagation as to explain how features are extracted. The forward propagation is mathematically expressed as :

A close up of a text

Description automatically generated

where 𝑥𝑘 𝑙 is defined as the input, 𝑏𝑘 𝑙 is defined as the bias of the 𝑘𝑡ℎ neuron at layer 𝑙, 𝑠𝑖 𝑙−1 is the output of the 𝑖𝑡ℎ neuron at layer 𝑙−1, 𝑤𝑖𝑘 𝑙−1 is the kernel from the 𝑖𝑡ℎ neuron at layer 𝑙−1 to the 𝑘𝑡ℎ neuron at layer 𝑙. This equation shows how a 1d kernel is passed across the 1d array to create 1d feature map. Due to the lack of padding we will have some information at the edges of each array which is helpful as it ignores and left over artifact like a QRS peak cut in half since we split the 60000-time stamp samples into 2000 regardless of QRS peak placement. This doubles the dimensions of the output array for example 32 to 64. The Maxpool layer had a pool size of 2 and so for 2 features the max number is taken and the other is discarded also halving the dimensions of the output e.g from 1000 to 500. After this a flatten and dense(fully connected) layer are used to flatten the output to 512 as this represent the latent space of 512 of the most important extracted from the signal(Kiranyaz *et al.*, 2021). For the decoding a dense layer dense layer at 512 size is used to expand the latent space to 64000 in which it is reshaped back to 125,512. Then we run a feature map of 125,512 through a transposed convolution 1d layer which essentially reverses the convolutional process and up samples the data so that we can up sample the feature map back into the required for the output(Lane, 2018). Finally a conv1d transpose layer of size 4 with a linear activation function, producing an output of any number negative or positive is used to upsample to the original 4 lead ecg . In the case of the reconstruction of the intracardiac ryhtym the same process is undertaken however the transpose layer is reduced to a size of 1 resulting in an output of 2000,1 as to represent one 2000 timestamped EGM signal .

CNN  
The alternative the primary objective of feature extraction is to use the ECGs to predict the sample entropy of the Intracardiac reading rather than to reconstruct it. To do this we use a standard CNN architecture to predict a sample entropy value from an input of 4 ecg leads. This architecture is adapted from (Khan *et al.*, 2023) who leverage a resnet architecture to classify different types of heartbeat from an ECG at a f1 score of 92.83 which shows improvements over other CNN. We felt that it would be helpful to utilize this model as it seemed amicable in detecting morphological areas of interest within an ECG in the hopes that it would be able to identify different morphologies in our 4 lead ECG and relate that to the sample entropy. The model is shown below.

For the fully connected dense layer was tried with a series of activation functions being sigmoid which ouputs values from 0 to 1 and soft plus which outputs positive values.

**Loss function**

*2* Pearson’s Chi-squared test; Fisher’s exact test

*3* Pearson’s Chi-squared test

*4* Fisher’s exact test; Pearson’s Chi-squared test

# Feature extraction, clustering and cluster analysis

**Feature Extraction and clustering**

Our secondary objective is to extract significant features and identify new af phenotypes through clustering upon those features. The features are extracted from the penultimate layer of the CNN as seen in the study pipeline each fold (1-5) will be used as a test set from one of these methods. To ensure we are assessing generalizability we only extract features from models used on the test fold for each set this will give us the feature of the 62,100 samples we have based on the test set predictions. We then cluster through a series of clustering techniques :Kmeans,GMMC and DBscan in which based on performance (silhouette score) we picked Kmeans based on this.Kmeans is a very common clustering computationally light algorithm that gained prominence in many fields during tha 1950-60s. Kmeans cluster worka by initially going by a user defined amount of cluster (k). We then need to minimise the sum of square error of the cluster itself to do this the kmeans algorithm picks a random Euclidian space to place (k) number of centroids. Each point of data is assigned to the nearest Euclidean distance centroid and the algorithm iteratively picks out these centroids location until there are stable clusters formed that are maximally separated from each other(Ikotun *et al.*, 2023). The silhouette score is then used as a graphical to assess the compactness and degree of separation between Euclidean distances caclculated using the proximities between data points (Rousseeuw, 1987). We used the elbow method to identify the most stable cluster number however we did go for 7 clusters as adding more clusters with make some clusters way too small with only 2 samples which would cause too much imbalance between the clusters.

**Cluster analysis**

**Data preparation**

A csv was supplied by the ecgcardiomaths department containing patient information some patients that were involved in the CNN and autoencoder had no patient info on them or they has an ablation before and therefore were considered invalid as we were only assessing patients who had an ablation done once, we cannot fairly assess patients who have had two ablation as we do not have the details of that ablation. Therefore, we removed the samples of these patients and there corresponding features resulting in 196 unique patient at 49,530 samples and sets of features. For these patients there was also data that was missing so we decided to remove any variables that where repeated other than hypertension and vascular disease which were kept for visual reason for the radar plot. The imputation process used here was miss Forest due to its adept ability to deal with different data types whether they were continuous or categorical in which it has performed better that other imputation algorithms like KNNimputation(Stekhoven and Bühlmann, 2012). Normally the threshold for imputing missingness is 30 percent but we needed to impute some important variables that could not be removed like LA volume which has a higher percentage of missingness above 45% so we used the out of bag error to determine if the imputation was acceptable. The out of bag error for numerical variables is represented by the normalized root mean squared error which essentially measure the numerical variance between the real data set and the imputed data set and can be measured between 0 and 1 here it was 0.1685 which is deemed acceptable. The categorical column imputation performance was the proportion of falsely classified (PFC) at 0.0098 which was also small and deemed acceptable for this data (Stekhoven, 2022).

**Cluster analysis**

Each set of features per sample was set to a 1 of the seven clusters. Each cluster would have a set of samples corresponding to a patient if the sample for the same patients was repeated that patient would be used in that cluster again depending on the number of sets of features in each cluster. The mean of the cluster was used to represent the distribution of certain characteristics within that cluster. To determine if the distribution of characteristics where truly different between clusters we used a series of statistical tests in which a p value under the Bonferroni correction (add a more stringent limitation to the p\_value) of 0.007 was a significant difference between the reference cluster (cluster 1) and the other clusters. Cluster 1 was considered the reference cluster as each patient had been assigned this cluster based on at least 10 sample from that patient, so it served as a baseline for the most common features among patients and the other clusters represent what distribution of features differentiate a set of samples from the baseline distribution. For numerical variables the statistical test used was the Wilcox rank sum test which is used to compare two samples from each other this is considered a non-paramteric test as we do not assume the samples are normally distributed(Xia, 2020). In this test the samples are combined and ranked in which the sum of ranks is calclculated for each sample and a test statistic (U) is used based on the lowest sum of ranks of between the samples . An expected value of the test statistic is calculated and variance between the real and expected is used to calculate the Z-score which is used to derive the P value. If the P value is under 0.007 then it is statistically significant and we can reject the null hypothesis that two values do have the same continuous distribution(*Wilcoxon Test: Definition in Statistics, Types, and Calculation*, 2023). For categorical values the Pearson chi squared test is used and for clusters that are smaller like cluster 2,5,3 and 6 the Fischer exact test was used. The Pearson chi squared test uses the calculated expected an observed values if there is large significant difference between the observed an expected values then the distribution of the categorical values are significant from each other. If the expected frequency of a categorical variable is less than 5. To assess the null hypothesis of independence we apply a hypergeometric distribution assessing the chance of observing the distribution of frequency seen without replacement . For both the fischer exact test and the chis squared test if the p value is lower than 0.007 then we can reject the null hypothesis of independence and assume that our categorical variables are differently distributed between two clusters.

**\textbf{Background:} As the burden of Cardiovascular disease continues to increase late-stage persistent atrial fibrillation has become more of a focus. The current gold standard for treating persistent atrial fibrillation is catheter ablation which provides use with informative intracardiac signals but only boosts success rates of 53.6\% at most. The inability to successfully treat longstanding and persistent AF has led to a focus on attempting to discover novel approaches to dealing with this disease. One looming issue is the lack of categorisation for a disease that is as heterogeneous as AF. A solution to this is AF phenotypes, the prevailing idea is to cluster patient information to develop new subgroups of AF that can better inform treatment. Therefore, the primary objective is to evaluate methods of predicting intracardiac morphologies of an intracardiac lead (CS3-4) from 4 ECG using ECG-AI models. This contributed to the secondary objective of extracting features from these predictions and analysing the characteristics of these cluster to evaluate their clinical relevance.\\**

**\textbf{Methods:} 62,100 samples were used to derive a prediction using a CNN for the sample entropy of the intracardiac lead based on 4 ECG leads. The features were then extracted for 49,530 of the samples that had valid corresponding patient info. K means clustering was used on these features to provide 7 clusters and a series of statistical tests were used to identify if the distribution of these clusters where significantly different to the firs cluster which was used as a reference cluster \\**

**\textbf{Results:}We have identified 4 statistically significant**