Utilizing Generative Adversarial Networks to Synthesize Electrocardiogram Data to Identify Various Arrhythmias with a Convolutional Neural Network Classifier

Iva Chee & Ella Chee

Yorktown High School

ABSTRACT

Arrhythmia identification from electrocardiograms through deep learning is a proven method for diagnosis. However, there is an imbalance of arrhythmia samples in existing databases, which can result in less accurate classification of more uncommon but still dangerous arrhythmias—which can go misdiagnosed, and result in a greater risk for stroke, cardiac arrest, and death. Convolutional neural network (CNN) arrhythmia classification via ECG datasets augmented through generative adversarial networks (GAN) can offer more accurate and effective diagnostic procedures. Our study aimed to create and evaluate the effectiveness of a one-dimensional CNN (1D-CNN) that automatically detects and identifies heart rhythms (sinus rhythm, sinus bradycardia, atrial fibrillation, sinus tachycardia, atrial flutter, sinus irregularity, supraventricular fibrillation, and atrial tachycardia) from short-term ECG recordings out of a dataset including over 10,000 real ECG samples and TimeGAN-synthesized data. ECG data underwent preprocessing and segmentation, and was split into training and testing sets conducted with 5-fold cross-validation with an 80/20 split. The 1D-CNN 8-layer model was developed and evaluated with several performance metrics, and resulted in inconsistent accuracies (95.576% for sinus bradycardia, but 5.514% for sinus irregularity). The TimeGAN was evaluated based on a visualization, which revealed completely unrealistic generated data. However, with further modifications and improvement, synthesized data could potentially be utilized to resolve the class imbalance issue, thus improving the performance of the CNN. Overall, the findings of this study, in tandem with previous research, could work to advance diagnostic procedures of arrhythmias through the use of GANs and CNNs for ECG analysis.

INTRODUCTION

Overview of Research Problem

Cardiovascular diseases are the main cause of death across the globe, as seen in Figure 1, with more than 17 million people dying each year (Ritchie & Roser, 2018). Furthermore, the prevalence and mortality of cardiovascular diseases continue to grow, underscoring the importance and necessity of regular heart rhythm monitoring to manage and curb cardiovascular diseases (World Health Organization: WHO, 2019). Cardiac arrhythmias are a class of cardiovascular diseases, characterized by an irregular rhythm of the heartbeat when electrical impulses in the heart fail to work properly. When those electrical impulses fail to function, arrhythmias can result in dangerous complications such as blood clots, stroke, heart failure, cardiac arrest, and even sudden death. One major issue with arrhythmia diagnosis is that various arrhythmias can present the same symptoms: palpitations, chest pain, weakness, shortness of breath, fainting (Mayo Clinic, 2017). Fortunately, electrocardiograms (ECGs) record the electrical activity of the heart, allowing for the discovery of specific abnormalities in the heart's rhythm. Specific arrhythmias are still difficult to identify, however, the signal processing and feature extraction of deep learning methods can allow for more accurate diagnosis--even better than human cardiologists (Hannun et al., 2017).

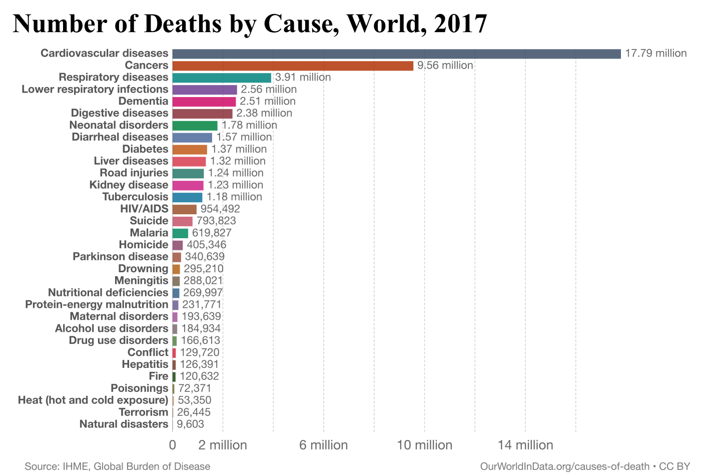

Figure 1. Cardiovascular diseases, such as arrhythmias, are the leading cause of death worldwide (Ritchie, 2018).

Review of Literature

Recently, deep learning techniques have accumulated much attention in biomedical engineering; specifically, the convolutional neural network (CNN) has grown in popularity due to its variety of successful applications. Several studies have applied CNNs to detect, predict, or classify arrhythmia from ECGs (Hannun et al., 2017; Krittanawong et al., 2019; Zheng et al., 2020). Because cardiac arrhythmias present severe health risks, techniques to improve deep learning models for diagnosis require more research (Krittanawong et al., 2019).

2D-CNN for Arrhythmia Classification via ECG

In a 2017 study, a dataset was built with over 500 times the number of patients than previous studies and was used to train a 34-layer two-dimensional CNN that maps ECG samples to a sequence of rhythm categories. The developed algorithm's performances in positive predictive value, sensitivity, and F1 score (.809, .827, .809, respectively) were superior to that of a board of certified cardiologists in the detection of several cardiac arrhythmias from electrocardiograms (Hannun et al., 2017). The findings of this study indicate the success in using a large dataset for an arrhythmia rhythm classifier.

ECG Arrhythmias Detection Using ACGAN and Residual Network

One 2019 study worked to investigate the use of a generative adversarial network (GAN) to overcome the issue of an imbalanced ECG dataset by constructing a deep learning ECG abnormality detection framework composed of two models: a data augmentation model and a classification model. Wang et al. created an auxiliary classifier GAN to generate artificial data (as seen in Figure 2) from the MIT-BIH and competition datasets, then used a stacked residual network with LSTM for single heartbeat detection and consecutive heartbeat detection. Despite successful data augmentation and network classification performance, the accuracy of the consecutive heartbeat detection (99.81) was not as high as that of the standard benchmark. Thus, Wang et al. established the usefulness of a GAN for data augmentation, while also highlighting the necessity for further research in terms of rhythm (consecutive heartbeat) classification.

Proposed ACE-GAN Arrhythmia Classification System

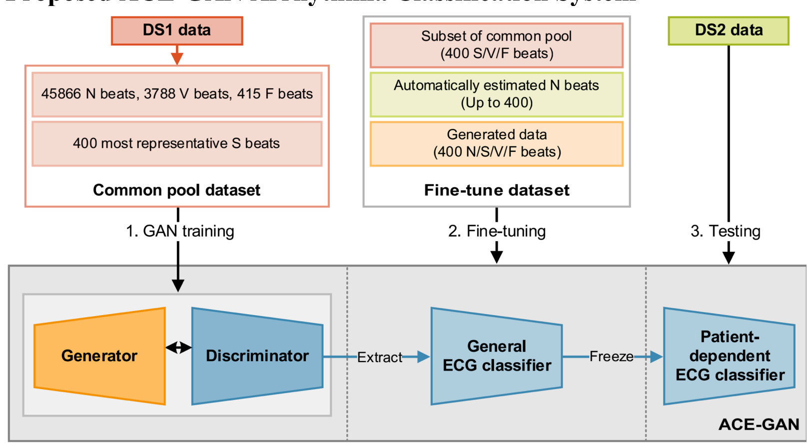

Figure 2. ACE-GAN first trained with DS1 dataset, then ECG classifier is extracted from trained discriminator and fine-tuned. ECG classifier then serves as final classifier to predict heartbeat labels for DS2 data (Zhou et al., 2021).

TimeGAN for Time-Series Data

GANs are mainly composed of generators and discriminators. The basis for the system is that the generator is trained to create fake data which the discriminator tries to classify as real or fake. Yoon et al. developed a TimeGAN (time-series generative adversarial network) consisting of four parts: an embedding function, recovery function, sequence generator, and sequence discriminator. The purpose of the first two components is so that the TimeGAN can encode features, generate representations, and iterate across time—which allows it to more efficiently and effectively generate time-series data (such as ECGs). They tested their GAN with time-series data, such as multivariate sinusoidal sequences, sequences of stock prices, energy prediction datasets, and sequences of lung cancer events, and compared their performance to previous types of GANs, as shown in Table 1 (Yoon et al., 2019).

Table 1. Results on Time-Series Datasets The TimeGAN performs with the lowest discriminative and predictive scores compared to several previously used GANs, establishing its success (Zheng et al., 2020).

Metric	Method	Sines	Stocks	Energy	Events
Discriminative Score (Lower the Better)	TimeGAN	**.011±.008**	**.102±.021**	**.236±.012**	**.161±.018**
	RCGAN	.022±.008	.196±.027	.336±.017	.380±.021
	C-RNN-GAN	.229±.040	.399±.028	.499±.001	.462±.011
	T-Forcing	.495±.001	.226±.035	.483±.004	.387±.012
	P-Forcing	.430±.027	.257±.026	.412±.006	.489±.001
	WaveNet	.158±.011	.232±.028	.397±.010	.385±.025
	WaveGAN	.277±.013	.217±.022	.363±.012	.357±.017
Predictive Score (Lower the Better)	TimeGAN	**.093±.019**	**.038±.001**	**.273±.004**	**.303±.006**
	RCGAN	.097±.001	.040±.001	.292±.005	.345±.010
	C-RNN-GAN	.127±.004	**.038±.000**	.483±.005	.360±.010
	T-Forcing	.150±.022	**.038±.001**	.315±.005	.310±.003
	P-Forcing	.116±.004	.043±.001	.303±.006	.320±.008
	WaveNet	.117±.008	.042±.001	.311±.005	.333±.004
	WaveGAN	.134±.013	.041±.001	.307±.007	.324±.006
	Original	.094±.001	.036±.001	.250±.003	.293±.000

Fully Automatic ECG ACE-GAN Classification System

Zhai, Zhou, and Tin developed a system that utilized a GAN to create more arrhythmic heartbeats for data augmentation and used the discriminator as a classifier (as shown in Figure 3). The system outperformed several state-of-the-art automatic systems in detecting supraventricular ectopic beats (greater accuracy of 99, recall of 87, F1 score of 86) and ventricular ectopic beats (greater recall of 93 and F1 score of 93). While their findings demonstrated the effectiveness of the implementation of a GAN in ECG classification, they utilized the MIT-BIH Arrhythmia database and classified beat type, whereas we hope a larger, more comprehensive dataset and classification based on rhythm type will produce better results (Zhou, Zhai & Tin, 2021).

Generated Data and Original Data of Atrial Premature Contraction

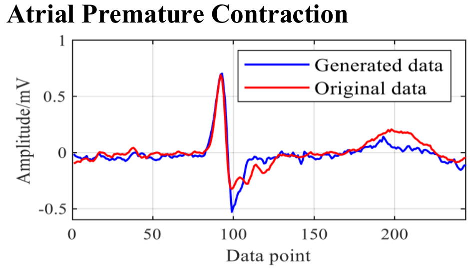

Figure 3. The GAN effectively synthesized several artificial beat types in order to enlarge the dataset (Wang et al. 2019).

Limitations of Previous Research

Past research has demonstrated the success of utilizing CNNs to classify various arrhythmias. However, one key issue is an imbalance of arrhythmia samples. In the United States, although approximately 16.41 million individuals have cardiac arrhythmias, some, like atrial fibrillation (AFIB), are more common, whereas others, like supraventricular tachycardia (SVT), are rarer (CDC, 2020; Rehorn et al., 2021). This results in lopsided datasets, with a significant amount of data for a more common arrhythmia and very little data for rarer arrhythmias, which in turn results in less effective neural networks. However, GANs can be implemented to synthesize data to allow for even larger datasets, as well as resolve the imbalance in data—like that of arrhythmias present in ECGs (Wang et al., 2019; Brophy et al., 2021). And similar to the use of CNNs and arrhythmia classification, many of the GANs that have been developed were used to augment specific beats, rather than whole rhythm arrhythmias. Combining a CNN and a GAN may result in an even more effective and useful system to advance diagnostic procedures for arrhythmias.

Problem Statements

Phase I

- PS1: Screening methods for atrial fibrillation from ECGs are lacking; low diagnostic yield of a single ECG to detect momentary arrhythmia, prolonged monitoring methods
- PS2: Few CNNs classify arrhythmias by rhythm type; past studies focus on beat type and those that did use rhythm type were not as accurate or used less realistic data

Phase II

- PS1: Few CNNs classify arrhythmias by rhythm type (i.e. atrial fibrillation and supraventricular tachycardia); past studies focus on beat type (i.e.: atrial premature beat, premature ventricular contraction) and those that did use rhythm type were not as accurate or used less realistic data (Erdenebayar et al. 2019).
- PS2: Class imbalance issue requires a solution—currently proposed techniques to resolve this issue (random resampling or cost-sensitive learning) could overfit the training data and reduce generalization when actually applied (Zhou et al., 2021).

Objectives

Phase I

- Obj1: Create and evaluate the effectiveness of a deep neural network that automatically detects and identifies atrial fibrillation rhythms from short-term ECG recordings, in order to advance diagnostic procedures.
- Obj2: Classify arrhythmias from ECGs by rhythm type: atrial fibrillation or not (normal sinus rhythm or other arrhythmias), from ECGs containing various types of rhythms.

Phase II

- Obj1: Create and evaluate the effectiveness of a deep convolutional neural network that automatically detects and identifies arrhythmias--sinus rhythm (SR), sinus bradycardia (SB), atrial fibrillation (AFIB), sinus tachycardia (ST), atrial flutter (AF), sinus irregularity (SI), supraventricular tachycardia (SVT), atrial tachycardia (AT)--by rhythm type from short-term ECG recordings.
- Obj2: Create and evaluate a GAN to synthesize more ECG data to be used as input for the CNN (solving the class imbalance issue).

Hypotheses

Phase I

- H1: A CNN can be trained to identify AFIB present in a short-term 30-second 12-lead ECG from a dataset of patients with various arrhythmias and healthy sinus rhythms.
- H2: ECGs from patients with various arrhythmias can be used to distinguish atrial fibrillation rhythm arrhythmia from normal sinus rhythm or other arrhythmias.

Phase II

- H1: A CNN can be trained to accurately and efficiently identify several types of arrhythmias present in a short-term 10-second 12-lead ECG from a dataset of patients with various arrhythmias and healthy sinus rhythms.
    - Hannun et al. developed a successful 1D-CNN that classified 12 classes of arrhythmias, sinus rhythm and noise from longer 30-second ECGs (2017).
- H2: A GAN can be used to augment existing short-term 10-second 12-lead ECGs of several types of arrhythmias and resolve class imbalance issue and enlarge the dataset.

- Yoon et al. created a TimeGAN that can synthesize accurate time-series data (i.e.: multivariate sinusoidal sequences and sequences of stock prices); this TimeGAN may be applicable for other time-series data—ECGs (2019).
- Wang et al. developed a GAN to create a larger dataset of ECG signals, and used the generated data to train/test an LSTM for arrhythmia classification; their model achieved high accuracy, sensitivity, and specificity (2019).

PHASE I (SUMMER 2020) METHODOLOGY

Role of Student vs. Mentor

Over the course of six weeks during the summer of 2020 (over 90 hours), we were responsible for preprocessing the ECG signals, constructing CNN, training/testing the model, and completing data analysis to evaluate its performance. Our mentor supervised the processes above and provided guidance and assistance for more specific programming, as well as troubleshooting errors.

Equipment & Materials

Data Collection

Short-term ECG samples were taken from two publicly available PhysioNet databases: the MIT-BIH Arrhythmia Database and the MIT-BIH Atrial Fibrillation Database. Many studies have used these data for arrhythmia detection via ECG, and methods for identification have been greatly improved. The MIT-BIH Arrhythmia Database consisted of 48 annotated, half-hour excerpts of two-channel ambulatory ECG recordings, taken from 47 subjects at 360 Hz. The MIT-BIH Atrial Fibrillation Database contained 25 annotated, 10-hour ECG records from atrial fibrillation patients taken with a Holter monitor at a sampling frequency of 250 Hz. A combination of data from the two databases allowed for a more realistic representation of arrhythmias—the classifier sorted between samples containing several types of arrhythmia, atrial fibrillation, and normal sinus rhythm.

CNN Implementation

The CNN was developed in Python using the Keras framework and Tensorflow backend. Python waveform-database, NumPy, and SciPy packages were also utilized for data analysis.

Procedures

Data Processing and Segmentation

Since Atrial Fibrillation records were at a sampling frequency of 250 Hz, they were resampled to 360 Hz using fast Fourier transform, in order to match the frequency of the Arrhythmia ECG records. The upper signal, modified limb lead II (MLII), was extracted for use while the lower signal was excluded, due to the fact that normal beats are often difficult to detect in the lower signal, and to simplify classification in the CNN (rather than having to extract information from two sets of data) (Moody & Mark, 2001). The ECG records were then segmented into samples of 30 seconds, resulting in a total of 30,480 short-term samples, which were used as the input for the neural network model. Of those samples, 60 contained noise, a change in signal quality, missed beats, pauses, or tape slippage, and 90 contained no data. Those defective samples were therefore removed from the dataset, resulting in 30,330 remaining segments. The samples were then organized into the following groups based on annotations: atrial fibrillation ('A'), atrial fibrillation and other arrhythmias ('B'), other arrhythmias ('O'), and normal sinus rhythm ('N'), as seen in Table 2. The group, 'other arrhythmias,' included atrial bigeminy, atrial flutter, ventricular bigeminy, nodal (AV junctional) rhythm, and ventricular flutter, among other arrhythmias. 5-fold cross-validation (CV) was carried out in order to obtain training and testing sets with an 80% to 20% proportion, respectively. This split was chosen as it was proven to be most effective, as shown in Figure 4 (Gholamy et al., 2018).

Table 2. Types of ECG Samples from MIT-BIH Databases
Samples containing AF, AF and other arrhythmias, other arrhythmias, and sinus rhythm found in the MIT-BIH databases, and the combined databases (Source: Student Researchers).

Type of Sample/Database	AF	AF & Other	Other	Normal
Arrhythmia DB	8	25	1048	1739
Atrial Fibrillation DB	11226	37	188	16059
Both DB	11234	62	1236	17798

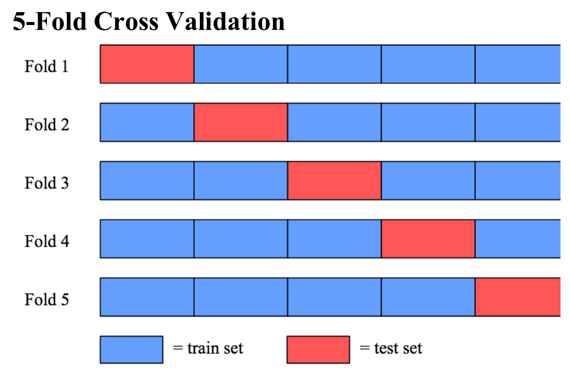

Figure 4. Process of 5-fold CV: with 5 folds, data is split into fifths, and the fifth used for testing is alternated each fold, maximizes data used for training and testing (Source: Student Researchers).

CNN Model Architecture

The CNN model was designed to identify AF from a short-term ECG (Figure 5). The CNN comprised three primary layers—convolutional layers, pooling layers, and a fully-connected layer. First, the convolutional layer created feature maps by merging the whole input signal using several kernels. Since the input ECG signal was a 1D time series, a one-dimensional convolution was used. The proposed CNN model contained four convolutional layers with a kernel size of 36, 100 filters, a stride of 1, an input size of 10,800 x 1, and a batch size of 24, as seen in Table 3. The kernel size and batch size were chosen based on the hyperparameter tuning experiments. Then, the pooling layer chose representative features by decreasing the size of the feature maps with max-pooling. The 1D max-pooling layers contained a kernel size of 36 and had a pool size of 3. Two more convolutional layers were added, followed by a global average pooling layer with a kernel size of 36. A dropout layer was also included to prevent overfitting, with a rate of 0.5. Finally, the fully-connected layer contained a dense layer and sigmoid activation function for the concluding differentiation of the model (binary classification between 'not atrial fibrillation' and 'atrial fibrillation'). The rectified linear unit (ReLU) activation function was used for the convolutional layers, as it is more computationally efficient, simplistic, and shows better convergence performance than other activation functions (Brownlee, 2019). The 1D-CNN model contained a total of 8 layers and 272,051 total trainable parameters.

Method for ECG Classification Using 1-D CNN

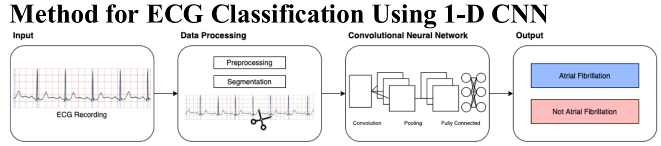

Figure 5. The proposed method for ECG classification. Input will be a segmented 30-second ECG recording, and the output of the 1D-CNN will be a classification of either AF or not AF (Source: Student Researchers).

Table 3. Overview of Proposed 1D-CNN Model
Types of layers and parameters used for the classifier (Source: Student Researchers).

Layer	Type	Kernel Size	Stride	Filters	Input Size
1	Conv1D	36	1	50	10800 x 1
2	Conv1D	36	1	50	10800 x 1
3	MaxPooling1D	36	1		
4	Conv1D	36	1	50	10800 x 1
5	Conv1D	36	1	50	10800 x 1
6	GlobalAveragePooling1D	36			
7	Dropout	36			
8	Fully-connected	36			

Training and Testing

The training of the CNN model was fully-supervised (learning on a labeled dataset rather than finding patterns independently) and model hyperparameters were optimized using the Adam optimizer with binary cross-entropy as the loss function. We manually altered the batch size, kernel size, and the number of filters to achieve the highest accuracy, specificity, recall, and precision. 5-fold cross-validation was completed to evaluate the performance of the CNN model.

Data Analysis

Accuracy, precision, sensitivity (or recall), specificity, and the F1 score were calculated to determine the performance of the CNN model. Where true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are labeled, accuracy refers to (TP + TN)/(TP + TN + FP + FN); Precision refers to TP/(TP + FP), recall refers to TP/(TP + FN); Specificity is TN/(TN + FP), as seen in Table 4. AUC curves were used to plot the TP rate against the FP rate in order to show the relationship between recall and specificity, as well as evaluate the performance of the CNN. These C-statistics, ROC curve, recall, and specificity are a standard for model evaluation with cardiovascular medicine (Krittanawong et al., 2019). The performance metrics were then compared to those of previous studies identifying AF from ECG recordings using a CNN.

Table 4. Confusion Matrix for ECG Classification
Confusion matrix for true positives, true negatives, false positives, and false negatives, and their corresponding meanings, for the metrics (Source: Student Researchers).

	Positive	Negative
True	AF classified as AF	normal classified as normal
False	normal classified as AF	AF classified as normal

PHASE II (SUMMER 2021) METHODOLOGY

Role of Student vs. Mentor

Over the course of six weeks during the summer of 2021 (over 90 hours), we were responsible for constructing the GANs and CNN, training/testing the models, and completing data analysis to evaluate its performance. Our mentor supervised and provided guidance on our project design, changes to processes, and programming.

Equipment & Materials

Data Collection

Jianwei Zheng et al. created a 12-lead electrocardiogram database (Zheng et al. database) covering 10,646 patients containing de-identified, annotated data sampled at 500 Hz. The data consists of 5,956 males and 4,690 females, and among the patients, 83% had at least one abnormality and 17% had sinus rhythm. A great variety of rhythms are represented in the data, including sinus bradycardia, atrial fibrillation, sinus tachycardia, atrial flutter, sinus irregularity, supraventricular tachycardia, and atrial tachycardia (as seen in Table 5). Due to noise contamination from power line interference, motion artifacts, electrode contact noise, muscle contraction, baseline wandering, and random noise, the developers of this database denoised the data. First, the Butterworth low pass filter was utilized to remove the signal with a frequency above 50 Hz. Next, LOESS (local regression) smoother was used in order to eliminate baseline wandering. Lastly, Non Local Means (NLM), typically used to smooth repeated structures in digital images, was implemented to erase remaining noise.

Compared to Phase I's project and other databases, this database was much more valuable. As shown in Table 6, the Zheng et al. database has the largest sample size and highest sampling frequency, which is important in detecting certain cardiac conditions (Zheng et al., 2020). The ECG records contained 12 leads, but only lead I was utilized as it has yielded good results and is most relevant—wearable devices that detect arrhythmias are becoming more common and effective at identifying arrhythmias (Raghunath et al., 2020).

Table 5. Rhythm Information
Rhythm information and baseline characteristics of participants in ECG dataset (Zheng et al., 2020).

Acronym Name	Full Name	Frequency, n(%)	Age, Mean ± SD	Male, n(%)
SB	Sinus Bradycardia	3,889 (36.53)	58.34 ± 13.95	2,481 (58.48%)
SR	Sinus Rhythm	1,826 (17.15)	54.35 ± 16.33	1,024 (56.08%)
AFIB	Atrial Fibrillation	1,780 (16.72)	73.36 ± 11.14	1,041 (58.48%)
ST	Sinus Tachycardia	1,568 (14.73)	54.57 ± 21.06	799 (50.96%)
AF	Atrial Flutter	445 (4.18)	71.07 ± 13.5	257 (57.75%)
SI	Sinus Irregularity	399 (3.75)	34.75 ± 23.03	223 (55.89%)
SVT	Supraventricular Tachycardia	587 (5.51)	55.62 ± 18.53	308 (52.47%)
AT	Atrial Tachycardia	121 (1.14)	65.72 ± 19.3	64 (52.89%)
AVNRT	Atrioventricular Node Reentrant Tachycardia	16 (0.15)	57.88 ± 17.34	12 (75%)
AVRT	Atrioventricular Reentrant Tachycardia	8 (0.07)	57.5 ± 16.84	5 (62.5%)
SAAWR	Sinus Atrium to Atrial Wandering Rhythm	7 (0.07)	51.14 ± 31.83	6 (85.71%)
All	All	10,646 (100)	51.19 ± 18.03	5,956 (55.95%)

Table 6. Comparison of ECG Databases
Compared to several other ECG databases, the Zheng et al. database contains the largest number of subjects with the greatest diversity in age (Zheng et al., 2020).

Name	Subjects	Records (length)	Sampling rate	Age	Male, n(%)	Lead, n
MIT-BIH	47	48 (30 min)	360 Hz	23–89	25 (52.08)	2
EDB	79	90 (120 min)	250 Hz	30–84	70 (88.61)	2
AHA	N/A	154 (180 min)	250 Hz	N/A	N/A	2
CU	35	35 (8 min)	250 Hz	N/A	N/A	2
NSD	2	12 (30 min)	360 Hz	51–69	1 (50)	2
St Petersburg DB	32	75 (30 min)	257 Hz	18–80	17 (53.13)	12
Proposed one	10646	10646 (10 second)	500 Hz	4–98	5956 (55.95)	12

GAN & CNN Implementation

The convolutional neural network was developed in Python using the Keras framework and Tensorflow backend. Python waveform-database, NumPy, and SciPy packages were also utilized for data analysis.

Procedures

Two experiments were run in order to determine the effects of a greater dataset to resolve the class imbalance issue: one CNN was run only with data from the Zheng et al. dataset (in order to create a baseline) and the other along with artificial data from the GAN (in order to compare).

CNN Model Architecture

The CNN model was designed to identify various arrhythmias from a short-term ECG. The proposed CNN model contained four 1D convolutional layers with a kernel size of 18, 125 filters, a stride of 1, an input size of 5,000 x 1, and a batch size of 24. The kernel size and batch size were chosen based on the hyperparameter tuning experiments. Then, the pooling layer chose representative features by decreasing the size of the feature maps with max-pooling. The 1D max-pooling layers contained a kernel size of 18 and had a pool size of 3. Two more convolutional layers were added, followed by a global average pooling layer with a kernel size of 18. A dropout layer was also included to prevent overfitting, with a rate of 0.5. Finally, the fully-connected layer contained a dense layer and softmax activation function for the concluding differentiation of the model (multi-class classification between eight rhythms: SR, AFIB, AF, SB, SI, SVT, ST, AT). The rectified linear unit (ReLU) activation function was used for the convolutional layers, as it is more computationally efficient, simplistic, and shows better convergence performance than other activation functions (Brownlee, 2019). The 1D-CNN model contained a total of 8 layers and 847,508 total trainable parameters (as seen in Table 7).

CNN Training and Testing

The training of the CNN model was fully-supervised (learning on a labeled dataset rather than finding patterns independently) and model hyperparameters were optimized using the Adam optimizer with binary cross-entropy as the loss function. We manually altered the batch size, kernel size, and the number of filters, and 5-fold cross-validation was completed to evaluate the performance of the CNN model.

Table 7. Overview of Proposed 1D-CNN

Layer	Type	Kernel Size	Stride	Filters	Input Size
1	Conv1D	18	1	25	5000 x 1
2	Conv1D	18	1	25	5000 x 1
3	MaxPooling1D	18	1	25	5000 x 1
4	Conv1D	18	1	25	5000 x 1
5	Conv1D	18	1	25	5000 x 1
6	GlobalAveragePooling1D	18			
7	Dropout	18			
8	Fully-connected	18			

Architecture of the ID-CNN model, including layer type and parameters (Source: Student Researchers).

CNN Evaluation & Data Analysis

Accuracy, precision, recall, specificity, and the F1 score were calculated to determine the performance of the CNN model.

TimeGAN Model

To correct the class imbalance, four TimeGANs were created for each class that required more data (atrial flutter, sinus irregularity, supraventricular tachycardia, and atrial tachycardia). The TimeGAN was composed of three losses. The reconstruction loss compares the reconstruction of the encoded data to the original one. The supervised loss determines the ability of the generator to approximate the next time step in the latent space. And finally, the unsupervised loss demonstrates the min-max game between the generator and discriminator networks. Given the architecture of the TimeGAN as well as the loss functions, there are three training phases. First, the autoencoder was trained on the original ECG data for reconstruction. Then, the supervisor was trained using the original data to learn the temporal behavior of the historical information. Lastly, the four components were trained altogether while minimizing the three aforementioned loss functions. Figure 6 demonstrates how the three loss functions work along with the four models (Yoon et al., 2019).

The Four Components of the TimeGAN Architecture

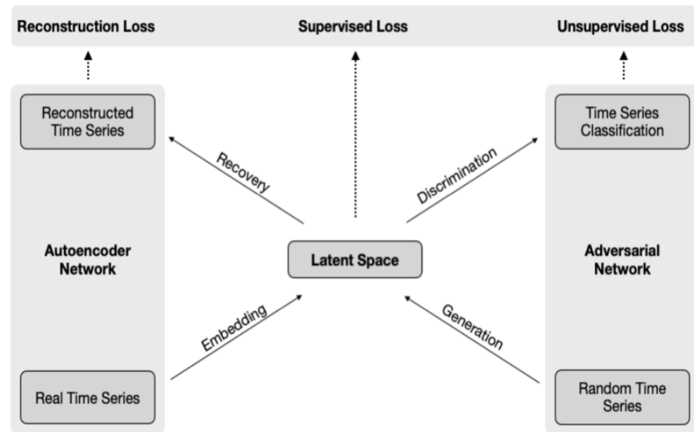

Figure 6. TimeGAN architecture utilizes an adversarial network and an autoencoder. The autoencoder consists of embedding and recovery networks and the adversarial network consists of a sequence generator and discriminator (Jansen, n.d.).

TimeGAN Evaluation

In order to evaluate the performance of the GAN, the generated data was examined with visualizations to provide a qualitative measure of diversity (whether the distribution of the generated data roughly matches that of the original data). To do this, a dimensionality reduction (principal components analysis (PCA) and t-SNE) was utilized to determine how similar the distribution of the generated data matches that of the real data.

Implementation of Generated ECG Data

Once the TimeGAN-generated ECG data was evaluated, it was combined with existing real data as the train set for the CNN, while the original data was utilized as the test set. The same architecture and parameters were applied with the CNN and the same performance metrics were calculated so that comparisons could be made.

PHASE I RESULTS

Data from the MIT-BIH Arrhythmia Database and MIT-BIH Atrial Fibrillation Database was used in Phase I in order to identify atrial fibrillation in ECGs. For the experiment, the data was divided into 80% for training and 20% for testing sets. Training and testing of the model were done with 20 epochs, dropout regularization at a rate of 0.5, and an early stopping function. Early stopping was utilized with a patience argument set to 2 to end training once the model performance stopped improving over 2 epochs and thus prevent overfitting. Several performance metrics were utilized in testing to measure the ability of the model. The accuracy and loss curves of the training model are shown in Figure 7, and with the early stopping function, training ended early at 14 epochs. The model demonstrated a very high average accuracy (98.89%) and low average loss (0.04), indicating that the model performed well and accurately after each iteration of optimization. Additionally, the model achieved a 98.89% average accuracy (as mentioned before), 99.34% average precision, 97.66% average recall, and 99.62% average specificity. The F1 score was calculated to be 98.50%, indicating an excellent balance between the precision and recall of the model. Area under the curve (AUC) measures the sensitivity and specificity tradeoff, showing the model's classification ability, and was found to be .9972, indicating an outstanding classification capability, as seen in Figure 7. Overall, the performance metrics indicate successful performance of the model.

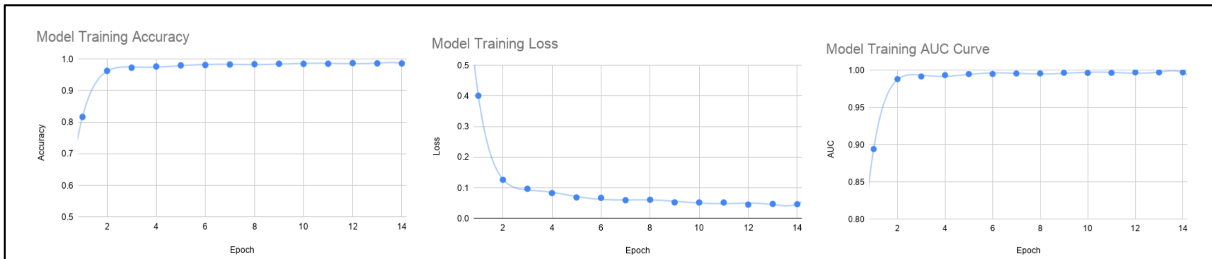

Figure 7. Accuracy, loss, and AUC of training model with 5-fold CV, using 20 epochs and patience of 2 (Source: Student Researchers).

PHASE I (SUMMER 2020) DISCUSSION

Support for Hypotheses

A model was successfully created to effectively automatically detect and identify atrial fibrillation rhythms from short-term electrocardiogram recordings, as indicated by the excellent performance metrics. The hypothesis that a CNN could be trained to identify atrial fibrillation present in short-term 30-second 12-lead ECGs from a dataset of patients with various arrhythmias and healthy sinus rhythms was supported by this study. Additionally, this study substantiated the hypothesis that ECGs from patients with various arrhythmias could be used to distinguish atrial fibrillation rhythm arrhythmia from normal sinus rhythm or other arrhythmias.

Potential Limitations

Although the results of the study indicate a success, several limitations exist. One potential oversight was that when hyperparameters were originally being altered, multiple parameters (such as number of filters, kernel size, and batch size) were changed at once, therefore complicating the determination of how each specific parameter was a factor of the model's performance, as seen in Table 8. Additionally, due to the extended duration of completing 5-fold cross validation training and testing, experiments were run with only one epoch potentially resulting in miscalculations in accuracies and losses for adjusted hyperparameters. Though the final 1D-CNN performed extremely well, alternative hyperparameter configurations may have provided for an even more accurate model. Additionally, much of the original hyperparameter tuning was conducted using the softmax activation function, but later it was found that a sigmoid function operated more effectively for binary classification. Perhaps, more research could be done to determine the efficacy of the softmax function, used for multi-class regression, over the sigmoid function, used for binary regression, which could be valuable in creating classifiers that identify multiple arrhythmias. Another limitation may have been sample size as numerous previous studies have utilized the MIT-BIH Arrhythmia Databases, but perhaps newer information with a larger more diverse dataset would train the model more effectively. Also, while the 1D-CNN identified only AF or not AF, a more useful neural network would also differentiate the 'not AF' category into either normal sinus rhythm or other arrhythmias (and could go further to identify those 'other arrhythmias').

Table 8. Hyperparameter Tuning Experiments
Several experiments were conducted adjusting parameters to determine optimal hyperparameters for the model before 5-fold CV (Source: Student Researchers).

Experiment	1	2	3	4	5	6	7
Filters	100	80	20	100	100	100	**100**
Batch Size	32	24	24	24	32	16	**24**
Kernel Size	18	18	360	9	36	36	**36**
Loss	0.5089	0.5129	0.6566	0.5158	0.4071	0.4852	**0.353**
Accuracy	0.7579	0.7574	0.6307	0.7511	0.8067	0.7796	**0.8452**

Comparison to Past Studies

As previously mentioned, several studies have utilized the MIT-BIH Arrhythmia database for various applications with different neural networks. The performance of our 1D-CNN model was compared to that of four previous studies, placing fairly well amongst the others. The performance comparison is not exact, due to the fact that all the CNNs are slightly different. While the studies by Zheng et al. (2020) and Huang et al. (2019) both produced CNNs, they were varied: a 2D-CNN LSTM and a STFT 2D-CNN, respectively. Also, though both identified different arrhythmias, they classified several beat types, among normal sinus rhythm (not atrial fibrillation), but the same MIT-BIH Database was used. On the contrary, research by Erdenebayar et al. (2019) and Hannun et al. (2017) both distinguished atrial fibrillation from ECGs using 1D-CNNs, and though rhythm classifications were made from 30s segments for both studies, Erdenebayar et al. only classified AF, while Hannun et al.'s CNN addressed 12 rhythm types. Additionally, Erdenebayar et al. used the MIT-BIH Database containing only normal sinus rhythm and atrial fibrillation, and may have achieved such successful performance due to the clear difference. Unlike Erdenebayar et al., not only did Hannun et al. classify 12 types of arrhythmia, but they also collected 64,121 ECG records from 29,163 patients for a diverse and extremely large dataset—possibly the reason for their model's slightly lower performance. Our 1D-CNN model distinguished AF from a combined MIT-BIH dataset containing 30s ECG samples of AF, several other arrhythmias, and normal sinus rhythm. As seen in Table 9, Zheng et al.'s 2D-CNN LSTM technology achieved highest accuracy (99.01%) and Erdenebayar et al.'s 1D-CNN achieved highest recall (98.60%). Our model achieved highest precision, specificity, F1 score, and lowest loss: 99.34%, 99.62%, 98.50%, and 0.0410, respectively, demonstrating its strong performance. Those successful metrics, however, may be accredited to the fact that our model distinguished between only two classes (AF and not AF), rather than 8 beat types, 12 different rhythms, or 5 beat types. In addition, some of the metrics from other studies were not

Table 9. Comparison of Results Across Several Studies
The performance metrics of this study, in comparison to those of four previous studies (Source: Student Researchers, Zheng et al., 2020, Erdenebayar et al., 2019, Hannun et al., 2017, Huang et al., 2019).

Model	Study	Rhythm Types	Accuracy	Precision	Recall	Specificity	F1 Score	Loss
2D-CNN LSTM	Zheng et al.	8	**0.9901**		0.9767	0.9957		
1D-CNN	Erdenebayar	2	0.9870		**0.9860**	0.9870		
1D-CNN	Hannun et al.	12		0.8090	0.8270		0.8090	
STFT 2D-CNN	Huang et al.	5	0.9900					0.0414
1D-CNN	**This Study**	2	0.9889	**0.9934**	0.9766	**0.9962**	**0.9850**	**0.0410**

available and may have been better than those of our study. Nonetheless, our 1D-CNN measured up fairly well in comparison to previous studies.

Implications and Applications

During this study, it was found that our 1D-CNN could effectively classify atrial fibrillation from short-term ECG recordings containing various arrhythmias and normal sinus rhythms. This finding indicates that deep learning methods could be applied to diagnose cardiovascular disease that manifest themselves in electrocardiograms. As current screening methods are challenging, due to issues with diagnosing momentary arrhythmia from a single electrocardiogram and inconvenient prolonged monitoring methods, this study could present a potential solution that isn't so intrusive, uncomfortable, or inefficient. As demonstrated by the findings in this study, neural networks have great potential to advance diagnostic methods for cardiovascular diseases, however, it was determined that additional procedures, such as generating synthetic data, could potentially improve diagnosis.

PHASE II RESULTS

Control CNN

This study utilized data from the Zheng et al. database in order to synthesize and detect eight different rhythm types (sinus rhythm, sinus bradycardia, atrial fibrillation, sinus tachycardia, atrial flutter, sinus irregularity, supraventricular tachycardia, and atrial tachycardia) in ECGs. For the control experiment, the data was divided into 80% for training and 20% for testing sets. Similar to the Phase I model, the CNN was trained and tested with 20 epochs, dropout regularization at a rate of 0.5, and an early stopping function set to 2. This was done as the early stopping function proved extremely effective as several metrics were tested in the previous phase.

To evaluate the ability of the model, several performance metrics were utilized, such as specificity, recall, and precision. As shown in Table 10, the model achieved an overall average accuracy of 80.14%, 98.11% average specificity, 74.99% average recall, and 85.00% average precision. The F1 score was also calculated to evaluate the balance between the precision and recall of the model, and was found to be .7969. The AUC, measuring the tradeoff between sensitivity and specificity, was found to be .9692, indicates its ability to discriminate between classes. Furthermore, the AUC and loss curves were found and are shown in Figure 8.

Table 10. Performance Metrics for 5-Fold CV of Control CNN
The metrics for 5-fold CV with Early Stopping at 20 epochs with a patience of 2 (Source: Student Researchers).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Loss	0.6228	0.6104	0.6103	0.6213	0.7135	0.6357
Accuracy	0.7937	0.8106	0.8121	0.8125	0.7780	0.8014
Precision	0.8561	0.8604	0.8630	0.8467	0.8240	0.8500
Recall	0.7395	0.7546	0.7654	0.7777	0.7125	0.7499
AUC	0.9703	0.9736	0.9704	0.9700	0.9617	0.9692
Specificity	0.9822	0.9825	0.9826	0.9799	0.9783	0.9811

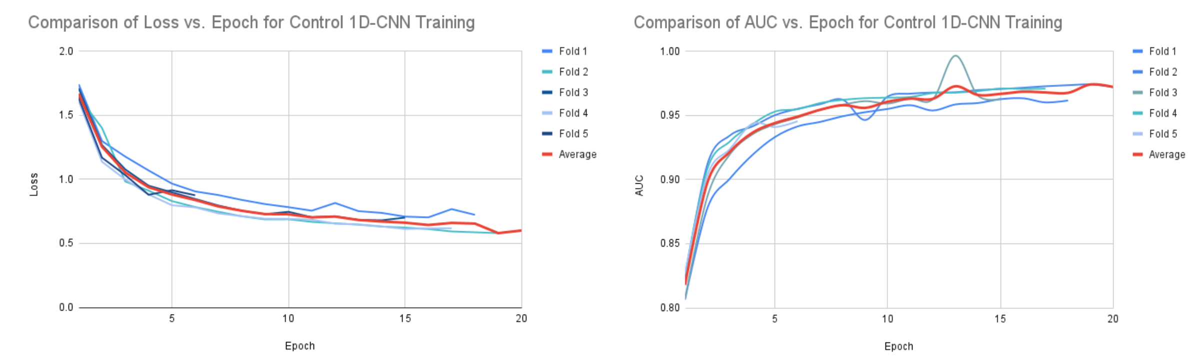

Figure 8. Loss and AUC curves for each fold of the 5-fold CV training, along with the average loss curve (Source: Student Researchers).

The model classified sinus bradycardia, atrial fibrillation, sinus tachycardia, supraventricular tachycardia, and sinus rhythm fairly well, with accuracies of 95.576%, 91.067%, 84.821%, 80.068%, and 73.494% respectively. However, it classified atrial tachycardia, atrial flutter, and sinus irregularity extremely poorly, with accuracies of 0.826%, 0.899%, and 5.514% respectively. The accuracy for each class of arrhythmias was found and a confusion matrix is shown in Table 11. As the class with the greatest amount of data

Confusion Matrix for Control CNN
Table 11. Confusion matrix of control CNN, where the CNN struggled most to classify AF, AT, and SI, as they had significantly less data (Source: Student Researchers).

Class	AF	AFIB	AT	SI	SB	SR	ST	SVT
AF	4	317	0	0	16	3	42	63
AFIB	3	1621	0	0	73	5	49	29
AT	2	64	1	0	6	16	20	12
SI	0	9	0	22	116	237	15	0
SB	0	56	0	4	3716	106	6	0
SR	0	28	0	3	340	1342	111	2
ST	2	120	0	1	10	77	1330	28
SVT	4	71	0	0	5	8	29	470

(sinus bradycardia) achieved the greatest accuracy (95.576%), and the class with the least amount of data (atrial tachycardia) achieved the lowest accuracy (0.826%), clearly a greater amount of data provided for more effective training and testing, and therefore, stronger performance. The CNN model's extremely inconsistent classification ability demonstrates the issue of class imbalance.

TimeGANs

Four TimeGANs were then created for each class that had significantly less data: atrial tachycardia, atrial flutter, sinus irregularity, and supraventricular tachycardia. The TimeGAN-generated data was initially assessed with visual inspection of random sample visualizations, which revealed a failure in the TimeGAN to synthesize accurate data. It was found that the TimeGAN data was not at all similar to the real ECG data. For example, the AF samples that were generated depicted straight-line signals, rather than the repetitive heartbeat rhythm standard of an ECG, as seen in Figure 9.

Comparison of Real and Generated AF Samples (at 500 Hz)

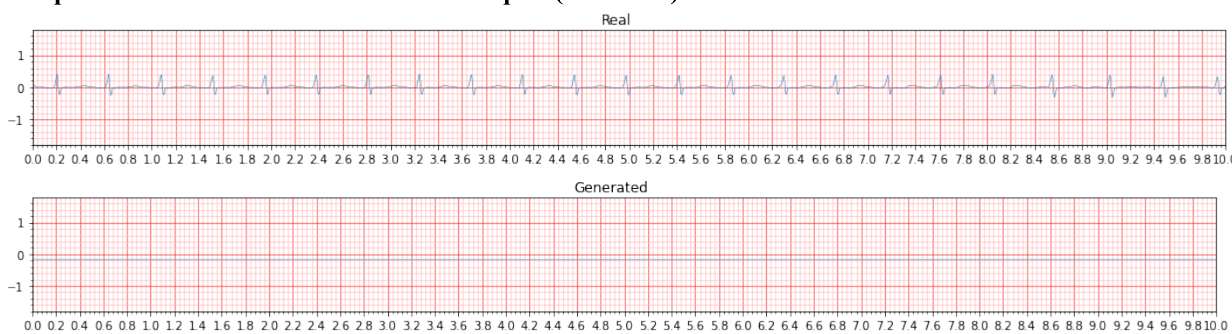

Figure 9. Random samples of real and TimeGAN-synthesized data were selected to evaluate the performance of the TimeGAN (Source: Student Researchers).

Due to the failure of the TimeGAN to produce accurate synthetic ECG data, we attempted to improve the TimeGAN with a few approaches. One of those was resampling our original ECG. Because the TimeGAN had previously only generated short sequence data, we resampled our data from 500 Hz to various lower frequencies (250 Hz, 100 Hz, 50 Hz), with the intention of shortening our sequence length (to 2500, 1000, and 500, respectively). This resampling attempt, however, yielded no progress in TimeGAN improvement (as shown in the visualizations in Figure 10).

Furthermore, we tried to use an alternate implementation of TimeGAN in Google Colab, but that also did not result in the synthesis of accurate TimeGAN data.

Comparison of Real and Generated AF Samples (100 Hz)

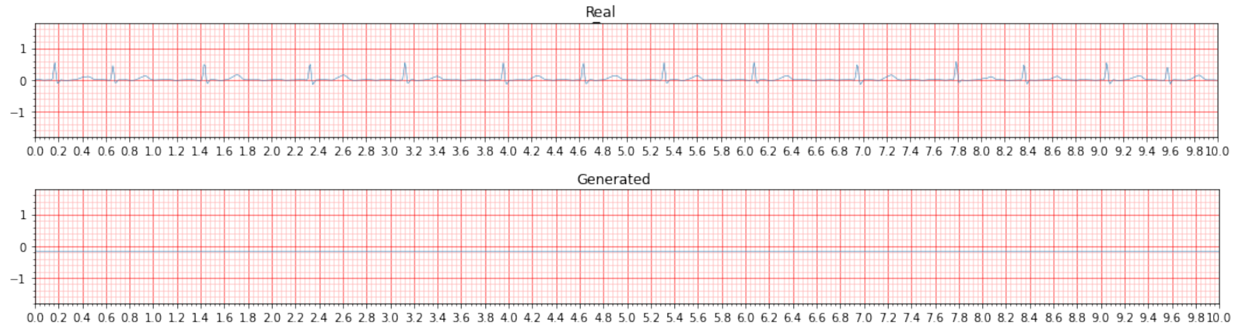

Figure 10. Random samples of real and TimeGAN-synthesized data were selected to evaluate the performance of the TimeGAN (Source: Student Researchers).

CNN with TimeGAN Data

Because we were unable to synthesize ECG data of authentic quality with the TimeGAN, we were unable to train and test the proposed CNN with the implemented data. Thus, we were unable to evaluate the performance of the CNN and compare it to previous studies.

PHASE II DISCUSSION

Support for Hypotheses

A model was created to effectively automatically detect and identify arrhythmia rhythms from short-term electrocardiogram recordings, however only a few of the eight classes were able to successfully classify arrhythmias with high performance metrics (sinus bradycardia, atrial fibrillation, sinus tachycardia, supraventricular tachycardia, and sinus rhythm). The hypothesis that a CNN could be trained to identify several types of arrhythmias present in short-term 10-second 12-lead ECGs was refuted by this study. Additionally, this study refutes the hypothesis that a TimeGAN could be implemented with a CNN to augment existing ECGs to enlarge the dataset and resolve class imbalance issues. As the TimeGAN failed to generate realistic data, we could not amplify the dataset to correct the class imbalance, and consequently, improve the performance of the CNN classifier.

Potential Limitations

While the results of the study indicated that TimeGAN was ineffective at synthesizing long sequence ECG data, several limitations exist. Because the TimeGAN has only been used to generate short sequence data, it couldn't perform with respect to long sequence data. An area of examination is the use of other types of GANs that may be more suitable for long sequence data. There has been success with countless generative neural networks like RC-GAN (Esteban, Hyland, Ratsch, 2017), C-RNN-GAN (Mogren, 2016), and WaveGAN (Donahue, McAuley, Puckette, 2019). Furthermore, perhaps CNN architecture can be adjusted with better parameters or more layers. Though the CNN underwent significant training and testing, the main focus was on the effect of data on the performance of the model, rather than on improving the model itself, and as such, more training and testing could further improve the accuracy of the CNN. Additionally, like the Phase I project, we could have examined other leads in order to improve the quality of the classifications for the CNN.

Implications and Applications

This study found that TimeGANs cannot effectively generate long sequence data for CNNs to classify several arrhythmias from short-term ECG recordings. Although this finding indicates that TimeGANs do not provide a solution to the class imbalance issue, perhaps adjustments can made that can allow it to more effectively generate ECG data for use as training/testing input. As noted in our Phase I study, the lack of a comprehensive, diverse ECG database is an issue that persists and continues to hinder progress in deep learning detection of cardiovascular disease. However, if TimeGANs can provide reliable, convincing synthetic data, they can work to improve the performance of arrhythmia classifiers, and thus, improve methods to detect cardiovascular disease. Despite the ineffectiveness of TimeGAN in this study, GANs still hold potential to synthesize data to assist in medical diagnosis.

FUTURE RESEARCH

There are several other ways deep learning could be developed to advance diagnosis and monitoring of heart issues. For instance, the difference between utilizing certain types of leads could be examined. Although we chose lead 1 based on current technology and its relevance due to its use in smart watches, other leads may be more effective as they can provide more information and thus provide a more accurate diagnosis. Similarly, utilizing multiple ECG leads and developing a non-1-dimensional CNN may be more effective. Raghunath et al. determined the mean AUC for several types of single leads and found that specific leads led to higher AUCs (Figure 11). Moreover, modifications to the TimeGAN could be made to improve the synthesis of long sequence time-series ECG data. Rather than using a recurrent neural network framework (as utilized in Yoon et al.'s TimeGAN), a temporal convolutional network could be implemented as it can shorten the training time and improve the quality of the synthesized data (Viklund, 2021; Lee, 2019). Alternatively, in the future, other types of GANs could be evaluated to determine which can most effectively generate data. Other types of neural networks could also be tested to determine the most effective classifier. In addition, the prediction of arrhythmias with the use of GAN-generated data may potentially reveal more findings that could advance diagnosis. Furthermore, using different length ECG samples, perhaps 30 seconds or 5 seconds rather than 10 may result in different, more successful outcomes, as they can provide a longer window to classify arrhythmias in. And lastly, using ECGs sampled at different frequencies, 360 Hz instead of 500 Hz may provide more insight into how to utilize data most effectively for classification, as well as develop the most successful neural networks.

Mean AUC for Several ECG Leads

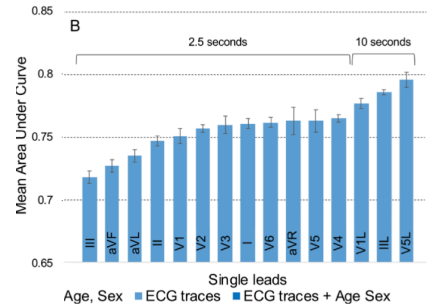

Figure 10. The average area under the curve for various different length ECG leads (placed at different areas of the body) (Raghunrath et al., 2019).

CONCLUSION

Cardiovascular diseases, such as arrhythmias are extremely prevalent, and will continue to become more common in the future. This study aimed to create an effective convolutional network with a TimeGAN that automatically classifies arrhythmias from short-term ECG recordings. ECG data was taken from the Zheng et al database, and split into training and testing sets conducted with 5-fold cross-validation. A control model was first created and found to have an overall average accuracy of 80.14%, 98.11% average specificity, 74.99% average recall, 85.00% average precision, F1 score of .7969, and AUC of .9692. Due to the inability of the TimeGAN to generate realistic ECG data, the class imbalance could not be resolved and the CNN could not be improved. Despite this, the findings of this study, along with previous research on the applications of deep learning, could make diagnostic procedures of arrhythmias via ECG more accurate, allowing us to detect disease more effectively and efficiently.

References

Brophy, E., Wang, Z., Lab, B., Qi, C., Bytedance, S., Lab, A., & Ward, C. (2021). Generative Adversarial Networks in Time Series: A Survey and Taxonomy. https://arxiv.org/pdf/2107.11098.pdf

Brophy, E., Wang, Z., & Ward, T.E. (2019). Quick and easy time series generation with established image-based GANs. *ArXiv, abs/1902.05624*.

Brownlee, J. (2019, April 20). A gentle introduction to the rectified linear unit (ReLU) for deep learning neural networks. Machine Learning Mastery. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

CDC. (2020, September 8). Atrial fibrillation. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/atrial_fibrillation.htm

Donahue, C., McAuley, J., & Puckette, M. (2019). Adversarial audio synthesis. https://arxiv.org/abs/1802.04208

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. *Undefined*. https://api.semanticscholar.org/CorpusID:7467506

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *CoRR*, (),. https://doi.org/

Hyland, S., Zurich, E., Switzerland, Esteban, C., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. https://arxiv.org/pdf/1706.02633.pdf

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H. W., Halperin, J. L., & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: a practical primer. *European Heart Journal*, *40*(25), 2058–2073. https://doi.org/10.1093/eurheartj/ehz056

Lee, C. (2019, March 14). [Tensorflow] implementing temporal convolutional networks. Veritable. https://medium.com/the-artificial-impostor/notes-understanding-tensorflow-part-3-7f6633fcc7c7

Mayo Clinic. (2017). Heart arrhythmia - symptoms and causes. Mayo Clinic; https://www.mayoclinic.org/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668

Mogren, O. (2019). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. https://arxiv.org/pdf/1611.09904.pdf

Raghunath, S., Ulloa Cerna, A.E., Jing, L., van Maanen, D. P., Hartzel, D. N., Good, C. W., Patel, A. A., Delisle, B. P., Haggerty, C. M., & Fornwalt, B. K. (2019). Deep neural networks can predict one-year mortality and incident atrial fibrillation from raw 12-lead electrocardiogram voltage data. *Journal of Electrocardiology*, *57*, S104–S105. https://doi.org/10.1016/j.jelectrocard.2019.08.033

Rehorn, M., Sacks, N. C., Emden, M. R., Healey, B., Preib, M. T., Cyr, P. L., & Pokorney, S. D. (2021). Prevalence and incidence of patients with paroxysmal supraventricular tachycardia in the United States. *Journal of Cardiovascular Electrophysiology*. https://doi.org/10.1111/jce.15109

Ritchie, H., & Roser, M. (2018). Causes of death. Our World in Data. https://ourworldindata.org/causes-of-death

Viklund, J. (2021). Synthesis of sequential data. https://www.diva-portal.org/smash/get/diva2:1558794/FULLTEXT01.pdf

Wang, P., Hou, B., Shao, S., & Yan, R. (2019). ECG arrhythmias detection using auxiliary classifier generative adversarial network and residual network. *IEEE Access*, *7*, 100910–100922. https://doi.org/10.1109/access.2019.2930882

World Health Organization: WHO. (2019, June 11). Cardiovascular diseases. Who.int; World Health Organization: WHO. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

Zheng, Z., Chen, Z., Hu, F., Zhu, J., Tang, Q., & Liang, Y. (2020). An automatic diagnosis of arrhythmias using a combination of CNN and LSTM technology. *Electronics*, *9*(1), 121. https://doi.org/10.3390/electronics9010121

Zhou, Z., Zhai, X., & Tin, C. (2021). Fully automatic electrocardiogram classification system based on generative adversarial network with auxiliary classifier. *Expert Systems with Applications*, *174*, 114809. https://doi.org/10.1016/j.eswa.2021.114809