Employing Adversarial Machine Learning and Computer Audition for Smartphone-Based Real-Time Arrhythmia Classification in Heart Sounds

Aditya Kendre
Cumberland Valley High School, Mechanicsburg, PA
ENBM035

April 22, 2021

Abstract

We propose a novel approach to detect arrhythmias in Phonocardiograms (PCGs). Typically, many arrhythmia conditions are unknown until a patient is suggested an ECG/EKG test. This method, despite being accurate, limits the use case to hospitals and clinics with specialized equipment; thus, limiting the portability of diagnosing. Implementation of Adversarial Machine Learning (ML) and Computer Audition (CA) in combination with heart sounds provide ease of access to everyone who has a device capable of recording audio. Ideally, allowing medical professionals to treat arrhythmias in the developmental stages. The new design is comprised of two subsystems: one is based on the relationship between Electrocardiograms (ECGs) and PCGs, and the other between PCGs and arrhythmias. The first subsystem uses a Generative Adversarial Networks (GAN), in which both generated and real PCG signals are fed into the discriminator for classification. In subsystem two, ECG spectrograms are dimensionally reduced, then constructed into PCG spectrograms using a transGAN. These constructed PCG spectrograms, when converted back into time series, should be identical to the ground truth. This novel approach allows for an increase in the number of cardiovascular pathologies classified in heart sounds. After testing, the GAN model (subsystem one) achieved an accuracy of 94.98%, a specificity of 90.30%, and a sensitivity of 99.52% on the testing set. Furthermore, the transGAN showed extremely promising results, in that the transGAN discriminator was able to construct the PCG spectrogram accurately. The proposed method's ease of use allows for simple integration in mobile devices, such as smartphones, making it feasible in medical and consumer applications.

***K*eywords** Arrhythmias · Phonocardiograms · Electrocardiograms · Biomarkers

Problem Current detection methods have limited performance in pathologies and lack real-time classification capabilities.

Motivation To create a fast and accurate model capable of detecting a variety of arrhythmias in heart sound recordings (PCGs) without the need for specialized equipment.

Question Is it possible to use Generative Adversarial Networks (GANs) to accurately detect arrhythmias in PCGs and surpass previous methods in detection tasks?

Hypothesis If a novel heart sound analysis system is developed to detect a variety of arrhythmias, then arrhythmias will have a decreased undiagnosed rate.

Solution The new semi-supervised approach is composed of two subsystems; the first subsystem uses a discriminator, in which PCG signals are classified into categories based on arrhythmias in the signal. The second subsystem – a generator, aims to generate data such that, when fed into the discriminator, the discriminator will classify the generated data as abnormal or normal. While the discriminator aims to simultaneously classify the generated data as fake and classify the real PCG signals to their respective categories. This adversarial approach allows the discriminator to not only extract PCG signal-specific features but also introduces the discriminator potential noisy signals that should not be classified.

Engineering Goals

1. **Increase the Number of Cardiovascular Pathologies** — Develop a model to construct heart sounds from pre-existing data.
2. **Develop a System for End-to-End Heart Sound Arrhythmia Detection** — Create a system that is able to record and analyze heart sounds for Cardiovascular modalities without specialized equipment.
3. **Real-World Testing** — Test the end-to-end system in a real-world environment to ensure practicality and generality of the system.

Constraints Constraints include the use of a small dataset and noisy recordings of heart sounds auscultated5 at different locations..

Contents

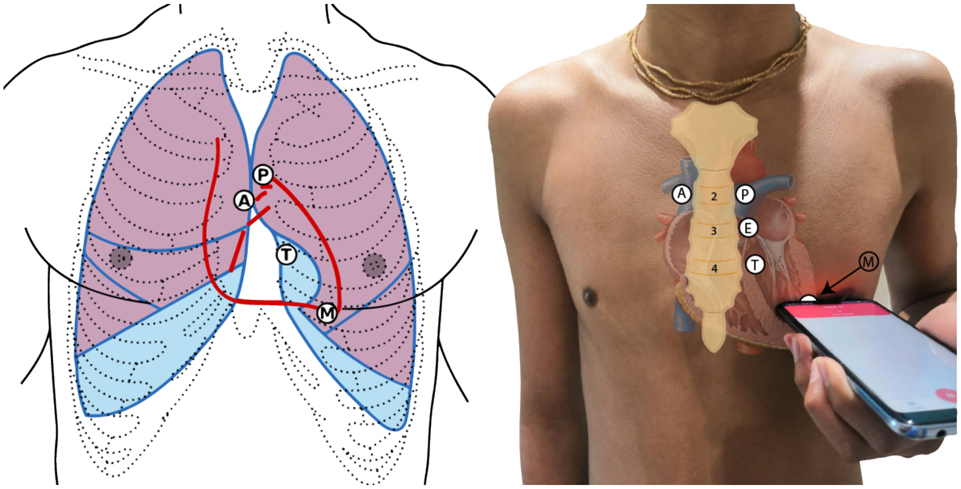


Figure 1: Representation of the source of heart sounds (left) and location of standard cardiac landmarks used for auscultation (right) with the following locations: **A**ortic area, **P**ulmonic area, **T**ricuspid area, **E**rb's point, **M**itral Area. Taken by Kendre, 2021

1 Introduction

An estimated three million cases of arrhythmia occur in the United States yearly (Mayo Clinic), with 300,000 sudden deaths per year – an incidence rather higher than stroke, lung cancer, or breast cancer (American Heart Association). Traditionally, non-invasive arrhythmia analysis is based on multiple electrodes that reflect the electrical activity on ECGs. This method, despite being accurate, limits the use case to hospitals and clinics with specialized equipment; thus, limiting the portability of diagnosing, let alone classification of the type of pathology.

Phonocardiograms (PCGs) are sounds that are created by the mechanical movement of the heart. This physical movement produces four distinct sounds: S1, S2, S3, S4, and murmurs. S1 and S2 are sounds created by a healthy heart; whereas, S3 and S4 (gallops) and murmurs refer to diseases or abnormalities. The first heart sound, S1, marks the start of Systole. Systole occurs when the heart muscle contracts and pumps blood from the chambers into the arteries. The second heart sound, S2, marks the end of Systole and the start of Diastole. Diastole is a phase of the heartbeat when the heart muscle relaxes and allows the chambers to fill with blood. Ideally, these heart sounds are recorded with digital stethoscopes. These tools use transducer technology to convert sound into an electrical signal. Over the past decade, this technology has grown immensely (by cause of speech recognition). Modern phones have the potential to record the sounds at a high resolution, given the microphone is located at the correct position relative to the heart. Such a device will prove to extremely beneficial in providing diagnosis without the need for specialized equipment.

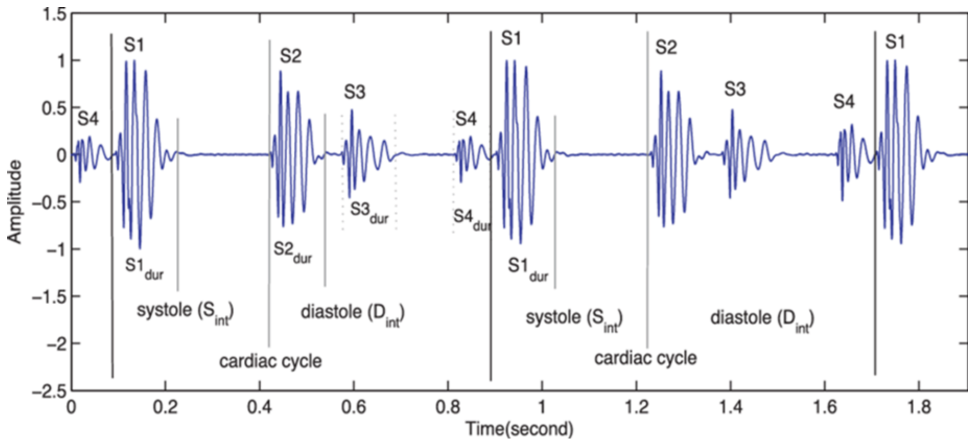


Figure 2: Illustrates the S1, S2, S3, and S4 biomarkers of heart sounds. Created by Ramachandran et al., 2014

Although heart sound databases do exist, these datasets are still limited by the number of pathologies that are collected, often having to divide the dataset into two categories: normal and abnormal. Currently, only three major supervised PCG datasets exist PhysioNet Classification of Heart Sound Recording Challenge dataset, PASCAL Heart Sound Challenge dataset, and the Heart Sound and Murmur Library. The presently available PCG datasets have a limited number of samples and do not cover the complete range of pathologies that are likely to be encountered in clinical settings.

In diagnosing heart sounds, two major challenges arise localization and classification. Localization aims to find the position of the aforementioned biomarkers in heart sounds. By doing this, heart sounds can be segmented into signals containing a single heart sound. Furthermore, classification attempts to categorize heart sounds into normal and abnormal groups by exploiting the information extracted from localization. Conventional heart sound localization and classification methods involve time, frequency, or both, and are typically dependent on machine learning algorithms to enhance the results. These algorithms typically include artificial neural networks (ANNs), support vector machines (SVMs), self-organizing maps (SOMs), and are limited to the number of samples and pathologies covered in a given dataset. This leads to a surface-level analysis of the heart sounds. The main challenge of effective heart sound detection stems from an analysis of noisy heartbeats and a lack of variety.

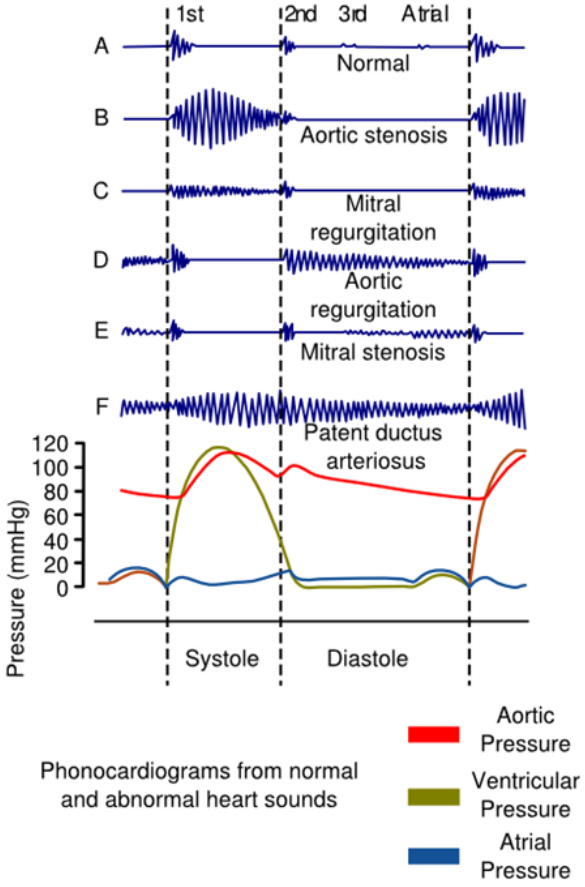


Figure 3: Representation of different abnormalities in sound and pressure. Created by Cruz-Correia et al., 2016

2 Related Work

For clean datasets, e.g., the PhysioNet Challenge dataset, a variety of time and frequency of methods converged on localization accuracy of 96.9% [1] and 86.02% classification accuracy [2]. [1] used bidirectional LSTMs with Attention for segmentation, while [1] employed AdaBoost and CNNs for classification. From the viewpoint of practical applications, the development of computationally efficient solutions is extremely important to the success of a model's

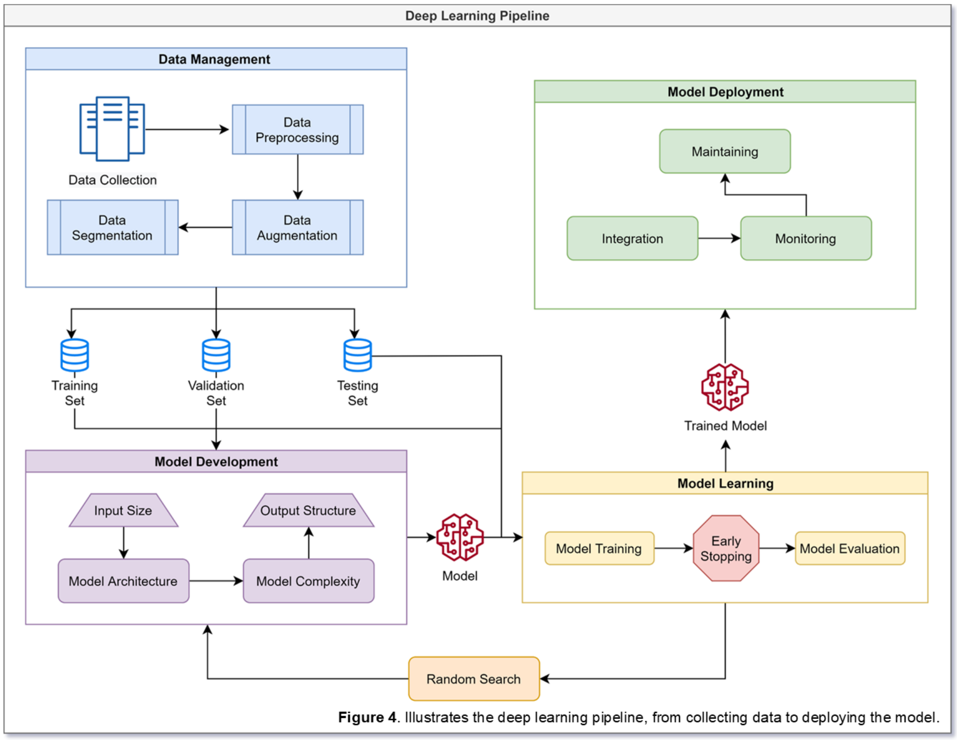


Figure 4: Illustrates the deep learning pipeline, from collecting data to deploying the model. Created by Kendre, 2021

deployment. Many studies have negated to comment on the practicality of their proposed methods. From our research, we have concluded only two studies have noted their time efficiency, [1] and [3]. However, both studies only addressed localization and not classification. Hence, the fastest localization model processed 1000 heart state classifications in 56.88 seconds [1], suggesting the model can process 1080 bpm. Keeping in mind [1] only localizes heart sounds, current models need severe optimization to achieve near to real-time analysis in the classification domain.

As mentioned earlier, PCG datasets are limited in cardiac pathologies, given this, no advancements have been made in increasing the number of arrhythmias in heart sound datasets.

Commercial products do exist for heart sound heart analysis; however, these products use proprietary hardware for such analysis. Steth IO attempts to convert a smartphone device into a digital stethoscope but requires an external case costing upwards of $300 and a client/patient fee. Even with these external accessories, the app can only detect heart murmurs. Similarly, the Eko by Littmann 3M digital stethoscope ($350) only detects heart murmurs with an 87% sensitivity and 87% specificity and has a monthly fee of $50.

Thus, the problem of computationally efficient and accurate classification of noisy heartbeats, especially with datasets with a variety of pathologies still remains a problem.

3 Data Management

3.1 Data Collection

Although PCG signals are analyzed less often than ECG signals, these signals are rather analyzed in real-time by physicians and healthcare workers. Preliminary studies done on PCG segmentation and classification primarily used private datasets. Hence, there existed no publicly available datasets until recently. Since then, many public datasets have been developed aiding researchers in their studies and creating open benchmarks for researchers to use in comparing similar findings. However, these datasets are still limited by the number of classes that are collected, when compared to ECG datasets.

Currently, only three major supervised PCG datasets exist: PhysioNet Classification of Heart Sound Recording Challenge dataset, PASCAL Heart Sound Challenge dataset, and the Heart Sound and Murmur Library. These datasets

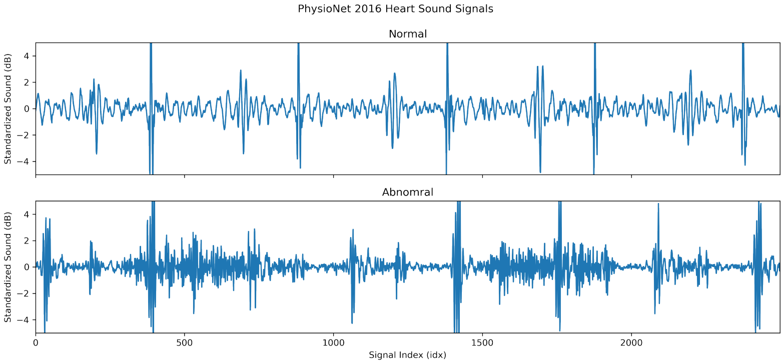


Figure 5: Plots the classes of the PhyioNet 2016 heart sound dataset (Normal and Abnormal). Created by Kendre, 2021

are all anonymized and de-identified for the safety of their subjects, and thus includes no personal information such as name, income, age, etc.

The PhysioNet Classification of Heart Sound Recording Challenge dataset was produced as a part of the 2016 PhyisoNet Computing in Cardiology Challenge. The heart sounds were collected from both clinical and non-clinical environments (in-home visits). The challenge focused on creating an accurate dataset of normal and abnormal heart sound recordings, especially in real-world (extremely noisy and low signal quality) scenarios. These recordings were sourced from nine independent databases and in total, contain 4,593 heart sound recordings from 1072 subjects, lasting from 5-120 seconds. Of which, 409 recordings that were collected from 121 patients contain one PCG lead and one simultaneously recorded ECG. Though, all recordings were resampled to 2,000 Hz using an anti-alias filter. Furthermore, the dataset is comprised of 3 classes: normal, abnormal, and unsure (this is due to poor recording quality), and have the following proportion respectively: 77.1%, 12,0%, 10.9%.

The PASCAL Classifying Heart Sounds Challenge dataset was released to the general public in 2011. The challenge consisted of two sub-challenges: heart sound segmentation, and heart sounds classification. These sub-challenges corresponded with dataset A, and dataset B respectively. Both datasets have recordings of varying lengths, between 1 second and 30 seconds. Dataset A was collected via the iStethoscope Pro iPhone app and contained 176 heart sound recordings. 124 of which are divided into four classes: Normal (31 recordings), Murmur (34 recordings), Extra heart sound (19 recordings), and Artifact (40 recordings); the rest of the records are unlabeled for testing purposes. Dataset B was collected using a DigiScope (a digital stethoscope), and included 656 heart sounds. All except 370 were separated into three classes: Normal (320 recordings), Murmur (95 recordings), and Extra-systole (46 recordings). Both datasets A and B vary in sound recordings between lengths of 1 second and 30 seconds.

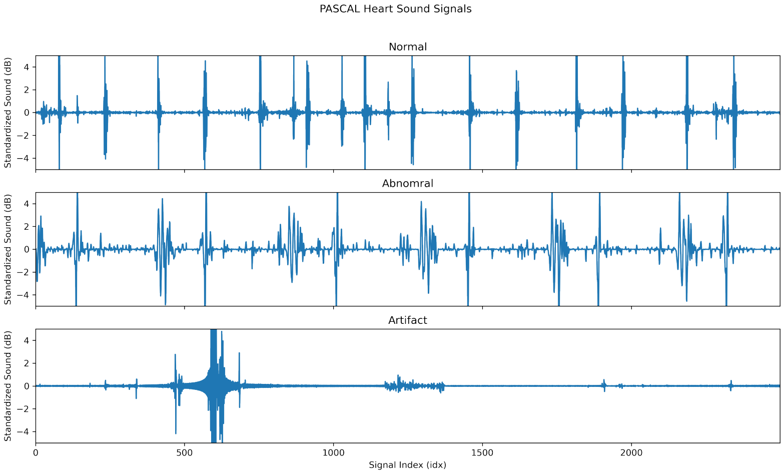


Figure 6: Plots the classes present in the PASCAL dataset (Normal, Abnormal, and Artifact). Created by Kendre, 2021

More than 300 million ECG recordings are analyzed yearly, and thus create an exceptional tool for arrhythmia classification. Coupled with the recent surge in research interest in 2015, many massive publicly available datasets have been published, notable by PhysioNet - the moniker of the Research Resource for Complex Physiologic Signals. Numerous, datasets ECG exist, however, many are limited to few classes (Normal and Abnormal). At present, three public datasets exist that have more than 4 classes: AF Classification Challenge 2017, PTB Diagnostic ECG, and PTB-XL dataset. Additionally, iRhythm Technologies have developed a semi-public dataset, that is available upon request, that contains 12 classes.

The PTB-XL is the largest publicly available dataset for ECGs and contains 21,837 clinical 12-lead ECG recordings from 18,885 patients of 10-second length. These recordings are separated into 5 super-classes: Normal, Myocardial Infraction, Hypertrophy, ST/T-Change, and Conduction Disturbance. These super-classes are further split into 71 sub-classes that range from AV Block to Posterior Myocardial Infraction. The raw signal data were downsampled to 100 Hz and annotated by up to two cardiologists, who assigned potentially multiple ECG statements to each record. iRhythm Technologies developed a large, 12 classes ECG dataset using raw single-lead ECG inputs. The 12 classes include Atrial fibrillation and flutter, AVB, Bigeminy, EAR, IVR, Junctional rhythm, Noise, Sinus rhythm, SVT, Trigeminy, Ventricular tachycardia, and Wenckebach. The dataset consists of 91,232 ECG recordings from 53,549 patients. This training dataset is available upon request under license from iRhythm Technologies, Inc. The publicly available test dataset contains 328 records collected from 328 unique patients, split between 6 classes. Both datasets were recorded using a Zio monitor, which monitors the heart through a single-lead sensor at 200 Hz. The annotation was done by a consensus committee of expert cardiologists.

The PhysioNet AF Classification database, presented in 2017 for the Computing in Cardiology Challenge, contains 8,528 ECG recordings, divided into 4 classes: Normal (5154 recordings), Atrial Fibrillation (771 recordings), Other arrhythmias (2557 recordings), and Noisy (46 recordings). The single-lead recordings last from 9 - 61 seconds, with a mean of 32.5 seconds and a standard deviation of 10.9 seconds. The ECG recordings were sampled to 300 Hz and provided in MATLAB V4 WFDB-compliant format.

3.2 Data Preprocessing

PCG recordings often are recording in non-ideal environments that are filled with unwanted background noise and interference. Data preprocessing is the process of altering the data in the signal, often by denoising, normalizing, standardizing, and transforming the signal. These steps are crucial for automatic localization and classification tasks. Preprocessing the data allows a model to extract meaning features efficiently and reveals the physiological structure of the heart sounds [Latif et al.]. Furthermore, preprocessing helps ensure that the data that is fed into the model is always in the same domain. This allows the model to generalize more easily.

We first resample the data to 500 Hz to decrease the spatial resolution of the heart sound recordings but still retain important features. Thus, helping the model to converge faster. The resampled data is standardized using the standard score equation. This scales the mean of the distribution to 0, artificially scaling all data into similar ranges, thus, helping combat the exploding gradient problem. The standardized data is then fed into a high/low band-pass filter that denoises data.

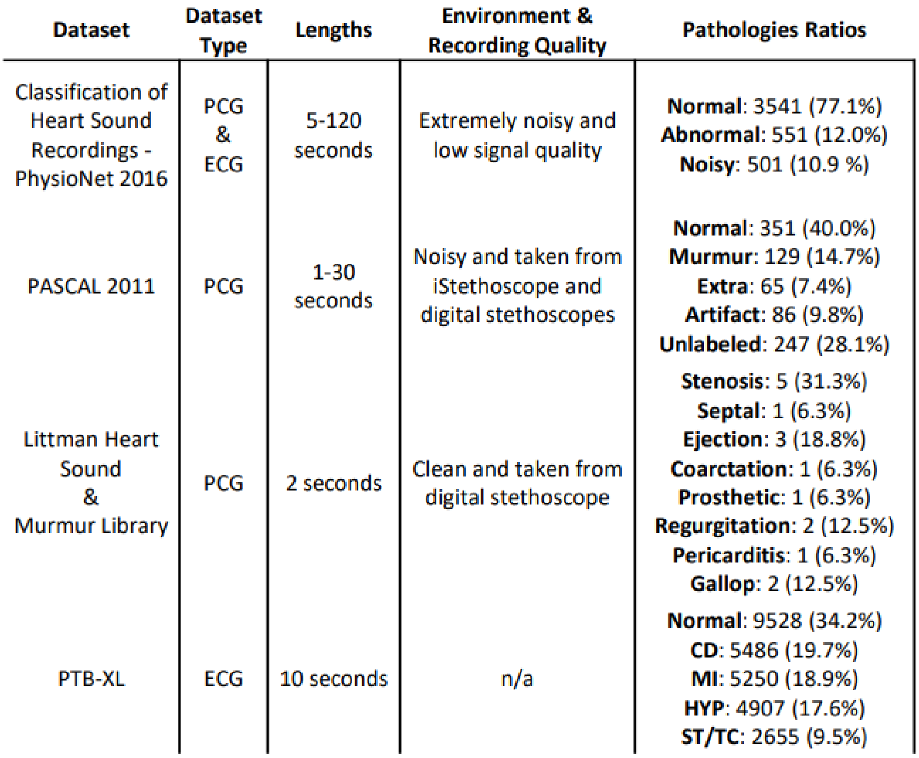
3.3 Data Segmentation

Data segmentation refers to the process of creating cross-validation datasets. This process assists in validating if the model is overfitting to the dataset. These datasets include the training set, validation set, and testing set. Typically, the training set is 70%-80% of the dataset, the reset of the dataset is split among the validation set and testing set. Here, we split the data 80% training, 10% validation, and 10% testing.

3.4 Data Augmentation

Data augmentation is a strategy that enables a significant increase in the diversity of data available while training a model, without actually collecting new data. Data augmentation techniques aim to slightly alter existing data to a point where the model cannot recognize the augmented data as one it has trained on before, but still retains the characteristics of the data's category. This helps in reinforcing important features within the data and is only done during the training portion of the workflow.

A common misconception arises when comparing preprocessing and augmentation. To be clear, preprocessing aims to clean the data of unwanted artifacts that are not meant for classification and is done in place. Augmentation, on the other hand, is solely done for expanding the dataset's size, often to combat overfitting. Augmenting the data before preprocessing further obscures the data unrealistically and beyond classification.



Dataset	Dataset Type	Lengths	Environment & Recording Quality	Pathologies Ratios
Classification of Heart Sound Recordings - PhysioNet 2016	PCG & ECG	5-120 seconds	Extremely noisy and low signal quality	**Normal**: 3541 (77.1%) **Abnormal**: 551 (12.0%) **Noisy**: 501 (10.9 %)
PASCAL 2011	PCG	1-30 seconds	Noisy and taken from iStethoscope and digital stethoscopes	**Normal**: 351 (40.0%) **Murmur**: 129 (14.7%) **Extra**: 65 (7.4%) **Artifact**: 86 (9.8%) **Unlabeled**: 247 (28.1%)
Littman Heart Sound & Murmur Library	PCG	2 seconds	Clean and taken from digital stethoscope	**Stenosis**: 5 (31.3%) **Septal**: 1 (6.3%) **Ejection**: 3 (18.8%) **Coarctation**: 1 (6.3%) **Prosthetic**: 1 (6.3%) **Regurgitation**: 2 (12.5%) **Pericarditis**: 1 (6.3%) **Gallop**: 2 (12.5%)
PTB-XL	ECG	10 seconds	n/a	**Normal**: 9528 (34.2%) **CD**: 5486 (19.7%) **MI**: 5250 (18.9%) **HYP**: 4907 (17.6%) **ST/TC**: 2655 (9.5%)

Figure 7: Dataset table that shows the dataset type (PCG or ECG), the lengths of the recordings (in seconds), the environment in which the signals were recorded, the recording quality, and the number of categories in the dataset (Table taken by Kendre, 2021).

We resample the heart sound recordings to different frequencies to simulate slower and faster beats per minute (bpm). The normal bpm for a human is between 60-100 bpm. Thus, measuring the sample distance between the first S1 (the start of systole) and the second S1, we calculate the bpm and resample accordingly.

Furthermore, we use noise injection directly to preprocessed PCG recordings [Messner et al.]. This process is identical to the process of synthetically adding noise to PCG recordings described in the preprocessing step. A variety of noises, like white noise, is added to the signal to increase the sample of recordings per class. This method is extremely beneficial for training on small datasets, like the PASCAL dataset.

4 Model Development

4.1 Model Architecture

Here we propose using Generative Adversarial Networks (GANs) for increased success in PCG heart sound detection. GANs pose a unique advantage over traditional machine learning and deep learning methods, in that a model learns to mimic a dataset by creating its own data, and tries to fool a discriminator into thinking the generated data is real. In a supervised approach, a GAN consists of two parts, a generator, and a discriminator. The generator is responsible for creating fake heart sound data, while the discriminator tries to predict where the incoming data is fake or real. In a semi-supervised approach, however, the data from a real dataset and the generator are fed into the model. Here, the discriminator tries to classify the generator's fake data, as well as predict the classes from the real dataset.

4.2 Model Complexity

Traditionally, generators are dense layers that slowly increase the dimensionality of the generated data to match that of the real dataset. Discriminators, on the other hand, are commonly CNNs because the majority of their applications work with images. However, it is possible to use a wide variety of architectures; such as LSTMs, RNNs, SVMs, DNNs, ANNs, and Transformers. As mentioned previously, there are many types of model architecture, some are used for classification, and others for feature extraction. Optimizing the combination of feature extraction layers and classification layers is extremely time-consuming and computationally taxing. This is because there exist many combinations of hyperparameters, thus making it difficult to optimize each parameter. To optimize hyperparameters, we used hyperparameter sweeps to make the optimization process more efficient. This method involves using one of three methods: grid search, random search, and Bayesian search. Grid search computes each possible combination of all hyperparameters and tests them all. Although this is very effective, it can be computationally costly. Random search selects a new combination at random, provided a distribution of values. This method is surprisingly effective and scales very well. Bayesian search creates a probabilistic model of metrics and suggests parameters that have a high probability of improving metrics. This works well for small-scale projects but scales poorly as the complexity of parameter relationships increases. Here, we used a random search to optimize our hyperparameters.

5 Model Learning

5.1 Model Training

During the training phase, the model is trained using backpropagation in conjunction with a cost function. Backpropagation attempts to calculate the gradient of the cost function with respect to the weight and biases of the model. This process involves an optimizer, which optimizes the model's parameters and a cost function that measure the correctness or incorrectness of the model. The goal of the optimizer is to minimize the cost function's error by adjusting the parameters to the given label. In this study, we used the Adam optimizer in union with Cross-Entropy Loss. The Adam optimizer uses a hyperparameter that dictated the change in the model's parameters on each backpropagation step, this is called the learning rate. Here we choose a learning rate of 0.0001.

The model is only trained on the training set; thus, backpropagation only occurs on the training set. Additionally, for each step in the training set, the optimizer backpropagates and optimizes the parameters and calculates metrics to further evaluate the model. The amount of steps in the training set is dictated by the batch size, the number of signals the model is trained on, in a single forward pass. Here we use a batch size of 32, meaning that the model is fed 32 signals per input. This significantly speeds up the process of training as more signals are passed through the model every time the model is optimized. A full pass of the training set is called an Epoch, here we train the model on 100 Epochs.

5.2 Model Training

To ensure the model is not overfitting, but generalizing to the training set, we use a validation set to track the metrics of the model. In theory, the metrics on the training set equal to that of the validation set. In practicality, after many epochs of training the metrics of the validation set become static, but the metrics of the training set still increase. This suggests that the model is overfitting. Thus, we stop training the model on the training set and test it as a testing set.

5.3 Model Evaluation

Testing sets or hold-out sets are used to validate the metrics of the model, this is because both the validation set and the testing set have been tested by the model; thus, the model has developed a latent bias to both sets. Therefore, a third set is needed to assess the model's ability to generalize on an independent dataset. The metrics calculated on the testing set include Accuracy, Sensitive, Specificity, Positive Predictive Value, Negative Predictive Value, F1 Score, and Time Complexity.

6 Model Deployment

Model Deployment is one of the last stages of any machine learning project and involves releasing the model to the public.

6.1 Integration

Integration consists of implementing the model in a system, whether it happens on the client-side or the backend. The most popular backend model integration tools involve Flask, Azure, and FastAPI. These tools create APIs that encapsulate the model prediction, given a GET request with the desired input.

6.2 Monitoring & Maintaining

Following model integration and deployment, we move onto the next phase, monitoring and maintaining the system. As more and more data passes through the model, it increases the opportunity for the model to learn from a more generalized dataset. Though such data would be unsupervised, we could use unsupervised techniques to categorize the data. Based on the improvement of the model, the model would be reintegrated and deployed. In essence, looping the whole process from data management to model learning.

7 Model Structure

The GAN model contains two sub-models, a Generator, and a Discriminator. The Generator is responsible for extracting relevant features from ECG signals and constructing a PCG signal from extracted latent features (in the case of VQGAN). The Discriminator is responsible for extracting relevant features from the PCG signal and creating predictions of whether or not the signal was generated by the Generator.

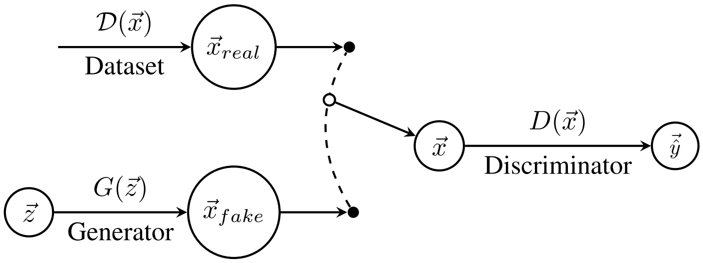


Figure 8: GAN Structure.

The PCG arrhythmia detection algorithm is a sequence-to-sequence Generative Adversarial Network, with the architecture that contains a Generator (G) and a Discriminator (D). The function of a Generator is to generate data such that, the Discriminator will classify the generated data as real, while the Discriminator aims to classify the generated data as fake. Hence, the Generator requires a series of noise signals $z_i \sim N\left(\mu, \sigma^2\right)$; such that $\mu = 0$ and $\sigma = 1$. Thus, $\overrightarrow{z} \in \{-1, 1\}$ where every element of z is between -1 and 1, with a fixed length z_{100}. The Generator then outputs $\overrightarrow{x}_{fake} = [x_1, ...x_n]$ biased on $\overrightarrow{z}$. Conversely, the Discriminator takes inputs $\overrightarrow{x} = [x_1, ...x_n]$ as features from a dataset, $\mathcal{D}$, and calculates the outputs $\overrightarrow{y} = [\hat{y_1}, ...\hat{y_c}]$ based on latent features extracted from $\overrightarrow{x}$, where $\hat{y_i} - 1$ represents each class in the dataset. Given $|\overrightarrow{x}_{fake}| \sim |\overrightarrow{x}|$, meaning the Generator output, $\overrightarrow{x}_{fake}$ is similar in structure and cardinality with the input of the Discriminator, $\overrightarrow{x}$.

7.1 Net-A

7.1.1 Dense Generator

The input consists of a 2-dimensional tensor (batch size x signal) with a length of 128. This input is randomly generated. This input is then fed into the Dense layers that upscale the length of the input vector until the length reaches that of a real input signal (2500). This allows for the output of the generator to be directly fed into the discriminator. Between each dense layer, a Rectified Linear activation function alters the range of the incoming data by setting all numbers below 0 to 0 and leaving all positive numbers intact.

7.1.2 Convolutional Discriminator

The input consists of a 3-dimensional tensor (batch size x channel size x signal) with a length of 2500. Then the convolution decreases the size of the input vector. A vector (with a size of 1x5) filters across the data by multiplying all values in the input vector by the filter vector. This method also assists the model in finding features. The Rectified Linear activation function, between each convolutional and dense layer, alters the range of the incoming data by setting all numbers below 0 to 0 and leaving all positive numbers intact. The liner function flattens the incoming result (batch size x 64 x 309) into a 2D tensor (batch size x19776). The Dense layers downscale the length of the input vector until the length reaches that of the number of classes. This allows for the output of the discriminator to be directly interpreted by the cost function. The Linear Output layer transforms the Linear layer output into a 1x3. Each column in the tensor represents a class's likelihood of being the correct class in the dataset. Thus, the column with the largest values is the model prediction for the input.

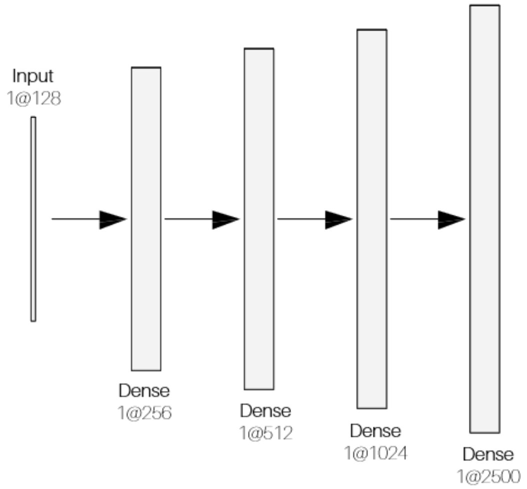


Figure 9: Representation of PCG-Net dense generator structure.

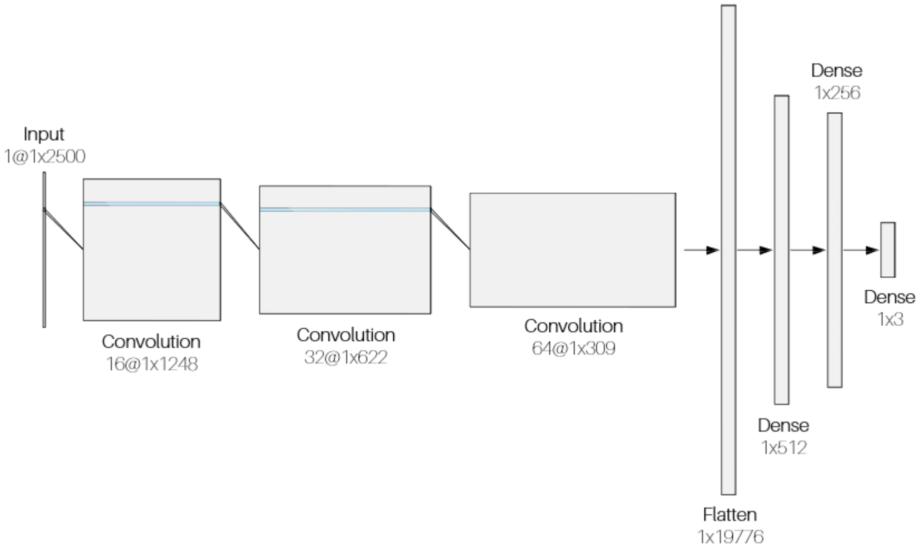


Figure 10: Representation of PCG-Net convolutional discriminator structure.

7.1.3 Architecture

Variational Autoencoders (VAE) is an architecture that attempts to map the initial data into an encoded space ($\hat{z}$) that stores latent features, the encoder is responsible for doing this task. The encoded space is regularized to avoid overfitting. In the context of ECG and PCG signals, both signals come from different latent spaces, which makes it difficult to travel from one domain to another. Regularizing this encoded latent space ensures that the latent space has good properties. This space is then reconstructed by a decoder into an encoded-decoded space, identical to the initial data in shape. With respect to ECG and PCG signals, the input, a spectrogram representation of the ECG signals, is encoded into the encoded space. The encoder consists of ResNet and Multi-head attention blocks that introduce a method to negate the vanishing gradient problem and weights for each pixel in the spectrogram respectively. The encoded space represents important structural elements, such as the positions of the PQRST complex. This space is then decoded using a revered structured encoder, into an encoded-decoded state that respectable a spectrogram of the PCG signal. These reconstructed PCG signals are sent to a convolutional discriminator, which attempts to classify the generated PCG spectrogram as real or fake (reconstructed). Of course, the goal of the discriminator is to classify the constructed spectrograms as fake and spectrograms from a PCG dataset as real. Whereas, the generator attempts to fool

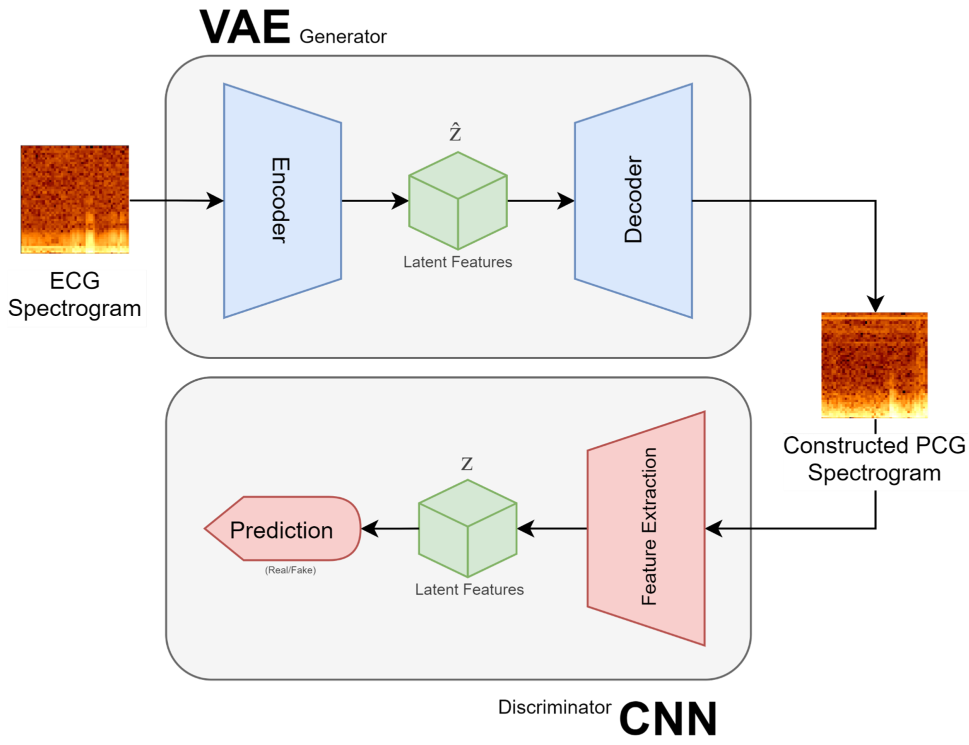


Figure 11: Diagram of VQGAN, composed of a VAE Generator and a CNN Discriminator.

the discriminator into classifying the generated spectrogram as real. This system ensures that the generated spectrogram looks authentic and genuine.

7.1.4 Data Transformation

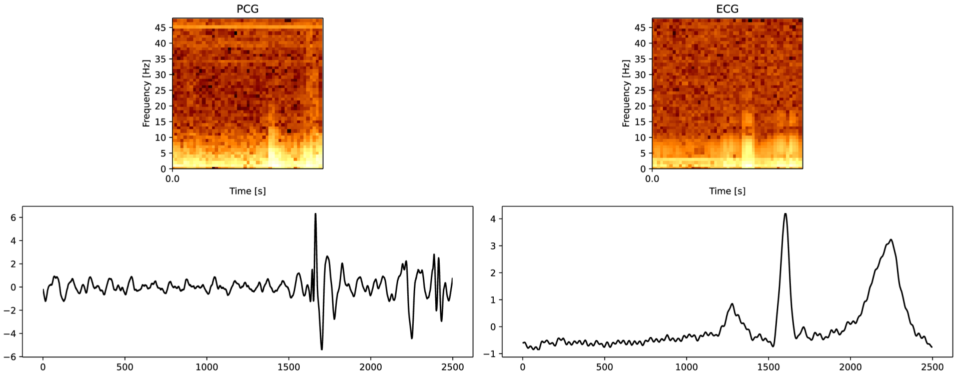


Figure 12: Comparison of PCG and ECG log spectrograms.

Traditional machine learning reconstruction techniques use images because they contain dense information in a low dimensionality. However, time-series data, such as heart sound signals, are one-dimensional. Thus, we use time and frequency analysis, specifically log spectrograms to represent heart sound recordings in a two-dimensional manner. These spectrograms show the frequency of the signal against time. This process is done on both the PCG and ECG signals and results in a 48 x 48 image.

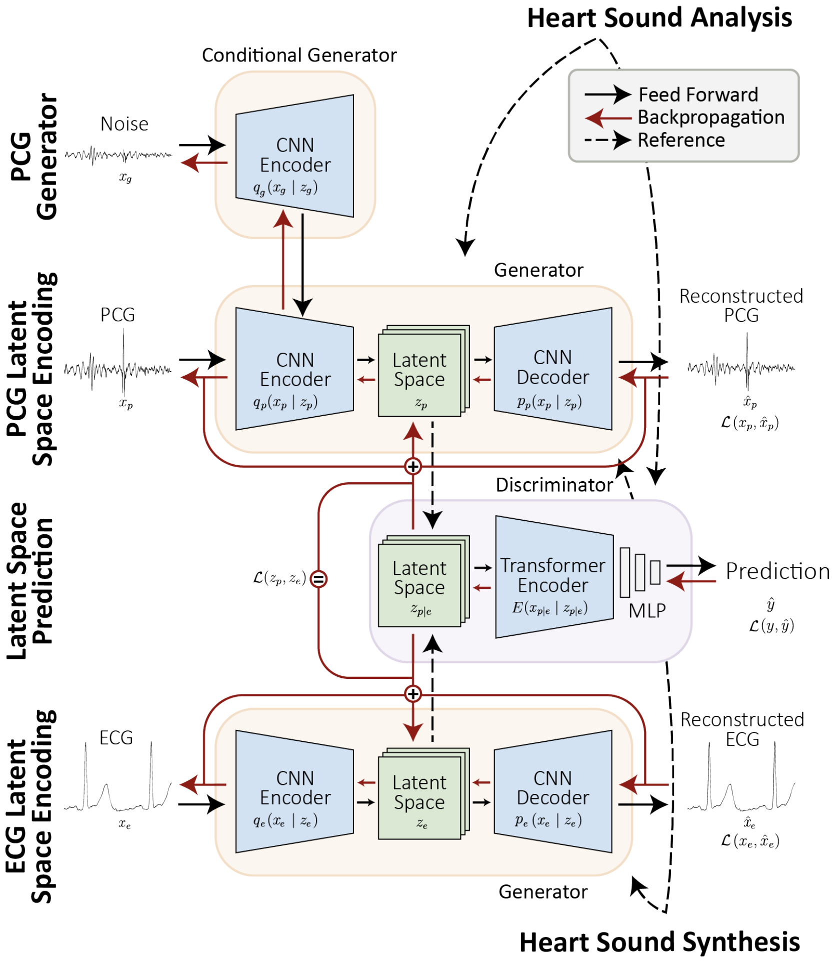


Figure 13: The pipeline of the CNN based generator and Transformer based discriminator. This shows the process of using both PCG and ECG datasets for heart sound (PCG) analysis and synthesis (Figure taken by Kendre, 2021).

7.2 Net-B

7.2.1 Heart Sound Analysis

In heart sound analysis, the PCG CNN Encoder is pre-trained, to compress the PCG signals into a latent space, and then reconstructed into PCGs using the PCG CNN Decoder. After pre-training, the PCG signal's latent spaces are fed into the GAN discriminator (Transformer Encoder). Here, the Transformer creates a prediction based on latent space. Additionally, the PCG Generator tries to mimic PCGs to fool the Transformer into predicting the generated PCG as normal/abnormal, rather than as noisy.

7.2.2 Heart Sound Synthesis

In heart sound synthesis, the process of ECG reconstruction is identical to that of PCG reconstruction. However, during pre-training, the ECG's latent spaces are optimized towards the PCG's latent space. This novel concept forces the Generators to produce equivalent latent spaces for ECG and PCG (given both signals were recorded simultaneously). After pre-training, the ECG CNN Encoder is fed ECG data from categorized arrhythmia datasets.

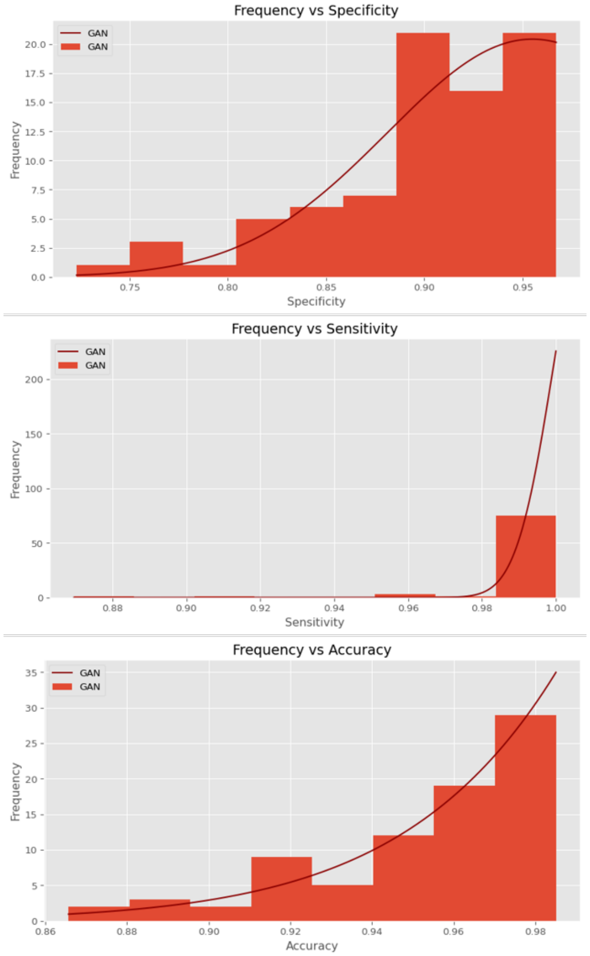


Figure 14: Plot of testing set results on specificity, sensitivity, and accuracy

8 Results

8.1 Net-A

8.1.1 Distributions

Figure 14 illustrates that the GAN's accuracy, specificity, and sensitivity are skewed to the left. This suggests that the model is extremely successful at differentiating between abnormal and normal heart sounds.

The specificity distribution has a mean of 0.9030, with a standard deviation of 0.0547. The best model reached a specificity of 0.9672. This suggests that 90.3% of normal heart sounds were correctly identified.

The sensitivity distribution has a mean of 0.9952, with a standard deviation of 0.0197. The best model reached a sensitivity of 1.0. This suggests that 99.5% of abnormal heart sounds were correctly identified. In detecting pathologies in medicine, we often attempt to maximize sensitivity, the rate at which a subject with a disease is correctly identified amongst other ill subjects. This is because we want to ensure that all potential subjects with a disease are sent for further examination. Essentially, weighting sensitivity higher than specificity, the rate of correctly identified normal or healthy patients from a sample of healthy patients. Thus, from the results, the proposed model is robust, in that it can detect abnormalities nearly 100% of the time.

The accuracy distribution has a mean of 0.9498, with a standard deviation of 0.0293. The best model reached an accuracy of 0.9875. This suggests that 95.0% of both abnormal and normal sounds were classified correctly.

8.1.2 Testing Sample Distributions

Figure 14 illustrates that the GAN's accuracy, specificity, and sensitivity are skewed to the left. This suggests that the model is extremely successful at differentiating between abnormal and normal heart sounds.

The specificity distribution has a mean of 0.9030, with a standard deviation of 0.0547. The best model reached a specificity of 0.9672. This suggests that 90.3% of normal heart sounds were correctly identified.

The sensitivity distribution has a mean of 0.9952, with a standard deviation of 0.0197. The best model reached a sensitivity of 1.0. This suggests that 99.5% of abnormal heart sounds were correctly identified. In detecting pathologies in medicine, we often attempt to maximize sensitivity, the rate at which a subject with a disease is correctly identified amongst other ill subjects. This is because we want to ensure that all potential subjects with a disease are sent for further examination. Essentially, weighting sensitivity higher than specificity, the rate of correctly identified normal or healthy patients from a sample of healthy patients. Thus, from the results, the proposed model is robust, in that it can detect abnormalities nearly 100% of the time.

The accuracy distribution has a mean of 0.9498, with a standard deviation of 0.0293. The best model reached an accuracy of 0.9875. This suggests that 95.0% of both abnormal and normal sounds were classified correctly.

From the p-values shown, we can conclude that all metrics were statistically significant because all p-values were less than 0.05. This implies that the null hypothesis can be rejected and the alternative hypothesis is accepted. The decrease in specificity is expected because we prioritized sensitivity over specificity to maximize the true positive rate, the percent of correctly identified abnormal heart sounds from a sample of only abnormal heart sounds.

8.1.3 Generalization Statistics

Generalization is important in creating accurate predictions, as it establishes that the model is learning meaningful features that are not just applicable to the training data, but signals overall. Figure **??** plots the loss and accuracy of the training and validation set over each epoch. The large fluctuations in the training set metrics are caused by logging the metrics after each step in the training set (on every change in the model parameters). Thus, the optimizer is bound to decrease gradients in the wrong direction, thus correcting for such variations cause those fluctuations. Both lines on the loss plot resemble an exponential curve, which suggests the model continues to learn as training progresses. The average deviation between the validation and training set for each epoch is 0.4817; though this deviation is high, it is due to the lack of meaningful surface-level features that would lead to accurate detection. Meaning, the model's deep feature extraction layers are responsible for the gap. Furthermore, the accuracy plots follow a logarithmic curve. In other words, as the training accuracy increasing, the validation correspondingly increases, though at a slower rate. Both graphs illustrate the model is reached convergence by the end of the training phase. This is confirmed by the static change in metrics in both datasets.

8.1.4 Dataset Feature Visualization

Dataset visualization is critical in understanding the dataset's complexity and model's effectiveness. Here, we use t-distributed stochastic neighbor embedding (t-SNE), a statistical method for visualizing multi-dimensional data in less computationally expensive dimensions. The method is presented with the raw prediction values for each input of the validation set and maps the corresponding predictions into a 2-dimensional space. Tracked over epochs, the visualization allows us to view the progression of the model's competence while training. The visualization highlights clear clustering within the dataset, which suggests the model is stable. Though, from epochs 70 and onwards, it is evident that there is overlapping between abnormal and normal signals. Assuming these signals as ground truth, this implies that additional feature engineering is required to adequately classify heart sounds.

8.1.5 Confusion Matrix

The confusion matrix (Figure 16) aids illustrates the performance for each class of the proposed method. Specifically, we evaluated the model's success on the grounds of average accuracy, specificity, and sensitivity of the classification. We calculated the average true positives, false positives, etc. for all 150 trials of the testing set. Using these floored values, we normalized each label along the y-axis. The matrix reveals that the most common misunderstanding occurs in predicting a normal heart sound; this is not surprising as the model is trained to be biased in detecting abnormal heart sounds. However, this occurs at the cost of specificity, which decreases the average normal misclassification rate to 8.1%. Overall, we conclude that the average accuracy of abnormal heartbeat detection is 95% with a misclassification rate of just 5%. Thus, the model is extremely accurate in detecting abnormalities in heart sounds and has the capabilities to further classify abnormal heart sounds into labeled arrhythmias.

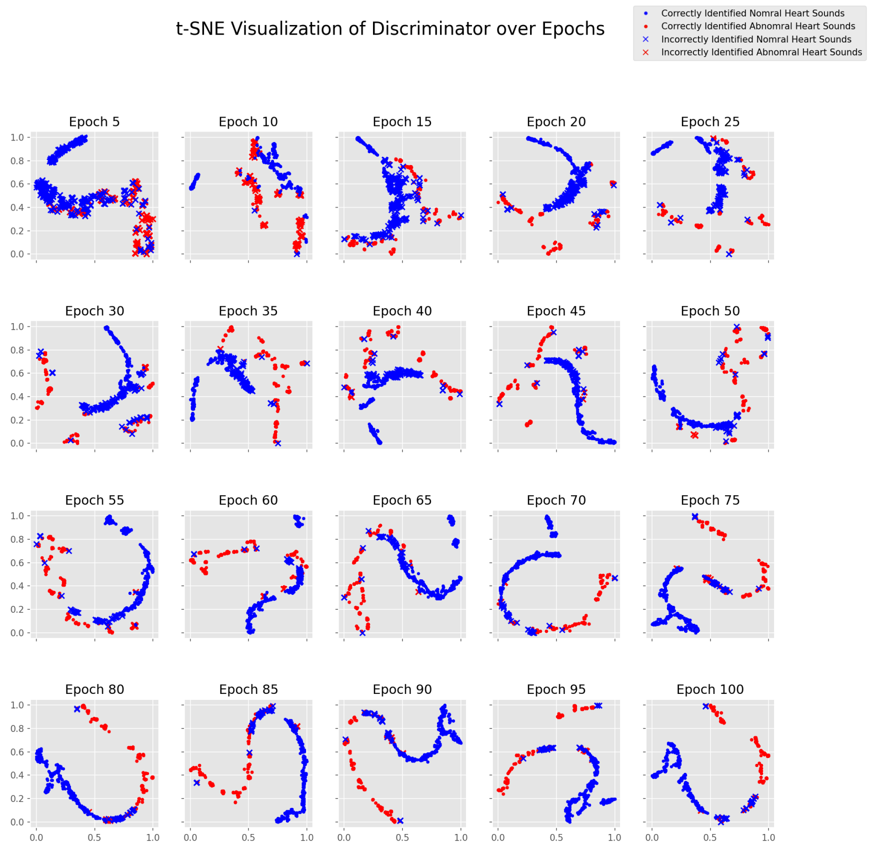


Figure 15: Plot of t-SNE visualizations over 100 epochs of the validation set. The coordinate points are based on the prediction vector of the model

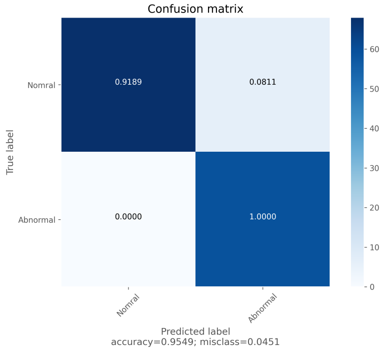


Figure 16: Matrix of accuracy between labels in the dataset.

8.1.6 Time Complexity

Model complexity is used to gauge and evaluate the efficacy of a model against an increase in data (n). We mainly focus on time complexity as it is most relevant to the problem at hand (space complexity is O(1)). Depending on model deployment and integration, the complexity can vary. For example, GPUs have parallel processing capabilities, which allow them to process multiple signals at once, efficiently decreasing the model complexity to O(1). For this reason, we use the worst-case scenario (a CPU), for analysis of the proposed method's time complexity. Figure 17implies the

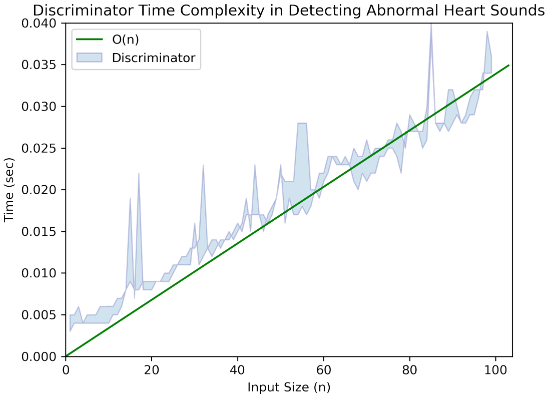


Figure 17: Time complexity of discriminator in classifying heart sounds.

model's time complexity is directly and linearly correlated to the input size, suggesting the complexity is O(n). Thus, the model on average can predict 2800 heart sounds in the worst-case scenario. This will prove to be greatly helpful in real-time detection.

8.1.7 PCG Construction Visualization

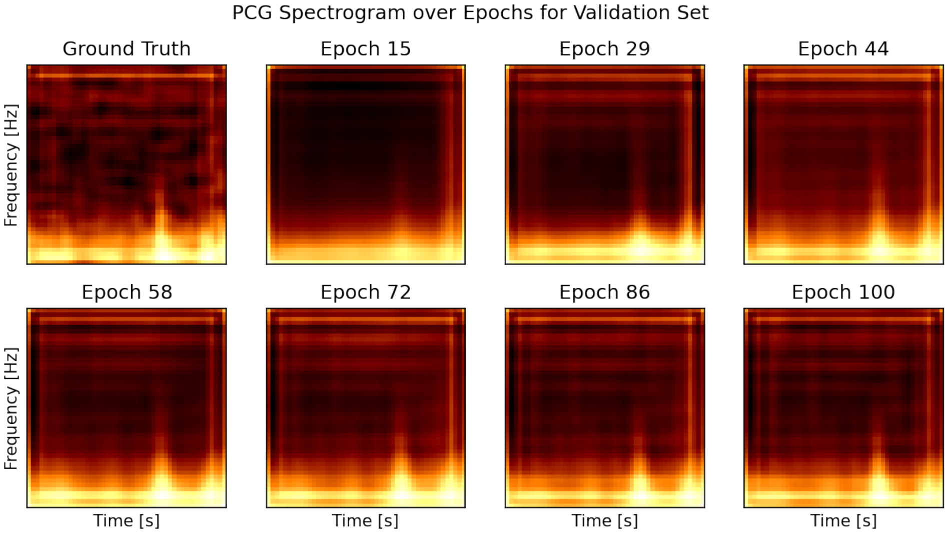


Figure 18: Plot of PCG spectrograms over epochs

Figure 18 show the progression of heart sound construction from ECGs over epochs of the validation set. Ideally, we would want the reconstruction of the PCG spectrogram to be identical to that of the ground truth. In practice, this doesn't occur, some features may be lost in the latent representation of the ECG spectrogram. These missing features will cause a spatial anti-aliasing effect, as the latent space doesn't have the dimensionality to extract pixel-to-pixel information. The series of spectrograms show the development of features in the latent space through the epochs. For example, the frequency of the first peak (S1) varies significantly until the 86th epoch. This feature is important because it determines the rate of the S1 or "lub" sound; thus, creating the illusion that the sound is occurring faster relative to the ground truth. Furthermore, the S2 marker is severely softened, this is due to the rapid change in frequencies surrounding the marker and the light vertical bars to the right in each spectrogram. This suggests that much of the information regarding S2 will be lost when converting the spectrogram into a wave signal.

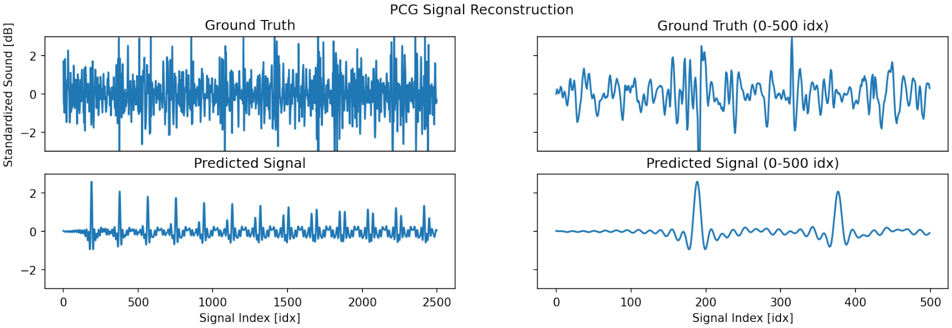


Figure 19: PCG construction accuracy.

8.1.8 Testing Sample Distributions

Figure 20 shows that the VQGAN's loss is skewed to the right, but still has a large amount of variability. This suggests that the model has the ability to construct PCG spectrograms, but has a difficult time accurately constructing all spectrograms. This is supported by both the mean loss - 1.34, and the standard deviation - 0.258, as both values convey a high level of variance. We believe the inconsistencies are caused by the narrow gap of meaningful data. Markers, such as S1 and S2, only occur for a narrow amount of time relative to the signal size. Thus, the majority of the signal's data is nonimportant and contains background noise.

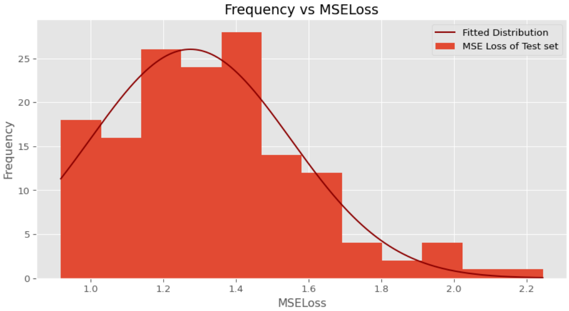


Figure 20: Distribution plot of testing results from PCG spectrogram construction.

8.1.9 Dense Generator Noise

Figure 21shows the input values (seeds) and corresponding generated noise. The variance in amplitudes and periods in the signal mimics that of real heartbeats changing. At a closer examination of the results, the generator appears to resample the seed values to that of the output (2500). This is expected as the generator consists of dense layers, which linearly increase the signal length and complexity.

8.2 Net-B

8.2.1 Distribution Table

The proposed method introduces significant features and architectures that aid in abnormality detection. Using such techniques, the classification metrics showed extremely promising results as shown in Figure 22. Specifically, the model achieved 73.7% accuracy on the PASCAL dataset and 89.5% accuracy on the PhysioNet dataset. Hence, the model proved to be very efficient for the classification of normal and murmur-filled heart sounds. Furthermore, we also propose creating heart sounds from ECGs for additional arrhythmia-specific training. Training the model on the synthesized heart sounds provided a mean accuracy of 93.0%, ECG to PCG conversion is a feasible approach. Additionally, training the model on all 3 datasets indicated excellent results, predicting 95.0% of abnormalities correctly.

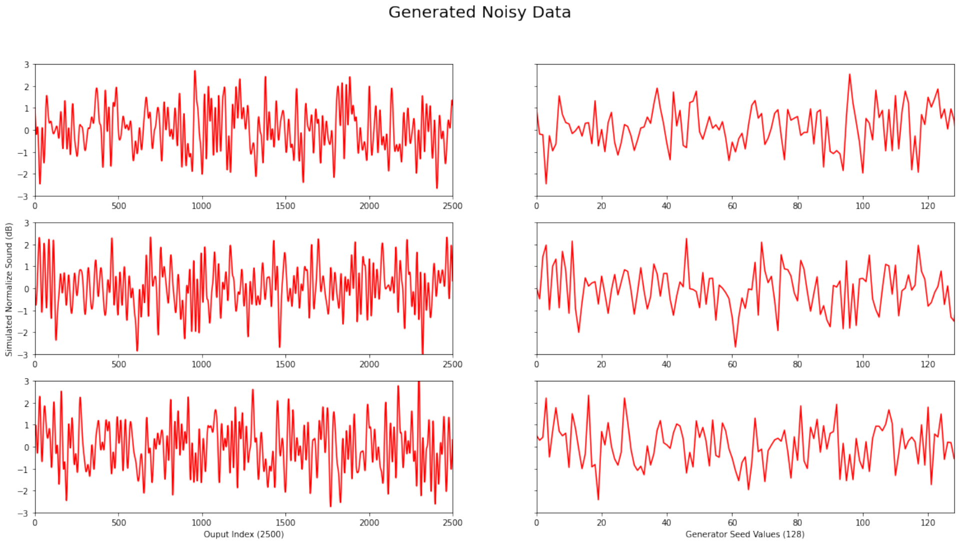
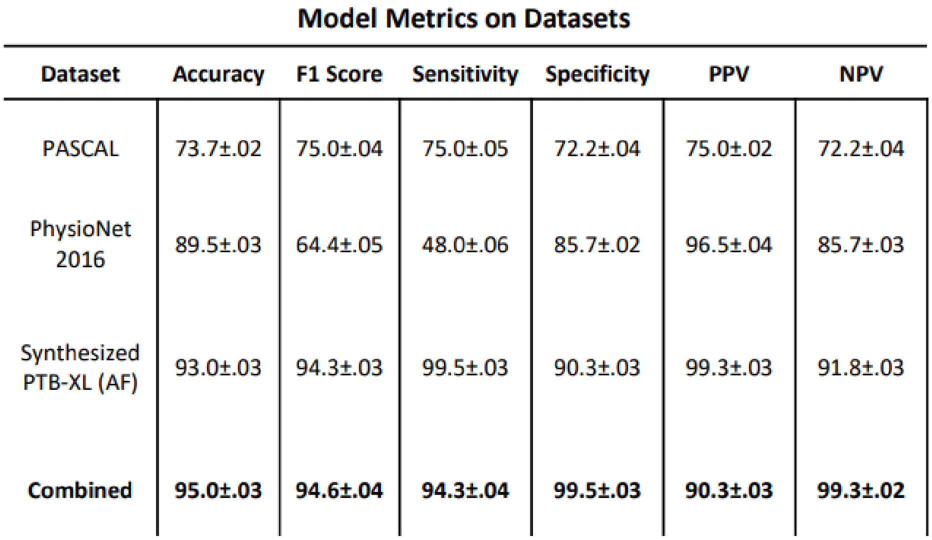


Figure 21: Plot of the noise (artifact class) generated from the generator.



Model Metrics on Datasets

Dataset	Accuracy	F1 Score	Sensitivity	Specificity	PPV	NPV
PASCAL	73.7±.02	75.0±.04	75.0±.05	72.2±.04	75.0±.02	72.2±.04
PhysioNet 2016	89.5±.03	64.4±.05	48.0±.06	85.7±.02	96.5±.04	85.7±.03
Synthesized PTB-XL (AF)	93.0±.03	94.3±.03	99.5±.03	90.3±.03	99.3±.03	91.8±.03
Combined	**95.0±.03**	**94.6±.04**	**94.3±.04**	**99.5±.03**	**90.3±.03**	**99.3±.02**

Figure 22: The table displays the metrics collected on the testing set for all classes in the dataset (Table taken by Kendre, 2021).

8.2.2 ROC/AUC

The receiver operating characteristic curve (ROC) aids in understanding the model's ability to predict the correct heart sound class. The curve plots the true positive rate against the false-positive rate at various prediction thresholds. Based on the ROC, we choose the optimal threshold for all classes. However, the optimal threshold depends on a subjective trade-off between the true and false-positive rates. Here, we choose to optimize for increased true positive rate, as we want to ensure all potential subjects with a disease are sent for further examination. Additionally, the graph illustrates the area under the curve (AUC), this analysis provides an aggregate measure of the model's performance over all classification thresholds. All AUC scores are above the 0.5 threshold; this suggests that the model's ability to distinguish between classes is high.

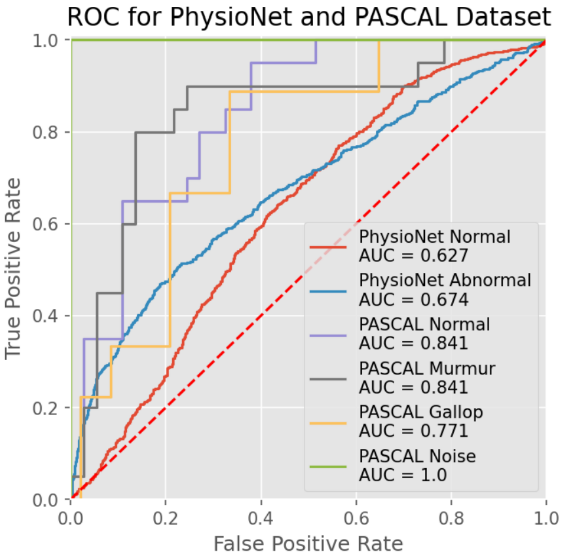
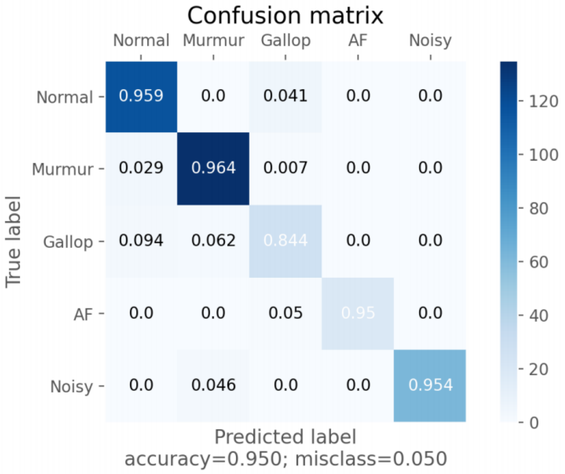


Figure 23: The ROC curve shows the TPR and FPR for different thresholds for the model's prediction and the AUC for each class, which shows the model's efficiency (Figure taken by Kendre, 2021).



	Normal	Murmur	Gallop	AF	Noisy
Normal	0.959	0.0	0.041	0.0	0.0
Murmur	0.029	0.964	0.007	0.0	0.0
Gallop	0.094	0.062	0.844	0.0	0.0
AF	0.0	0.0	0.05	0.95	0.0
Noisy	0.0	0.046	0.0	0.0	0.954

Figure 24: Matrix of accuracy between categories (pathologies) classified by the model (predicted label) and the true categories (true label) (Figure taken by Kendre, 2021).

8.2.3 Confusion Matrix

The confusion matrix in Figure 24 illustrates the performance for each class of the proposed method. Specifically, we evaluated the model's success on the grounds of accuracy, specificity, and sensitivity of the classification. We calculated the average true positives, false positives, false negatives, and false positives for all testing sets. Using these floored values, we normalized each label along the y-axis. The matrix reveals that the most common misunderstanding occurs between Gallops and Normal rhythms. This is expected as Gallops are heart sounds that contain an extra sound like S3 or S4, which are often lacking in amplitude. Overall, we conclude that the average accuracy of abnormal heartbeat detection is 95% with a misclassification rate of just 5%. Thus, the model is extremely accurate in detecting abnormalities in heart sounds and displays the capability to classify abnormal heart sounds into arrhythmia and abnormality types.

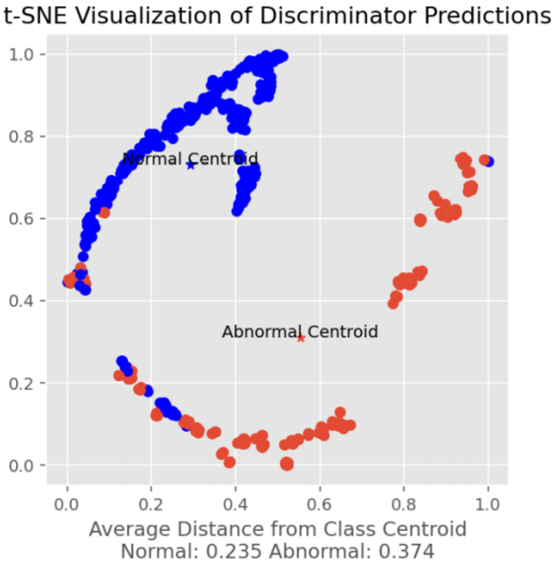


Figure 25: t-SNE visualization of testing dataset after model training (Figure taken by Kendre, 2021).

8.2.4 t-SNE Visualization

Dataset visualization is critical in understanding the dataset's complexity and model's effectiveness. Here, we use t-distributed stochastic neighbor embedding (t-SNE), a statistical method for visualizing multidimensional data with less computational expense. The method is presented with the raw prediction values for each input of the validation set and maps the corresponding predictions into a 2-dimensional space (x, y). Tracked over the best epoch, the visualization allows us to view the differentiation between heart sounds from the model's prediction. The visualization highlights clear clustering within the dataset, which suggests the model is stable. Though, it is evident that there is overlapping between abnormal and normal signals in the y=0.4-0.6 range. Assuming these signals as ground truth, this implies that additional feature engineering is required to adequately classify heart sounds.

8.2.5 Model Interpretation

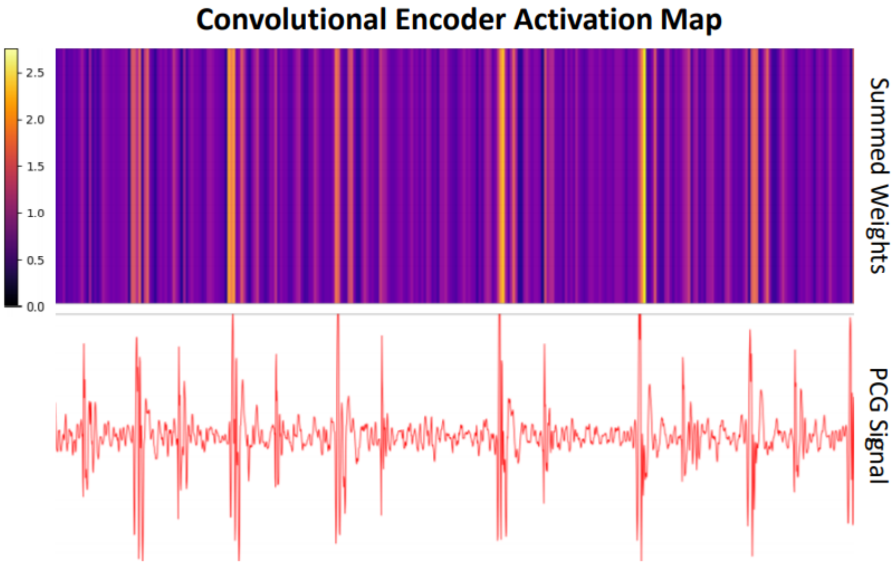


Figure 26: Visualization of CNN encoder when fed a heart sound (Figure taken by Kendre, 2021).

Figure 26 visualizes the channels of the Convolutional Encoder layers which are responsible for extracting features from heart sounds. The color of the line represents the importance of the feature relative to other features. Meaning, the lighter the color, the more important the feature. The map illustrates that the model pays more attention to peaks of higher amplitude in the heart sound. This indicates that the layers are extracting latent features from the signal. Specifically, the plot depicts the extraction of important biomarkers such as S1 and S2. Furthermore, the extractions make it clear that the S1 sound is more important, as those are the brightest throughout all layers. This parallels medical knowledge, as most cardiovascular anomalies occur in Systole, or at the start of S1.

9 Real-World Test

9.1 End-to-End System

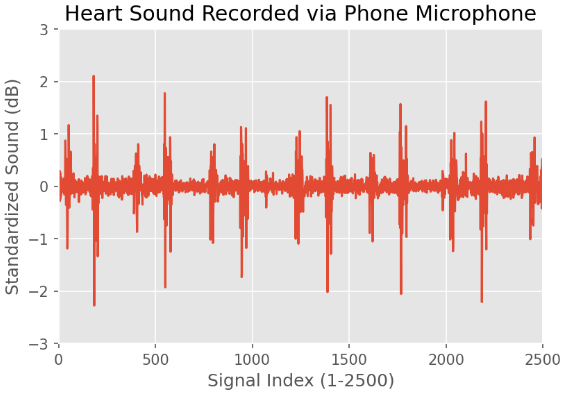


Figure 27: Plot of heart sound recording taken from a phone microphone.

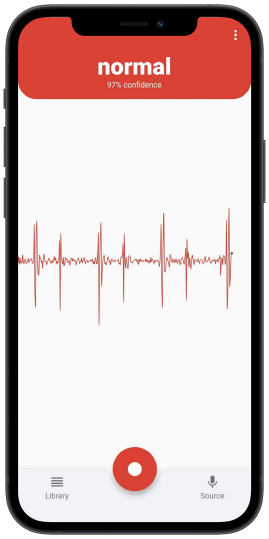
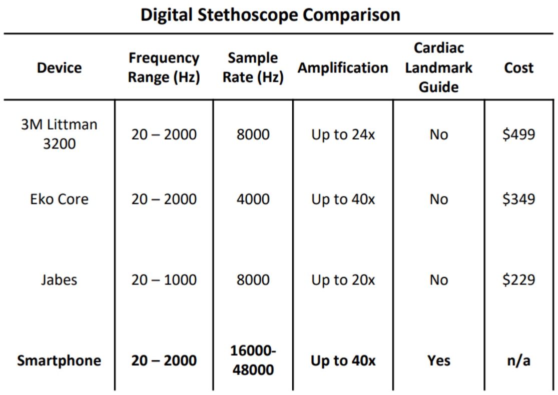


Figure 28: Screenshot of app while recording a heart sound from the built-in microphone.

Testing the model's viability is crucial for ensuring the model's success in the real world. Ideally, recording heart sounds are recorded with digital stethoscopes. These tools use transducer technology to convert sound into an electrical signal. Over the past decade, this technology has grown immensely (by cause of speech recognition). Modern phones have the potential to record the sounds at a high resolution, given the microphone is located at the correct position relative to the heart. Such a device will prove to extremely beneficial in providing diagnosis without the need for specialized equipment. Figure 28 shows a heart sound recording from our smartphone app. The plot shows that import biomarkers, like S1 and S2, remain visible. This ensures that the recording doesn't contain excessive amounts of noise that may hinder the performance of the detection system. Hence, feeding it into the proposed method resulted in a normal classification. The smartphone app not only allows anyone to record their heart sounds with any device with a microphone but does so without needing any external equipment; such as a case or a stethoscope. Additionally, the app allows the users to download and email the sound recordings, providing physicals and cardiologists with deeper insight. This approach allows the app to also track heart activity over time, allowing for more awareness of your fitness level, heart health, and emotional health.

9.2 Comparative System Evaluation



Digital Stethoscope Comparison

Device	Frequency Range (Hz)	Sample Rate (Hz)	Amplification	Cardiac Landmark Guide	Cost
3M Littman 3200	20 – 2000	8000	Up to 24x	No	$499
Eko Core	20 – 2000	4000	Up to 40x	No	$349
Jabes	20 – 1000	8000	Up to 20x	No	$229
Smartphone	**20 – 2000**	**16000-48000**	**Up to 40x**	**Yes**	**n/a**

Figure 29: The table displays different devices that record heart sounds digitals with their respective features (Table taken by Kendre, 2021).

In a clinical setting, physicians often use digital stethoscopes for listening to heart sounds. By monitoring PCG characteristics such as amplitude, pitch, and cyclic patterns, physicians differentiate between abnormalities in the heart sound. Additionally, these devices allow their users to significantly amplify heart sounds by eliminating ambient noises by filtering background noise and amplifying the sounds recorded by the sensor. However, the average heartbeat is between 60 and 100 bpm, meaning a heart sound can occur anywhere from once a second to twice a second. Hence, digital stethoscopes require high sampling rates for collecting heart sounds. Table 4 conveys most digital stethoscopes have a sampling rate of 4000-8000 Hz; however, modern phones have sampling rates of 16000-48000. This high sampling rate provides more accurate and granular control over traditional digital stethoscopes.

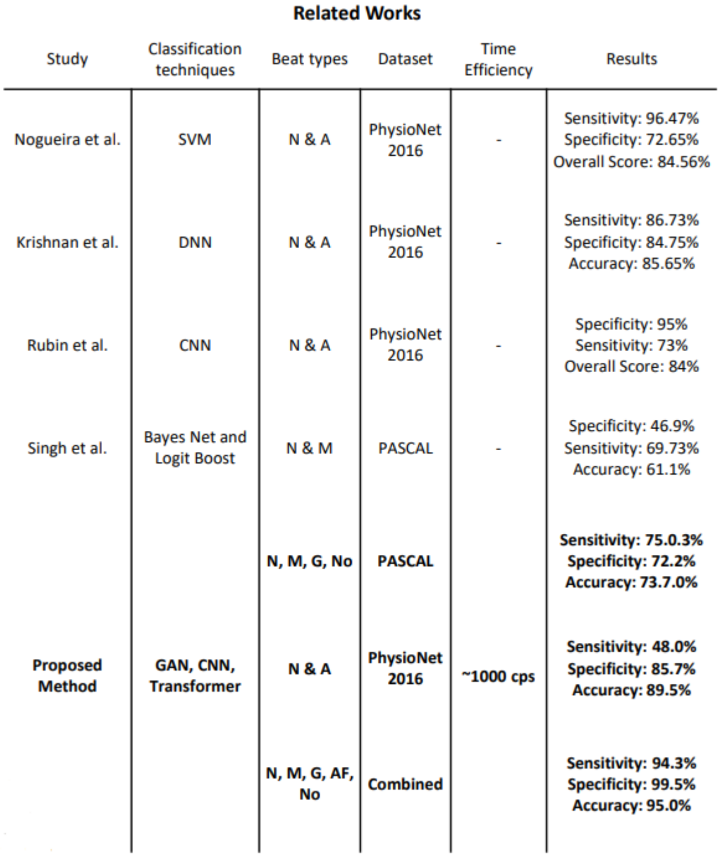
10 Conclusion

10.1 Comparative Model Evaluation

Comparing different methods, we observe the 85.7% was the best accuracy reached on the PhysioNet dataset and 61.1% was the best accuracy reached on the PASCAL dataset. Conversely, our proposed method archived 89.5% accuracy on the PhysioNet dataset and 73.7% accuracy on the PASCAL dataset. The increase in performance is attributed to our semi-supervised approach, where we used adversarial Conditional Generators to generate heart sounds. This exposed the Discriminator to a wide range of heart sounds which assisted the model in optimizing for generalized features. Additionally, we conducted a statistical significance test (t-test) to show the probability our proposed method's results are due to random chance. The test concluded the results were statistically significant as all p-values were less than 0.05. This implies that the null hypothesis can be rejected, and the results are statistically significant.

10.2 Discussion

We proposed a Generative Adversarial Network (GAN), composed of a Convolution Transformer Generator and a Transformer Discriminator to detect abnormal heart sounds in a recording. The results from model testing and evaluation, along with results from the t-test revealed the proposed method reached better performance than the previous state-of-the-art methods. The introduction of heart sounds analysis with ECGs allowed for increased arrhythmia labels for classification and in a time-efficient manner. Furthermore, the proposed method showed real-world deployment capabilities for autonomous heart sound abnormality detection with recordings collected from a phone microphone. In



Related Works

Study	Classification techniques	Beat types	Dataset	Time Efficiency	Results
Nogueira et al.	SVM	N & A	PhysioNet 2016	-	Sensitivity: 96.47% Specificity: 72.65% Overall Score: 84.56%
Krishnan et al.	DNN	N & A	PhysioNet 2016	-	Sensitivity: 86.73% Specificity: 84.75% Accuracy: 85.65%
Rubin et al.	CNN	N & A	PhysioNet 2016	-	Specificity: 95% Sensitivity: 73% Overall Score: 84%
Singh et al.	Bayes Net and Logit Boost	N & M	PASCAL	-	Specificity: 46.9% Sensitivity: 69.73% Accuracy: 61.1%
Proposed Method	**GAN, CNN, Transformer**	**N, M, G, No**	**PASCAL**	**~1000 cps**	**Sensitivity: 75.0.3%** **Specificity: 72.2%** **Accuracy: 73.7.0%**
		N & A	**PhysioNet 2016**		**Sensitivity: 48.0%** **Specificity: 85.7%** **Accuracy: 89.5%**
		N, M, G, AF, No	**Combined**		**Sensitivity: 94.3%** **Specificity: 99.5%** **Accuracy: 95.0%**

Figure 30: The table displays studies with their proposed classification techniques, beat types (types of heart sounds), dataset, time efficient, and results. N(ormal), A(bnormal), M(urmur), G(allop), No(isy) (Table taken by Kendre, 2021)

terms of future development, we propose conducting prospective clinical trials with patients that have different types of arrhythmias. This will allow us to truly test the generalization capabilities of the model and smartphone app in the real world. Depending on these results, we may opt to develop a low-cost DIY and clinical solution for increased sensitivity in heart recordings. Also, applicable fields include medical emergencies that are time constraint (ER) and developing rural communities that don't have access to arrhythmia expertise. Applications with the development of the multiview approach include language and time series processing. Specifically, we can train models to convert language to speech and speech to language without the need for a supervised dataset of language A and language B. Rather, the model can be trained to convert language A to an intermediary language (language C), this language can then be converted into language B. Moreover, we can train the model to reconstruct speech recording directly from electrical signals (EEGs) from the auditory cortex or reconstruct vision from the visual cortex. The object of this study was to create a fast and accurate end-to-end heart sound arrhythmia detection system, capable of detecting abnormalities in real-time without specialized equipment. While also increasing the number of cardiovascular pathologies classified. Our proposed method accomplishes exemplary statistics in abnormality detection and shows promising results in increased heart sound synthesis. Hopefully, this study will shed light on abnormality detection techniques and give birth to applications with signal construction.

11 Further Exploration and Application

1. Deploying the model with an app that is available to 3rd world countries that can't afford to conduct in-depth testing regularly
   (a) Integrate and serve the model using FastAPI

2. Use abnormal heart sound unsupervised datasets as a basis of categorical arrhythmia classification
   (a) Using low dimensional visualization techniques like t-SNE or UMAP
   (b) Cluster data using methods like K means and hierarchical clustering
3. Create a classifiable latent representation of PCG signal biomarkers that can be represented with accuracy and precision
   (a) Created by VAE that are fed the PCG signals directly instead of a spectrogram
4. Investigate training a heart sound discriminator from generated PCG data from ECG datasets
5. Reconstructing speech (wav) recordings from the human auditory cortex (EEG) using techniques used for PCG construction

12 Acknowledgment

I thank Professor Lifang He of Leigh University for helpful feedback on the experiments and references, along with Dr. Venkatesh Murthy at the University of Michigan, Dr. Tarun Sahamd from UPMC CCP Heritage, and Dr. Patel.

13 Code Availability

All code used in this project is available at *https://git.io/JtAuU* and *https://git.io/JtAuO*. All models were run on Google's Colab service with PyTorch and PyTorch Lightning as the framework.

References

[1] Tharindu Fernando, Houman Ghaemmaghami, Simon Denman, Sridha Sridharan, Nayyar Hussain, and Clinton Fookes. Heart Sound Segmentation using Bidirectional LSTMs with Attention. *IEEE Journal of Biomedical and Health Informatics*, 24(6):1601–1609, June 2020. arXiv: 2004.03712.

[2] Cristhian Potes, Saman Parvaneh, Asif Rahman, and Bryan Conroy. Ensemble of Feature:based and Deep learning:based Classifiers for Detection of Abnormal Heart Sounds. September 2016.

[3] Elmar Messner, Matthias Zohrer, and Franz Pernkopf. Heart Sound Segmentation—An Event Detection Approach Using Deep Recurrent Neural Networks. *IEEE Transactions on Biomedical Engineering*, 65(9):1964–1974, September 2018.

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. *arXiv:1811.10597 [cs]*, December 2018. arXiv: 1811.10597.

[5] Prakash D, Uma Mageshwari T, Prabakaran K, and Suguna A. *Detection of Heart Diseases by Mathematical Artificial Intelligence Algorithm Using Phonocardiogram Signals*.

[6] PhysioNet/CinC Challenge 2016: Training Sets.

[7] Nabina N Rawther and Jini Cheriyan. Detection and Classification of Cardiac Arrhythmias based on ECG and PCG using Temporal and Wavelet features. 4(4):6, 2015.

[8] Shahid Ismail, Imran Siddiqi, and Usman Akram. Localization and classification of heart beats in phonocardiography signals —a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2018(1):26, December 2018.

[9] Omer Deperlioglu, Utku Kose, Deepak Gupta, Ashish Khanna, and Arun Kumar Sangaiah. Diagnosis of heart diseases by a secure Internet of Health Things system based on Autoencoder Deep Neural Network. *Computer Communications*, 162:31–50, October 2020.

[10] Dinesh Surukutla, Karan Bhanushali, and Trupti Patil. Cardiac Arrhythmia Detection Using CNN. *SSRN Electronic Journal*, 2020.

[11] Abdulhamit Subasi. Biomedical Signals. In *Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques*, pages 27–87. Elsevier, 2019.

[12] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Las Vegas, NV, USA, June 2016. IEEE.

[13] Palani Thanaraj Krishnan, Parvathavarthini Balasubramanian, and Snekhalatha Umapathy. Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. *Physical and Engineering Sciences in Medicine*, 43(2):505–515, June 2020.

[14] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *arXiv:1606.03556 [cs]*, June 2016. arXiv: 1606.03556.

[15] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *arXiv:1707.01836 [cs]*, July 2017. arXiv: 1707.01836.

[16] A. Almasi, M. B. Shamsollahi, and L. Senhadji. A dynamical model for generating synthetic Phonocardiogram signals. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5686–5689, Boston, MA, August 2011. IEEE.

[17] V. Kalaivani. DIAGNOSIS OF ARRHYTHMIA DISEASES USING HEART SOUNDS AND ECG SIGNALS. *Russian Journal of Cardiology*, (1-ENG):35–41, January 2014.

[18] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11):10.23915/distill.00007, November 2017.

[19] M Finlay. The "mobile-phonocardiogram", a new tool in the arrhythmia clinic. *Heart*, 92(7):898–898, May 2006.

[20] Varsha Garg, Arpit Mathur, Nishant Mangla, and Aman Singh Rawat. Heart Rhythm Abnormality Detection from PCG Signal. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5, Noida, India, August 2019. IEEE.

[21] L. Jordaens. A clinical approach to arrhythmias revisited in 2018: From ECG over noninvasive and invasive electrophysiology to advanced imaging. *Netherlands Heart Journal*, 26(4):182–189, April 2018.

[22] Serkan Kiranyaz, Morteza Zabihi, Ali Bahrami Rad, Turker Ince, Ridha Hamila, and Moncef Gabbouj. Real-time phonocardiogram anomaly detection by adaptive 1D Convolutional Neural Networks. *Neurocomputing*, 411:291–301, October 2020.

[23] P. Lubaib and K.V. Ahammed Muneer. The Heart Defect Analysis Based on PCG Signals Using Pattern Recognition Techniques. *Procedia Technology*, 24:1024–1031, 2016.

[24] Lars-Jochen Thoms, Giuseppe Collichia, and Raimund Girwidz. Real-life physics: phonocardiography, electrocardiography, and audiometry with a smartphone. *Journal of Physics: Conference Series*, 1223:012007, May 2019.

[25] Preecha Yupapin, Wardkein, Preecha Yupapin, Phanphaisarn, Koseeyaporn, Roeksabutr, Roeksabutr, Wardkein, and Koseeyapon. Heart detection and diagnosis based on ECG and EPCG relationships. *Medical Devices: Evidence and Research*, page 133, August 2011.

[26] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*, November 2013. arXiv: 1311.2901.

[27] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016. arXiv: 1609.03499.

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825 [cs, stat]*, June 2017. arXiv: 1706.03825.

[29] Siddique Latif, Muhammad Usman, Rajib Rana, and Junaid Qadir. Phonocardiographic Sensing using Deep Learning for Abnormal Heartbeat Detection. *arXiv:1801.08322 [cs]*, July 2020. arXiv: 1801.08322.

[30] Stephanie Ger and Diego Klabjan. Autoencoders and Generative Adversarial Networks for Imbalanced Sequence Classification. *arXiv:1901.02514 [cs, stat]*, August 2020. arXiv: 1901.02514.

[31] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular Depth Estimation Based On Deep Learning: An Overview. *Science China Technological Sciences*, 63(9):1612–1627, September 2020. arXiv: 2003.06620.

[32] Amy T. Dao. Wireless laptop-based phonocardiograph and diagnosis. *PeerJ*, 3:e1178, August 2015.

[33] Sumair Aziz, Muhammad Umar Khan, Majed Alhaisoni, Tallha Akram, and Muhammad Altaf. Phonocardiogram Signal Processing for Automatic Diagnosis of Congenital Heart Disorders through Fusion of Temporal and Cepstral Features. *Sensors*, 20(13):3790, July 2020.

[34] Md. Khayrul Bashar, Samarendra Dandapat, and Itsuo Kumazawa. Heart Abnormality Classification Using Phonocardiogram (PCG) Signals. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 336–340, Sarawak, Malaysia, December 2018. IEEE.

[35] S1 and S2 Heart Sound Recognition Using Deep Neural Networks. *IEEE Transactions on Biomedical Engineering*, 64(2):372–380, February 2017.

[36] Grzegorz Redlarski, Dawid Gradolewski, and Aleksander Palkowski. A System for Heart Sounds Classification. *PLoS ONE*, 9(11):e112673, November 2014.

[37] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair E W Johnson, Zeeshan Syed, Samuel E Schmidt, Chrysa D Papadaniil, Leontios Hadjileontiadis, Hosein Naseri, Ali Moukadem, Alain Dieterlen, Christian Brandt, Hong Tang, Maryam Samieinasab, Mohammad Reza Samieinasab, Reza Sameni, Roger G Mark, and Gari D Clifford. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181–2213, December 2016.

[38] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv:1904.06037 [cs, eess]*, June 2019. arXiv: 1904.06037.

[39] Mohammed Nabih Ali, EL-Sayed A. El-Dahshan, and Ashraf H. Yahia. Denoising of Heart Sound Signals Using Discrete Wavelet Transform. *Circuits, Systems, and Signal Processing*, 36(11):4482–4497, November 2017.

[40] K. Ajay Babu, Barathram Ramkumar, and M. Sabarimalai Manikandan. S1 and S2 heart sound segmentation using variational mode decomposition. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 1629–1634, Penang, November 2017. IEEE.

[41] Wenjie Zhang, Jiqing Han, and Shiwen Deng. Heart sound classification based on scaled spectrogram and partial least squares regression. *Biomedical Signal Processing and Control*, 32:20–28, February 2017.

[42] M. Abo-Zahhad, Mohammed Farrag, Sherif N. Abbas, and Sabah M. Ahmed. A comparative approach between cepstral features for human authentication using heart sounds. *Signal, Image and Video Processing*, 10(5):843–851, July 2016.

[43] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841 [cs]*, February 2021. arXiv: 2012.09841.

[44] Photo Posters - Create Custom Photo Posters | Walgreens Photo.

[45] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv:1904.06037 [cs, eess]*, June 2019. arXiv: 1904.06037.