THESIS

THE SELECTIVE DE-IDENTIFICATION OF ECGS

Submitted by

Musamma Akhtar

Department of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2022

Master's Committee:

Advisor: Steven Simske

Zhijie Wang
Marie Vans

ABSTRACT

THE SELECTIVE DE-IDENTIFICATION OF ECGs

Biometrics are often used for immigration control, business applications, civil identity, and healthcare. Biometrics can also be used for authentication, monitoring (e.g., subtle changes in biometrics may have health implications), and personalized medical concerns. Increased use of biometrics creates identity vulnerability through the exposure of personal identifiable information (PII). Hence an increasing need to not only validate but secure a patient's biometric data and identity. The latter is achieved by anonymization, or de-identification, of the PII. Using Python in collaboration with the PTB-XL ECG database from Physionet, the goal of this thesis is to create "selective de-identification." When dealing with data and de-identification, clusters, or groupings, of data with similarity of content and location in feature space are created. Classes are groupings of data with content matching that of a class definition within a given tolerance and are assigned metadata. Clusters start without derived information, i.e., metadata, that is created by intelligent algorithms, and are thus considered unstructured. Clusters are then assigned to pre-defined classes based on the features they exhibit. The goal is to focus on features that identify pathology without compromising PII. Methods to classify different pathologies are explored, and the effect on PII classification is measured. The classification scheme with the highest "gain," or (improvement in pathology classification)/ (improvement in PII classification), is deemed the preferred approach. Importantly, the process outlined can be used in many other systems involving patient recordings and diagnostic-relevant data collection.

ACKNOWLEDGEMENTS

I would first like to acknowledge my advisor Dr. Steven Simske for being the best advisor anyone could ever ask for. I am truly grateful and blessed to have had such a kind mentor to help and guide me through this whole process. Thank you for being so patient. I am also very thankful for the other members of my committee, Dr. Zhijie Wang and Dr. Marie Vans, and the energy put forth to review my work. I would also like to acknowledge the Colorado Cybersecurity B.71 SB18-086 grant for funding this work.

My dad for setting an example for me, being proud of me and always encouraging me. My mom for always believing in me, being proud of me, and reminding me everything would work out. My sister Laiba for always helping me out no matter how busy she was. My brother Saad for always cheering me on and randomly facetiming me to share some nonsense.

My baby brother Sinan for being my comic relief is this stressful time and the best baby brother ever anyone could ask for. My dear friend Sam Wolyn for making this all possible with his coding help, sorry for dragging you into this. Dr. Simske's Lab for always supporting and giving suggestions. Sara Mattern for being so kind from the beginning of my journey at CSU and for helping me with all the forms I needed. Also, the close friends I have made here in Fort Collins, specifically Huma, Shay, and Madi for always being there for me especially on the hard days to cheer me up and for making me feel less alone by keeping me company. Lastly, CSU's amazing recreation center for giving me a place to take time focus on myself to get stronger mentally and physically.

DEDICATION

To Mami, Dadi, Laiba, Saad, and Sinan. I love you all more than you'll ever know, thank you for being my biggest supporters. Also, for the engineering students that may seem like it is impossible now, it is not, so keep on moving forward. Believe in yourself and it will eventually work out in the end as it always does.

TABLE OF CONTENTS

Chapter 1

Literature Review and Goals of the Thesis

1.1 Introduction

Biometrics are often used for immigration control, business applications, civil identity, and healthcare. Biometrics can also be used for authentication, for monitoring (e.g., subtle changes in biometrics may have health implications), and for personalized medical concerns. Increased use of biometrics creates identity vulnerability, and greater opportunities for this personal identifiable information (PII) to be exposed, exploited, and exported. Hence, there is an increasing need to not only validate but secure a patient's biometric data and identity. This chapter presents an overview of the research on biometrics, classification, de-classification, and previous classification methods analyzing ECG time series data. The purpose and goals of the research in this thesis, which are related to the needs identified in the literature review, are presented at the end of this chapter, with the objective of creating a stronger method of protecting patient data.

1.1.1 What is an ECG, its importance, and applications

An electrocardiogram, or ECG, is a bioelectrical recording taken from the surface of the thoracic cavity. The ECG is associated with the electrical activity of the heart. Electrical signaling mediated by the cardiomyocytes (heart cells) controls the timing and location of the heart's beating. An ECG records this electrical activity, and it flows throughout the heart. Healthy patients typically have normal ECGs, indicating a healthy beating heart. However, patients with diseases or heart muscle damage have abnormal ECGs. Depending on the heart disease or damage, some or all parts of the recorded ECG signal can be affected. Patients with related heart health conditions logically are expected to have similar ECG anomalies and signals. An ECG is an important tool to assess the health of the heart though the detection and diagnosis of possible arrhythmias due to the heart beating too slowly, quickly, or irregularly.

1.1.2 Biometric data and ECG as a Biometric

Biometric data is used for the identification of an individual based on the physical, chemical, or behavioral attributes of the person [1]. Examples of biometric traits that can be used for identification are facial features, iris pattern, fingerprints, voice, and keystroke patterns [1]. A partial advantage of biometric qualities is that they cannot be inferred unless granted access to them. This provides a false sense of security because “they cannot be guessed.” However, once access is gained, biometrics become compromised ever after since they are relatively easy to replicate and cannot be changed by the individual (acting as a permanent, unchangeable password). Thus, biometric identification is a security risk under certain contexts because they have unique features associated with the individual that they purportedly identify. An ECG is one example of a biometric that is at risk of data piracy and loss of privacy. ECGs have been used by many researchers in the biometric identification system, since they have feature types that are unique to everyone, such as statistical, morphological, and wavelet features [2]. This leads to biometric identification existing in the same continuum as diagnosis by using both AI (artificial intelligence) and ML (machine learning) for biometrics and for authentication [2]. Diagnosis of an ECG is specific to that certain individual’s reading. For example, the peaks of the ECG {P, Q, R, S, and T waves} or the spatiotemporal attributes of the individuals are altered in identifiably different ways in each individual.

For the purpose of this study, several classification goals are considered to exist in the same broad analysis domain, with one (accurate diagnosis and classification) being defined as the "good" or intended accurate output and the other (individual identification, or biometric analysis) as the "bad" or non-intended output.

An ECG signal with feature patterns that have been recorded from a patient with ventricular escape is likely to match another ECG signal from patient also with ventricular escape in certain signal (time series) aspects. Similarities could lie in waveform morphology, wave frequency comparison before and after normalization or electrostatic discharge (ESD) of the amplitude from the signal. Gregg et al. used two methods to find similar ECGs in a database, the first being pair-wise template matching and the second being fast query with the use of a "k-dimensional tree architecture" along with a feature vector representing a processed rendering of the ECG signal [3]. Query methods are different methods to find data from the database. The conclusion was that low complexity with fast query could be utilized to search a huge database for 12-lead ECGs that are similar. In other words, waveform morphology was the easiest method to be able to compare patients with similar ECG readings.

ECG signals change in measurable and at least partially predictable ways with specific cardiac pathologies, allowing individual ECGs to be assigned to relevant clusters in the pathology analysis process. It is therefore possible to identify heart diseases and even patients by evaluating ECG signal patterns. This is using an ECG as a diagnostic. If, however, the ECG contains idiosyncratic information that helps us to identify the specific (named) patient from whom the signal has been recorded, this is using an ECG as a biometric and may constitute a breach of privacy. For the purposes of anonymization, diagnostic value is meant to be retained after data collection, while biometric identification is meant to be obfuscated (de-identified).

1.1.3 ECG Classification and Selective De-identification

The three main types of traditional time series classification methods are model-based, feature-based, and distance-based. However, deep learning models are also becoming more popular and have been effectively used in the time series classification problem as a result of rising interest in graphics processing unit (GPU)-based computing [4].

When dealing with time series classification problems, re-mapping, or transformation with AlexNet has become a more commonly adopted process. AlexNet is a CNN (convolutional neural network). There are two main approaches for time series classification with a CNN. In one approach, a traditional CNN approach is modified to accept 1-dimensional time series as input, in the other approach, time series are converted into a 2-dimensional image to be used with a conventional CNN [5]. In previous studies [4] about time series classification with the use of CNN, time series associated with individuals were inserted into various algorithms, such as a 1NN classifier with DTW (dynamic time warping), a Cross Translation Error (CTE) process, and a 1NN classifier with Euclid distance. The outcomes of these experiments were compared to time series classification using fully convolutional neural network (FCN) and Residual network (ResNet) approaches. For the majority of the data sets, it was discovered that the 2-dimensional recurrence plot representation of input data with a CNN yields a higher classification accuracy than the 1-dimensional raw time series data with a 1NN classifier and Euclidean distance [4]. De-identification is the method of removing specific identifiable information from data, i.e., a way of protecting an individual’s identity by clustering their data with that of other individuals in such a way as to prevent individual-level identification.

For this research, the Python programming environment was used to read the ECG signals and convert them into images. Next, the pre-existing AlexNet classifier was used to classify the ECG signals as 2-dimensional images. A CNN was used so that the output layer and the pre-output layers are retrained for the specific ECG problem which is focused on classifying ages and pathologies. The end goal of this research is to create “selective de-identification” using a time series ECG database.

1.2 The Literature

According to a search of the related scientific literature, there is existing work focused on the classification of time series data in general, and ECG based time series classification, de-identification, and anonymization efforts in specific. However, due to the variability and individuality of ECG signals, fixed feature classification could not be applied to all patients, resulting in faulty performance [6].

Processing/postprocessing and feature extraction approaches – for example, binarization, segmentation, feature selection, normalization, and dimensionality reduction – are likely to affect the classification process and cause it to be unreliable for smaller health monitoring devices. Binarization is one of the early processing steps in image analysis and is the modification of a document image into bi-level document image [6]. Image pixels are separated into dual collection of pixels, such as black and white [6]. The primary objective of picture binarization is to separate the background from the foreground text in a document. Segmentation is used to process and analyze images to extract information. Feature selection often incorporates a process of reshaping a large set of repeating data and turning it into a set of features or feature vectors of reduced dimension. Feature selection was used in this research since the optimal features were selection in the methods process of maximizing pathology accuracy. Normalization is used to scale the features for use together. Lastly, dimensionality reduction is used to reduce the number of input variables in training data [6], These concepts were employed in the methods of this research in an attempt to anonymize age accuracy while keeping pathology accuracy as high as possible.

To reduce the limitations mentioned previously, a deep learning ECG classification-based system incorporating convolutional neural networks (CNNs) is used. CNNs have become increasingly popular for deep learning processes due to their vast capabilities with speech and image. They have the capability to identify the important patterns of signals, and in specific AlexNet with its ability to use image-based classification [4].

Extracting features is a key step in ECG classification. The elemental steps of ECG pattern classification are preprocessing, feature extraction and classification. There are many methods to go about this. A handful of researchers have tested different feature extraction and processing approaches for ECG classification. Previous researchers have focused on automatic classification and detection of ECG signals leading to a variety of proposed algorithms for individual heartbeat pattern classification. The most recent works are those described in the following paragraphs.

ECG heartbeat classification using an adequate patient adaptable algorithm was found in [7]. A linear discriminant classifier was included within this patient adjustable algorithm. This classifier, along with an independently working clustering algorithm, was used to remove RR interval (the time difference between two consecutive R waves in the ECG) series to perform classification and clustering on those features. The algorithm's performance was assessed using several ECG databases for comparison purposes and the results showed the algorithm did slightly improve performance for heartbeat classification. The algorithm achieved a mean increase for all databases (6.9% for accuracy A, 6.5% for global sensitivity S, and of 8.9% for global positive predictive value P +). Heartbeat classification is an important aspect of ECG analysis and the first step to identifying arrhythmias; however, there is room for further improvement.

A different classifier was utilized in an automatic diagnostic system for classification of ECG arrhythmias [8]. This classifier, Type-2 Fuzzy Clustering Neural Network (T2FCNN) in combination with a neural net, achieved a classification accuracy of 99%. The researchers classified ten various types of arrhythmias from the MIT-BIH database, by using a combined fuzzy clustering neural networks algorithm. When using the T2FCNN architecture, decisions were made in two steps: first, a new training set was created by choosing the optimal arrhythmia for each arrhythmia class using T2FCM, and second, classification using a neural network trained on the new training set [8].

A similar method of fuzzy clustering, specifically, Fuzzy C-means clustering along with neutral networks to classify cardiac arrhythmia [9]. The researchers worked with ECG filtering, extraction of RR interval using wavelet transform, pre-classification based on fuzzy c-mean clustering technique, and a final classification stage based on neural networks. Their output classification accuracy range was between 98.5% and 99.6% with average accuracy 99.05% [9].

A classification algorithm using the combination of fuzzy and artificial networks for cardiac rhythms [10] achieved a classification accuracy of 80 to 85%. Two different feature extraction approaches for the classification of ECG beats using a system based upon MLP-NN classifier incorporated two approaches: “S-transform based features along with temporal features” and the “mixture of ST and WT based features along with temporal features” to classify five classes of ECG beats [11]. The mean sensitivity results from the two approaches for the five ECG beat classes: “normal beat (N), ventricular ectopic beat (V), supraventricular ectopic beat (S), fusion beat (F), and unknown beat (Q)” are 95.70%, 78.05%, 49.60%, 89.68%, and 33.89%,” respectively [11]. The results improved upon previous feature extraction techniques [11].

To classify ECG cycles, Zidelmal et al. [12] used a support vector machine-based algorithm. Zidelmal et al. used the MIT-BIH database to test the algorithm and they achieved a classification accuracy of 98.9%. The MIT-BIH arrhythmia database was also used to test the classification of ECG signals into normal and arrhythmia groups in Vijayavanan et al. [13]. They used probabilistic neural networks, and their accuracy was 96.5%.

Researchers have used several different feature extraction programs to analyze and classify three classes of ECG signals with the use of four different artificial neural networks. The extraction programs are Discrete Fourier Transform, Principal Component Analysis, Discrete Wavelet Transform, and Discrete Cosine Transform [14]. The most accurate from the list was the Discrete Cosine Transform with 98.36% accuracy.

Further research into classification explores systems meant to classify patients using heartbeat data rather than focusing on pathology classification with the same data. This is focused on in Ince et al. [15]. The researchers extracted ECG features using wavelet transform and principal component analysis (PCA). Using feed forward and fully connected artificial neural networks, they also classified the ECG signals.

A collection of ECG databases, preprocessing techniques, feature extraction methods and classifiers is surveyed in [16]. Researchers examined ECG signal preprocessing, heartbeat segmentation methods, feature description techniques, and learning algorithms in-depth with a focus on current approaches to ECG-based automated abnormalities heartbeat classification. They explained some limitations and disadvantages with these techniques.

Among all classifiers used for ECG classification, artificial neural networks were found to be the most suitable and widely utilized [17]. Artificial neural networks have recently become popular due to their strong learning capacity. They produced improved results working with applications such as image object recognition [19], face recognition [20], classification [18, 21], time series data [22] and medical image analysis [23].

The focus of the approach in Kiranzyaz et al. [24] is deep learning for active classification of ECGs. The INCART, SVBD and again the MIT-BIH arrhythmia database was used to test the evaluation performance. The proposed deep learning methods achieved two objectives: (i) create an automatic learning method for suitable feature representation of the ECG data and (ii) inducing the DNN (deep neural net) classifier by using AL (active learning) criteria to select the most useful ECG beats. The accuracy was improved significantly compared to state-of-the-art methods based on shallow architectures and handcrafted features.

Real time patient specific ECG signal classification based on 1-D convolutional neural networks (CNNs) is the method utilized in [25]. The CNN is tested with a common and patient-specific dataset. Again, the MIT-BIH arrhythmia database was used for evaluation performance of the classifier which resulted in higher classification results. This was compared to other state-of-the-art methods for classification of ectopic beats.

Utilizing 1-dimensional CNNs for ECG arrhythmia classification in [26] shows that CNNs can apply to 1-dimensional ECG signals. This study tests various 1D-CNN configurations on the MIT-BIH database to see which one performs the highest at classifying arrhythmias from ECG data. The aim of this work was to evaluate the behavior of the neural network as well as the results using confusion matrix analysis in relation to the various arrhythmia classifications. The four primary 1D-CNN experiment configurations tested were named Net1, Net2, Net3 and Net4, with

Net4 being the most complex in terms of set parameters. The test resulted in Net4 achieving the highest accuracy of 95% for training and 98% for testing groups with the proposed technique.

A deep 2-dimensional CNN for ECG arrhythmia classification was introduced in [27]. Each ECG beat was turned into a 2-dimensional grayscale image for input data into the CNN classifier. They optimized the CNN classifier by various deep learning techniques such as “batch normalization, data augmentation, Xavier initialization, and dropout.” Jun et al. compared their classifier to AlexNet and VGGNet. They found that the proposed method of data augmentation helped the model reach the highest sensitivity average. Again, the MIT-BIH arrhythmia database was used for evaluation performance of the classifier. They reached a mean accuracy of 99.05%.

The researchers tested security and privacy of time series using a k-anonymization approach of time series data [28]. The main object with this approach was to anonymize the data while minimizing the loss of information caused from this approach. They used the technique CATs (Clustered k-Anonymization of Time Series Data). Which clusters the time series data to guarantee anonymization while also minimizing the loss of information within venerable utility. WEKA and ARX were the anonymization tool used. Results showed that CATs did have loss of information that ranged from a “18% to 24%" reduction rate when compared with previous TSA (Time Series Anonymization) approaches. [28]

Lastly, [29] also focused on ECG time series data and its security risks using a k-anonymization approach to prevent reidentification. The prospective method was focused on an algorithm for k-anonymization of Ngram models of time series. Their algorithm updated the Ngram frequency to mask Ngrams that have smaller frequency values by inserting void Ngrams into a non-k-anonymous Ngram model to increase the frequency of Ngrams [29]. Using this algorithm on time series data Zare-Mirakabad et al. found that there was a two percent maximum information loss compared to approaches that suggest algorithms that cluster, classify, or index time series without employing quasi-identifier attributes taken from the series.

1.3 Clusters, Classes, and Categories

When dealing with data and de-identification, the definitions for clusters, classes, and categories must be addressed since they will be used in this thesis. Clusters are groupings of data with similarity of content and location and do not need metadata. Metadata is a collection of data that labels and/or provide contextual information about other data. Classes are groupings of data with content matching that of a class definition within a given tolerance and are assigned metadata. Lastly, categories are sets of descriptors, etc. generated by mining data from a class or sometimes clusters. These descriptors constitute metadata about the items. Clusters start without derived information, i.e., metadata, that is created by intelligent algorithms, and are thus considered unstructured. Clusters are then assigned to pre-defined classes based on the features they exhibit.

For example, clusters, classes, and categories, we view the data as a vector, <a,b>, with "a" labeled as pathology, while "b" is ID of patient. The goal is to focus on features that identify pathology and ailment only while protecting the patient's ID. The data is presented with the goal of de-identifying channel b. It is grouped into different ailments and then tested to see if identifies pathology only, identifies both pathology and ID or only identifies ID only. The goal is to identify pathology only, allow for the data to determine or diagnose the protected patient's ID.

This will lead to distinguishing which features are relatively useful for classification of a patient from those relatively useful for classification of the disorder. There may be other features that are not useful for either of these two tasks. Determining feature utility incorporates ratio methods to assess relative utility for patient versus disorder identification.

1.4 Specific Aims

This goal of this research is to functionally de-identify ECGs such that they cannot be classified with accuracy above random guessing but will still provide diagnostic value similar to their pre-anonymized version. The means of assessing the success of this "selective de-identification" as seen in Equation 1, involves a method in which the gain (change in classification accuracy divided by the change in diagnostic value) is optimized based on efforts to obscure features of the ECGs that lead to classification (identification) accuracy while not affecting the features needed for diagnostics.

Equation 1:

$$\text{Gain} = \left(\frac{\Delta\ \text{Classification Accuracy}}{\Delta\ \text{Diagnostic Value}}\right)$$

Various approaches will be explored, each focused on these three principles:

1. Removing features that lead to classification (identification) of the ECG source individual
2. Retaining and/or accentuating features that lead to diagnostic accuracy (e.g. identify atrial fibrillation, ventricular escape, etc.)
3. "Fuzzing" of the features that are useful for both identification and diagnosis so that they dissuade identification but do not harm diagnosis (this is an area for very interesting work)

The goal of this thesis is to provide privacy in data (literally, in the manner in which the data is archived) through selective filtering, amplification and other augmentation of bioelectrical data. We introduce the clustering centered definitions of obfuscation, de-identification, and anonymization. A visual of each concept is shown in Figure 1.

1. Obfuscation: aligns with unstructured data with occasional confusion of pairs of individuals. (Full knowledge about individual)
2. De-identification: aligns with clusters (Partial knowledge individual)
3. Anonymization: aligns with identification of individuals with similar accuracy to no clustering or random classification. It is our objective but is likely unattainable. (No knowledge about the individual)

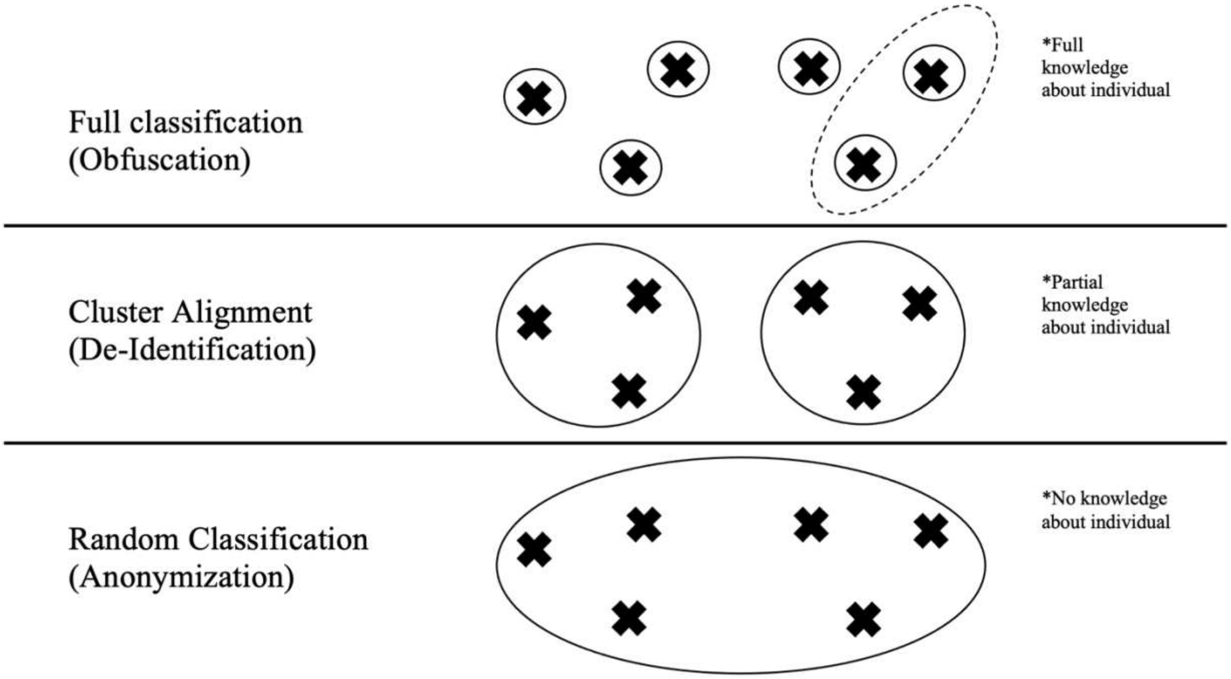

Figure 1: Visualization of obfuscation, de-identification, and anonymization

Chapter 2

Approach: Experimental Procedure and Methods

2.1 Methods and Materials

For this research, the raw signal data that was stored in a compressed format from the PTB-XL dataset (version 1.0.1) [30] was used. For all signals recorded, a standard set of 12 leads: I, II, III, AVL, AVR, AVF, V1, ..., V6 was included with the reference electrode being the right arm. Metadata such as age, sex, weight, and height were also collected from each patient. A report string created by the cardiologist or automatic interpretation by ECG-device was used to annotate each record. This was then transformed into a "standardized set of SCP-ECG statements" [30]. Most patients also had their heart axis and infarction stadium recorded. All patient records were confirmed "by a technical expert focusing mainly on signal characteristics" along with a second cardiologist confirming most of the patient records [30].

2.1.1 Data Preprocessing

Wagner et al. [30] performed data preprocessing by labeling the ECGs and patients with unique identifiers to be used in place of other patient identifiers. The metadata containing personal information was pseudonymized. When data is pseudonymized, the specific markers or information that can be used to identify the subject is changed with ""pseudonyms" or identifiers [30]. To further state this, all "ECG recording dates were shifted by a random offset for each patient" [30]. The ECG statements utilized for annotating "the records follow the SCP-ECG standard" [30].

2.2 Data Specifics

The dataset consists of 18,885 patients with male records comprising 52% of the total and female records comprising the residual 48%. The age range is "0 to 95 years" with the median being "62" and an interquartile range of 22 [30]. From the group of patients, there were "21,837 clinical 12-lead ECG records of 10 seconds length" [30].

The dataset contains 5,158 samples with co-occurring pathologies, as well as a wide dispersion of 9,528 healthy control samples. The diagnosis distribution can be seen in Table 1. For the sake of simplicity, the diagnostic statements have been collected into "super" classes (i.e., closely related classes are pooled together). The sum of statements surpasses the number of records due to the possibility of "multiple labels per record" [30]. The numbers in the parentheses indicate the number of patients in each class after removing co-pathology patients for the experiments done.

Table 1: Distribution of diagnosis

#Records	Superclass	Description
9528 (9083)	NORM	Normal ECG
5486 (2538)	MI	Myocardial Infarction
5250 (2406)	STTC	ST/T Change
4907 (1709)	CD	Conduction Disturbance
2655 (536)	HYP	Hypertrophy

The "Normal ECG" class consists of normal patients (no ECG pathologies), while myocardial infarction consisted of patients who had experienced heart attacks. For ST/T change patients, anomalies such as ventricular escape or ventricular arrhythmias occurred. Conduction disturbance meant AV block while hypertrophy was thickening of the right and left ventricle heart muscle. The waveforms were stored in a WFDB format (Waveform Database) with “16-bit precision at a resolution of 1μV/LSB and a sampling frequency of 500Hz” [30]. The signals were down-sampled to 100Hz, and these 50 Hz Nyquist frequency time series were used for the analyses herein. Due to the size of the data set, only the first 10,000 samples were used, 5,825 of which are NORM for pathology for each experiment. Age was classified between “young” and “old” with 65 being the cut off for the groups.

2.3 Methods

2.3.1 Signal Noise - Adding Sinusoids

One of the early methods for classification focused on timing methods to extract features relevant to achieving high classification accuracy. It is important to note that, in an effort to reduce accuracy, gender accuracy was chosen as the identifying data rather than age accuracy in this initial experiment. This process was simply to add noise to the signals by adding sinusoid functions. The dataset was read in at 100Hz. Lead 1 was derived from Lead 2 and Lead 3, since two leads can be used to generate the missing one, much like the pathogen theorem. Single FFTs, which were regarded as features of the signal, were used. Data from the patients was then randomly assigned to training and testing sets. They were assessed for balance and cleared of any potential repeat patients. However, to ensure that training and testing were equal, the residual data, or the repeated patients, were used for validation. Then an objective function J as seen in Equation 2 was developed, with 0.2 being the amplitude, phase angle randomly chosen and t(a) as the time.

Function J was a mathematical function that added random sinusoids to the raw signal to act as "noise."

Equation 2:

$$\mathrm{J(a)} = \mathrm{J(a)} + 0.2 * \cos(\mathrm{freq} * \mathrm{t(a)} + \mathrm{phase})$$

The last step was to identify the best accuracy by running the dataset through various classifiers such as: K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Decision Tree, Random Forest, AdaBoost, Quadratic Discriminant and Neural Net. This technique was repeated several times and applied to all the classifiers to accurately identify which classifier produced the best accuracy results.

2.3.2 AlexNet

The data was low-pass filtered before employing AlexNet. For this experiment, 10,000 signals were used. Each signal had a frequency of 100Hz and lasted 10 seconds for a total of 1000 samples per signal. AlexNet requires three color channels and there are three leads in an ECG, so each lead was assigned a color channel. Color channel red was assigned to Lead 1, which lies between the RA (right arm) and LA (left arm) quadrant. Lead 2, which is between the RA (right arm) and LL (left leg) quadrant, was assigned green. Finally, color channel blue was assigned to lead 3 which is between the LA (left arm) and LL (left leg) quadrant. This is represented in Figure 2 A red pixel indicates a large lead 1 component with a low amount of lead 2 and lead 3. The same applies for the other color pixels. The value of each sample was then converted to an intensity for a certain color. A value of 0 mV leads to a pixel intensity of 128 for an 8-bit color scheme. The

ECG voltage of -2 mV was assigned a pixel intensity of 0, and 2 mV assigned to a pixel intensity of 255 for the 8-bit color scheme (all ECG voltages were between -2 mV and +2 mV).

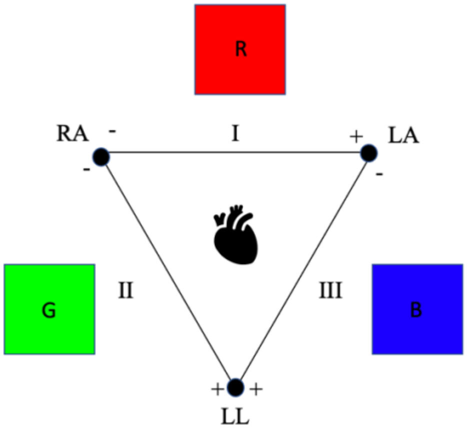

Figure 2: Visualization of leads and their assigned color channels

Once the 1-dimensional signal was turned into a 2-dimensional image, the dataset was reshaped into a 32 x 32 image with the last 24 pixels being empty. This is seen as a black line at the bottom of Figure 3. Figure 3 represents one patient's signal. In Table 2 below, starting from the upper left, each sample is assigned to a pixel, with the top row containing samples 1 through 32, the second row containing 33 through 64, etc. Since there are 1000 samples and 1024 pixels in an image the last 24 pixels in the bottom row are empty indicated by "N/A."

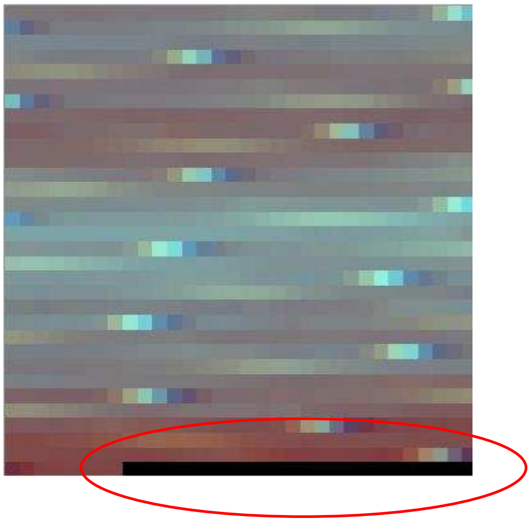

Figure 3: 32 x 32 image with the last 24 pixels empty (black pixels in lower right).

Table 2: Pixel Samples Grid

1	2	3	…	31	32
33	34	35	…	63	64
…	…	…	…	…	…
969	970	971	…	991	992
993	994	995	…	**N/A**	**N/A**

The last step for data interpretation was sorting the dataset into balanced testing and training sets. All the even samples were sent to training while all the odd samples were sent to testing resulting in a simple 50/50 split. Transfer learning on AlexNet was used to classify the images. Binary classification using the NORM and MI pathology classes was conducted. The training and testing were trimmed/balanced so there was equal representation of class.

2.3.3 Feature Defining, Extracting, and Filtering

For this method, defining, extracting, and filtering the features was employed to anonymize the data. The GNB classifier was used and the data for this test was balanced. The median frequency of the FFT (Feature 0), and the amplitude of the FTT, which represents the total amount of energy in the FFT (Feature 1), were the two features that were extracted from each signal, respectively. There was a total of 8 Features.

The median frequency of the Fast Fourier Transform (FFT) of the ECG was designated as Feature 0. The total energy (magnitude) of the FFT was Feature 1. The mean time between R Peaks was Feature 2. Feature 3 was the mean time between P Peaks. The typical time between T Peaks was Feature 4. Features 5-7 represented the R, P, and T peaks' respective mean amplitudes. The features of each signal were filtered to move it 96% of the distance to the centroid of the cluster for each pathology. The last step was taking the adjusted FFT and turning it back into a signal and attempting to use it for classification.

2.3.4 Feature Extraction and Selection

For the final process, determining optimal features for improving pathology accuracy while minimizing age accuracy was the main objective. There was a total of 8 Features. The features used in this methodology were the same set used in the previous extraction and filtering experiment and the data for this test was balanced. To determine the optimal features, the features were sorted (ranked) starting with the feature that most increased the pathology accuracy.

Additional features were sequentially included based on their ranks to see how the accuracies would be affected. The sequential feature addition order was [6,1,5,7,0,4,3,2]. When performing this test, the data was randomly assigned to training and testing groups using a random seed. There was a total of 10,000 samples in this initial exploration, with 5,000 samples in each testing and training group. The random seed was set to a constant for debugging because when the random seed does not change the data split will not change. Random seed repetition was only used in debugging.

After this was conducted, all the trials incorporated the same final set of features, so the next step in the process was removing features rather than adding them to see how the accuracies would be affected. For example, a feature set containing [6, 0, 4, 2, 1, 7, 5, 3] can be changed to [6, 0, 4, 2, 1, 5, 3] by removing feature "7." The logic was that, given there were possible correlations and interactions between features, removing features from the current set to investigate possible further improvement of results through pruning was used. As the number of features increases, the odds that removing a feature from a previous iteration may affect overall accuracy positively increases since there are more opportunities for interaction.

Next, pathology accuracy was maximized for the testing and training set and then compared. The comparison was done to check for overfitting of the data. Overfitting is when a machine learning algorithm memorizes a process of steps to solve a specific problem and then applies it to the problem without alteration. In the case of overfitting, the algorithm would memorize the input of the training data rather than learning patterns that can be applied to a more general data set. If the training accuracy is substantially higher than testing, then it would indicate overfitting. There was no overfitting in the data because the training accuracy was not substantially higher than the testing accuracy throughout, (ex. 0.5432 >0.5282).

After running this test, the optimal features were determined using the GNB classifier again for both age and pathology accuracy. The same test was performed with different classifiers such as KNN and AdaBoost as well.

2.4 Gain Ratio

To calculate the gain ratio a baseline for pathology and age was determined first. Afterwards, the original pre-classified accuracy was found. This varied from method to method. For instance, with the addition of the sinusoid method, the original accuracy was the accuracy prior to adding sinusoids. For the AlexNet method, the original accuracy was the accuracy before the signal was low-pass filtered. The feature defining, extracting, and filtering method's original accuracy was the pre-filtered accuracy. Finally, for the feature extraction and selection method, the original accuracy was the optimal feature accuracy in relation to the optimal feature set. Next was finding the maximum accuracy for age and pathology, which again varied for each method. Lastly, incremental age and pathology accuracy was found. The ratio was calculated using the gain equation that was previously mentioned.

Chapter 3

Results

3.1 Signal Noise - Adding Sinusoids (10,2)

Figure's 4 and 5 represent the mean accuracy in relationship to the number of random sinusoids added. The number of samples for this experiment was 2000. The range of sinusoids tested and added was between 1 and 10,000 using increments of 5. The graphs conditions were set to a mean frequency of 10 with a standard deviation of 2 and a sampling rate of 100Hz. The graphs data displays multiple peaks and lows before it eventually asymptotes to guessing.

As shown, the addition of extra sinusoids did affect accuracy to an extent. Ultimately, adding sinusoids should bring gender classification to 0.50 (0.52 if unbalanced) so the classifiers cannot differentiate between M or F, and 0.20 for pathologies. However, the AdaBoost classifier performed well at this goal compared to the other classifiers. The gender and pathology accuracies can be seen in Table 3. The lowest classifier was Neuralnet. The gender and pathology accuracies can be seen in Table 4.

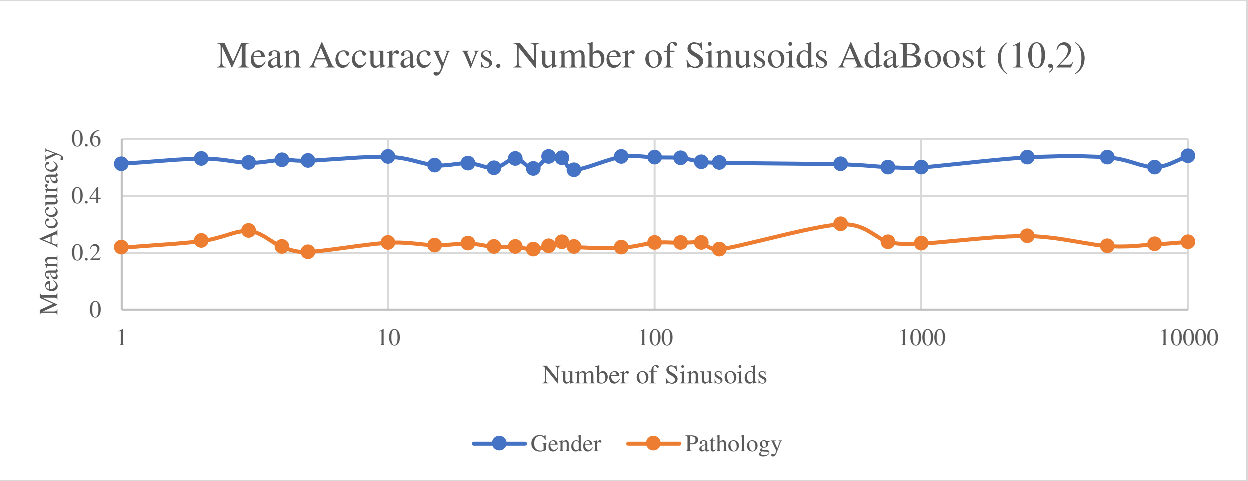

Figure 4: AdaBoost Mean Accuracy vs. Number of Sinusoids (10,2)

Table 3: AdaBoost Accuracies (10,2)

Gender accuracy w/o J func. (10,2)	**Pathology Cluster accuracy w/o J func. (10,2)**
0.527	0.226

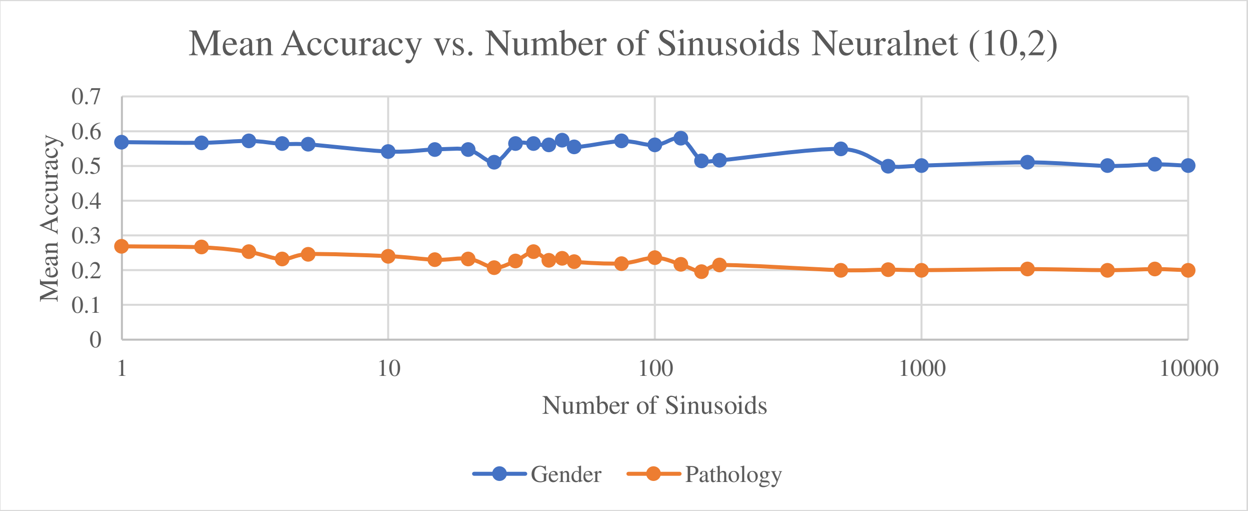

Figure 5: Neuralnet Mean Accuracy vs. Number of Sinusoids (10,2)

Table 4: Neuralnet Accuracies (10,2)

Gender accuracy w/o J func. (10,2)	Pathology Cluster accuracy w/o J func. (10,2)
0.564	0.299

3.2 Signal Noise - Adding Sinusoids (20, 5)

Figure's 6 and 7 represent the mean accuracy in relationship to the number of random sinusoids added. The number of samples for this experiment was 2000. The range of sinusoids tested and added was between 1 and 10,000 using increments of 5. The graphs conditions were set to a mean frequency of 20 with a standard deviation of 5 and a sampling rate of 100Hz. The graphs data displays multiple peaks and lows before it eventually asymptotes to guessing.

As shown, the addition of extra sinusoids did affect accuracy to an extent. Again, ultimately adding sinusoids should bring gender classification to 0.50 (0.52 if unbalanced) so the classifiers cannot differentiate between M or F, and 0.20 for pathologies. However, the AdaBoost classifier performed well again at this goal compared to the other classifiers. The gender and pathology accuracies can be seen in Table 5. The pathology accuracy was the same as seen in Table 3. The lowest classifier was again Neuralnet. The gender and pathology accuracies can be seen in Table 6. Table 7 shows the gain ratios for (10, 2) and (20, 5) across all classifiers.

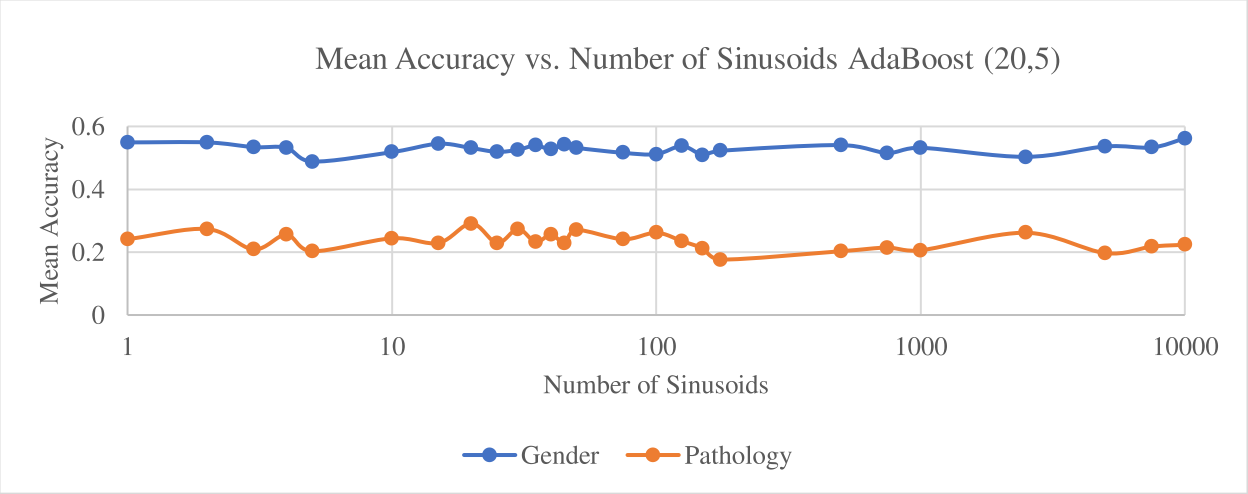

Figure 6: AdaBoost Mean Accuracy vs. Number of Sinusoids (20,5)

Table 5: AdaBoost Accuracies (20,5)

Gender accuracy w/o J func. (20,5)	Pathology Cluster accuracy w/o J func. (20,5)
0.527	0.226

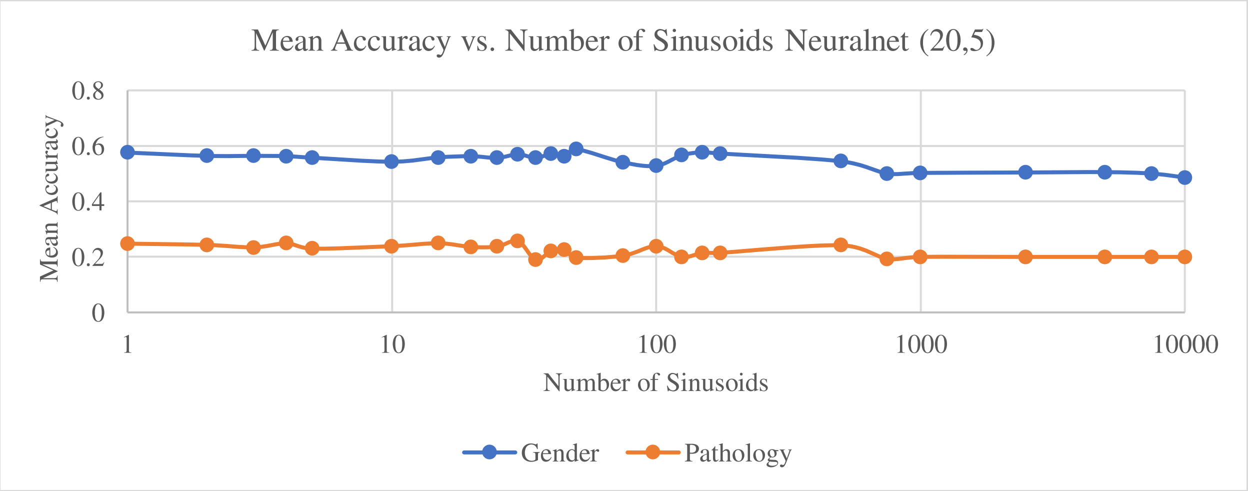

Figure 7: Neuralnet Mean Accuracy vs. Number of Sinusoids (20,5)

Table 6: Neuralnet Accuracies (20,5)

Gender accuracy w/o J func. (20,5)	**Pathology Cluster accuracy w/o J func. (20,5)**
0.556	0.278

Table 7: Gain ratios for (10, 2) and (20, 5) across all classifiers

Classifiers	**Gain Ratio (10,2)**	**Gain Ratio (20,5)**
KNN	3.55	3.43
GNB	15.45	2.09
DecisionTree	8.88	13.85
Random Forest	5.32	4.05
AdaBoost	36.11	19.44
Quadratic Discriminant	2.49	5.16
Neuralnet	0.652	0.989

3.3 AlexNet

Table 8 below represents the age and pathology accuracies at each filter frequency cut off while Table 9 shows the gain ratio.

Table 8: Age and pathology accuracies at each filter frequency

Frequencies	**All**	**30Hz**	**25Hz**	**20Hz**	**15Hz**	**10Hz**	**5Hz**	**4Hz**	**3Hz**	**2Hz**	**1Hz**
Age Accuracy	0.635	0.619	0.656	0.579	0.623	0.597	0.587	0.576	0.583	0.581	0.522
Pathology Accuracy	0.734	0.733	0.719	0.732	0.702	0.687	0.686	0.682	0.671	0.670	0.585

Table 9: Alexnet gain ratio

AlexNet Gain Ratio	1.54

3.4 Feature Defining, Extracting, and Filtering

Each pathology group is represented by a color in Figures 8 and 9 where NORM is black, MI is blue, STTC is green, CD is magenta and HYP is red. Figure 8 represents the features before filtering the FFT, while Figure 9 represents the features after filtering the FFT. The x-axis represents the median frequency (Feature 0), and the y-axis represents the amplitude (Feature 1). Table 10 shows the age and pathology accuracy before and after filtering while Table 11 shows the gain ratio.

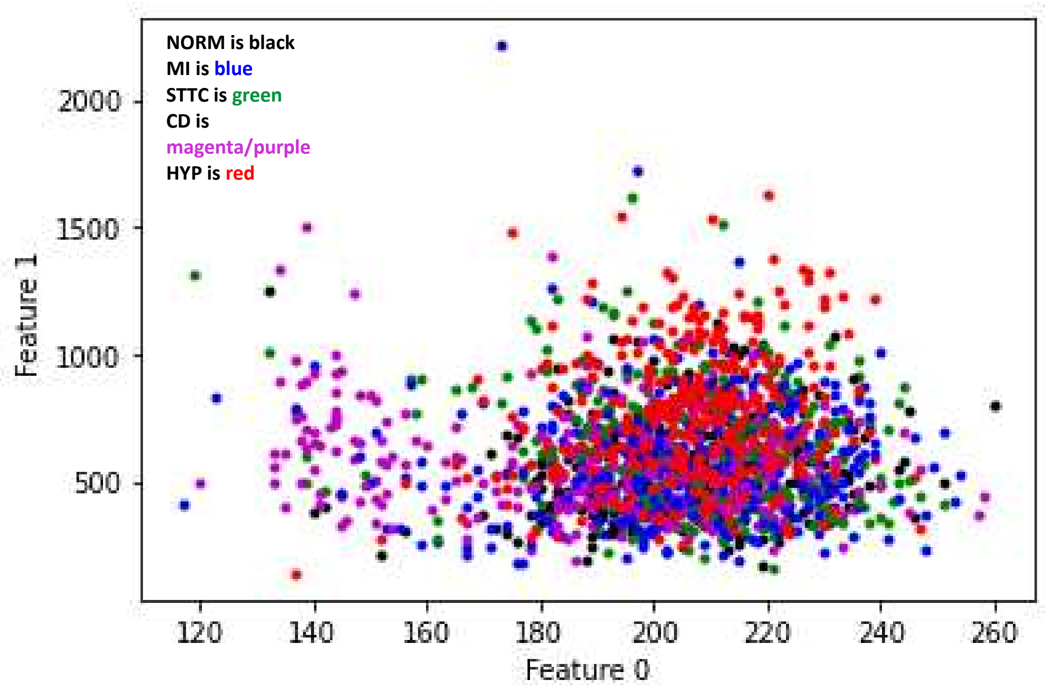

Figure 8: Before filtering FFT

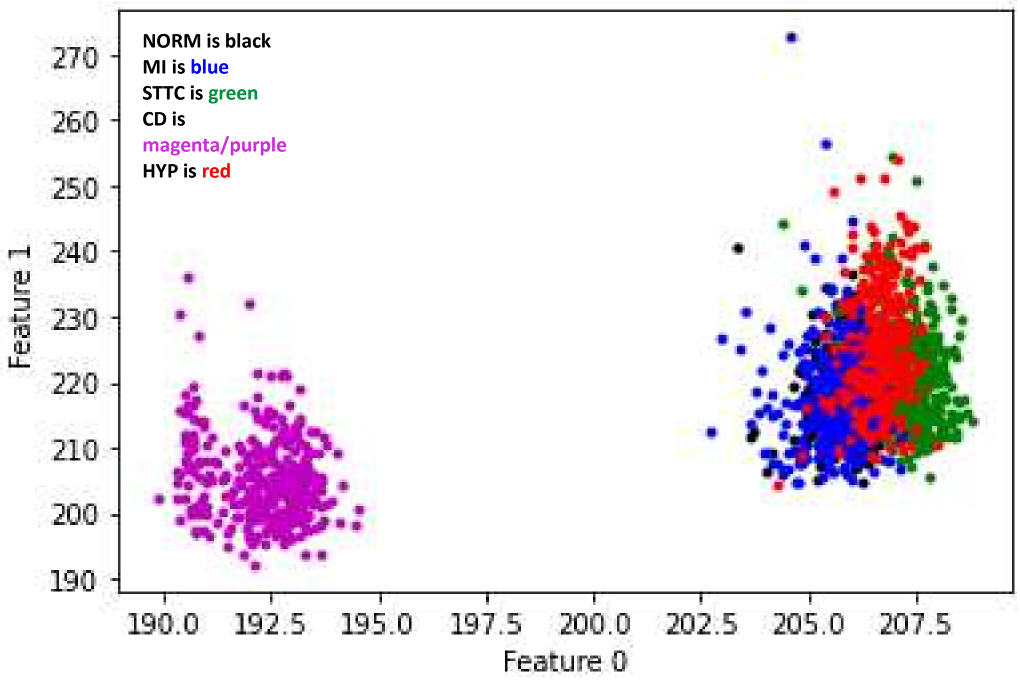

Figure 9: After filtering FFT

Table 10: Age and pathology accuracy before and after filtering

	Age	Pathology
Before filtering accuracy	0.607	0.371
After filtering accuracy	0.559	1.00

Table 11: Gain ratio

Feature Defining, Extracting and Filtering Gain Ratio	8.48

3.5 Feature Extraction and Selection

This feature combination, when added sequentially using GNB, was the most successful for maintaining high pathology accuracy while keeping age accuracy as low as possible for the testing group. As shown in Table 12, the best feature set had an age accuracy of 0.586 while the pathology accuracy was 0.388, employing a subset of 5 of the 8 features. This specific feature set contained Features: [6, 1, 5, 7, 0].

The poorest performing feature set for the testing group had an age accuracy of 0.496 while the pathology accuracy was 0.255, out of a set of 8 features to be selected. This specific feature set was just Feature 6 by itself. Figure 10 shows the accuracies plotted in relation to the number of features added.

Table 12: Testing Group Accuracies for Ranking of Features (GNB)

Feature set (GNB)	**Age Accuracy**	**Pathology Accuracy**
[6]	0.496	0.255
[6, 1]	0.516	0.298
[6, 1, 5]	0.528	0.311
[6, 1, 5, 7]	0.592	0.347
[6, 1, 5, 7, 0]	0.586	0.388
[6, 1, 5, 7, 0, 4]	0.597	0.382
[6, 1, 5, 7, 0, 4, 3]	0.601	0.369
[6, 1, 5, 7, 0, 4, 3, 2]	0.607	0.371

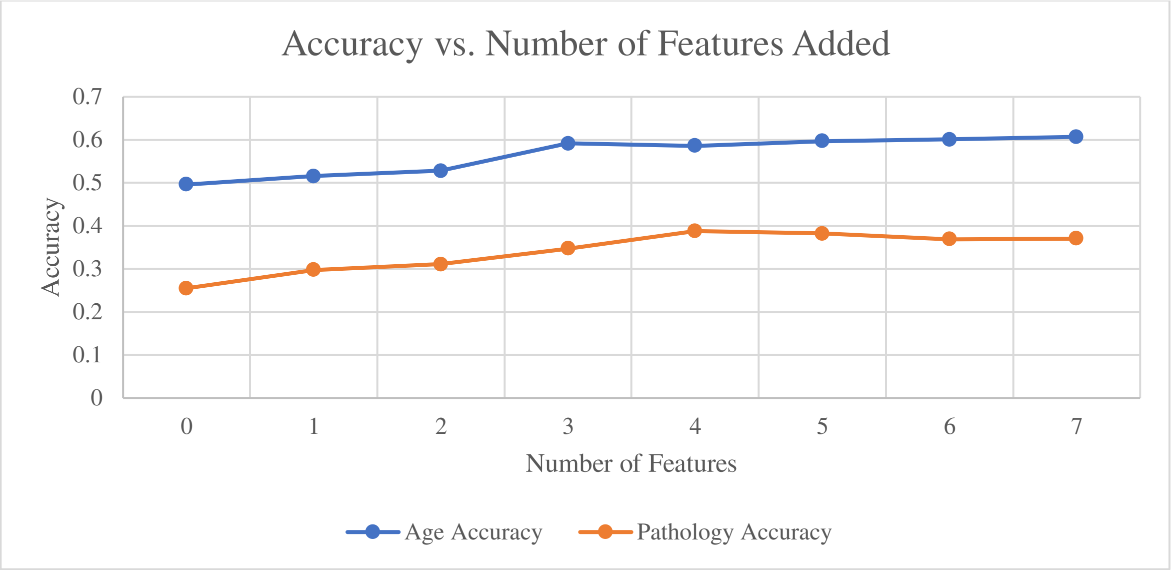

Figure 10: Accuracy vs. number of features added for testing group

Table 13 below represents the optimal feature sets for age and pathology using the GNB, KNN and AdaBoost classifiers. For KNN, the feature set [7, 6, 5] was optimal for age while [5, 6] was optimal for pathology. For GNB, the feature set [7, 5, 6] was optimal for age while [7, 5, 0, 6] was optimal for pathology. Lastly, for AdaBoost, the feature set [7, 6, 5] was optimal for age while [7, 5] was optimal for pathology. Table 14 displays the pathology class and age class distributions while Table 15 shows the gain ratio across each classifier respectively.

Table 13: Age and Pathology optimal feature sets

Classifier	**Age**	**Pathology**
KNN	[7, 6, 5]	[5,6]
AdaBoost	[7, 6, 5]	[7,5]
GNB	[7, 5, 6]	[7, 5, 0, 6]

Table 14: Pathology class and age class distributions

	Young (65>X)	**Old (65<X)**
NORM	306	47
MI	142	211
STTC	144	209
HYP	163	190
CD	143	210

Table 15: Feature set gain ratios

Classifier	**Feature Set**	**Gain Ratio**
KNN	[1, 2, 3]	4.73
AdaBoost	[2, 3, 4, 5, 6]	32.1
GNB	[0, 1, 6]	1.84

3.6 Histogram of Distribution of Features

Below, Figure’s 11-18 are the balanced histograms for the distribution of features. The x-axis represents the value of features, and the y-axis represents the number of signals that have the features within that range.

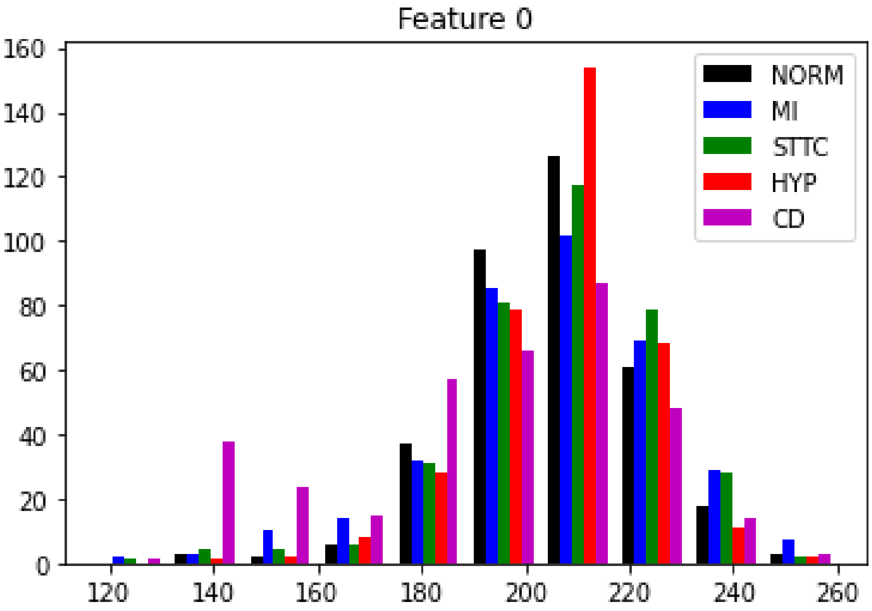

Figure 11: Feature 0 histogram

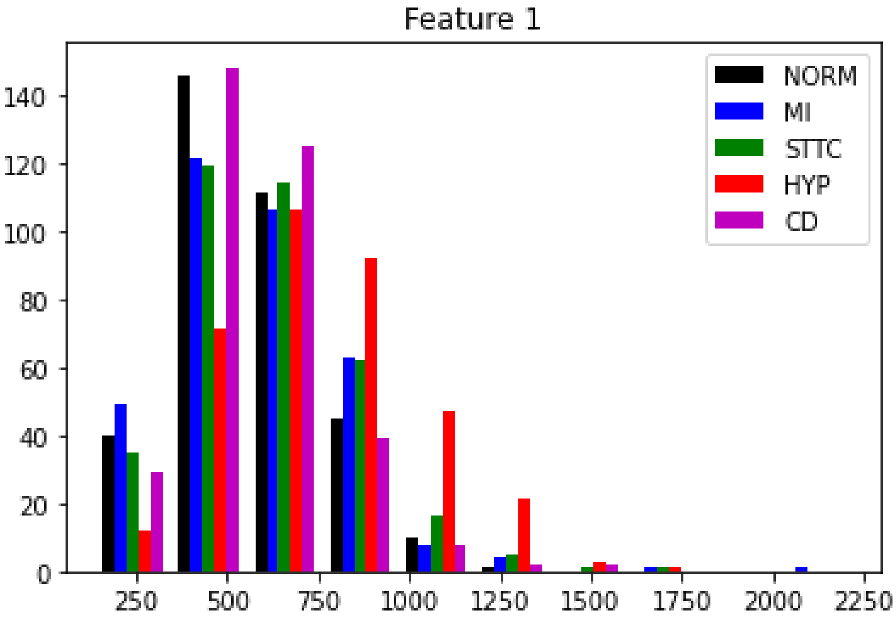

Figure 12: Feature 1 histogram

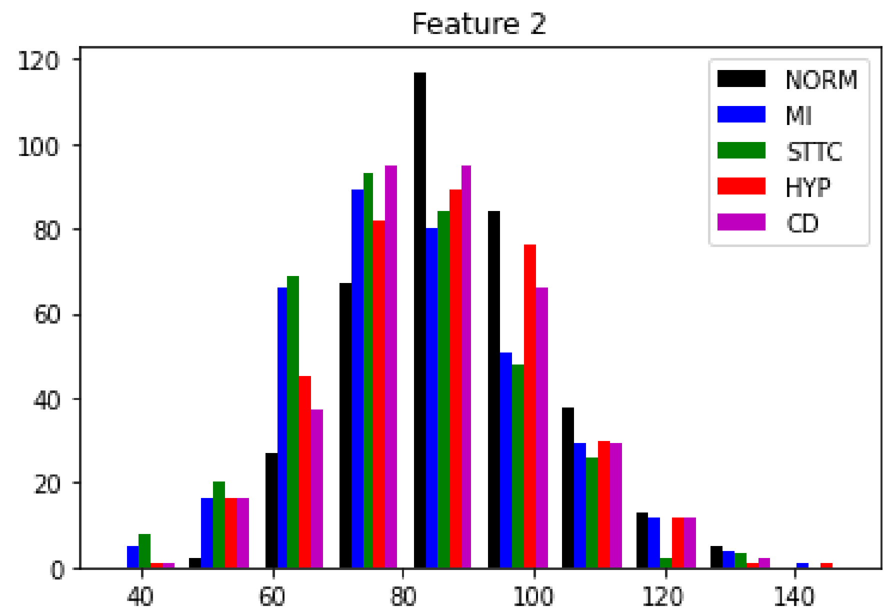

Figure 13: Feature 2 histogram

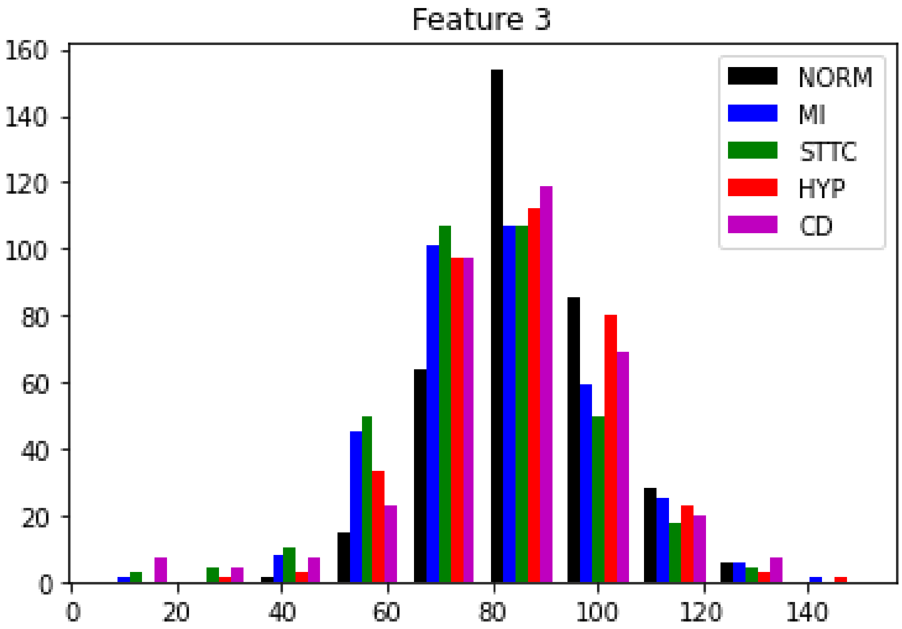

Figure 14: Feature 3 histogram

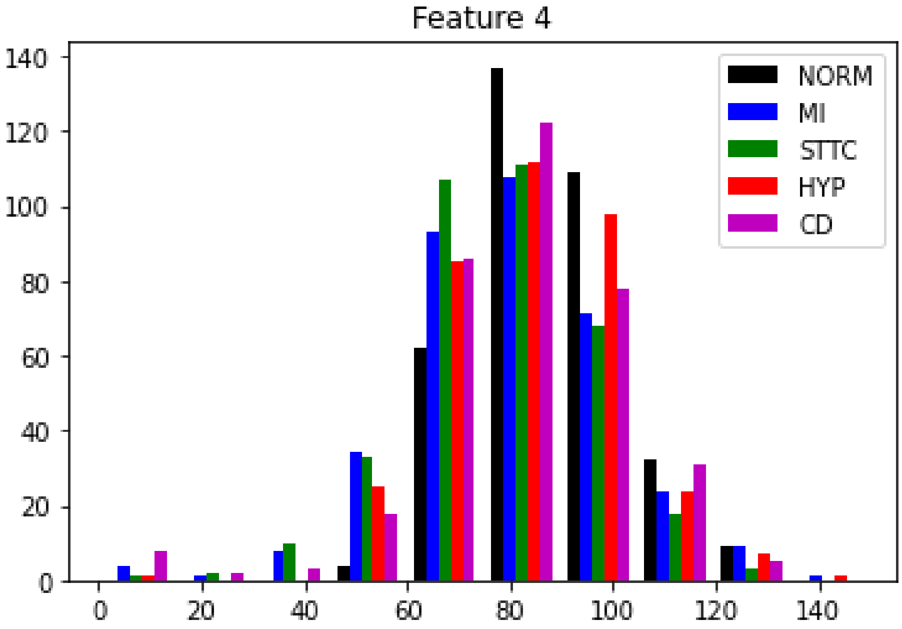

Figure 15: Feature 4 histogram

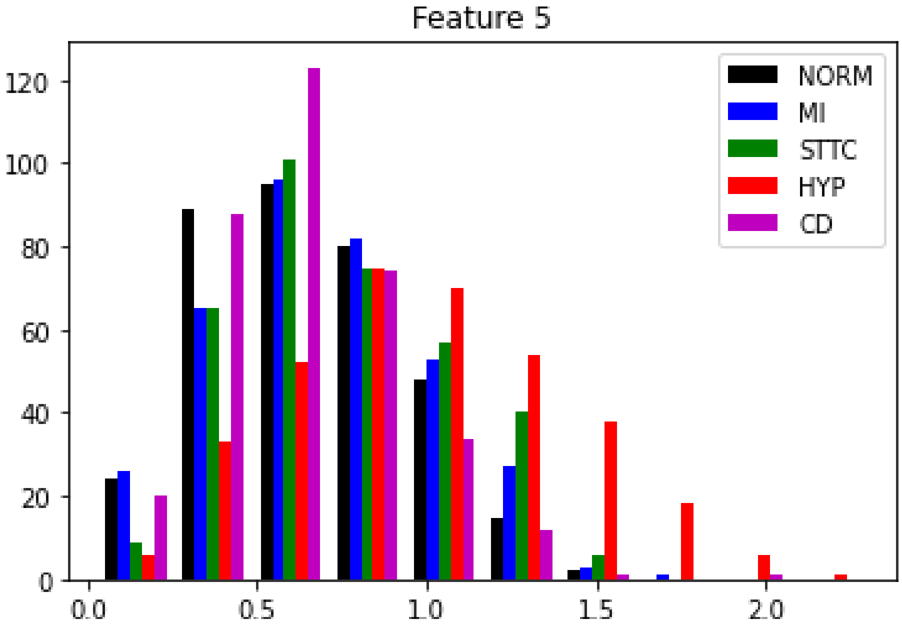

Figure 16: Feature 5 histogram

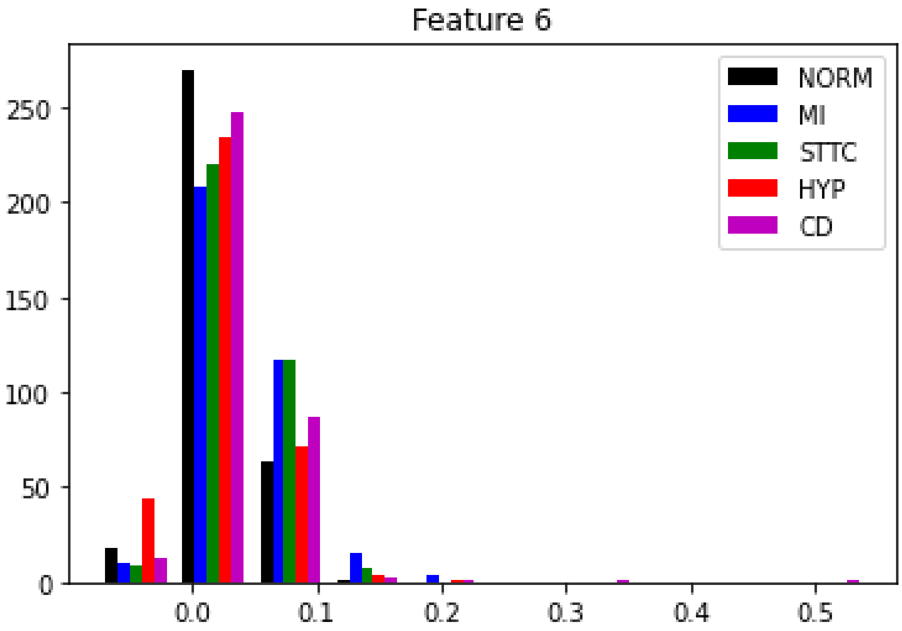

Figure 17: Feature 6 histogram

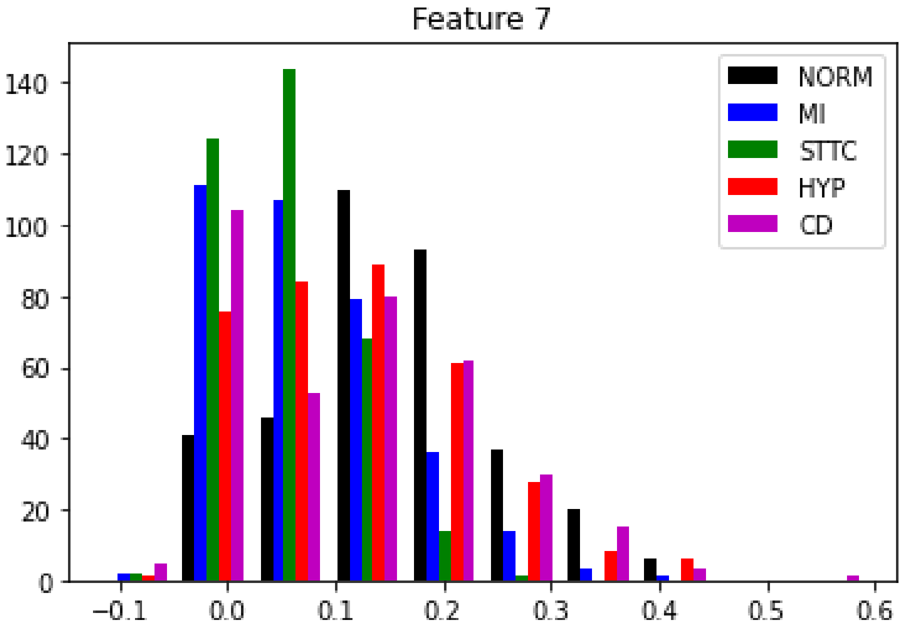

Figure 18: Feature 7 histogram

3.7 Gain Ratio

A baseline accuracy was found by using what percent of samples were from the largest class. The baseline for pathology accuracy was 0.200 while the baseline for age accuracy was 0.500 for this balanced data set. However, for the binary classification with AlexNet, the baseline accuracy for both age and pathology was 0.500. The baseline accuracy does not use any features and represents what the accuracy would be if every sample was "guessed" to be from the largest class pathology (NORM). Table 16 below compares the gain ratios calculated using the various methods tested.

Table 16: Comparative ratios

Experiment/Classifiers	**KNN**	**GNB**	**DecisionTree**	**RandomForest**	**AdaBoost**	**Quadratic Discriminant**	**Neuralnet**	**AlexNet**
Adding sinusoids (10,2) Gain Ratio	3.55	15.45	8.88	5.32	36.11	2.49	0.652	
Adding sinusoids (20,5) Gain Ratio	3.43	2.09	13.85	4.05	19.44	5.16	0.989	
AlexNet Gain Ratio								1.54
Feature Defining, Extracting and Filtering Gain Ratio		8.48						
Feature Extraction and Selection Gain Ratio	4.73	1.84			32.1			

3.8 Hamming Distances

The hamming distances between each optimal feature set for age and pathology are shown in Table 17 below.

Table 17: Hamming distances

Classifier	**Age**	**Pathology**	**Hamming Distance**
KNN	[7, 6, 5]	[5,6]	1
AdaBoost	[7, 6, 5]	[7,5]	1
GNB	[7, 5, 6]	[7, 5, 0, 6]	1

Chapter 4

Discussion

4.1 Signal Noise - Adding Sinusoids

The results show that the sinusoids were added, but that their presence had little effect on the accuracy of the results. Since accuracy tended to fluctuate erratically, there did not appear to be a consistent correlation between the number of sinusoids added and accuracy. In several instances, the accuracy was slightly improved by the sinusoids, but the effect was minor and inconsistent. There was a very minimal decrease in gender accuracy when comparing the accuracy results before and after adding the sinusoids, indicating that the addition of sinusoids did alter the accuracy results to some extent.

4.2 AlexNet

According to AlexNet data, the frequency range that gave the best pathology accuracy after applying low-pass filters was between 30Hz and 20Hz. After filtering frequencies over 30Hz, the maximum pathology accuracy was 0.733 and the lowest age accuracy was 0.619. Between 30Hz and 20Hz, pathology accuracy appeared to stay high, suggesting that pathology was favored over age. However, after filtering frequencies over 25Hz, the highest age accuracy was 0.656, whereas pathology accuracy was still high at 0.719. Throughout all filtering frequency values, pathology accuracy remained higher than age accuracy. After filtering out frequencies over 1Hz, the lowest accuracy values for both age and pathology were 0.585 for pathology and 0.522

for age. This relation implies that for each filtering frequency value, age and pathology were influenced similarly.

4.3 Feature Defining, Extracting, and Filtering

Referencing the results before filtering, a cluster of ECGs of the CD pathology in the median frequency range of 130 to 160 was observed. Many of the HYP clusters can be seen in the 180 to 230 median frequency range along with the other pathology classes. After filtering, CD is farther than the other pathology clusters with the main median frequency range being 190 to 195. Stacked clustering in the 205 to 207.5 median range frequency for the other pathology classes implies they share similar traits with their pathologies. ST/T change and normal would have similar aspects since ST/T change is a shifted frequency reading of a normal ECG. Hypertrophy and myocardial infarction are similar since typically most patients that have myocardial infarction are seen to have left ventricular hypertrophy. This similarity is due to the fact that hypertrophy can lead to myocardial infarction. Before filtering, the pathology accuracy was 0.371. However, after filtering, it was 1.00, which is unrealistic. Since the signals were down-samples to 100Hz, that may have been a reason for lower overall pathology accuracy. However, there is a possibility that down-sampling to 100Hz may improve anonymization at the same time.

4.4 Feature Extraction and Selection

The optimal feature sets for age across all three classifiers were the same set of features, [7,6,5]. This shows that the R, P, and T peaks mean amplitudes are better at distinguishing age than pathology. Between the age and pathology optimal feature sets for KNN, the difference was Feature 7. This means, by removing Feature 7, pathology can be maximized, whereas by adding

Feature 7, age is optimized. For Adaboost, removing the feature allowed for the pathology accuracy to increase. With GNB, the addition of Feature 0, median frequency, the pathology accuracy was optimized. When removing Feature 0, age accuracy was maximized. Across all six feature sets, the main set of reoccurring Features were 5, 6, and 7. However, this is interesting, since the histograms suggest that the combination of Features 0, 1, and 5 should be useful for classifying pathology classes. The remaining Features (2, 3, 4, 6, 7), based on rank, should be less useful for classifying pathologies. The same set of features, except for Feature 1, are also useful for age classification. The pathology class and age distributions show that many of the young patients were normal while the number of old patients was similar across MI, STTC, and HYP. There were very few healthy old patients.

4.5 Gain Ratio

When examining the ratio numbers across all methods, the larger the ratio, the more successfully the age was de-identified. The starting point of the ratio is one. According to the data, age accuracy is approaching fifty as compared to the initial baseline value. The ratios for those would therefore be infinite. This means age was able to be fully anonymized while also increasing pathology. The ratio for the addition of the sinusoid's method showed that for both the (10, 2) and (20, 5) iterations, the AdaBoost classifier produced the highest gain ratio of 36.11 and 19.44. The lowest gain ratio for both the (10, 2) and (20, 5) iterations was using Neuralnet producing a gain ratio of 0.652 and 0.989. The second highest ratio for (10, 2) was GNB with 15.45 and for (20, 5) it was DecisionTree with 13.85. The gain ratio for the AlexNet approach was 1.54, while for the feature filtering technique it was 8.48.

Lastly, for the feature selection method, the best ratio was AdaBoost with a ratio of 32.1 while the lowest ratio was GNB with 1.84. KNN had a ratio of 4.73. This indicates anonymization was achieved sufficiently so that someone may know which pool of people the patient belongs to but not identify the actual individual patient. It also shows that the pathology and features were improved. Overall, the two highest similar gain ratios achieved across the various methods were the addition of sinusoids with the use of the AdaBoost classifier for the (10, 2) iteration and the feature selection method using the AdaBoost classifier to find the optimal feature set. This suggests that these two methods would be a good metric for the selective anonymization algorithm. This ratio allows us to select which types of classification to anonymize and which to selectively improve the accuracy of—including age, gender, race, height, pathology, location, diet, exercise regime, and any combination thereof.

4.6 Suggestions for Future Work

As seen through the results, the process works regardless of not having the perfect set of features. ECGs can be difficult to classify since their frequency histograms (FFTs) are bimodal, with high frequencies shared regardless of the pathology. Future tests could include shape templating so the higher frequencies can be dropped out easier. Instead of AlexNet, using text CNN on the time series is also an option. Using other pathology classes for binary classification on AlexNet other than NORM and MI to see how the accuracies differ is a possibility. A future possibility is investigating feature ratios or impact features on pathology accuracy versus age accuracy.

Chapter 5

Conclusion

Biometrics are often used for immigration control, business applications, civil identity, and healthcare. Biometrics can also be used for authentication, monitoring (e.g., subtle changes in biometrics may have health implications), and personalized medical concerns. Increased use of biometrics creates identity vulnerability through the exposure of personal identifiable information (PII). Hence there is an increasing need to not only validate but secure a patient's biometric data and identity. The latter is achieved by anonymization, or de-identification, of the PII. Using Python in collaboration with the PTB-XL ECG database from Physionet, the goal of this thesis was to create "selective de-identification."

A various number of methods were tested, however, the two that successfully worked in relation to our objective were the addition of sinusoids with the use of the AdaBoost classifier for the (10, 2) iteration and the feature selection method using the AdaBoost classifier to find the optimal feature set. These two methods produced the highest gain ratio of 36.11 and 32.1 indicating that anonymization was achieved sufficiently so that someone may know which pool of people the patient belongs to but not identify the actual individual patient. It also shows that the pathology and features were improved. This suggests that these two methods would be a good metric for the selective anonymization algorithm.

References

1. Dantcheva, P. Elia and A. Ross, "What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics," in IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 441-467, March 2016, doi: 10.1109/TIFS.2015.2480381.
2. I. Natgunanathan, A. Mehmood, Y. Xiang, G. Beliakov and J. Yearwood, "Protection of Privacy in Biometric Data," in IEEE Access, vol. 4, pp. 880-892, 2016, doi: 10.1109/ACCESS.2016.2535120.
3. Gregg RE, Yang T, Smith SW, Babaeizadeh S. ECG reading differences demonstrated on two databases. J Electrocardiol. 2021 Nov-Dec;69S:75-78. doi: 10.1016/j.jelectrocard.2021.09.005. Epub 2021 Sep 10. PMID: 34544590.
4. Nakano, K., & Chakraborty, B. (2019). Effect of Data Representation for Time Series Classification - A Comparative Study and a New Proposal. *Mach. Learn. Knowl. Extr., 1*, 1100-1120.
5. Morrone, Adam & Anderson, Wes & Simske, Steven. (2021). Occluded Image Function: A Novel Measure for Evaluating Machine Learning Classifiers for Biometrics. Journal of Imaging Science and Technology. 66. 10.2352/J.ImagingSci.Technol.2022.66.1.010501.
6. Habib, Md. Ahsan & Karmakar, Chandan & Yearwood, John. (2021). Learning post-processing for QRS detection using Recurrent Neural Network.
7. Llamedo et al., “A automatic patient-adapted ECG heartbeat classifier allowing expert assistance”, IEEE Transiction on Biomed. Eng., vol. 59, no. 8, pp. 2312-2320, 2012.
8. R. Ceylan, Y. Ozbay, and B. Karlik, "A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network," Expert Syst. with Applicat., vol. 36, no. 3, pp. 6721-6726, 2009.

9. A. Dallali, A. Kachouri, and M. Samet, "Classification of Cardiac Arrhythmia Using WT, HRV, and Fuzzy C-Means Clustering," Signal Processing: An Int. J. (SPJI), vol. 5, no. 3, pp. 101-109, 2011.

10. R. Acharya et al., "Classification of cardiac abnormalities using heart rate signals," Medical and Biological Eng. and Computing, vol. 42, no. 3, pp. 288-293, 2004.

11. M. K. Das and S. Ari, "ECG Beats Classification Using Mixture of Features" Int. Scholarly Research Notices, 2014.

12. Z. Zidelmal, A. Amirou, D. O. Abdeslam, and J. Merckle, "ECG beat classification using a cost sensitive classifier," Comput. methods and programs in biomedicine, vol. 111, no. 3, pp. 570-577, 2013.

13. M. Vijayavanan, V. Rathikarani, and P. Dhanalakshmi, “Automatic Classification of ECG Signal for Heart Disease Diagnosis using morphological features,” Int. J. of Comput. Sci. and Eng. Technology (IJCSET), vol. 5, no. 4, pp. 449-455, 2014.

14. R. Ghongade et al., “Performance analysis of feature extraction schemes for artifical neural network based ECG classification”, International conference on Computational Intelligence and Multimedia Applications 2007.

15. Ince et al., “A generic and robust system for automated patient-specific classification of ECG signals”, IEEE Trans. On Biomed. Eng., vol. 56, no. 5, pp. 1415-1426, 2009.

16. Jambukia, Shweta H., Vipul K. Dabhi, and Harshadkumar B. Prajapati. "Classification of ECG signals using machine learning techniques: A survey." Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in. IEEE, 2015.

17. B. Mohamed et al., “ECG image classification in real time based on the harr-like feature and artificial neural networks”, In procedia Computer Science vol. 73, pp. 32-39, 2015.

18. W M. Hayat , M. Bennamoun , S. An , “Deep reconstruction models for image set classification,” IEEE Trans. Pattern Anal. Mach. Intell. Vol.37, pp. 713-727, 2015.

19. J. Bai , Y. Wu , J. Zhang , F. Chen , “Subset based deep learning for RGBD object recognition,” Neurocomputing vol.165, pp.280–292, 2015.

20. Z. Huang , R. Wang , S. Shan , X. Chen , “Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning,” Pattern Recognit. Vol. 48, pp. 3113–3124, 2015.

21. T. Brosch , R. Tam , Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2D and 3D images, Neural Comput. Vol.27, pp. 211–227, 2015

22. J. Deng , Z. Zhang , F. Eyben , B. Schuller, “Autoencoder-based unsupervised domain adaptation for speech emotion recognition, IEEE Signal Process. Lett.” Vol. 21,pp. 1068–1072, 2014.

23. AL Rahhal, M. M., et al. “Deep learning approach for active classification of electrocardiogram signals”, Information science vol. 345, pp. 340-354, 2016.

24. Rahhal, M. A., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F., & Yager, R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, *345*, 340-354. https://doi.org/10.1016/j.ins.2016.01.082

25. S. Kiranzyaz et al., “ Real time patient-specific ECG classification by 1- D convolutional neural networks”, IEEE Trans. On Biomed. Eng., vol. 63, no. 3, 2016.

26. Ferretti, Jacopo & Randazzo, Vincenzo & Cirrincione, Giansalvo & Pasero, Eros. (2021). 1-D Convolutional Neural Network for ECG Arrhythmia Classification. 10.1007/978-981-15-5093-5_25. Rajpurkar P, Hannun AY, Haghpanahi M et al (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836

27. Jun, T.J., Nguyen, H.M., Kang, D., Kim, D., Kim, D., & Kim, Y. (2018). ECG arrhythmia classification using a 2-D convolutional neural network. *ArXiv, abs/1804.06812*.

28. Johnsana, J. & Appusamy, Rajesh & Verma, Kishore. (2016). CATs-Clustered k-Anonymization of Time Series Data with Minimal Information Loss and Optimal Re-identification Risk. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i47/101081.

29. M. Zare-Mirakabad, F. Kaveh-Yazdy and M. Tahmasebi, "Privacy preservation by k-anonymizing Ngrams of time series," *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, 2013, pp. 1-6, doi: 10.1109/ISCISC.2013.6767335.

30. Wagner, Patrick, et al. “PTB-XL, a Large Publicly Available Electrocardiography Dataset.” *PTB-XL, a Large Publicly Available Electrocardiography Dataset v1.0.1*, 24 Apr. 2020, https://physionet.org/content/ptb-xl/1.0.1/#files-panel.