

**A Study on Classifying Fetal Distress from Large-Scale
Cardiotocographic (CTG) Data Using Different Machine
Learning Approaches**

Doctoral Thesis

March 2022

Tokyo University of Technology

Graduate School of Bionics, Computer, and Media

Mohannad Alkanan

Abstract

Obstetricians use Cardiotocography (CTG) during labor to look for vital information affecting the health and safety of mother and fetus. However, CTG faces various challenges such as high noise, interpretation inconsistency, and the need for continuous expert presence. This study aims to provide different artificial intelligence approaches where each one acts as supportive guidance to help overcome the present challenges of CTG and predict fetal with potential complications. In the first part of the study, we denoised CTG signals and used an algorithm for extracting important features using Japan Society of Obstetricians and Gynecology (JSOG) guidelines and applied four machine learning methods: SVM, RF, DT, and ANN on extracted features to test the performance of the algorithm in detecting high-risk birth on large data gathered under clinical conditions. The process was tested on pH alone, then on Apgar 1 and 5 only, and finally on pH, Apgar scores 1 and 5. The best result achieved was of Apgar 1 and 5 only using RF with an area under the curve (AUC) of 0.89. The second part discusses our multi-input convolutional neural network (CNN) model which bypasses the need for CTG guidelines and extracts features directly from CTG images and gestational age. The model predicts infants with potentially low Apgar scores and achieved an AUC of 0.958 when classifying infants with Apgar score 5 minutes < 7 and an AUC of 0.955 when Apgar score 1 or 5 minutes < 6 . In the third part of the study, we used anomaly detection generative adversarial networks (ANOGAN) to analyze and rate normal and abnormal CTG images based on anomalies found in them. This study overcomes data imbalance, a major challenge in artificial intelligence, by training only on the majority class which is the patients with a pH of 7.1 or higher. Moreover, the original ANOGAN uses DCGAN architecture so we also implemented ANOGAN on WGAN and WGAN-GP architectures. The trained model is tested on a dataset that includes an identical number of normal and abnormal CTG where it generates each image in the test dataset and provides a score based on the similarity between real and generated images. The model is trained solely on negative class. We found that CTG images with pH < 7.1 tend to have a higher anomaly score than CTG with pH ≥ 7.1 . Based on the results of the aforementioned studies, we conclude that our approaches could support each other in guiding medical teams in mitigating risks they encounter during childbirth.

Contents

1	General Introduction	1
1.1	Research Background	1
1.2	Previous Studies	2
1.3	Research Objective	3
1.4	Thesis Organization	5
2	Application of Machine Learning Techniques to Classify Fetal Hypoxia Using Japan Society of Obstetricians and Gynecology Guidelines	6
2.1	Introduction	6
2.2	Related Works.	7
2.3	Study Objective.	7
2.4	Method.	7
2.5	Data Set and Processing.	8
	2.5.1 Description of the Collected Data.	8
	2.5.2 Denoising Process.	8
	2.5.3 Feature Extraction.	10
	2.5.4 Applying Machine Learning Methods.	13
2.6	Experimental Results.	14
2.7	Discussion and Conclusion.	16

3	Predicting high risk birth from real large-scale cardiotocographic data using multi-input convolutional neural networks	18
3.1	Introduction.	18
3.2	Related works.	20
3.3	Study Objective.	20
3.4	Data Set and Processing.	21
3.5	Target tasks for prediction.	21
3.6	Prediction using deep CNN model.	22
	3.6.1 Image processing block.	22
	3.6.2 Meta-data processing block.	23
	3.6.3 Merging blocks.	23
3.7	Results, ML vs DL comparison and discussion.	25
3.8	Conclusion.	26
4	Analyzing the Relationship between Abnormality in Fetal Heart Rate and low pH test using Anomaly Detection Generative Adversarial Networks model	28
4.1	Introduction.	28
4.2	Related Works.	30
4.3	Study Objective.	31
4.4	The GAN Models.	31
	4.4.1 Deep Convolutional Generative Adversarial Networks (DCGAN).	32
	4.4.2 Wasserstein Generative Adversarial Networks (WGAN).	33
	4.4.3 Wasserstein Generative Adversarial Networks Gradient Penalty (WGANGP).	34
4.5	Method.	34

4.5.1	Data Set and Processing.	34
4.5.2	Define the GAN models for training and the applied hyperparameters.	35
4.5.3	Generated Image Evaluation.	37
4.5.3.1	Fréchet Inception Distance (FID).	37
4.5.3.2	Probability Density Function (PDF).	38
4.5.3.3	Precision and Recall (PR).	40
4.5.4	Compare the generated images and anomaly scores.	41
4.6	Results and discussion.	48
4.7	Conclusion.	55
5	Conclusion	56
5.1	Research Results.	56
5.2	Future works.	57
	Acknowledgments	59
	References	60
	Letter from OBGYN department at Fukuoka University	70

List of Figures

1	Results of removing noise from FHR & UC where first image represents the original signal and the second image represents the denoised signal.	9
2	A summary of the process for extracting features from CTG signals and assigning target labels.	12-13
3	The architecture of the proposed model.	24
4	GAN model.	32
5	DCGAN architecture.	33
6	A plot of the achieved FID through all the training process for DCGAN, WGAN, and WGANGP.	38
7	A plot of the generated distribution compared to the real distribution of DCGAN, WGAN, and WGANGP.	39
8	Precision and recall in GAN models.	40
9	Generated CTG images compared to the real and the recorded anomaly score.	42
10	Samples of generated CTG images compared to the real images for DCGAN, WGAN, and WGANGP.	43-47
11	All positive and negative CTG images after dividing each record into 4 parts	52
12	A bee swarm plot of Table 14 where (A) represents the part 1 column and (B) the part 4 column.	53
13	The above subplots show that positive records tend to have a higher anomaly score in part 1 compared to part 1 in negative records and a lower anomaly score in part 4 than its counterpart in negative records.	53
14	A plot of Generator and Discriminator/Critic loss for DCGAN, WGAN, and WGANGP.	54

List of Tables

1	The institutes that provided the dataset.	8
2	JSOG five-tier fetal heart rate classification.	11
3	Each row in the dataset should have the below features for each 10 minutes. Since the size of each CTG is limited to 30 minutes, the total features equal to 44.	11
4	The number of positive and negative samples in training and testing data for the ML experiment.	14
5	ML using three feature sets on pH only as target label. In each test, extra features are being added to experiment the performance variation.	15
6	ML using three feature sets on both Apgar score 1 and 5 only as target labels. In each test, extra features are being added.	15
7	ML using three feature sets on pH, Apgar score 1 and 5 as target labels. In each test, extra features are being added.	16
8	The number of positive and negative samples in training and testing data for the CNN model.	22
9	The chosen parameters for the model.	25
10	AUC after training the model for different targets.	25
11	A summary of DCGAN, WGAN, and WGANGP networks.	36
12	The achieved precision and recall for DCGAN, WGAN, and WGANGP.	41
13	A summary of the number of positive CTG compared to negative for all the four parts when using DCGAN.	48
14	A summary of the number of positive CTG compared to negative for all the four parts when using WGAN.	48
15	A summary of the number of positive CTG compared to negative for all the four parts when using WGANGP.	49

16	16.1 a summary of all positive CTG images sorted based on the highest anomaly score. DCGAN.	50
	16.2 a summary of all negative CTG images sorted based on the highest anomaly score. DCGAN.	50
17	17.1 a summary of all positive CTG images sorted based on the highest anomaly score. WGAN.	51
	17.2 a summary of all negative CTG images sorted based on the highest anomaly score. WGAN.	51
18	18.1 a summary of all positive CTG images sorted based on the highest anomaly score. WGANGP.	51
	18.2 a summary of all negative CTG images sorted based on the highest anomaly score. WGANGP.	51

Chapter 1

General Introduction

1.1 Research Background

The implementation of machine learning (ML) and neural network in the medical field sought an ever grown in the last decade. This interest emerges from the coinciding growth of artificial intelligence research and the need for better health care, lower costs, and higher efficiency. It is usually challenging to have professional practitioners and the challenge becomes more apparent if it is a complicated medical condition or a remote place in the world. There are many ways in which machine learning and neural network can support medical staff such as data collection, recordkeeping, medical imaging, diagnoses, outbreak predictions, and more which mitigate some difficulties facing specialists and help them in making informed decisions, focusing on more critical conditions and expanding their availability for more patients. Medical researchers utilized machine learning in different ways to benefit their needs. For example, the use of deep learning in medical imaging to identify cancerous tumors in early-stage to raise the probabilities of finding a treatment [1] or as a decision-making tool to support the experts in assessing kidney allograft biopsies transplant to diagnose allograft rejection [2].

Cardiotocography (CTG) is the process of reading fetal heart rate and uterine contraction in the third trimester of pregnancy and during birth. It is an important assessment tool used by obstetricians worldwide to assess the general health condition of the fetus [3]. CTG consists of fetal heart rate (FHR) and uterine contractions (UC). Interpreting both FHR and UC is a challenging task since it relies on the visual assessment of heartbeats and requires the continuous presence of experts. This can be difficult especially in hospitals with many patients and few obstetricians. Continuous monitoring of fetal heart rate is essential to detect the potential of fetal

hypoxia and other serious health risk issues [4]. It is difficult to make a consensus interpretation when interpreting CTG due to the inconsistency between obstetricians' judgment [5]. Moreover, it is not uncommon for an obstetrician to have a different opinion on the same CTG at separate time periods [6]. CTG signal also suffers from high noise due to the nature of the environment it is used in. The transducers attached to the mother monitor FHR and UC signals and the continuous movement of the mother during labor cause artifacts to CTG signals and false readings. FHR abnormalities are significantly influenced by the occurrence of hypoxic-ischemic encephalopathy (HIE) which involves an unusual change in the oxygen supplied to the fetal through the umbilical cord and the well-being of the placenta and uterus [7]. Abnormalities are found when a fetal suffers from oxygen deficiency which gives low arterial cord hydronium ions concentration (pH) a lower reading. This suggests the urgency of medical team intervention.

Another major indicator of fetal abnormalities is the Apgar score. After the first minute of delivery, the medical team checks how the fetal tolerated the birth process and rates it based on the Apgar rating system, which rates fetal skin color, pulse rate, reflex, muscle tone, and respiratory effort. Each is rated from 0 to 2 where a total of ≥ 7 out of 10 is the average. This procedure is repeated after 5 minutes, hence they are named Apgar score 1 and 5. The use of FHR monitoring leads to a substantial decrease in the risk of early neonatal mortality and morbidity which cause infant mortality [8]. However, even though it's been used for a long time, the potential of continuous FHR monitoring is yet to be completely explored [9]. Artificial intelligence showed significant advancement in the medical domain and many studies applied it to computer-based intrapartum fetal monitoring and CTG analysis. Georgieva et al., a review of a biannual workshop on the latest advancement of signal processing and monitoring in labor, discussed multiple approaches and techniques done in CTG and how AI, especially machine learning (ML) and deep learning (DL), can further improve this field and propose potential solutions to its problems [10].

1.2 Previous Studies

By looking through research projects implementing machine learning and deep learning in CTG, we can outline the general ideas of the studies into the following:

Some studies build or use an algorithm based on one of CTG guidelines to extract important features and input these features into different ML models to classify abnormal from normal births [11,12]. Obstetricians rely on guidelines to interpret FHR and UC patterns and based on the meaning of the CTG's patterns they make decisions. These researches depend on the CTG guidelines to extract important features and use ML to classify risks to newborns. However, the

previous experiments still use guidelines built by humans and there are several guidelines to implement and the guidelines have inconsistencies between them [5,13]. Other researchers utilized deep learning approaches to extract features from CTG directly and interpret the newborn condition based on the features. Some studies applied tweaks to some of the popular neural network architectures and some built new models specifically for the researched task [14,15,16].

Additionally, we noticed the variety in datasets is quite limited. Many types of research projects were done on the CTU-UHB Intrapartum Cardiotocography dataset [17]. The dataset contains only 552 cardiotocography (CTG) recordings with a gestational age of 37 weeks or more and suffers from data imbalance. Some studies worked on improving the quality of the studied data. For example, I. Linardos used Generative Adversarial Networks (GAN) to generate non-anomalous data [18] and R. D. I. Puspitasari et al. used Time Series GAN to generate more data in the CTU-UHB CTG database which, according to the paper, could solve data imbalance and improve classification when using convolutional neural networks [19].

1.3 Research Objective

By reviewing the outcomes of previous studies that focused on utilizing machine and deep learning in classifying high-risk birth we think this problem could be solved through the use of multiple models where each model confirms the outcome of the other. Our goal is to contribute different artificial intelligence approaches in providing supportive guidance to help overcome the present challenges of CTG and predict fetal with potential complications. This was also the suggestion and conclusion of some of the recent studies [14, 20]. Relying on the judgment of only one model may not give a complete perception of the occurring issue. For instance, establishing an opinion based only on the Apgar score may be misleading [21]. This is why we think each study we did so far can be used to some extent to help a medical team to further confirm the health and safety of the fetus.

We attempt to achieve our goal through several objectives. First, is an improved CTG denoising process that removes high noise without affecting the overall form of the original signal. Our dataset was collected under a clinical trial. The signals suffer from relatively high noise. This is why we need a reliable denoising process that deals with high noise. Second, we wanted to study if machine learning models can learn features extracted from CTG guidelines and make a prediction about whether the fetus is at risk or not. We built an algorithm to extract important features from CTG using Japan Society of Obstetricians and Gynecology (JSOG) guidelines for CTG interpretation and use machine learning methods to evaluate the performance of the algorithm

in detecting high-risk birth on large data obtained under clinical conditions. The classification was tested using Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Artificial Neural Network (ANN). The process was tested on three target labels: 1- pH, 2- Apgar score 1 and 5, and 3- Both pH and Apgar score. Using Area Under Curve (AUC) to evaluate the performance, the highest AUC achieved was 0.89. The target label was Apgar 1 and 5 only using RF. Obstetricians use CTG guidelines to detect normal and abnormal patterns such as the International Federation of Gynecology and Obstetrics (FIGO) and the National Institute of Child Health and Human Development (NICHD) and several studies applied machine learning methods to these guidelines [11,22]. However, to the best of our knowledge, we found no study that applied machine learning using JSOG guidelines 5-level FHR patterns classification, which is the official practiced guidelines in Japan.

Third, contrary to the latter approach, we wanted to extract CTG features without depending on CTG guidelines. We built a deep neural network model to learn features directly from CTG images and classify fetal with the risk of a low Apgar score, bypassing the need for CTG guidelines. We used Efficient-Net with transfer learning on ‘image-net’ dataset weights concatenated with a simple neural network made solely for training on gestational age. The CNN model performance scored an Area Under Curve (AUC) of 0.958 when classifying infants with Apgar score 5 minutes < 7 and an AUC of 0.955 if Apgar score 1 or 5 minutes < 6 without an algorithm for feature extraction and without using data augmentation or synthesis. We noticed improvement in AUC performance when training on CTG images with longer signal lengths.

Lastly, we implemented anomaly detection using generative adversarial networks (ANOGAN) [23] to analyze anomalies in abnormal CTG images which may cause complications to a fetus and result in a low pH test score of lower than 7.1. The model is trained on a dataset containing CTG images of patients with non-acidosis pH, which is $\text{pH} \geq 7.1$, and tested on a dataset where 50% of its records are $\text{pH} < 7.1$ and 50% are $\text{pH} \geq 7.1$. Unlike the previous experiments, we don’t modify the dataset size for data imbalance as we use all of the possible records which are equal to 13723 after the data processing step. Because training contains only negative classes, training shouldn’t struggle with class imbalance, which is a major challenge in studying CTG using artificial intelligence. O’Sullivan ME et al. studied the challenges artificial intelligence faces when dealing with CTG. In their discussion section, they stated “class imbalance is a major concern, and perhaps an anomaly detection approach may be best suited” [24].

In the original ANOGAN paper, the model was built upon DCGAN [25] architecture. However, in this study, we also implemented ANOGAN on WGAN [26] and WGANGP [27] architectures. Both WGAN and WGAN-GP have shown success in solving some of DCGAN's issues such as the occurrence of exploding gradient, vanishing gradient, or mode collapse. In ANOGAN, a trained discriminator is treated as a feature extractor, not as a classifier. The discriminator learns the feature representation of the training set and maps new images to the latent space and finds an image with similar features. Based on these features the model generates CTG images from the test set and both the generated images and the real images are then compared with each other using the anomaly score algorithm [23]. The anomaly score is given to real images and the value is an indication of the difference between the real and the generated CTG image. A high anomaly score suggests a high difference between the two images. We found no study that applied GAN or ANOGAN to generate CTG images or detect abnormalities in FHR patterns, especially on big data sets and without data augmentation. We implemented ANOGAN using DCGAN, WGAN, and WGANGP architectures, then compared their performance in generating images using Fréchet Inception Distance (FID) [28], Probability Density Function (PDF) from probability theory, and Precision and Recall (PR).

1.4 Thesis Organization

In this thesis, different artificial intelligence approaches were applied to study classification using CTG where each approach touches upon a challenge in CTG to provide the medical team with prognosis tools to help identify infants with potential risk. Chapter 2 focuses on two parts, first: a description of the dataset used in this work with the processes applied to it and the denoising process needed to prepare the dataset for the study. Second, an algorithm for extracting important CTG features using Japan Society of Obstetricians and Gynecology Guidelines (JSOG) guidelines/rules and applied multiple machine learning methods to classify fetus hypoxia. Classification is tested using three labels: pH, Apgar score 1 and Apgar score 5. Chapter 3 introduces our multi-input convolutional neural networks (CNN) to predict neonatal with possibly low Apgar score. The CNN model accepts denoised cardiotocography (CTG) images and gestational age and uses binary classification to detect if a fetus may have a low Apgar score before birth. Chapter 4 examines the capability of Anomaly Detection Generative Adversarial Networks (ANOGAN) to analyze the relationship between CTG images with low and normal pH by generating fake CTG images and comparing them with real images. The quality of the generated images is evaluated using three different methods for evaluating GAN's generated images. Chapter 5 is a summary of the thesis.

Chapter 2

Application of Machine Learning Techniques to Classify Fetal Hypoxia Using Japan Society of Obstetricians and Gynecology Guidelines

2.1 Introduction

FHR interpretation relies on a set of rules or guidelines to understand them. These guidelines are not universal. There are multiple attempts by experts in the field to standardize the management of FHR patterns [29]. The International Federation of Gynecology and Obstetrics (FIGO) published the only international consensus so far. Still, health institutes usually use guidelines developed regionally or locally but it is not uncommon to find different health institutes in one country using different guidelines. For instance, in the United States, CTG is often interpreted based on the American College of Obstetrics and Gynecology (ACOG) or National Institute of Child Health and Human Development (NICHD) and in the United Kingdom usually interpreted using the institute for Health and Care Excellence (NICE) guidelines. In general, all the guidelines agree and disagree with each other in some aspects and have shown to have different accuracy [30].

In Japan, the regularly used guidelines are made by the Japan Society of Obstetricians and Gynecology (JSOG). Unlike other guidelines, JSOG uses a 5-level classification of CTG patterns where 1 suggests a normal pattern and 5 as the highest risk variant pattern as shown in Table 2. JSOG guidelines proved that it can predict early neonatal outcomes that usually result from a low Apgar score, and low pH, thus it is medically acceptable [31]. In this study, we built an algorithm that follows JSOG guidelines to detect important CTG features and used multiple ML methods to classify high-risk birth.

2.2 Related works

This work shares similarity with [11], [12], and [22]. The focus of [11] was comparing the performance of the most known ML techniques applied to FHR whereas Hasan et al. used FIGO guidelines as bases for their extracted features [12] and Esteban-Escano J et al. used NICHD guidelines [22]. So far, to the best of our knowledge, no study applied machine learning methods using JSOG 5-level FHR patterns classification on a realistic big dataset.

2.3 Study Objective

The objective of this study is to apply machine learning methods to test the performance of our locally built algorithm using JSOG guidelines' 5-level pattern judgment to find high-risk fetuses through the use of Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Artificial Neural Network (ANN). To achieve this objective, the algorithm must extract important features from each CTG record in the dataset. Those features are extracted by using JSOG guidelines, which assign each fetal with a level. Level 1 is the lowest level which indicates a normal pattern and level 5 is the highest level and indicates a severe variant pattern.

2.4 Method

The work can be summarized into three main steps: 1- Data processing which includes removing noises from both FHR and UC, dropping records with short or bad signals, and removing unnecessary or empty columns. 2-Creating the algorithm for extracting essential features and rating them with a level based on JSOG guidelines. 3-Apply SVM, RF, DT, and ANN on all records to measure the performance using the area under the curve of receiver operating characteristics (AUC- ROC) curve.

2.5 Data Set and Processing

2.5.1 Description of the Collected Data

The dataset used in this experiment is a restricted internally available 38,073 CTG records from 2012 to 2017 gathered from several hospitals (Table 1). In addition to FHR and UC data, the dataset metadata includes the delivery date, time and type, maternal and gestational age, fertility treatments, fetal weight and gender, pH, and Apgar score 1 and 5. All deliveries are singleton, extracted at the second stage of labor and nearly all are low-risk pregnancies.

Table 1: The institutes that provided the dataset

Name of Institute	Number of records
Fukuoka University Hospital	2193
Kyushu University Hospital	3539
Izuchi Hospital	4929
Fukuda Hospital	21,138
Kumamoto University Hospital	1,231
Total	38,073

2.5.2 Denoising Process

CTG records are prone to noise and signal drop. This is common to occur because the transducers attached to the mother's abdomen may not be well fastened or because of the physical activities of both the mother and the fetus, which affect the stability of the readings. Our dataset's CTG recordings were collected under clinical conditions and some records include excessive noise, long signal drop, and short signal recording.

The process of noise removal was inspired by Zhao, Zhidong et al. [15]. However, the noise in the dataset of this work suffers from high noise, so some tweaks were made to the [15] approach. If any ≥ 15 seconds in FHR is filled with 0s, then it is dropped. Any other 0s are removed (changed to null). Any FHR that ≥ 200 bpm or ≤ 50 bpm is removed. If any point within 15 seconds is $> |25|$ bpm than the mean bpm of that 15 seconds, this point is removed. Finally, linear

interpolation is applied to all removed values and 0s. UC noises were reduced using a moving average window of 60. An example of the CTG denoising process can be seen in Figure 1. Also, the time length of each record varies as some are more than 3 hours and some others are less than 30 minutes. We included only the last 60 minutes before birth of every CTG.

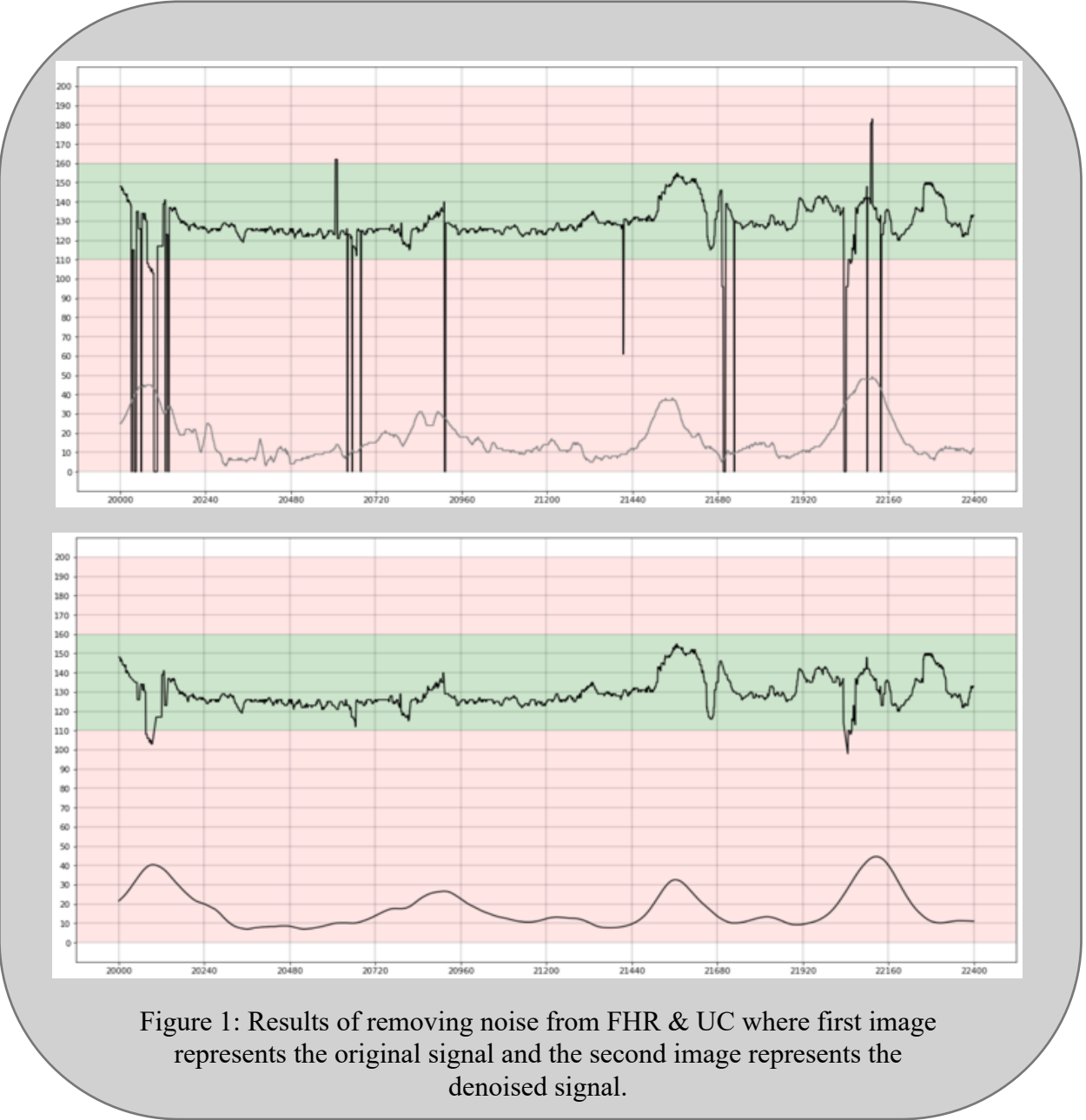


Figure 1: Results of removing noise from FHR & UC where first image represents the original signal and the second image represents the denoised signal.

2.5.3 Feature Extraction

Following JSOG guidelines definitions, the extracted features from CTG are *baseline*, *baseline variability*, *accelerations*, *early decelerations*, *mild/severe variable decelerations*, *mild/severe late decelerations*, and *mild/severe prolonged decelerations*. The baseline is defined as follows: first, rounding FHR in every quarter second within 10 minutes period to the nearest 5 bpm, and then, the mode (most frequent number) of that 10 minutes is taken and set as the baseline value. Baseline variability is counted by calculating the difference in FHR peaks per minute. Baseline variability is called “absent” if the difference is 0 bpm, “minimal” if less than 5, “moderate” if between 5 and 25, and “marked” if more than 25.

Regarding the definitions of accelerations and decelerations, in general, acceleration refers to when FHR increases rapidly, and deceleration refers to when FHR decreases rapidly or slowly. If FHR is 15 bpm or more above the baseline for 15-120 seconds, this is counted as acceleration. Variable deceleration is classified as “mild” when FHR is 15 bpm or more below the baseline for 15-120 seconds. It is classified as “severe” when the lowest point is less than 70 bpm and the duration is 30 seconds or more, or the lowest point is within 70-80 bpm and the duration is 60 seconds or more. Early deceleration happens if FHR, which decreases gradually for 30 seconds or more, coincides with the occurrence of the UC where both the FHR and the UC start and end at the same time. Late deceleration is classified as “mild” when the decrease in FHR continues gradually for more than 30 seconds, begins near the highest point of UC, and ends after UC ends, and the late deceleration becomes “severe” when FHR’s lowest point is more than 15 bpm from the baseline. Prolonged deceleration is classified as “mild” when FHR is more than 15 bpm below the baseline for 2-10 minutes, and the prolonged deceleration becomes “severe” when FHR’s lowest point is less than 80 bpm. Within a window of 10 minutes, when accelerations or any type of decelerations occur, they are counted and assigned to their respective column.

In JSOG guidelines, the extracted features are rated with a corresponding level from the 5-level FHR patterns classification, also called the judgment table (Table 2), which indicates the health risk condition of the FHR. Instead of using the judgment table, we apply machine learning methods to our collected dataset for more accurate prediction results.

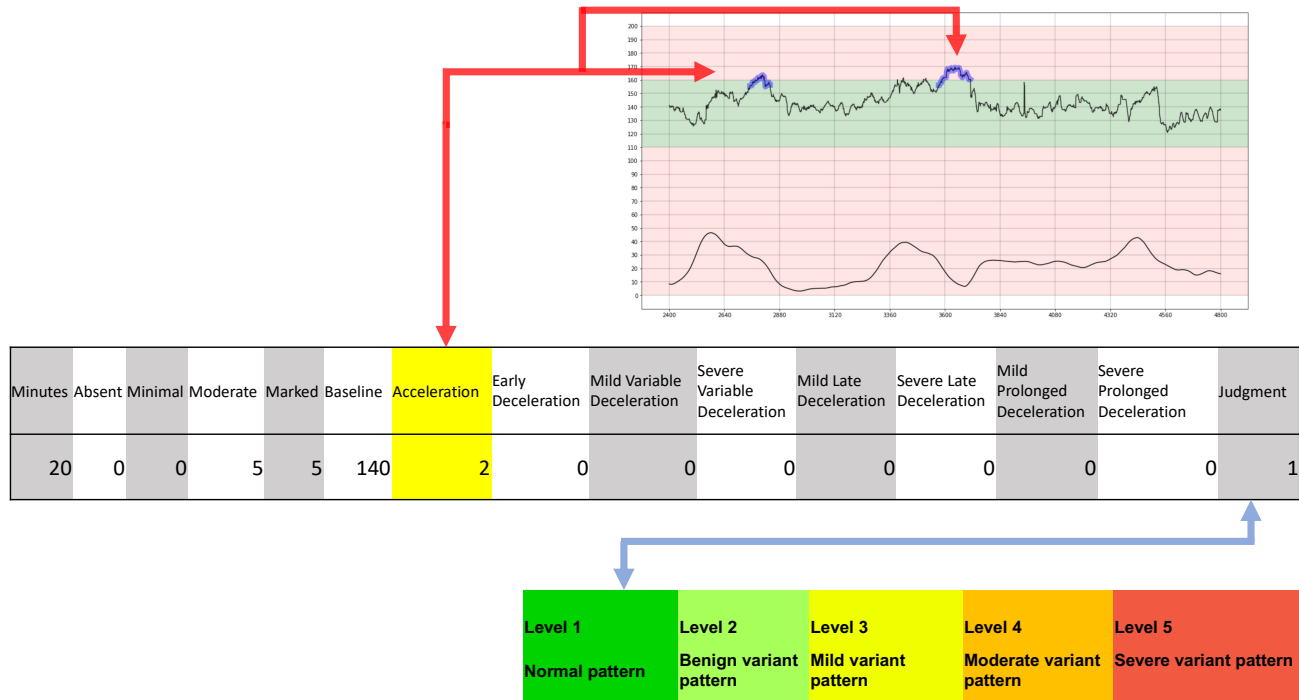
Table 2: JSOG five-tier fetal heart rate classification

FHR Pattern Classification Levels	
FHR pattern levels & Classification	
Level 1	Normal pattern
Level 2	Benign variant pattern
Level 3	Mild variant pattern
Level 4	Moderate variant pattern
Level 5	Severe variant pattern

For better data consistency, the time period chosen for the study is 30 minutes before 20 minutes of giving birth. The last 20 minutes are more prone to artifacts due to increased mother movement so they are excluded for better signal quality. The algorithm is applied to every 10 minutes of the chosen 30 minutes wherein every 10 minutes, features were extracted and judged independently resulting in 42 features; 14 features for every 10 minutes. Fetal weight and gestational age which are provided by the dataset are also included. The final form of extracted features for each record is shown in Table 3 and Figure 2 shows a summary of all the above-mentioned processes.

Table 3: Each row in the dataset should have the below features for every 10 minutes. Since the size of each CTG is limited to 30 minutes, the total features equal $(14*3) + 2 = 44$

Feature name		Description
Gestational age		At the time of delivery
Fetal Weight		At the time of delivery
Variability	Absent	No. per 10 min
	Minimal	No. per 10 min
	Moderate	No. per 10 min
	Marked	No. per 10 min
Baseline		Of 10 min
Acceleration		No. per 10 min
Early Deceleration		No. per 10 min
Mild Variable Deceleration		No. per 10 min
Severe Variable Deceleration		No. per 10 min
Mild Late Deceleration		No. per 10 min
Severe Late Deceleration		No. per 10 min
Mild Prolonged Deceleration		No. per 10 min
Severe Prolonged Deceleration		No. per 10 min
Judgment		1-5 rating for features



A) The above 10 minutes CTG signal has two accelerations. They are added to the table and judged as Level 1: Normal pattern based on JSOG guidelines.

Minutes	Absent	Minimal	Moderate	Marked	Baseline	Acceleration	Early Deceleration	Mild Variable Deceleration	Severe Variable Deceleration	Mild Late Deceleration	Severe Late Deceleration	Mild Prolonged Deceleration	Severe Prolonged Deceleration	Judgment
10	0	0	6	4	145	0	0	1	0	0	0	0	0	2
20	0	0	5	5	140	2	0	0	0	0	0	0	0	1
30	0	0	8	2	140	0	0	0	0	0	0	0	0	1
40	0	0	7	3	140	0	0	0	0	0	0	0	0	1
50	0	0	6	4	135	0	0	2	0	0	0	0	0	2
60	0	0	2	8	130	1	0	3	0	0	0	0	0	3

30 minutes before 20 minutes of giving birth.

Minutes	Absent	Minimal	Moderate	Marked	Baseline	Acceleration	Early Deceleration	Mild Variable Deceleration	Severe Variable Deceleration	Mild Late Deceleration	Severe Late Deceleration	Mild Prolonged Deceleration	Severe Prolonged Deceleration	Judgment
20	0	0	5	5	140	2	0	0	0	0	0	0	0	1
30	0	0	8	2	140	0	0	0	0	0	0	0	0	1
40	0	0	7	3	140	0	0	0	0	0	0	0	0	1

+
features from metadata

Gestational age	Birthweight
-----------------	-------------

B) Step A is applied to the highlighted rows where each row is 10 minutes of the chosen 30 minutes, then gestational age and birthweight of that record are added along the extracted features

Repeat same steps for all records in the dataset

Minutes Absent	Minimal	Moderate	Marked	Baseline	Acceleration	Deceleration	Deceleration	Deceleration	Deceleration	Deceleration	Deceleration	Deceleration	Deceleration	Judgment
20	0	0	5	5	140	2	0	0	0	0	0	0	0	1
30	0	0	8	2	140	0	0	0	0	0	0	0	0	1
40	0	0	7	3	140	0	0	0	0	0	0	0	0	1

- Original dataset has 38,073 records
- Only 3519 records were used for data balance.
- Three tests:
 1. Extracted features.
 2. Extracted features and Gestational age.
 3. Extracted features, Gestational age and Birthweight.

Three features tested on three target label pH, Apgar score 1 and 5

Target Label	No. positive samples	No. negative samples
pH < 7.1	299	3239
Apagr1 Apgar5 < 6	534	2985
pH Apgar1 Apgar5	775	2744

C) Step B is applied to the rest of records. The resulting dataset is prepared for three tests on three target labels.

Figure 2: A summary of the process for extracting features from CTG signals and assigning target labels.

2.5.4 Applying machine learning methods

Four ML methods were applied followed by their chosen hyperparameters: Support Vector Machine (SVM) (kernel = “linear”, C = 0.01, class weight = “balanced”), Random Forest (RF) (n-estimators = 100, max depth = 40), Decision Tree (DT) (max depth = 6, criterion = “gini”, splitter = “best”), and Artificial Neural Network (ANN) (activation function = “relu”, layers size = (45,45), max iteration = 1000, optimization = “adam”).

All of the aforementioned are applied using the Scikit-learn library for Python and the choices of the hyperparameter were based on the use of GridSearchCV in Scikit-learn which is a process of hyperparameter tuning for finding the optimal values for the model. Features in Table 3 are used as input data and pH (< 7.1), Apgar score 1 and 5 (<6) as binary target labels where 1 denotes high-risk birth and 0 as low-risk birth.

The classification accuracy was quantified and compared using ROC-AUC. All three processes were tested multiple times using 5-fold cross-validation. To solve class imbalance, random undersampling (for negatives) and SMOTE oversampling (for positives) were applied during training, which both are well-known methods used for handling data imbalance. Additionally, for better class balance, 3519 records in total were used in the experiment, making sure most of the positive classes in the original 38,073 records are included. This changes the ratio of positive to negative samples to ~11% for pH, ~18% for Apgar score 1 or 5, and ~28% for both (see Table 4).

Table 4: The number of positive and negative samples in training and testing data for the ML experiment. Positive classes refer to classes that satisfy the conditions of a label whereas negative classes refer to the ones that don't

Label	No. positive samples	No. negative samples
pH	299	3239
Apagr1 Apgar5	534	2985
pH Apgar1 Apgar5	775	2744

2.6 Experimental results

Both the processes of denoising CTG signals and the algorithm accuracy to predict the classic CTG features such as baseline, baseline variability, acceleration, and decelerations have been tested several times under the supervision of expert obstetricians which helped produce satisfying results. ML classification performance outcomes for 5-fold cross-validation are summarized in Table 5,6, and 7.

As can be seen, adding gestational age to the extracted features significantly improves the performance when the Apgar score is the target label. A slight improvement also occurs when birthweight is added as a feature. Almost none of the aforementioned happens when pH (Table 5) is the target label. It is also clear that AUC in Table 7 is less than the AUC in Table 6 because of the addition of pH as a target label. In all cases, using the Apgar score instead of pH produced much better classification performance. This could be related to the findings of [31] where they found a significant difference between Apgar score 1 in high-risk birth cases using JSOG guidelines classification compared to low-risk birth cases. On the other hand, Apgar score 1 did

not differ significantly between high and low-risk birth cases in the subjective classification methods (three-level risk classification). Another reason is the overall nature of pH analysis since pH, when obtained by obstetricians, does not discriminate respiratory acidemia from metabolic acidemia and the pH sampling fails with a ratio between 11% to 20% [32].

Table 5: ML using three feature sets on pH only as target label. In each test, extra features are being added to experiment with the performance variation

Features	method	5-fold CV AUC
Extracted features only (#features = 44)	SVM	.64 ± .06
	RF	.62 ± .05
	DT	.57 ± .05
	ANN	.63 ± .04
Extracted feature + Gestational age	SVM	.65 ± .02
	RF	.64 ± .03
	DT	.56 ± .04
	ANN	.63 ± .03
Extracted feature + Gestational age + birthweight	SVM	.65 ± .05
	RF	.68 ± .02
	DT	.61 ± .02
	ANN	.60 ± .03

Table 6: ML using three feature sets on both Apgar score 1 and 5 only as target labels. In each test, extra features are being added

Features	method	5-fold CV AUC
Extracted features only (#features = 44)	SVM	.71 ± .03
	RF	.73 ± .03
	DT	.68 ± .01
	ANN	.72 ± .02
Extracted feature + Gestational age	SVM	.88 ± .03
	RF	.89 ± .03
	DT	.82 ± .03
	ANN	.88 ± .03
Extracted feature + Gestational age + birthweight	SVM	.89 ± .02
	RF	.91 ± .02
	DT	.82 ± .02
	ANN	.88 ± .02

Table 7: ML using three feature sets on pH, Apgar score 1 and 5 as target labels. In each test, extra features are being added

Features	method	5-fold CV AUC
Extracted features only (#features = 44)	SVM	.65 ± .02
	RF	.67 ± .04
	DT	.64 ± .02
	ANN	.64 ± .02
Extracted feature + Gestational age	SVM	.78 ± .01
	RF	.80 ± .01
	DT	.76 ± .01
	ANN	.79 ± .02
Extracted feature + Gestational age + birthweight	SVM	.78 ± .01
	RF	.82 ± .01
	DT	.76 ± .02
	ANN	.77 ± .01

2.7 Discussion and Conclusion

So far, we successfully improved the accuracies for detecting high-risk births by introducing various methods such as denoising, constructing effective features, and choosing appropriate labels. Our method achieves 0.91 in prediction accuracy (see in Table 6, the row of RF with Extracted features + Gestational age + birthweight). However, birthweight is considered an unreliable feature in this study since it can only be accurately obtained after birth. Birthweight was used for the purpose of testing how much it would affect classification performance. The extracted features + Gestational age can be obtained before birth which is why they are usable features, thus AUC of 0.89 is the best achieved AUC in this experiment. We think there is still room to enhance it as described below.

While the dataset size provided flexible testing, the actual CTG recordings had plenty of noise to the point that may affect the signal quality and the ability to extract features from it. One of the causes that limit the performances of the extracted features is related to the length enforced on the signals. Some signals, when checked manually, were found to have significant information that could cause high-risk birth. However, the critical information time in the signal is not always consistent. For example, some serious decelerations were found within > 90 minutes before birth in some signals and < 30 minutes in some others. Forcing the length to 30 minutes before 20 minutes of giving birth may miss some serious information.

Still, using CTG guidelines in ML experiments may not be the best method to approach this problem. We are making the machine form its decision based on guidelines which their accuracy is disputed [30]. Rather than trying to find which guideline is the most accurate, we can input the CTG images directly to the CNN model for the classification using minimum features or without the need for interpretation. CNN takes in images and detects (learn) important features that could or couldn't be perceived by human vision.

Chapter 3

Predicting High-Risk Birth from Real Large-Scale Cardiotocographic Data Using Multi-Input Convolutional Neural Networks

3.1 Introduction

Neonatal death, hypoxic-ischemic encephalopathy (HIE), and respiratory distress are a few examples of risks fetuses may face the moment they are born. Such risks can be mitigated or prevented with knowledge about the fetus's condition inside the womb alongside speed and intensive care by the medical team. Obstetricians perform different tests to assess the safety of mother and fetus during birth such as umbilical-artery blood pH and Apgar score.

Apgar score is a rapid standardized assessment test that is performed on babies after one minute of birth to assess their health. The test measures five categories: appearance, pulse, grimace, activity, and respiratory. An Apgar score of 10 points means the newborn scored 2 in all categories in the Apgar scale, which is the highest. The same test procedures are reapplied after 5 minutes of birth resulting in the Apgar score 1 minute and Apgar score 5 minutes. The test may again be repeated, if needed, after 10, 15, or minutes of birth. Generally, an Apgar score of 7 or more is the

average and it means the newborn is healthy. A low Apgar score can be an indicator of the risk of fetal compromise. For example, an Apgar score 1 or 5 minutes of value between 0 to 6 is associated with a higher risk of cerebral palsy and epilepsy. The lower the score is, the higher the possibility can be [33]. This doesn't necessarily mean the infant will surely develop cerebral palsy, however, based on population studies, Apgar score after 5 and 10 minutes that are lower than 5 signifies a higher risk of cerebral palsy [34]. Also, an Apgar score 5 of less than 7 has shown an association with the occurrence of neurologic disability and low cognitive functions in early adulthood [35], increased the risk of neonatal respiratory distress and hypoxic-ischemic encephalopathy (HIE) [36], and increased the risk of Autistic Disorder [37]. Casey BM et al. and Li F et al. have shown that a very low Apgar score 5 (0-3) is correlated with neonatal mortality and the risk of neonatal death in term infants with (0 - 3) Apgar score 5 was eight times the risk in term infants with umbilical-artery blood pH values of 7.0 or less [38,39]. For identifying the risk of neonatal mortality, "Apgar score is a valid predictor of neonatal mortality. In fact, the Apgar score better predicted outcome than umbilical-artery pH of 7.0 or less." [40]. There is no considerable correlation between Apgar score at 1 and 5 minutes and umbilical cord pH in low-risk pregnancies, however, the correlation is more apparent between them in high-risk pregnancies [41].

Cardiotocography (CTG) is a monitoring tool used to look for vital information affecting the health and safety of mother and fetus during labor. CTG signals are interpreted by obstetricians using guidelines and there are multiple guidelines such as FIGO, ACOG, JSOG, etc. that tried to interpret the meaning of CTG signals and their consequence on mother and fetus. Nevertheless, these guidelines have disparity among them [13] and even obstetricians' opinions vary in classifying CTG [5]. According to the Each Baby Count report of 2019, more than 70% of stillbirths, neonatal death, and brain injuries cases would have had a different outcome with different care, and out of 420 delivery cases, 236 were linked to CTG false interpretation [42]. There has been a debate if visual-aided analysis is better than computer analysis or vice versa and some studies tried to compare them on a big scale [43,44]. Both studies saw no significant advantage of using one over the other. Berglund, S et al. showed that the care for two-thirds of newborns with a low Apgar score was substandard due to misinterpretation of cardiotocography (CTG) or poor response to abnormal CTG at the right time [45]. Abnormal CTG is usually related to fetal distress. There is an association between a low Apgar score and pathological CTG [46] and almost a fifth of pathological CTG may have a low Apgar score 5 [47].

Convolutional Neural Network (CNN) is an artificial intelligence deep neural network technique mostly applied to the image recognition field. CNN achieved the state of the art to

predict objects, faces, tumors, and more. CNN takes in images and detects (learns) important features that could or couldn't be perceived by human vision. Based on the aforementioned, we propose a CNN model that takes in CTG images and other major characteristics of fetuses available pre-labor to predict birth with low Apgar. If it is possible to predict a low Apgar score using minimum features or without the need to interpret CTG or have background knowledge about it, this would assist the medical team and help them prepare for the labor outcome in a quick and timely manner.

3.2 Related Works

Podda, M. et al. applied neural networks to predict the survival of preterm infants and achieved an AUC of 0.91. The study was applied to infants with <30 gestational weeks and needed features that were available within 5 minutes after birth [48]. In our previous work, we built an algorithm to extract some of the main features obstetricians look for in CTG such as accelerations, decelerations, baseline, etc. Using the algorithm, we trained a random forest ML model to classify Apgar score 1 or 5 of <6 and achieved an AUC of 0.89. Zhao, Zhidong et al. applied CNN to CTG to predict cord acidemia at birth, and Petrozziello, Alessio et al. are the first to do so [14,15]. One of their interesting conclusions is their model will not perform well if cord acidemia is absent and CTG may better be diagnosed with different models estimating different risks which might cause fetal compromise [15].

This work's intentions are to provide the medical team with a prognosis tool to predict infants with a high risk of a low Apgar score so they can be ready to take the required medical procedures before birth. We use a large dataset that is gathered under clinical conditions. So far, to the best of our knowledge, no study used deep learning methods to analyze Apgar score classification. No study, we are aware of, applied Convolutional Neural Network on CTG to classify fetuses with a low Apgar score.

3.3 Study Objective

This study aims to build a deep learning convolutional neural network model that acts as a prognosis tool to predict neonatal with possibly low Apgar score, which is one of the criteria that assess possible risks on neonatal, so they can be ready to take the required medical procedures.

This is done without using algorithms to extract CTG features such as baseline, accelerations, decelerations, etc.

The CNN model architecture used in this work is a multi-input model of two blocks, a block that accepts denoised CTG images and a block that accepts gestational age. The two blocks' outputs are concatenated to be trained together to predict a low Apgar score as binary classification. Figure 3 shows a detailed visualization of the built model and Table 9 describes the applied parameters. This work, however, doesn't claim to predict the implication, the meaning, or the condition of infants that causes Apgar score to become low. It gives the medical team a heads up or foresight of whether the newborn Apgar score will be high or low and the medical team prepares the necessary procedures for such concurrence. This is work intended to rely on real data only. The dataset was not augmented nor synthesized. The chosen metric to measure the model performance was the area under the curve of receiver operating characteristics (AUC-ROC) because it measures the model's ability to classify positive, represented as 1 in this work, from negative classes, represented as 0.

3.4 Data Set and Processing

The dataset used in this work is the same dataset from the previous study in Chapter 1. Similarly, the denoising process applied to the CTG is also applied here. All datasets share the same size used in our previous study and that is 3519. This size was chosen for better comparison with the previous work and to reduce class imbalance.

3.5 Target tasks for prediction

After the denoising process, the same dataset was replicated to test its performance on two targets of Apgar score:

1. Dataset where Apgar score 1 or $5 < 6$ as suggested by our medical researchers.
2. Dataset where Apgar score $5 < 7$ as several guidelines and obstetricians label it as low

Creating a dataset where Apgar score $5 < 4$ was disregarded because of the number of samples being too small. As seen in Table 8, in the dataset, the number of positive classes is limited compared to the negative classes. Almost all positive classes from the original 38,073 records are included. All negative classes were picked at random using Pandas DataFrame.sample, which is a

data analysis and manipulation tool in Python. Gestational age is provided by the dataset and it is the only non-image feature used. The mean of gestational age is 38 weeks, max is 42 and min is 22.

Table 8. The number of positive and negative samples in training and testing data for the CNN model.

Label	No. positive samples	No. negative samples
Apagr1 Apgar5 <6	534	2985
Apgar5 < 7	288	3231
Apgar5 < 4	57	3462

3.6 Prediction using deep CNN model

We want our CNN model to be able to accept mixed data where the inputs are not of the same type. In our case, the inputs are two: CTG images and gestational age. Thus, our model consists of two blocks, each of which processes CTG images or embedded gestational-age vectors. This work uses Keras, a deep learning API written in Python, and functional API in Keras library which helps manage multi-input models.

3.6.1 Image processing block

The first block is made solely for handling CTG images as inputs. The block is a pre-trained EfficientNet CNN architecture [49]. The reason to use EfficientNet is because of its high accuracy and is more computationally efficient than the best existing CNN/ConvNets. There are multiple architectures of EfficientNet such as EfficientNetB0 and EfficientNetB1 where the difference among each architecture is the number of layers in it. We applied transfer learning; the model is pre-trained on ‘imagenet’ dataset weights [50]. While the weights in imagenet are trained to classify objects, our model can benefit from the weights of lower layers, for it can detect low-level features such as edges and curves in images, rather than training the model from scratch. We can then edit or add on top layers with our own layers and fine-tune all the layers. On top of

EfficientNet, we added two dense layers, which are hidden layers connected to each neuron in the following hidden layer, with one dropout layer after each dense layer. Dropout is used to reduce the likelihood of model overfitting. The extra added dense layers serve as fine-tuning to the model and the number of layers was chosen based on experiments and following best practices in the field.

3.6.2 Meta-data processing block

The second block is made to take in gestational age. Gestational age is one of the most important fetal factors affecting the heart rate curve of CTG [51]. In fact, Tables 6 and 7 show that the prediction accuracies for the Apgar scores are clearly improved by adding gestational age as a feature. Our dataset contains both low and high-risk pregnancies with the majority being low risk. The values of gestational age are indifferent in both cases. In their study on low and high-risk pregnancies, Ahmadpour-Kacho, Mousa et al. [41] found gestational age was the same in low and high-risk mothers. The second block is a simple neural network that includes two dense layers.

3.6.3 Merging blocks

The output of both blocks is concatenated into one merged input of 64 dimensions vector (32 dim from the first block and 32 from the second). The last layer is a single unit that uses a sigmoid activation function, which is suitable for our binary classification. The values of gestational age are scaled by preprocessing, and CTG images are also scaled to [0,1]. The dataset is imbalanced, and this affects the performance. One of the methods to solve the class imbalance is to use a class-weight method which provides each class with a weight. The class with the fewer sample will have a heavier weight so it can be given more emphasis when training the model. The chosen optimizer and loss function were Adam and binary cross-entropy since Adam is the most recommended optimizer to be used in CNN and the latter is made for binary classification. Each target label was tested on CTG for 60 minutes and again on CTG for the first 30 minutes from the 60 minutes to test if CTG length has an impact on the performance. The model performance was tested using 5-fold stratified cross-validation (CV). CV makes multiple model predictions on the dataset. This provides us with more confidence in the performance of the model. Stratified CV is a CV variant that creates folds with a similar percentage of samples for every class. It is useful for our imbalanced dataset because the number of negative samples is more than the number of positive samples.

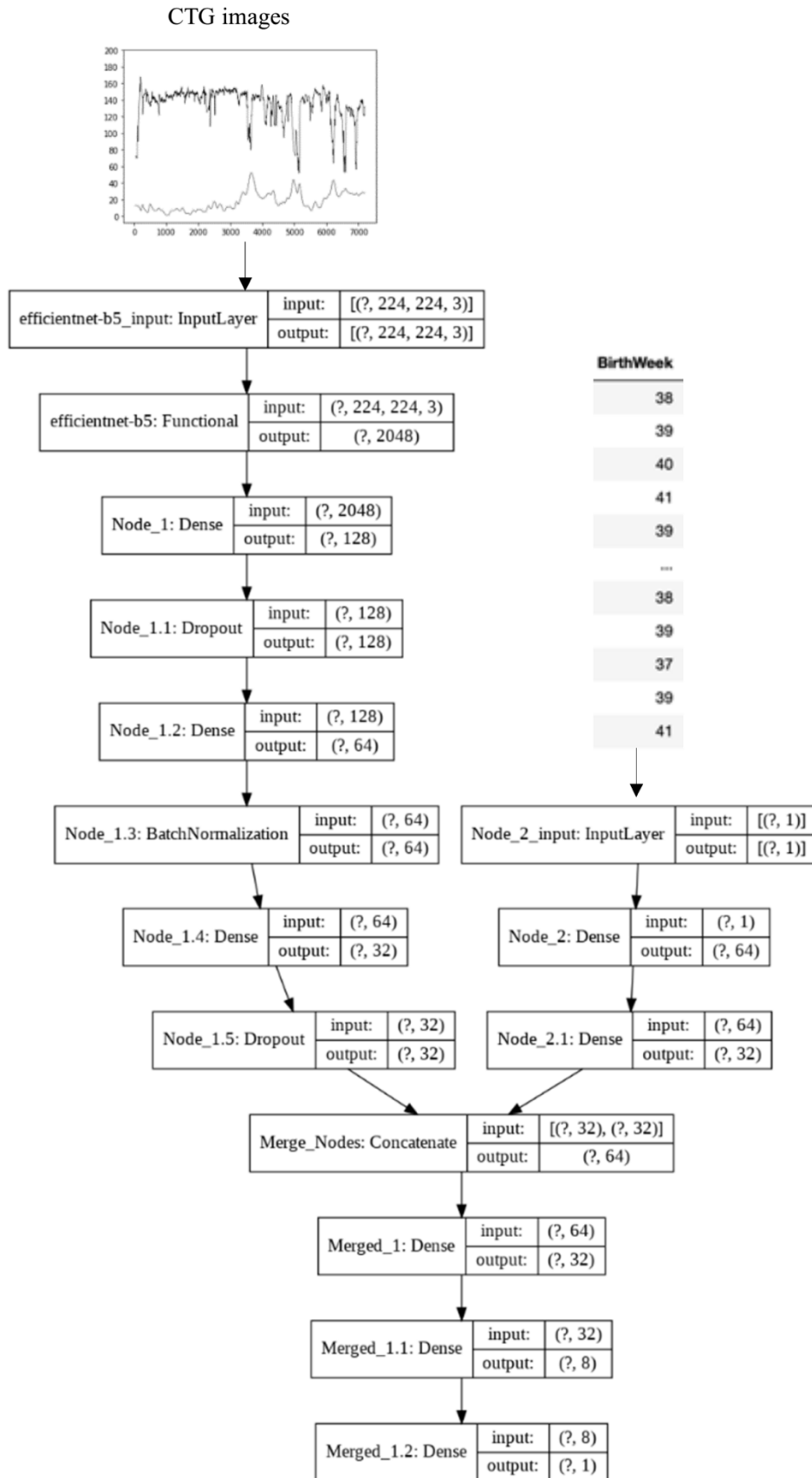


Figure 3: The architecture of the proposed model.

Table 9: The chosen parameters for the model.

Layer Name	Parameters
efficientnet-b5: Model	Max pooling
Node 1: Dense	Relu Activation function, L1 regularization
Node 1.1: Dropout	0.3
Node 1.2: Dense	Relu Activation function
Node 1.3: Batch Normalization	Relu Activation, L1 regularization
Node 1.4: Dense	Relu Activation function
Node 1.5: Dropout	0.3
Node 2: Dense	Relu Activation function
Node 2.1: Dense	Relu Activation function
Concatenate	Output of Both nodes
Merged 1: Dense	Relu Activation function
Merged 1.1: Dense	Relu Activation function
Merged 1.2: Dense	Sigmoid Activation function

3.7 Results, ML vs DL comparison, and discussion

Our CNN distinguishes CTG patterns with a low Apgar score from patterns with a high Apgar score and achieves a higher AUC than the extracted features+ML as described in our previous study. Moreover, it is possible that the model learns more features from CTG with longer length as shown in Table 10. AUC was higher in both labels with CTG with a length of up to 60 minutes.

Table 10: AUC after training the model for different targets.

Label	5-fold CV AUC for 30 minutes CTG	5-fold CV AUC for up to 60* minutes CTG
Apagr1 Apgar5 <6	0.949	0.955
Apgar5 < 7	0.953	0.958

The model was trained with a batch size of 8 and 10 epochs of training. More epochs caused the model to start overfitting. For the first block, all versions of EfficientNet were tested. The performance stopped improving after EfficientNetB5, so it is the version we based our results on.

Our ML model in chapter 2 used a total of 44 features, most of which required an algorithm to be extracted. This required human interference throughout the whole process. Additionally, building an algorithm to extract CTG features can be prone to inaccuracies or may miss critical information, and choosing which CTG guidelines to base the algorithm upon makes it an intricate task. The current CNN model relied only on the pre-existing data without the need for a feature extraction algorithm. It means that it can bypass one layer of human interference. Relatively, it required less steps than the ML approach and was able to learn distinctive features from CTG images alone paired with gestational age outperforming our previous RF model with AUC 0.89. The training process depended on real data only; no data were augmented or synthesized since we believe that the number of samples is sufficient in our dataset.

Gestational age is an important factor in both the ML and the CNN experiments because of its relationship with the Apgar score [41,51]. The diversity of gestational age in our dataset is one of the reasons our model achieved that performance in predicting a low Apgar score. The drop in performance could be connected to Cnattingius, Sven & Norman et al. findings where Apgar score prediction performance was better when used in infants with ≤ 31 gestational weeks than in full-term infants [52].

3.8 Conclusion

The proposed CNN model predicted fetuses with low Apgar score using pre-existing features. The ML approach in chapter 2 used an algorithm based on JSOG guidelines to extract important features from CTG signals for interpretation. Different typical ML methods are applied to test the performance of the algorithm in detecting high-risk birth on large data gathered under clinical conditions.

CTG signals were denoised for both the CNN and ML methods. Unlike the ML methods, the CNN method achieved higher AUC even though its features were fewer and available. It learned to distinguish features directly from the CTG images whereas ML relied on the features extracted from the algorithm.

The model acts as a pre-labor tool to guide the medical team during labor. On the other hand, we think one model is not enough to determine fetal safety. Relying on the outcome of only one test such as Apgar score or pH could be misleading [53]. In future work, to progress this work in the right direction, it is desirable to create multiple models and tools for different tests each of which serves a specific purpose. For instance, a combination of models that predicts Apgar score and umbilical cord pH can be used as a better indicator to detect complications in the newborn [20,40]. Apgar score works well as a short-term prognosis tool to assess expected risks on neonatal specifically among preterm infants [38,54,55]. There is also a need for a better method to solve data imbalance in the dataset. Finally, Our CNN is pre-trained and uses Efficientnet as an architecture which makes it relatively lightweight, still, CNN models, in general, require extensive time to train and they are computationally demanding. However, after a model is trained, it is quick in analyzing new samples, making them practical in a real-life situation.

Chapter 4

Analyzing the Relationship between Abnormality in Fetal Heart Rate and low pH test using Anomaly Detection Generative Adversarial Networks model

4.1 Introduction

Obstetricians perform umbilical arterial cord blood pH testing immediately after birth and if fetal undergo oxygen deficiency they get a low score on the arterial cord hydronium ions concentration (pH) test. Generally, an umbilical arterial cord blood pH test score of 7.1 or more is considered non-acidosis. A low umbilical arterial cord blood pH test score is acidosis and is considered an indicator of risk to fetal health safety. Monitoring FHR is important to identify changes to oxygen supplied to fetal since any serious change to it may cause abnormal FHR and it is an indication of probable health risk to both the mother and infants [4]. Abnormality in FHR can also be a sign of the prolonged second stage of labor or the presence of meconium-stained amniotic fluid both of which are major causes for the infant to develop Hypoxic Ischemic Encephalopathy (HIE) [56]

Through various research, it has been shown that artificial intelligence has the capability of experts in classifying abnormal health compromises such as cancer, and retinal diseases [57,58]. The idea of classification, also known as anomaly detection, is to identify unusual or rare data from other common ones. It is usually difficult to find abnormal data. In fact, this is one of the

main challenges facing ML and DL which cause datasets to be imbalanced or limited on minority class such as fewer fraudulent credit card transactions compared to real ones or phishing emails compared to real emails, and so on.

Training a ML or DL model to detect minority class is difficult if the difference between the number of minorities versus majority classes is huge as this will overclassify the majority class and misclassify the minority class, thus resulting in incorrect performance [59]. However, in medical diagnosis, minority class scarcity is far more apparent. For example, classifying malignant from a benign tumor, abnormal ECG signals from normal signals, etc. Many attempts tried to solve data/class imbalance such as giving more weight to minority class during training, oversampling, downsampling, weight class, etc. [60,61,62]. While every method did well in solving parts of class imbalance, in computer vision, synthetic images generated by Generative Adversarial Networks (GAN) [63] have been successful in mitigating high complexity class imbalance problems [64]. GAN is a subset of the DL framework for generating synthetic data instances similar to the training data using deep learning approaches. The vanilla GAN architecture consists of two deep networks, a discriminator (D) and a generator (G). The two networks are trained together where the generator generates a batch of data and the discriminator, which is trained to classify real from fake data, determines if the batch is real or fake. Since then, several versions of GAN have emerged where they fix some of the issues in the original GAN and concentrate on more specific problems. GAN achieved state of the art on different anomaly detection tasks such as manufacturing defect detection [65], breast cancer detection [66], Covid19 detection [67], and more.

In this work, we choose to detect anomalies in CTG images by using the ANOGAN approach [23]. The idea of ANOGAN is to map new images to the latent space of the trained model to find the most comparable image and based on the similarity between the two images, the new images receive an anomaly score. In our case, a high anomaly score means the new image has an abnormal CTG signal and it is more likely to be of a fetus with a pH < 7.1. The anomaly score is computed using the weighted sum of residual loss and discrimination loss [23]. ANOGAN is built on DCGAN architecture [25]. However, we're also going to apply it to two other GAN architectures; WGAN [26] and WGANP [27]. If it is possible to predict abnormal CTG with a low pH score with no or little knowledge about CTG interpretation and minimum obstetricians' intervention, it would help the medical team to prevent serious complications for the mother and fetus.

4.2 Related Works

One of the major parts of this work is to train and generate CTG images using multiple GAN models on a dataset of CTG images. To the best of our knowledge, no study used GAN to train and generate CTG images. Puspitasari et al. used Time Series GAN (TSGAN) for data augmentation to solve data imbalance in the CTU-UHB CTG database [17] to improve classification when using DL models [68]. As mentioned earlier, the size of our dataset benefits us in avoiding the need for data augmentation especially that data augmentation in time series signals such as CTG is yet to be a standard procedure [69].

GAN has been successful in generating Electrocardiogram (ECG) images [70,71,72]. Hossain et al proposed a GAN model that generates and detects arrhythmia in Electrocardiogram (ECG) and achieved a state-of-the-art performance [73]. However, to the best of our knowledge, no study we know used GAN image generation to study or analyze anomalies in CTG images. Anomaly detection using GAN is the process of modeling the normal pattern using the adversarial training process where anomalies are identified and given an anomaly score [23]. Anomaly detection using GAN has grown exponentially. Geiger et al. made TADGAN for detecting anomalies in time series using GAN and achieved a higher F1 score compared to 8 different anomaly detection methods [74]. However, it concentrates solely on time series data whereas our focus is on anomaly detection in CTG images. GANomaly [75] and ANOGAN [23] are two of the most well-known studies in this field. Our work is inspired by the ANOGAN method where it treats the trained discriminator as a feature extractor, unlike traditional GAN's discriminator which acts as a classifier and this is key to our experiment because we don't use labeled data in training as we want the discriminator to learn the feature representation of training set and use backpropagation to map new images to the latent space and find images with similar features. However, the original ANOGAN was implemented on DCGAN [25] architecture only so, additionally, we are also implementing the experiment on two more GAN models: WGAN [26] and WGANGP [27].

So far, to the best of our knowledge, this is the first study that applied an anomaly detection approach to CTG using generative models. No study, we are aware of, applied GAN or ANOGAN to generate CTG images or detect abnormality in FHR patterns. Another key factor in our study is that we use three generative model evaluation metrics to evaluate the quality of generated images, we use a big and extensive CTG dataset collected under clinical trials, and train our models only on real data; no data synthesize or augmentation was used.

4.3 Study Objective

The aim of this study is to use ANOGAN to analyze FHR signals and their relationship with normal and low pH score which could cause complications for fetuses. pH score is a vital factor to designate the well-being of newborns and a low pH test score of < 7.1 is considered serious. Without using any guidelines for CTG interpretation, we use ANOGAN to train on a dataset containing CTG images of the last 120 minutes before birth for patients with non-acidosis pH (7.1 or more), we also refer to them as normal or negative CTG images. Each record is divided into four parts of 30 minutes.

After training, the generative model goes through an anomaly identification process tested on a dataset containing %50 acidosis CTG images and %50 non-acidosis CTG images. The expected result is that the generator performs well when generating common FHR patterns but fails or performs badly when generating uncommon patterns since it was trained only on normal CTG images. The generated images are compared with real images and receive an anomaly score based on FHR similarity. A small anomaly score implies that the difference between real and generated images is small and a big anomaly score implies a big difference between real and generated images. Finally, we analyze the received anomaly score for CTG images in the test set and mark the results of our observation.

To make sure that the quality of the generated images is reliable, they are evaluated using three different metrics: Fréchet Inception Distance (FID) [28], Probability Density Function (PDF), Precision, and Recall (PR). This study is performed on real data only. Any form of data synthesis or augmentation was avoided as we rely on the size of our big dataset. Further, we're using only FHR since UC is subject to maternal ECG interference. This study by no means aims to provide an interpretation of the medical implication of the cause, the meaning, or the condition of newborns that receive low pH test score. This work, however, aims to study the relationship between abnormal FHR patterns and low pH and identify FHR patterns that are likely to have low pH, and prepare for any needed procedure.

4.4 The GAN Models

GAN uses two models a generator and a discriminator. The generator is a convolutional neural network whereas the discriminator is a backward convolutional neural network. The generator

network is a feed-forward neural network that continuously learns over time to create reliable fake data, such as fake cat images. It uses discriminator feedback to gradually improve its output until the discriminator can't ideally distinguish its output from real data. The discriminator has a classification identification role where it learns a decision boundary for predicting which class a data point belongs to. Both models work together in such a way that the generator tries to outsmart the discriminator and the discriminator tries to not be tricked by the generator's fake images. Whenever the discriminator correctly detects the synthesized work, through the use of backpropagation, it informs the generator on how to change its output so that it can be more realistic in the next epochs, this step is called weight updates. Figure 4 is a summarized image of the GAN system.

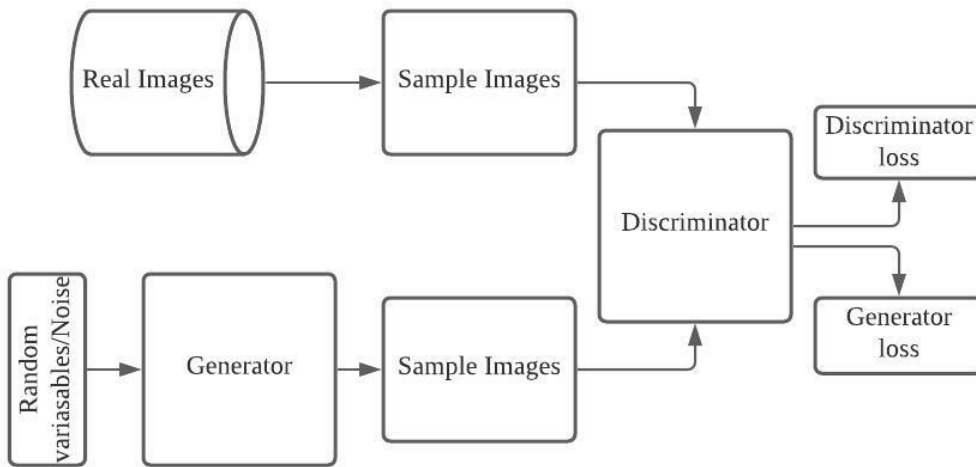


Figure 4: GAN model.

Since the emergence of the first GAN in 2014, There have been many forms of GANs. In this research, three GANs were picked: DCGAN, WGAN, and WGANGP. Below we explain the reason for choosing the aforementioned GANs:

4.4.1 Deep Convolutional Generative Adversarial Networks (DCGAN)

DCGAN is one of the most popular GAN variations. Fundamentally, it focuses on employing the deep convolutional approach. It is similar to GAN, however, in regular GAN, the aggregation levels in the generator and discriminator are replaced by transposed convolution to upsample images, also called full convolution or fractional-strided convolutions, (in the case of a generator)

and strided convolutions (in the case of a discriminator). When training the discriminator (D), the generator (G) tries to maximize the error probability of the discriminator (D). The discriminator's role is to check if an image is a real training image or a false image from a generator, while the discriminator tries to improve its classification capability and distinguish real and fake images correctly. DCGAN utilizes Conv nets which look for areas of correlation within an image. It also includes batch norm, a technique that improves the training stability of the neural network, which allows higher learning rates and lower chances of overfitting. This makes it appropriate for image datasets in addition to its system being easy to fully comprehend. DCGAN of Radford et al. has been improved in many aspects [76]. In this study, we're using the latest version of DCGAN.

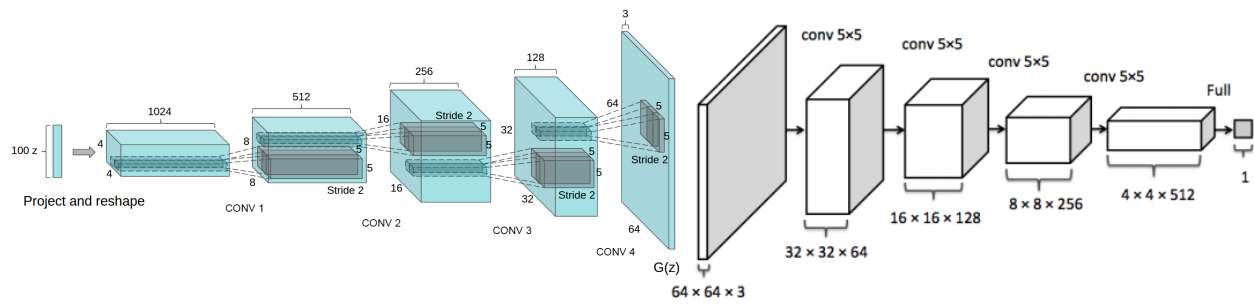


Figure 5: DCGAN architecture. Figures source [25][77]

4.4.2 Wasserstein Generative Adversarial Networks (WGAN)

The main idea of using GAN is to minimize the distance between real data distributions and generated data distribution. The original GAN and multiple GAN variations use Jensen Shannon (JS) and Kullback-Leibler (KL) divergence for measuring the distance between real data distributions and generated data distribution. JS and KL divergence have drawbacks such as vanishing gradient and mode collapse. In GAN, a vanishing gradient is when the loss function quickly falls to zero because the generator doesn't receive enough information from the discriminator. Mode collapse is when the generator generates only limited variations of the training samples in the dataset. Solving mode collapse is crucial, especially in our study, because the number of some FHR patterns is limited. WGAN replaces the JS and KL divergence with Wasserstein Distance, hence the name Wasserstein GAN. WGAN and WGAN-GP have shown success in eliminating the occurrence of vanishing gradient and mode collapse [26] [78]. Unlike

other variations of GAN, there is a relationship in WGAN between the convergence of discriminator (called critic in WGAN) loss function and generator loss function and the quality of generated images. A convergence in loss or close to convergence indicates a higher image quality. According to the WGAN paper “this is the first time in GAN literature that such a property is shown, where the loss of the GAN shows properties of convergence. This property is extremely useful when doing research in adversarial networks as one does not need to stare at the generated samples to figure out failure modes and to gain information on which models are doing better over others” [26]. WGAN architecture is similar to DCGAN in Figure 5 except 1. the discriminator is named critic and more iterations are being added to train it more than the generator. 2. the optimizer is modified from Adam to Root Mean Squared Propagation (RMSProp). 3. the loss is changed to Wasserstein loss. 4. weight clipping is enforced which clip every weight parameter in the network to between $[-0.01, 0.01]$ to force the critic to be within a certain range.

4.4.3 Wasserstein Generative Adversarial Networks Gradient Penalty (WGANGP)

While WGAN has great training stability and eliminates mode collapse, weight clipping in the critic needs careful tuning for it may cause the gradient of the model to vanish or explode [27]. Exploding gradient, in contrast to vanishing gradient, is when the generator cannot be updated because of receiving too much information from the discriminator and WGAN-GP has shown success in eliminating the occurrence of vanishing gradient and exploding gradient in addition to the original WGAN ability to element mode collapse [27]. WGAN-GP replaces the weight clipping parameter in WGAN with gradient penalty to help enforce Lipschitz constraints which makes the discriminator perform much better [26] and avoid weight clipping problems.

4.5 Method

This study can be summarized into the following steps: 1- Data Set and Processing which discuss the changes applied to the data set. 2- Define the GAN models for training and the applied hyperparameters. 3- Evaluate the trained models. 4- Compare generated images and anomaly score.

4.5.1 Data Set and Processing

The dataset is the same dataset used in Chapters 2 and 3. However, in this experiment, the last 120 minutes of a CTG signal is divided into 4 parts where each part is a CTG image with a length of 30 minutes. Part 1 is the first 30 minutes and part 4 is the last 30 minutes before giving birth. Any

signal that is shorter than 120 minutes is dropped. Additionally, this study focuses only on low-risk pregnancies. Records with pre-term and post-term pregnancy are removed. This is why only gestational ages of <42 and >35 weeks are included. moreover, any pregnancy that resulted in a planned C-section was not included and any record with a null blood pH test is removed. After applying all of the aforementioned processes, the total number of positive records is 142 positive(pH<7.1) records which were added to the test set. We added an equal number of negative(pH>=7.1) records to the test set and the total became 284 where 50% are positive and 50% are negative. The negative classes in the test set were chosen randomly using *Pandas DataFrame.sample*, which is a data analysis and manipulation tool in Python. The total number of CTG images in the test set after dividing each record into four parts is 1136 images. Because our model is trained only on negative records, all positive records are used in the test set. The training set consists of 13723 negative records. The total number of CTG images in the training set after the divide is 54892.

4.5.2 Define the GAN models for training and the applied hyperparameters.

After denoising CTG and filtering the dataset, the same dataset was trained on three different models: DCGAN, WGAN, and WGAN-GP. We want the three GAN models to receive CTG images and learn or estimate the data distribution of the training set to be able to generate fake samples from that learned distribution. The model would be representing normal non-acidosis CTG images. The test set is a collection of unseen images from both normal and abnormal CTG images. Data imbalance in the training set is not a problem since it contains only negative classes. The three models were built and trained using Pytorch, a deep learning API written in Python. The three model share the same network which is summarized in Table 11. The common hyperparameters between the three models are: `latent_size = 100`, `image_size = 64`, `batch_size = 64` and `channel = 3`. The distinctive hyperparameters for DCGAN are: `learning_rate = 1e-5`, Adam for optimization, and a sigmoid activation function as the last layer.

The distinctive hyperparameters for WGAN are: `critic_iterations = 5`, a hyperparameter that adds more iterations when training the critic to train it more than the generator, `weight_clip = 0.01`, to force the critic to be within a certain range, `learning_rate = 5e-5` and the optimizer was changed to RMSProp. As for WGAN-GP, we added the gradient penalty function from the original work and applied it to every iteration in the training to satisfy Lipschitz constraints and prevent vanishing or exploding gradients. The `critic_iterations = 5`, `learning_rate = 1e-4` and the optimizer was switched back to Adam. All of the hyperparameters applied to WGAN and WGAN-GP were taken directly from the original work. The image size of 64 was chosen for computation speed

purposes while keeping the CTG images as clear as possible. The latent space size, batch size, and the number of epochs were chosen based on experiments and following best practices in the field.

Table 11: A summary of DCGAN, WGAN, and WGANGP networks. The three models share similar networks except for the last layer in DCGAN which is a sigmoid activation function

Layer Name	Parameters
Generator	
ConvTranspose2d	(100, 512, kernel_size=(4, 4), stride=(1, 1), bias=False)
BatchNorm2d	(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
ReLU	Relu Activation function
ConvTranspose2d	(512, 256, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
ReLU	Relu Activation function
ConvTranspose2d	(256, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
ReLU	Relu Activation function
ConvTranspose2d	(128, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
ReLU	Relu Activation function
ConvTranspose2d	(64, 3, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
Tanh()	Tanh Activation function
Discriminator	
Conv2d	(3, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
LeakyReLU	(negative_slope=0.2, inplace=True)
Conv2d	(64, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
LeakyReLU	(negative_slope=0.2, inplace=True)
Conv2d	(128, 256, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
LeakyReLU	(negative_slope=0.2, inplace=True)
Conv2d	(256, 512, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1), bias=False)
BatchNorm2d	(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
LeakyReLU	(negative_slope=0.2, inplace=True)
D_final	
Conv2d	(512, 3, kernel_size=(4, 4), stride=(2, 2), bias=False)
Sigmoid	Sigmoid Activation function (For DCGAN only)

4.5.3 Generated Image Evaluation

In general, evaluating the quality of generated images is still an evolving field of study and a perfect evaluation metric is yet to exist. In fact, it is one of the main drawbacks of GAN since there is no certain tool to compare two GAN models [79]. Comparing the performance of different GAN models is difficult since GAN lacks an objective evaluation function [76] and quoting Borji's 'Pros and Cons of GAN Evaluation Measure: New Developments' "GAN evaluation is not a settled issue and there is still room for improvement" [80]. In essence, the evaluation measures rely on the quality of generated images of GAN models. Some Researchers studied the capability of human perception as an evaluation metric [81]. However, while the use of visual examination is intuitive and understandable in early training iterations and epochs, it is not practical to assess all images, especially in huge datasets. Also, the image analyst needs to be an expert in the type of analyzed images and need to be unbiased. Below we applied three metrics for evaluating generated images:

4.5.3.1 Fréchet Inception Distance (FID)

Proposed by Heusel et al. FID is a metric for evaluating images generated by GAN [28]. So far, it is the most popular metric for evaluating generated images [80]. It calculates the feature distance between the generated and real images. The resulting score implies the similarity of the compared images distribution and the lower the FID score the better the quality of the generated images is. It is an improved version of the Inception Score (IS) [76] as it solves IS drawbacks. Both techniques depend on the use of a pre-existing classifier inception v3 model (InceptionNet) trained on The ImageNet dataset [50]. Using the pool_3 layer of the Inception model to extract 2048-dimensional activations for both real and generated images, the function (calculate_activation_statistics) models the data distribution for the features using a multivariate Gaussian distribution. "The difference of two Gaussians is measured by the Fréchet distance also known as Wasserstein-2 distance" [28]. For this study, the best achieved FID score was 47 for DCGAN, 35 for WGAN, and 43 for WGANGP.

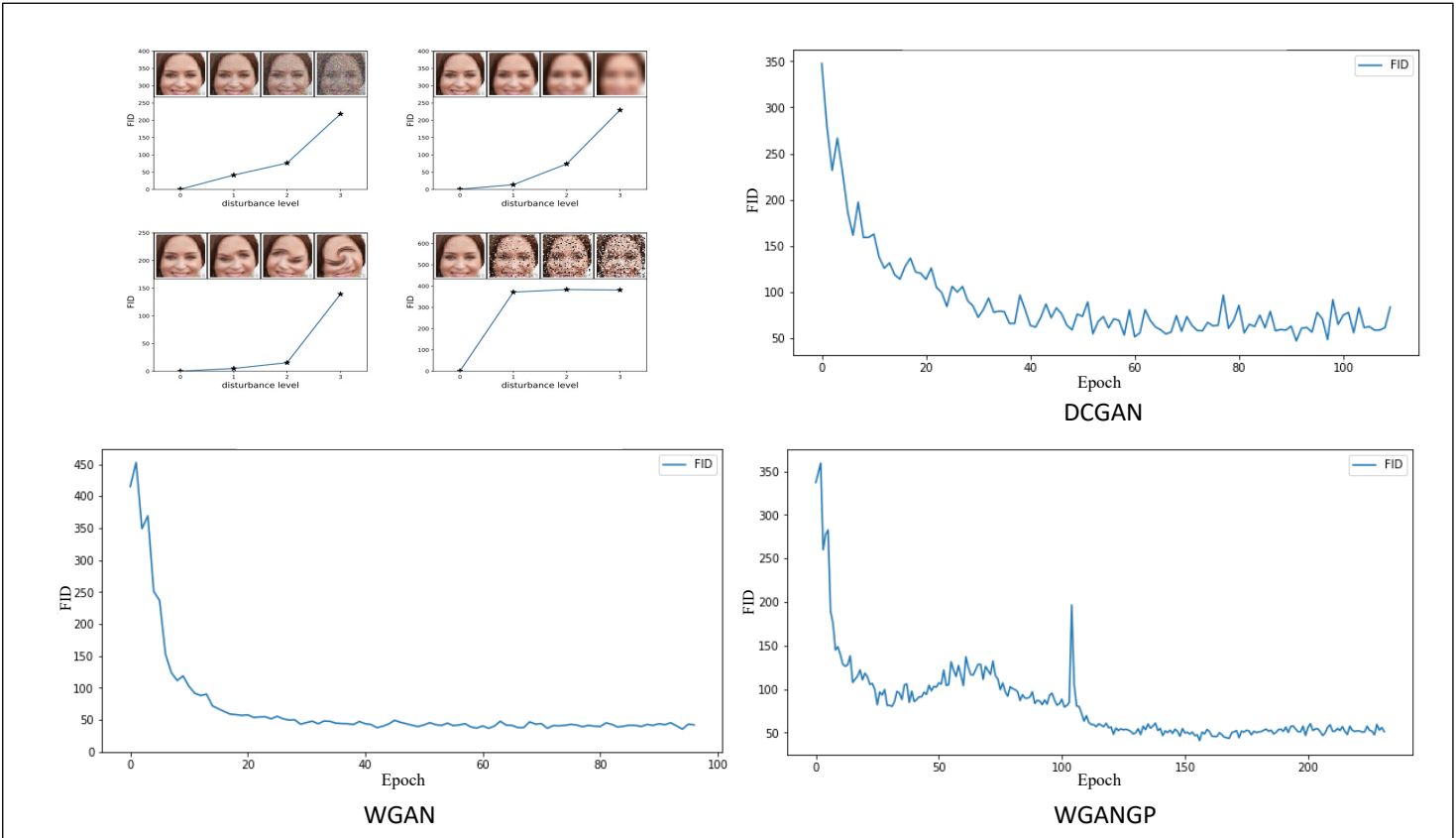


Figure 6: A plot of the achieved FID through all the training process for DCGAN, WGAN and WGANGP. The upper-left is a demonstration of how FID is affected when noises or distortions were applied on high quality image [28].

4.5.3.2 Probability Density Function (PDF)

By using PDF from probability theory, we can get an indication of the likelihood that the generated images are using the same or similar features of real images features. Generally, based on the learned features, the generator should be able to map features from latent space to the targeted data distribution. This is can be seen when plotting the probability density of GAN’s generated images compared to the real probability density of the data distribution. During training, the generator is expected to learn to create a distribution similar to the real probability density and this only happens after a number of epochs until the two distributions become as similar as possible. In earlier attempts, the generator starts with random noise and through continuous epochs, the generated distribution improves and becomes more similar to the real data distribution. The upper-

left images in Figure 7 are for a plot of probability density after 60 epochs. The generated distribution has taken the shape of the real distribution but the similarity gets better after more training. After training for about 120 epochs, the generated distribution is more similar to the real one. The X axis in Figure 7 denote the extracted features from the real test set (blue) and the extracted features from generated images of the test set (green). The features were extracted using the trained discriminator and then reshaped to vectors. The Y axis, according to PDF, is the probability density of that feature. Basically, if it is high, it means the number in the X axis was repeated many times, hence the spikes in the plot.

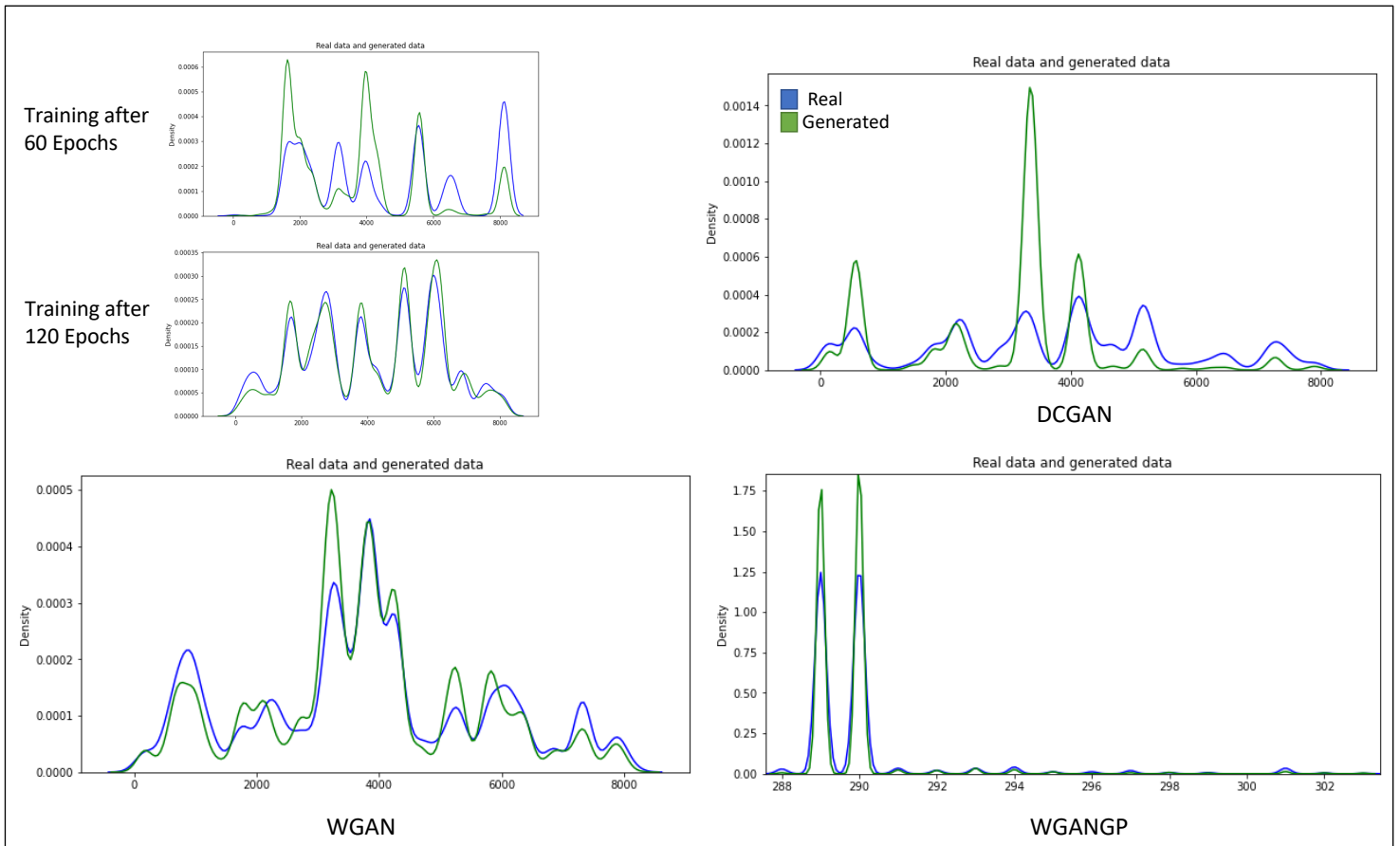
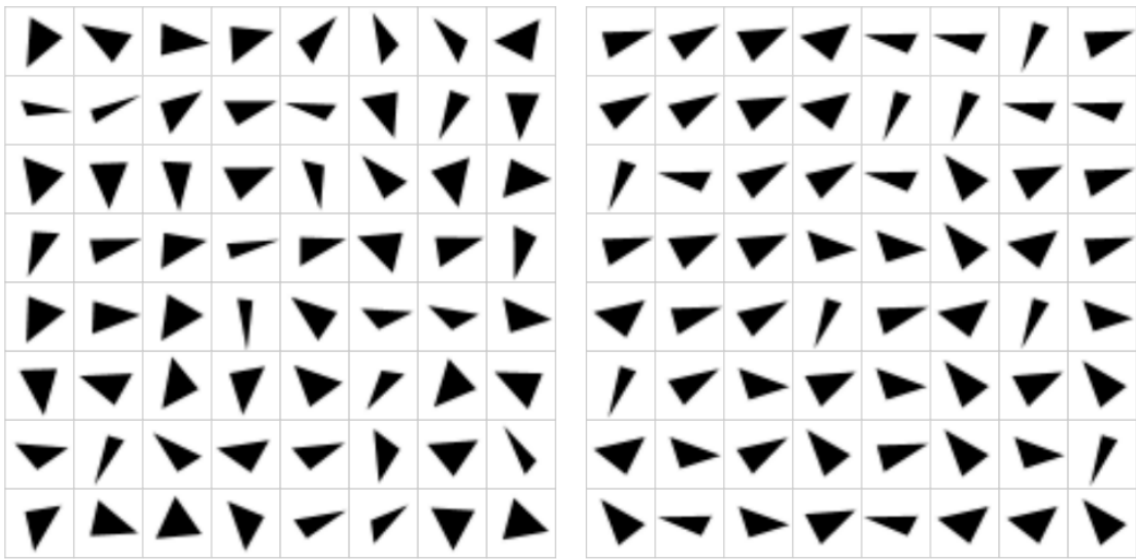


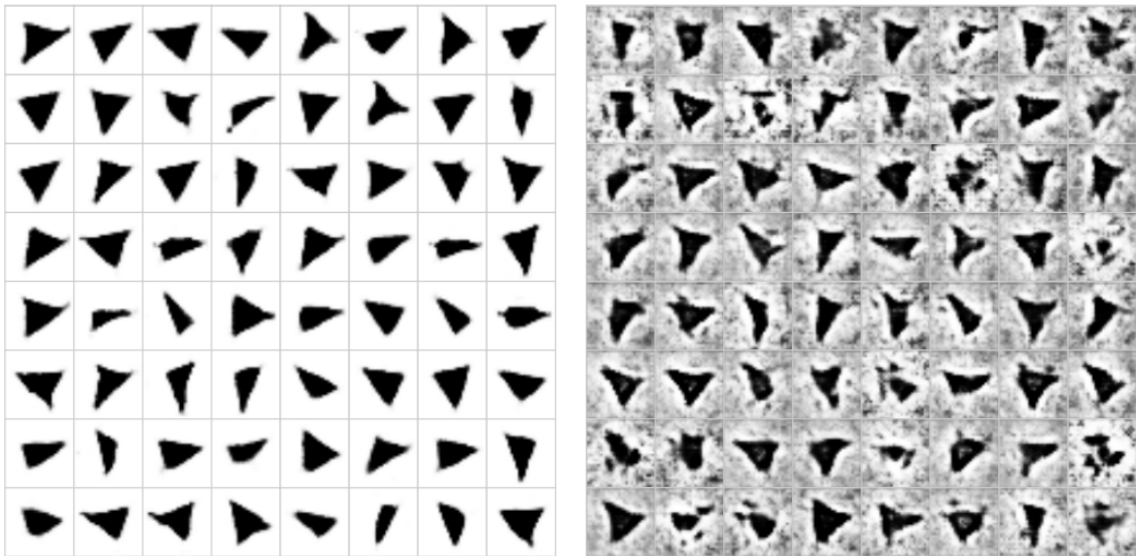
Figure 7: A plot of the generated distribution compared to the real distribution of DCGAN, WGAN and WGAN-GP.

4.5.3.3 Precision and Recall (PR)

PR is a key evaluation metric in ML and DL. However, applying it in GAN is still an open area of research. As seen in Figure 8, we can have an intuition about the quality of generated images. For example, we can assume high precision when the generated image is very similar to the real image. The recall can be evaluated by the variety of the generated images. If the generator is producing many similar images, this is an indication of low recall.



(a) High precision, high recall (b) High precision, low recall



(c) Low precision, high recall (d) Low precision, low recall

Figure 8: Precision and recall in GAN models. Figure source [79]

There are multiple approaches that tried to calculate PR for generated images [82,83,84]. We use [84] code for calculating precision and recall and Table 12 shows the achieved precision and recall for DCGAN, WGAN, and WGANGP models. While all models have a high recall, we think a major factor affecting the precision score is the size of the CTG images (3x64x64) inputted into the models. The small size causes CTG image quality to drop and this was done to reduce the immense computation time needed to complete the training process since increasing image size will intricate this issue even more.

Table 12: The achieved precision and recall for DCGAN, WGAN, and WGANGP.

Model	Precision	Recall
DCGAN	0.81	0.91
WGAN	0.83	0.99
WGANGP	0.86	0.96

4.5.4 Compare the generated images and anomaly scores

We use the trained model to generate the images in the test set and assign each real image in the test set with an anomaly score based on the difference between the real and the generated image. The anomaly score is calculated using the weighted sum of residual loss and discrimination loss where residual loss measures the visual dissimilarity between real and generated images and discrimination loss measures the feature dissimilarity between real and generated images. The formula is described in detail in the ANOGAN paper [23]. If the two images are very similar, the real image gets a low anomaly score. A higher Anomaly score indicates a bigger difference between the two images.

Figure 9 shows an example of how the model performs poorly in generating less occurring or infrequent CTG patterns and perform generally well in generating images with frequently recurring patterns. The pattern in image A is bradycardia, which is a condition that occurs when the FHR baseline is below the average. This is not a common pattern, especially in normal CTG records since bradycardia is linked to an increased risk of acidemia [85]. The generated image fails to generate the real image and we see that the difference between the two images is big, thus the anomaly score is 191. The same applies to image B where the real image contains high variability in FHR. The real and generated images show no similarity so the anomaly score is relatively high

which is equal to 299. The level of anomaly score changes if the experiment is repeated, however, the difference gap between all images still persists; the image will receive a relatively high anomaly score compared to the other CTG images in the same set. Images C and D show FHR patterns which are common in the training set. The difference between the generated and real image is small so they both receive relatively lower anomaly score.

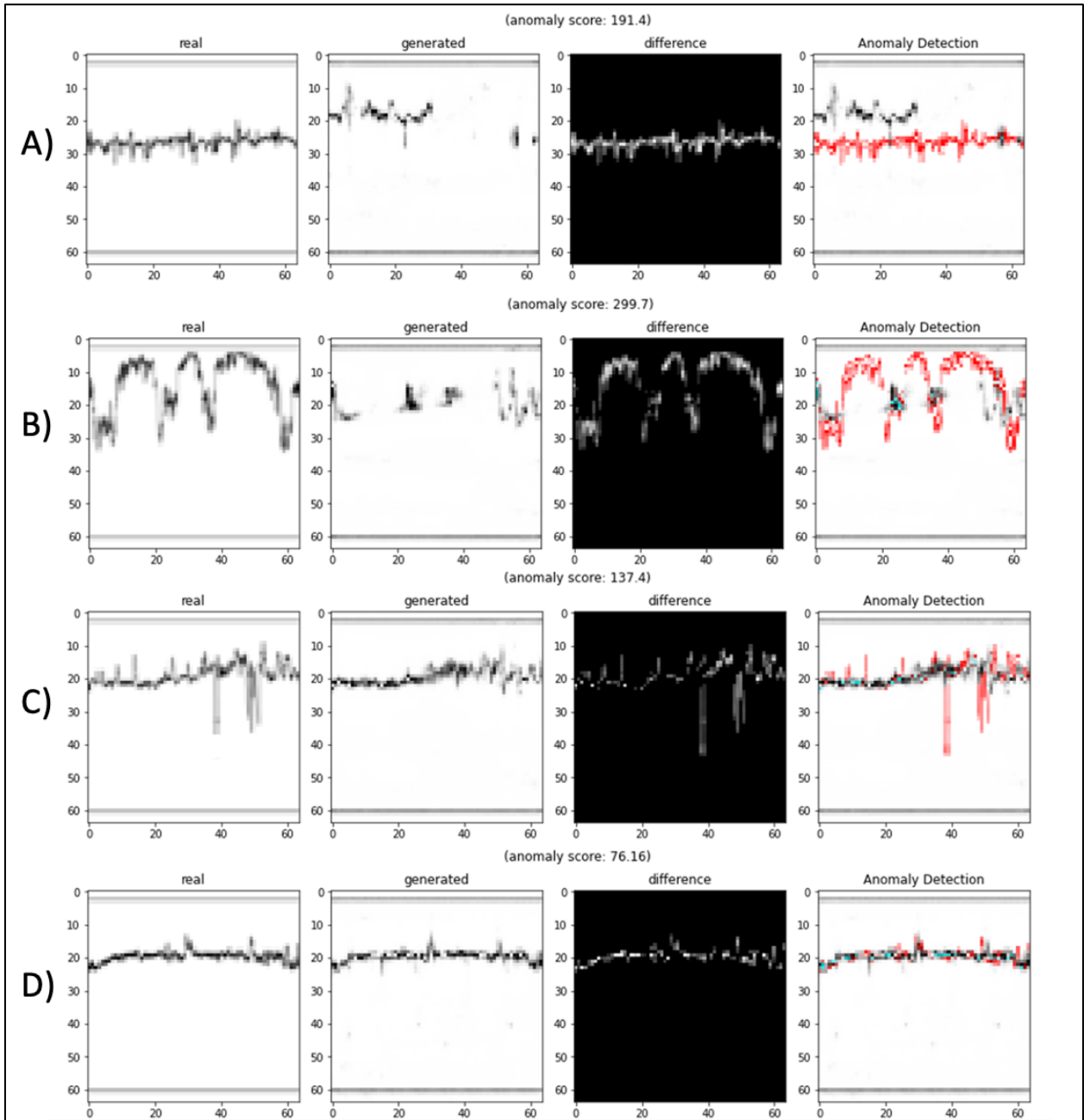
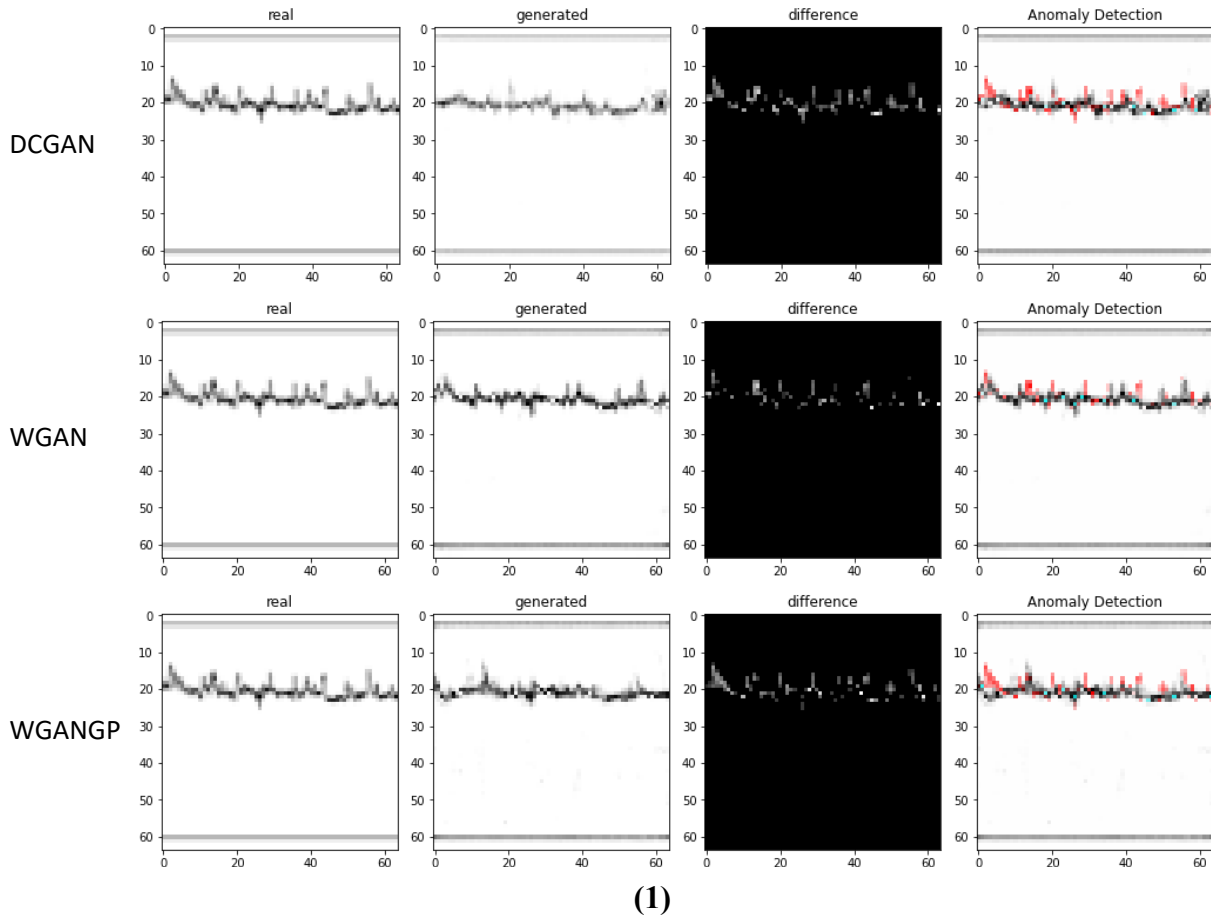
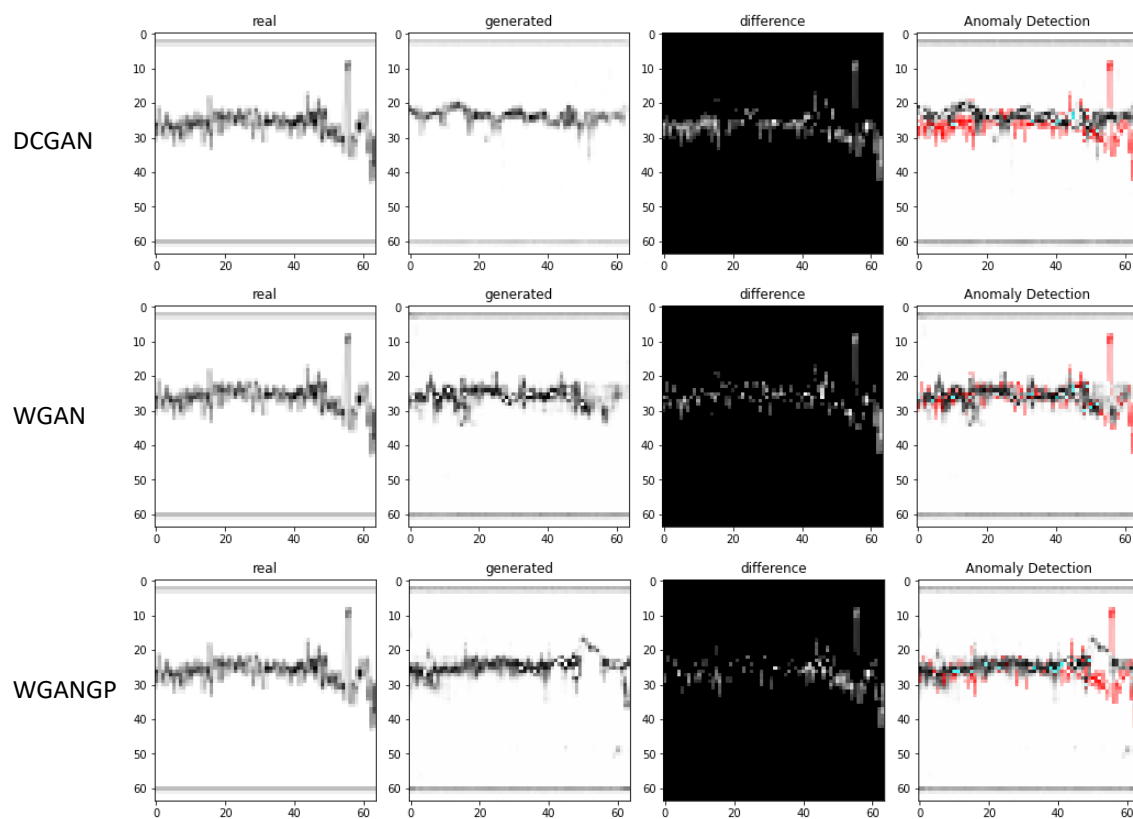


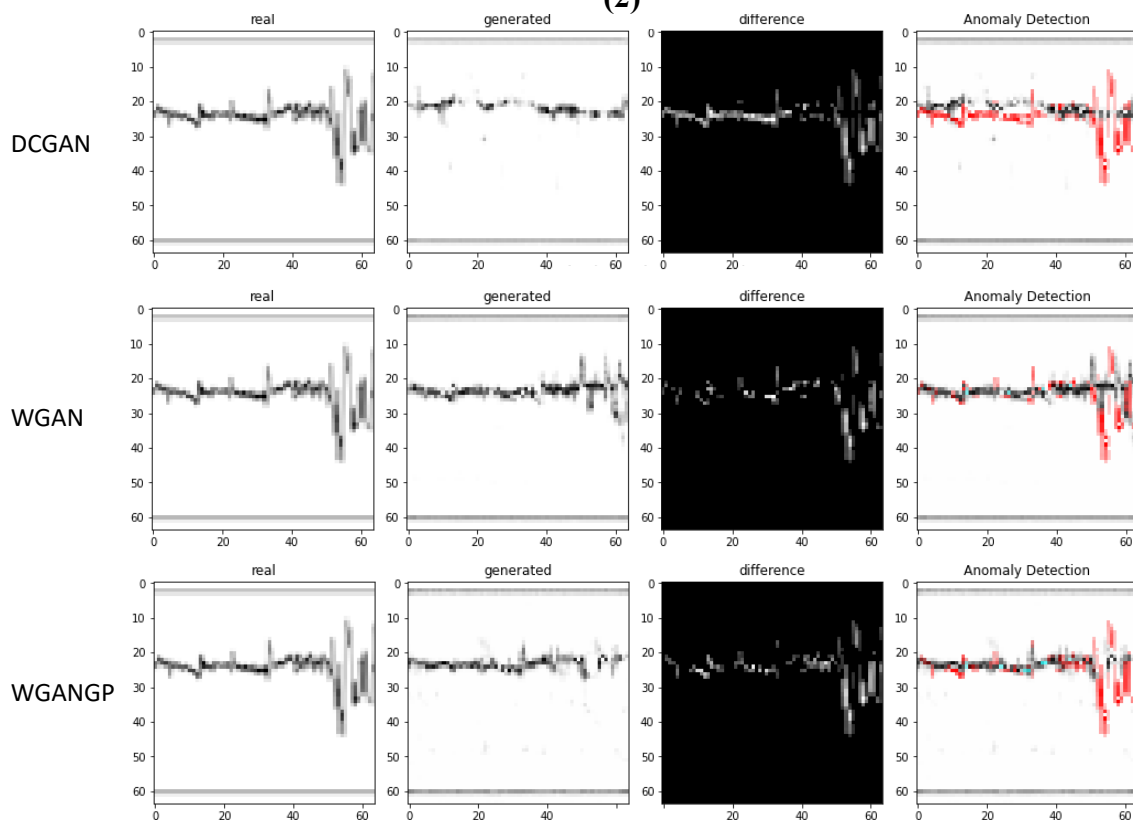
Figure 9: Generated CTG images compared to the real and the recorded anomaly score.

Figure 10 demonstrates varied samples of generated images for DCGAN, WGAN, and WGANGP. By looking at the generated images, we can see that the images generated by DCGAN are the least similar compared to the other models. It generates the general form of CTG but tends to fail in generating the details in a signal. WGAN and WGANGP performance are close as they both generate the majority of CTG signals. We noticed when parts of a CTG include abnormalities such as high variability or a sudden artifact, both models generate the CTG without the abnormal part as can be seen in samples numbers 2,3, and 6. We also noticed that WGAN is less likely to fail than WGANGP. It generates some parts of a signal even in some abnormal CTG signals such as tachycardia or bradycardia and sample number 8 is an example of that. This could be related to WGAN's high recall. As discussed earlier, a high recall in GAN means higher diversity of generated images. WGANGP has slightly higher precision than WGAN, however, judging the difference between the two by relying on visual assessment is challenging.

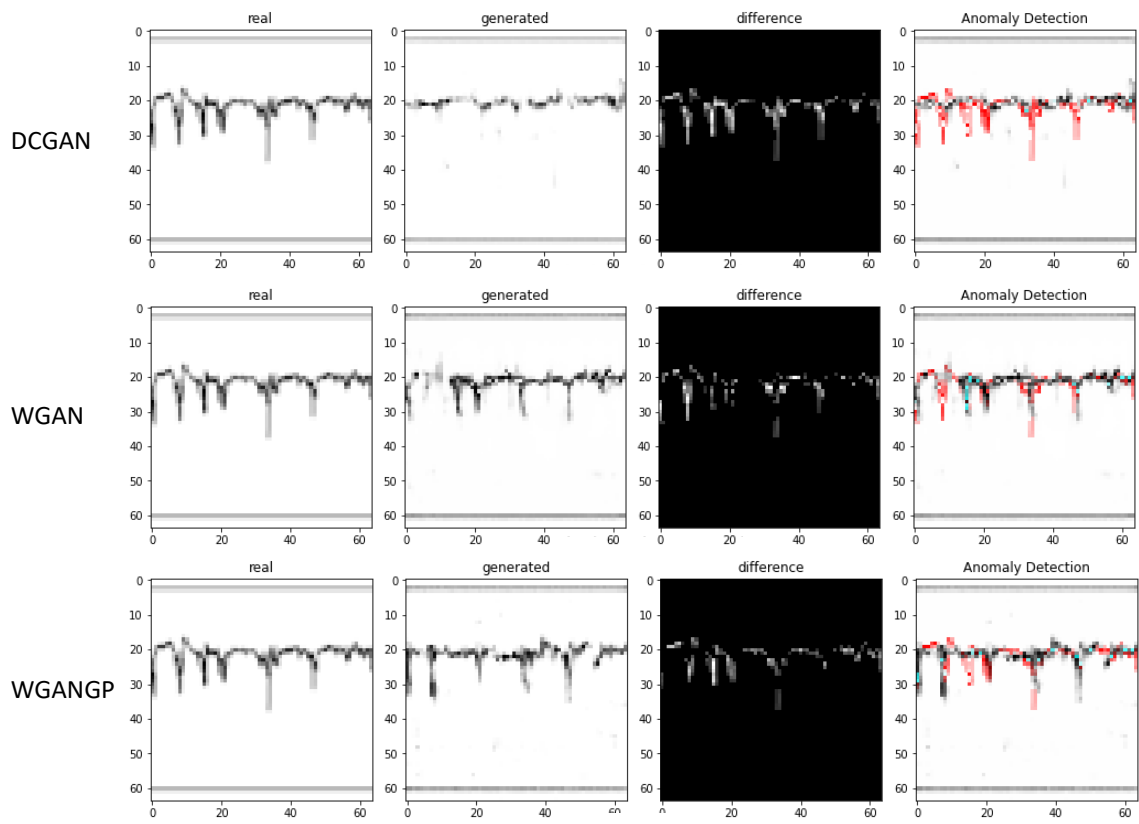




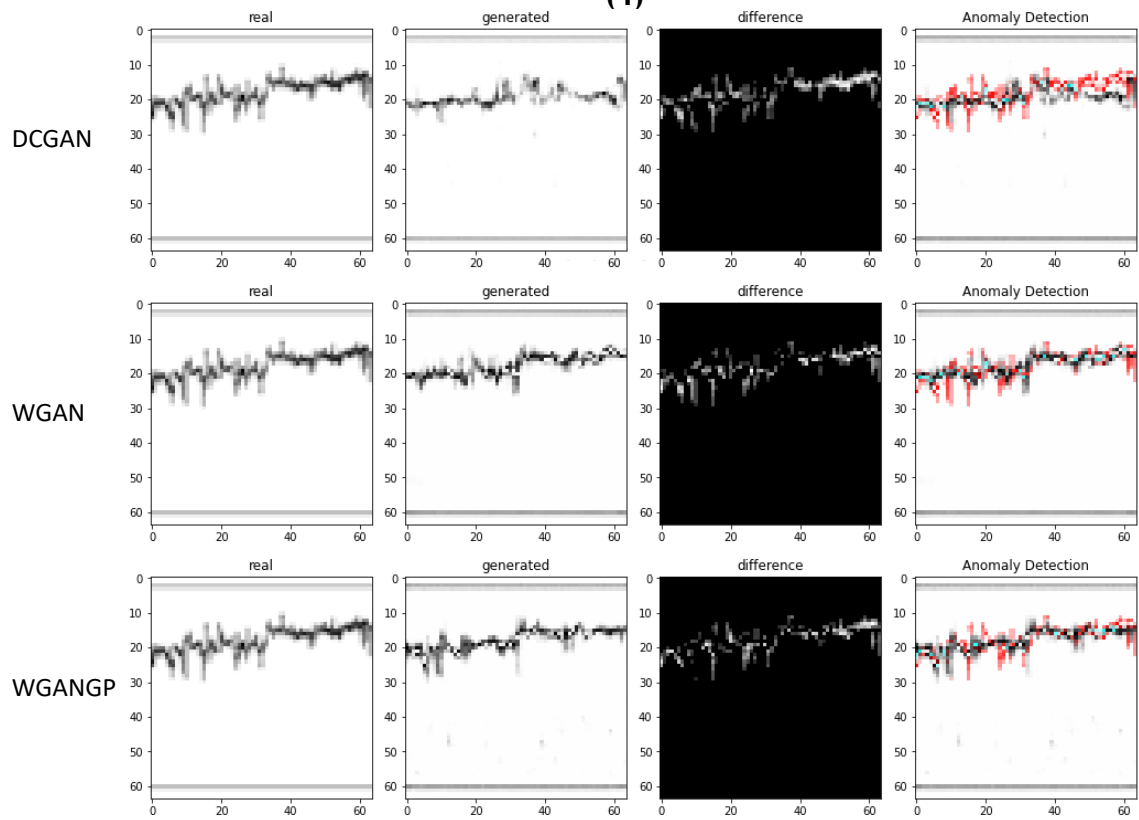
(2)



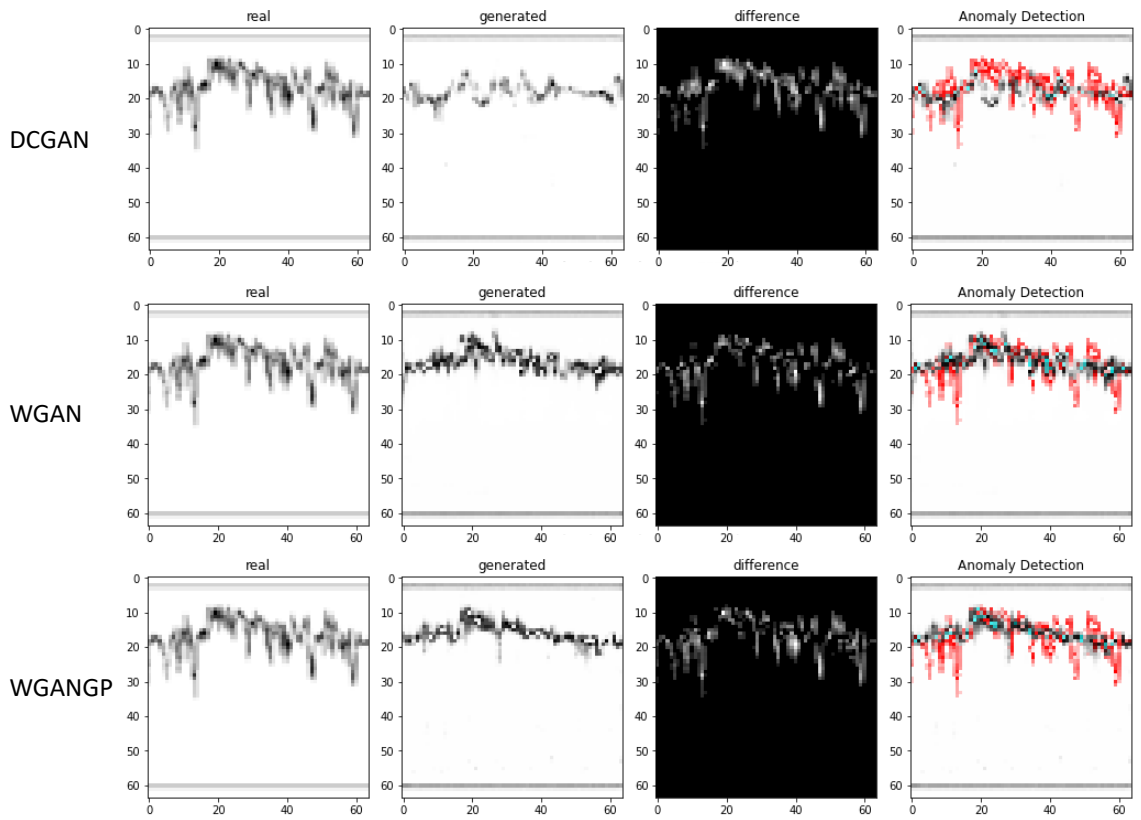
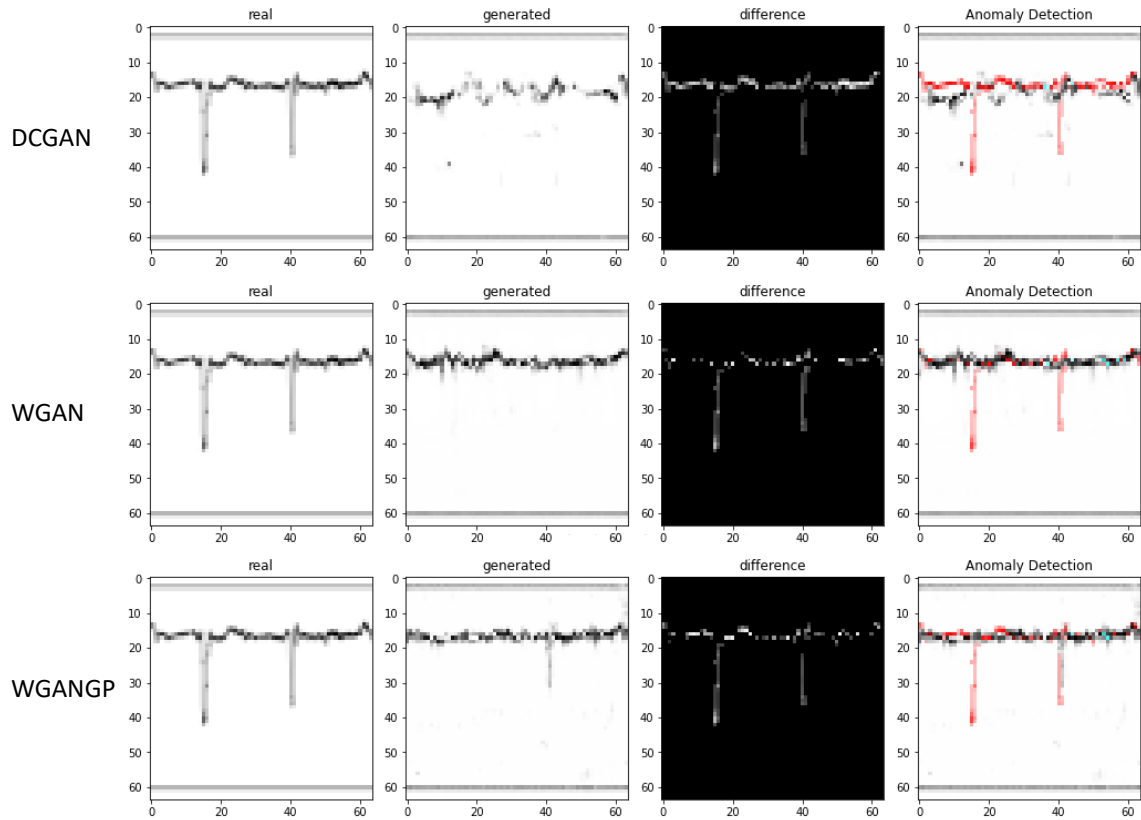
(3)

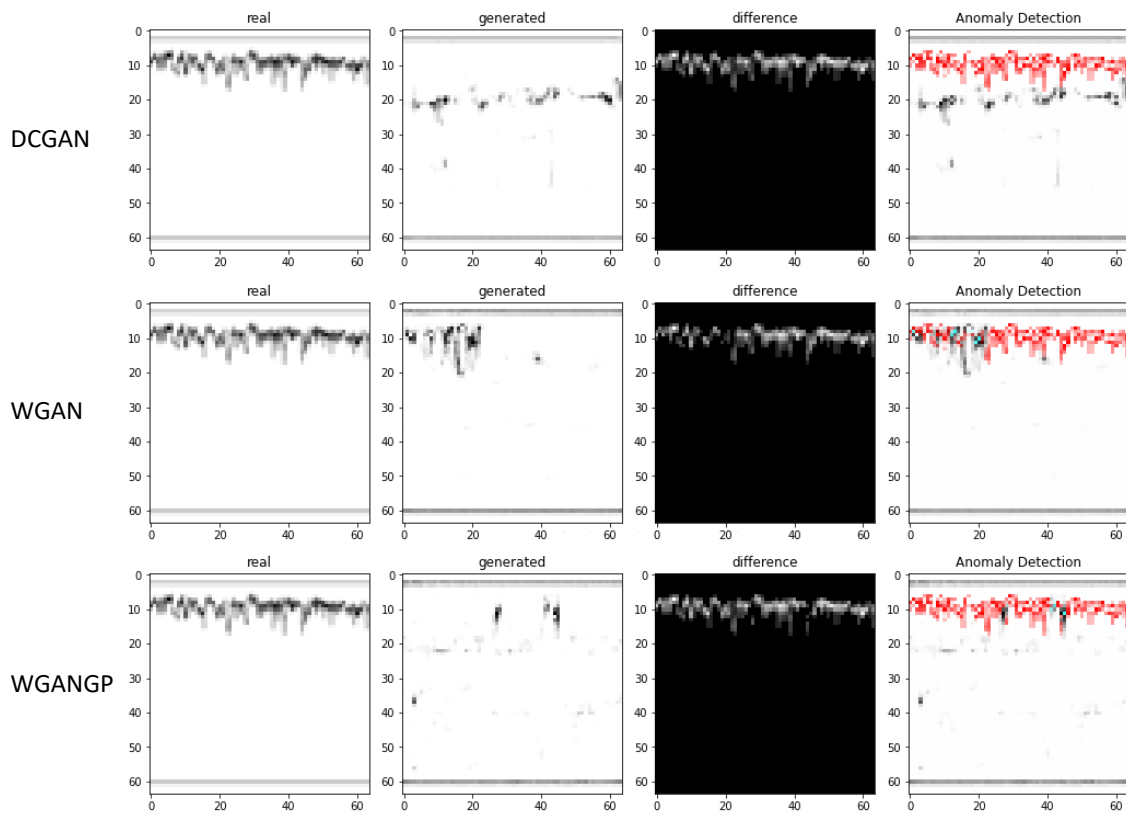


(4)

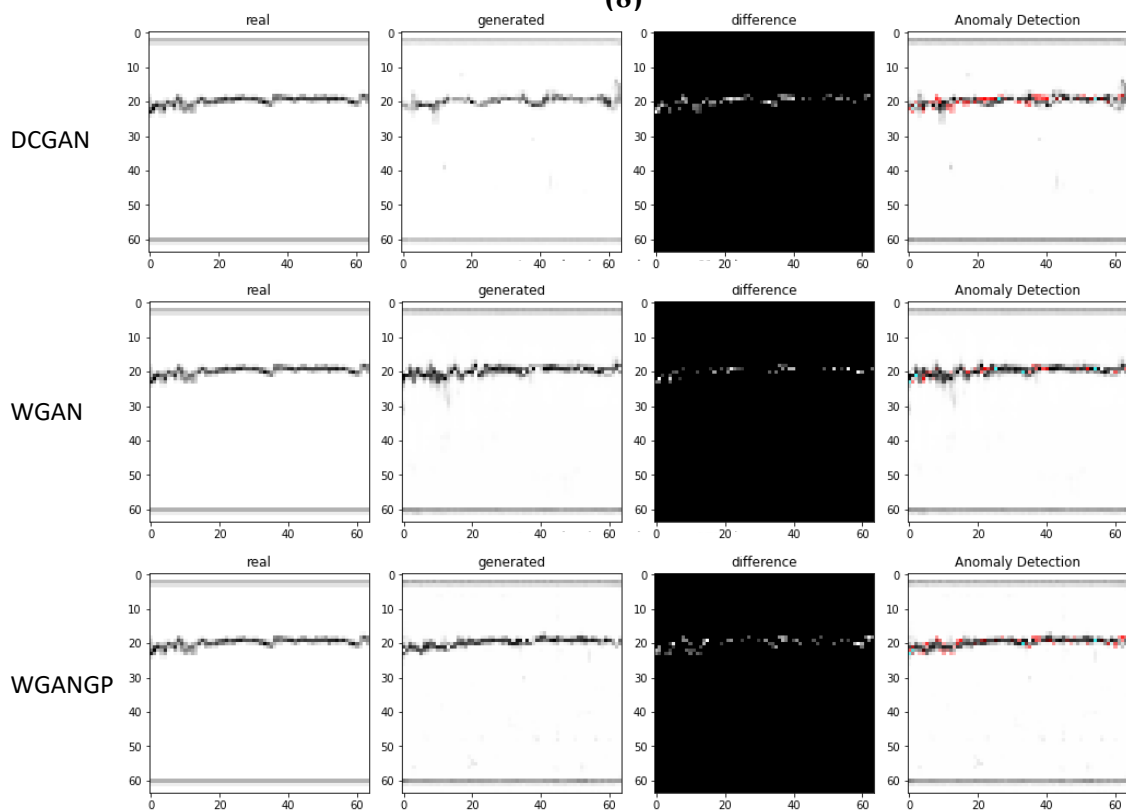


(5)





(8)



(9)

Figure 10: Samples of generated CTG images compared to the real images for DCGAN, WGAN and WGANGP.

4.6 Results and Discussion

After getting an anomaly score for each image in the test set, we created a spreadsheet and listed all positive and negative parts 1 of the CTG image in one column. The same is applied to parts 2,3 and 4 where each part is in a separate column. As mentioned in the data processing step, we divided the last 120 minutes of each record in the data set into 4 parts where each part is a CTG image with a length of 30 minutes. Part 1 is the first 30 minutes and part 4 is the last 30 minutes before giving birth. The total number of each part is 284 (142 positive and 142 negative). Next, all parts were sorted based on the highest anomaly score. Table 13 shows the percentage of positive parts compared to negative parts for DCGAN. For example, the highlighted cell in Table 13 means that 85% of the highest 20 part 1 of CTG images are positive. And the highlighted 25% means that 25% of the lowest 20 are positive part 1.

Table 13: A summary of the number of positive CTG compared to negative for all the four parts when using DCGAN

CTG images 120 divided into 4 parts	Number of positive compared to negative in part 1	Number of positive compared to negative in part 2	Number of positive compared to negative in part 3	Number of positive compared to negative in part 4
Percent in the Highest 20	85%	70%	60%	50%
Percent in the Highest 50	80%	66%	56%	52%
Percent in the Highest 100	66%	53%	55%	55%
Percent in the Lowest 20	25%	35%	20%	55%
Percent in the Lowest 50	38%	40%	40%	42%
Percent in the Lowest 100	38%	44%	44%	47%

Table 14: A summary of the number of positive CTG compared to negative for all the four parts when using WGAN

CTG images 120 divided into 4 parts	Number of positive compared to negative in part 1	Number of positive compared to negative in part 2	Number of positive compared to negative in part 3	Number of positive compared to negative in part 4
Percent in the Highest 20	90%	75%	60%	45%
Percent in the Highest 50	84%	60%	54%	54%
Percent in the Highest 100	65%	54%	54%	53%
Percent in the Lowest 20	40%	45%	50%	55%
Percent in the Lowest 50	34%	42%	42%	64%
Percent in the Lowest 100	37%	43%	47%	51%

Table 15: A summary of the number of positive CTG compared to negative for all the four parts when using WGANGP

CTG images 120 divided into 4 parts	Number of positive compared to negative in part 1	Number of positive compared to negative in part 2	Number of positive compared to negative in part 3	Number of positive compared to negative in part 4
Percent in the Highest 20	85%	70%	45%	40%
Percent in the Highest 50	80%	60%	52%	52%
Percent in the Highest 100	65%	53%	52%	53%
Percent in the Lowest 20	40%	45%	45%	65%
Percent in the Lowest 50	40%	42%	50%	56%
Percent in the Lowest 100	35%	45%	52%	52%

Looking at Table 13, it can be seen that positive CTG signals tend to have a higher anomaly score in part 1 compared to negative part 1, which is the first 30 minutes of the last 120 minutes before giving birth. For example, if we look at the highest 100 in Table 13, we see the number of positive parts 1 is 66%. This is consistent even if we look at the highest 20, 50, and 100 in DCGAN, WGAN, and WGANGP; all of them resulted in a higher anomaly in part 1 compared to negative part 1. The same can be observed in parts 2 and 3 but it is more prominent in part 1. The results of the three models are close, particularly in the highest 100 part. The experiment has been repeated several times and the results persist except the percentage varies between 65% and 70%. This applies to the three models.

Based on the above, we can see that positive CTG records tend to have higher anomaly score than negative records except for part 4. Excluding DCGAN, the number of positive parts 4 in both the highest 100 and the lowest 100 is close whereas positive parts 4 tends to be more in the lowest as shown in the lowest 20 and 50 for WGAN in Table 14 and WGANGP in Table 15. In fact, compared to other parts in the lowest records, we can clearly see positive parts 4 are more than other positive parts, especially part 1 being the fewest. We also noticed that the anomaly score in multiple positive records was big in part 1 but drop significantly in part 4, unlike negative records which a little bit more constant. In order to confirm this, Tables 16.1 and 16.2 summarize a spreadsheet where one column has all the positive parts and another column has all the negative parts regardless of the parts number. The two columns were sorted based on the highest anomaly score, then the number of each part in the highest 20, 50, and 100 and the lowest 20, 50, and 100 was counted.

Note: each part in a record was counted as a separate image, there are 1136 CTG images; 568 of them are positive and 568 are negative. If divided by the parts, 284 are part 1, 284 are part 2, 284 are part 3, and are 284 part 4. The highlighted row in Table 16.1 means in the top 100 positive rows with the highest anomaly score, 51% of them are part 1, 23% of them are part 2, 16% part 3, and 10% part 4.

Table 16.1: a summary of all **positive** CTG images sorted based on the highest anomaly score. DCGAN

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	55%	15%	15%	15%
Percent in the Highest 50 records	46%	22%	22%	10%
Percent in the Highest 100 records	51%	23%	16%	10%
Percent in the Lowest 20 records	10%	20%	10%	60%
Percent in the Lowest 50 records	10%	26%	22%	42%
Percent in the Lowest 100 records	15%	24%	25%	36%

Table 16.2: a summary of all **negative** CTG images sorted based on the highest anomaly score. DCGAN

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	25%	25%	35%	15%
Percent in the Highest 50 records	32%	26%	24%	18%
Percent in the Highest 100 records	31%	30%	23%	16%
Percent in the Lowest 20 records	5%	30%	25%	40%
Percent in the Lowest 50 records	10%	30%	30%	30%
Percent in the Lowest 100 records	16%	27%	27%	30%

In Table 16.1, the number of positive parts 1 in the highest 100 and the number of parts 4 in the lowest 100 are both high; 51% for part 1 and 36% for part 4. On the other hand, Table 16.2 shows that the number of negative parts 1 in the highest 100 is 31% and the number of parts 4 in the lowest 100 is 30%. This confirms that positive records tend to have a higher anomaly score in part 1 compared to part 1 in negative records and a lower anomaly score in part 4 than its counterpart in negative records. Again, this outcome is more noticeable in WGAN in Table 17 and WGANGP in Table 18 and we further clarify this in Figure 11, 12, and 13.

Table 17.1: a summary of all **positive** CTG images sorted based on the highest anomaly score. WGAN

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	60%	15%	5%	20%
Percent in the Highest 50 records	56%	20%	14%	10%
Percent in the Highest 100 records	51%	23%	17%	9%
Percent in the Lowest 20 records	15%	10%	30%	45%
Percent in the Lowest 50 records	10%	18%	28%	44%
Percent in the Lowest 100 records	8%	19%	28%	45%

Table 17.2: a summary of all **negative** CTG images sorted based on the highest anomaly score. WGAN

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	40%	30%	25%	5%
Percent in the Highest 50 records	38%	18%	24%	20%
Percent in the Highest 100 records	31%	33%	21%	15%
Percent in the Lowest 20 records	10%	25%	20%	45%
Percent in the Lowest 50 records	12%	26%	32%	30%
Percent in the Lowest 100 records	12%	26%	31%	31%

Table 18.1: a summary of all **positive** CTG images sorted based on the highest anomaly score. WGANGP

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	55%	30%	10%	5%
Percent in the Highest 50 records	58%	18%	14%	10%
Percent in the Highest 100 records	54%	22%	16%	8%
Percent in the Lowest 20 records	0%	30%	25%	45%
Percent in the Lowest 50 records	4%	20%	26%	50%
Percent in the Lowest 100 records	9%	23%	32%	36%

Table 18.2: a summary of all **negative** CTG images sorted based on the highest anomaly score. WGANGP

Positive CTG images 120 minutes divided into 4 parts	Number of parts1	Number of parts2	Number of parts3	Number of parts4
Percent in the Highest 20 records	40%	30%	20%	10%
Percent in the Highest 50 records	34%	18%	32%	16%
Percent in the Highest 100 records	35%	27%	22%	16%
Percent in the Lowest 20 records	10%	15%	40%	35%
Percent in the Lowest 50 records	10%	28%	30%	32%
Percent in the Lowest 100 records	14%	27%	30%	29%

Another thing to note is the order of the anomaly score of the four parts of the CTG signal. For instance, in Table 16.1, Table 17.1, and Table 18.1, in the highest 100, we see the order of the majority of the four parts goes part 1 > 2 > 3 > 4. For example, the average anomaly score of the highest 100 in Table 17.1 is 242 for part 1, 209 for part 2, 197.5 for part 3, and 183 for part 4. This is almost the same for Tables 16.2, Table 17.2, and Table 18.2 except the density is more disturbed between the four parts. We repeated this experiment multiple times and the results are consistent for both parts 1 and 4. Nevertheless, the anomaly scores for parts 2 and 3 are usually close to each other so the variability between these two parts changes a bit between experiments.

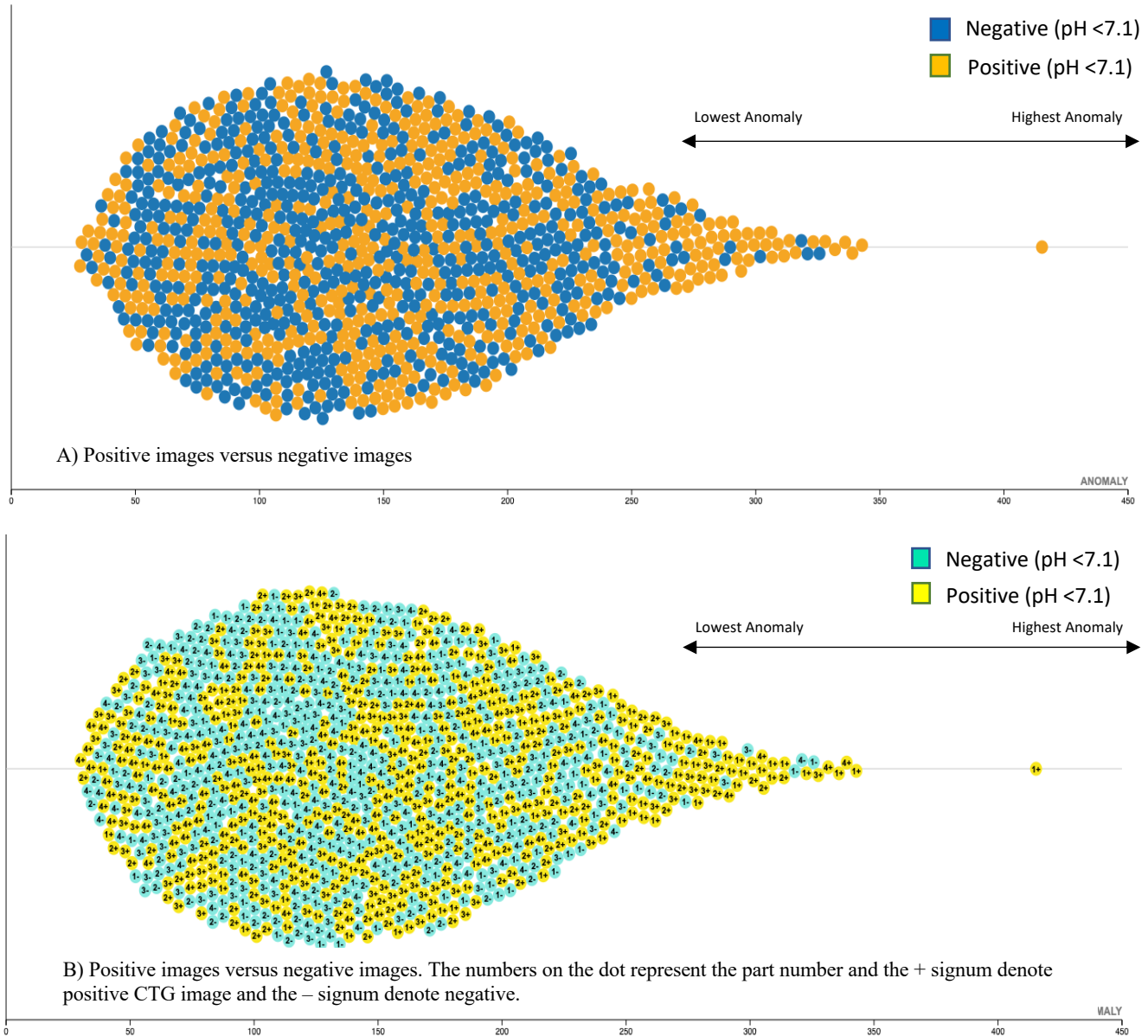


Figure 11: All positive and negative CTG images after dividing each record into 4 parts. Each dot represents a CTG image and all of the images are sorted based on the highest anomaly score of WGAN model. The number of images is 1136; 568 are positive and 568 are negative.

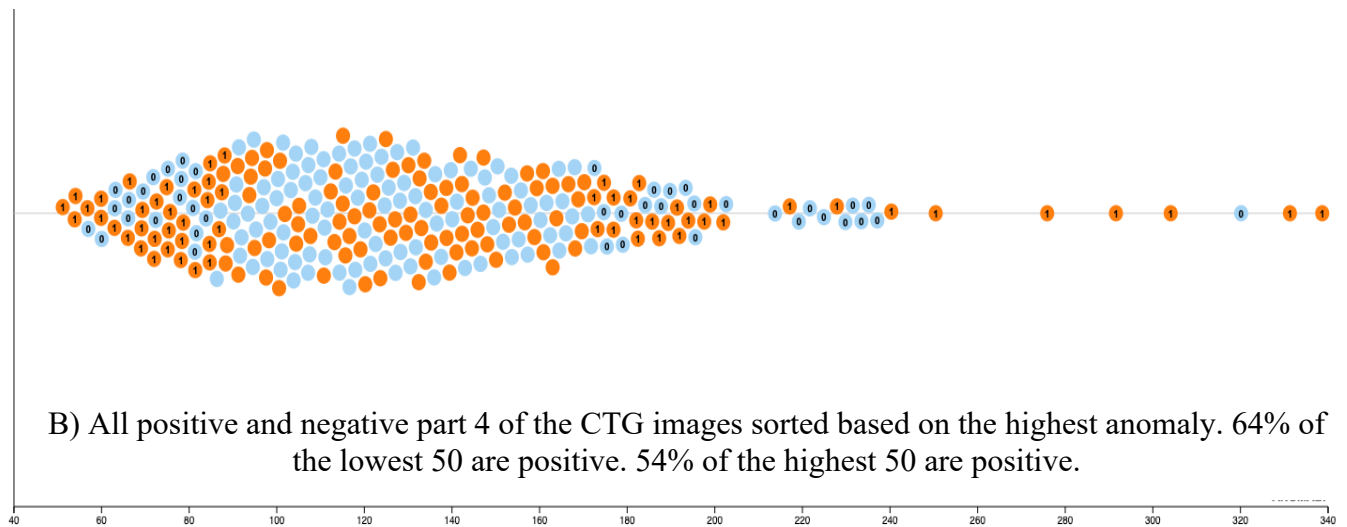
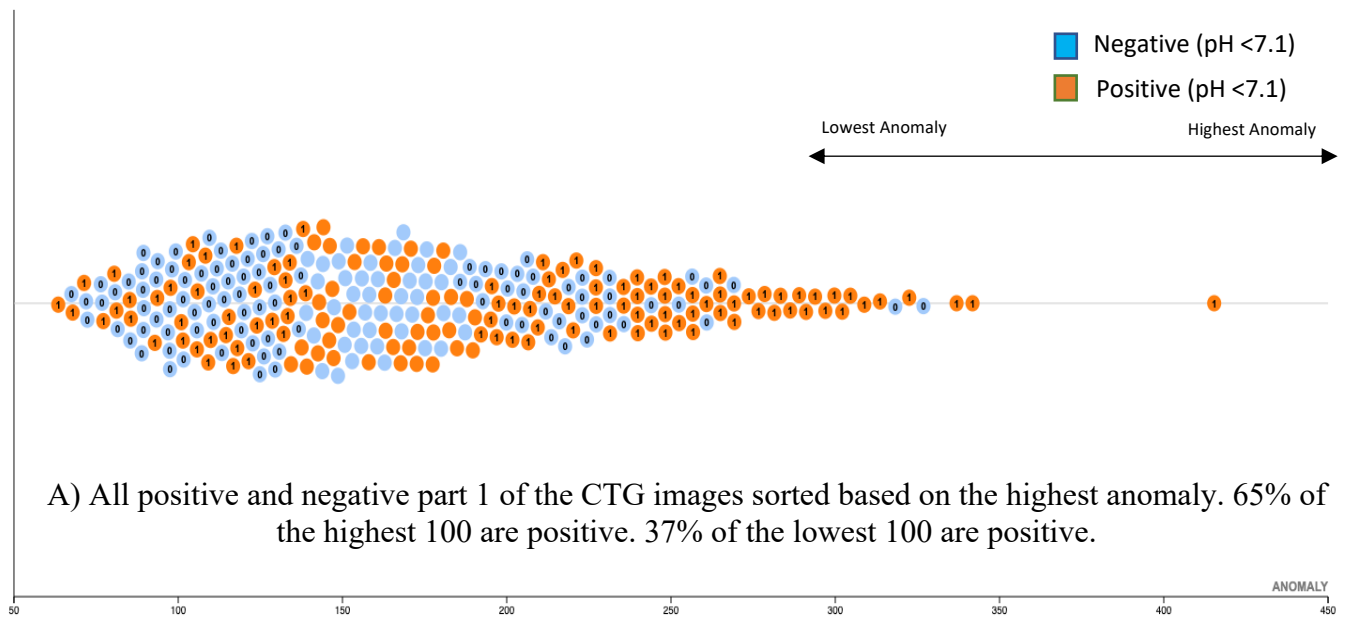


Figure 12: A bee swarm plot of Table 14 where (A) represents the part 1 column and (B) the part 4 column.

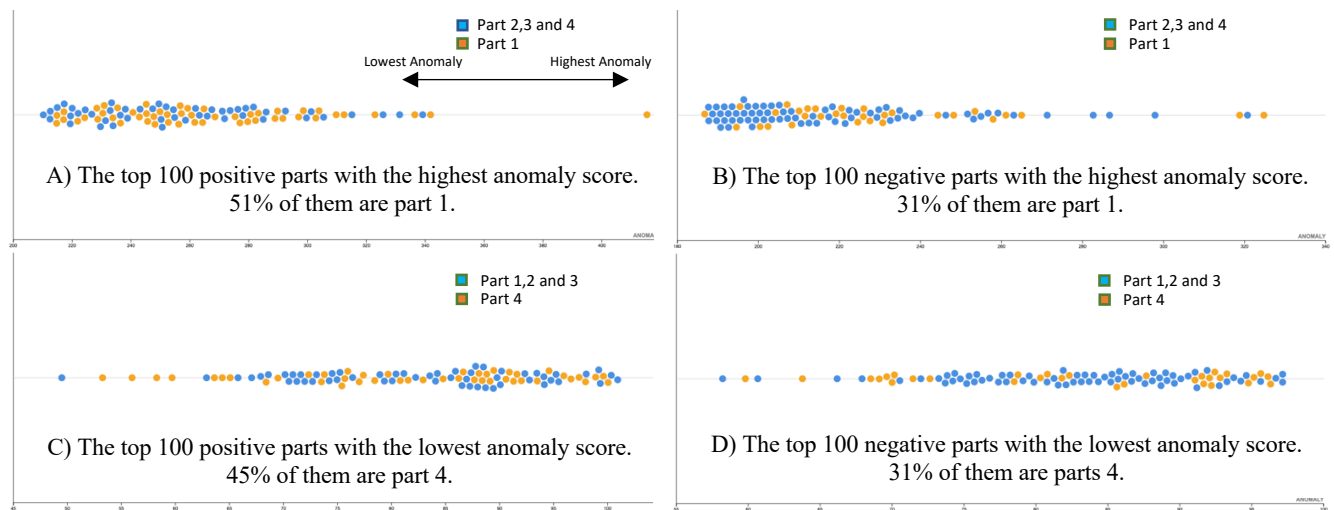


Figure 13: The above subplots show that positive records tend to have a higher anomaly score in part 1 compared to part 1 in negative records and a lower anomaly score in positive part 4 than its counterpart in negative records.

Lastly, the number of chosen epochs for training varied between models. For example, after training the DCGAN model for many epochs such as 150 epochs and more, the generated images have shown signs of mode collapse and the FID score curve, which was giving lower values between epochs, start to give higher values. This is why we trained DCGAN for about 100 epochs. As mentioned earlier in this chapter, WGAN uses the Wasserstein loss function which has a correlation between the divergence of generator and discriminator loss and the quality of generated images. For this reason, the WGAN model was trained until convergence occurred which was 86 epochs. As for the WGANGP model, it was able to be trained for many epochs without suffering from mode collapse, vanishing, or exploding gradient. We relied on the performance of FID as an indicator of when to stop WGANGP training. We trained the model for 400 epochs but the FID performance stopped improving after 200 epochs, thus we stopped training at about 200 epochs.



Figure 14: A plot of Generator and Discriminator/ Critic loss for DCGAN, WGAN and WGANGP.

4.7 Conclusion

In this study, we processed the dataset to have only low-risk birth. all pre-term pregnancies, post-term pregnancies, and C-sections were excluded from the study, unlike the two previous studies in chapters 2 and 3. Because of the nature of the experiment, the training set includes just negative records which help in preventing data imbalance. The results are not to be compared with the previous studies, rather, it utilizes anomaly detection in studying the difference between CTG images of low pH patients versus normal pH patients.

Using ANOGAN, we found that CTG images with $\text{pH} < 7.1$ tend to have higher anomaly score compared to CTG images with $\text{pH} \geq 7.1$. The model gives a higher anomaly score to uncommon FHR patterns and we think this can be helpful for a medical team to support their decision during birth. The performance of three trained models: DCGAN, WGAN, and WGANGP was evaluated using FID, PDF, and PR. The quality of the generated images can be further improved; however, it requires higher computation capabilities.

Chapter 5

Conclusion

Below we describe the summary of the thesis and how the results can be better further improved in future studies.

5.1 Research Results

Cardiotocography (CTG) is the practice of reading fetal heart rate and uterine contraction in the third trimester of pregnancy and during birth. CTG has drawbacks such as high noise, diagnosis contradictions, and the need for the continuous presence of an expert to read CTG signals. Many researchers see that the current progress in artificial intelligence has the capability in finding potential solutions to some of CTG challenges.

We proposed different machine and deep learning approaches which could act as a prognosis tool in predicting high-risk birth. The contributions of this work can be summarized as the following:

(1) An improved CTG denoising algorithm

The denoising algorithm has the capability to reduce artifacts from high noise signals with minimum alterations to the original form of the signal. Our dataset is collected under clinical condition and include signals with high noise. Our denoising process can be helpful in such scenarios.

(2) Algorithm for extracting features using JSOG 5-level FHR patterns and applied ML methods to classify high-risk birth using pH and Apgar score 1 and 5.

We built an algorithm to extract important features from CTG signals such as baseline, accelerations, decelerations, etc. based on JSOG guidelines. We tested the extracted features using SVM, RF, ANN, and DT and achieved an AUC of 0.89 using 5-fold cross-validation.

(3) A multi-input CNN model to classify neonatal with low Apgar using real data only.

Our CNN model is intended to be very light and uses features that are easily available. It is built upon EfficientNet CNN architecture and pre-trained on imagined dataset weights which give it both an advantage in accuracy and computation speed. Using 5-fold cross-validation, the model achieved an AUC of 0.949 when classifying fetal with the risk of a low Apgar score. Additionally, we found that performance got better when we trained the model with CTG images with longer signal lengths.

(4) Implemented Anomaly Detection using Generative Adversarial networks (ANOGAN) to generate CTG images and compare normal and abnormal image anomalies.

We implanted the ANOGAN approach using three GAN architectures: DCGAN, WGAN, and WGANGP, and evaluated their ability in generating images through recognized evaluation metrics in the field. The three models were trained on CTG images of the last 120 minutes before birth and all of the records were for patients with $\text{pH} \geq 7.1$. Class imbalance was not an issue since training was only on one class. The achieved results show that the majority of CTG images with $\text{pH} < 7.1$ have a higher anomaly score than CTG with $\text{pH} \geq 7.1$. To the best of our knowledge, no study applied GAN or ANOGAN to generate CTG images.

5.2 Future Works

This study uses multiple AI approaches to classify high-risk childbirth. We believe there is more that can be done to achieve better feature extraction, data balance, accurate classification, quality CTG image generation, and anomaly score. One of the issues when implementing AI on CTG is determining the length of the signals needed for the study. Understandably, not all CTG signals share the same length so they have to be tailored to a specific length for consistency when training and testing the AI model. The critical information that causes fetal to be at risk, for example, might be found in a signal within the last 20 minutes before birth and some may be

found in the last 120 minutes before birth. Another challenge is even after classifying abnormal CTG, there are other factors that need to be present before medical intervention such as childbirth progression and the risk profile of the mother [24]. In conclusion, this study can be further improved by including more pre-obtained features, more accurately gathered records, and involving obstetricians and medical experts in the trials of AI experiments. Our approach which focused on incorporating interrelated artificial intelligence decision support tools could support the medical team with guidance in mitigating the risks they confront in childbirth.

Acknowledgments

This thesis would not have been possible without the grace and blessings of Allah. I would like to acknowledge and give my deepest gratitude to Dr. Chihiro Shibata, Dr. Kazuya Tago, and Dr. Terumasa Aoki. Dr. Shibata is my supervisor. She welcomed me into her lab and provided me with ideas, books, and tools for my research. Her guidance and suggestions carried me through all the phases and taught me the right approaches to becoming a researcher. She continued to be my thesis advisor even after moving to a different university. I am grateful to all of the members of Shibata lab for their kindness and cooperation. Dr. Tago helped me move to his lab and made sure I have all what I needed to continue my work before he retired. I wish him a happy retirement. A big thanks to Dr. Terumasa Aoki for welcoming me into his lab and assisting me in completing all the requirements for submitting my thesis despite his busy schedule. I am always in his debt.

I am also extremely grateful to the Department of Obstetrics and Gynecology at Fukuoka University, especially Dr. Shingo Miyamoto, the chairman, and Dr. Kohei Miyata. For their collaboration and support. Dr. Miyata gave me the honor to use and study the unique dataset that he and his team prepared and worked on. It is a major foundation of this thesis. Additionally, both Dr. Toshiro Imamura from clinic of Shonan Kugenuma Obstetrics and Genecology and Dr. Miyata provided me with invaluable information about the Obstetrics field and how to read CTG guidelines which fully helped me to work better on my research without struggling with medical details. They constantly answered my emails and double-checked my results from the obstetrician's viewpoint. Their expertise and knowledge were exceptional. I cannot appreciate this enough.

I am thankful to the chief examiner Dr. Hiroyuki Kameda and the rest of committee members Dr. Chihiro Shibata, Dr. Hiroaki Fukunishi, Dr. Kazuya Tago, and Dr. Terumasa Aoki for spending their valuable time reviewing my thesis and providing me with constructive feedback.

A special thanks to my wife Munerah and my daughter Farah. They are the joy and the love of my life and the motivation that kept me pushing hard. They accompanied me through my journey to Japan and my wife was always encouraging, caring and supportive.

Last but not the least, I would like to give my sincere thanks to my father and mother who encouraged me to continue my higher studies. I am forever grateful for them.

References

- [1] M. A. Gulum, C. M. Trombley, and M. Kantardzic, “A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging,” *Applied Sciences*, vol. 11, no. 10. MDPI AG, p. 4573, May 17, 2021.
- [2] K. i. m. Solez et al., “International standardization of criteria for the histologic diagnosis of renal allograft rejection: The Banff working classification of kidney transplant pathology,” *Kidney International*, vol. 44, no. 2. Elsevier BV, pp. 411–422, Aug. 1993.
- [3] D. Ayres-de-Campos, C. Y. Spong, and E. Chandrharan, “FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography,” *International Journal of Gynecology & Obstetrics*, vol. 131, no. 1. Wiley, pp. 13–24, Sep. 30, 2015.
- [4] A. Pinas and E. Chandrharan, “Continuous cardiotocography during labour: Analysis, classification and management,” *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 30. Elsevier BV, pp. 33–47, Jan. 2016.
- [5] M. Rei et al., “Interobserver agreement in CTG interpretation using the 2015 FIGO guidelines for intrapartum fetal monitoring,” *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 205. Elsevier BV, pp. 27–31, Oct. 2016.
- [6] S. Schiermeier, G. Westhof, A. Leven, H. Hatzmann, and J. Reinhard, “Intra- and Interobserver Variability of Intrapartum Cardiotocography: A Multicenter Study Comparing the

FIGO Classification with Computer Analysis Software,” *Gynecologic and Obstetric Investigation*, vol. 72, no. 3. S. Karger AG, pp. 169–173, 2011.

[7] Ayres-de-Campos D. *Obstetric Emergencies*. Berlin, Springer, 2016.

[8] H.-Y. Chen, S. P. Chauhan, C. V. Ananth, A. M. Vintzileos, and A. Z. Abuhamad, “Electronic fetal heart rate monitoring and its relationship to neonatal and infant mortality in the United States,” *American Journal of Obstetrics and Gynecology*, vol. 204, no. 6. Elsevier BV, p. 491.e1-491.e10, Jun. 2011.

[9] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. Devane, “Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour,” *Cochrane Database of Systematic Reviews*, vol. 2019, no. 5. Wiley, Feb. 03, 2017.

[10] A. Georgieva et al., “Computer-based intrapartum fetal monitoring and beyond: A review of the 2nd Workshop on Signal Processing and Monitoring in Labor (October 2017, Oxford, UK),” *Acta Obstetricia et Gynecologica Scandinavica*, vol. 98, no. 9. Wiley, pp. 1207–1217, Jun. 18, 2019.

[11] Z. Cömert and A. F. Kocamaz, “Comparison of Machine Learning Techniques for Fetal Heart Rate Classification,” *Acta Physica Polonica A*, vol. 132, no. 3. Institute of Physics, Polish Academy of Sciences, pp. 451–454, Sep. 2017.

[12] B. Hasan, Z. Hoodbhoy, M. Noman, A. Shafique, A. Nasim, and D. Chowdhury, “Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data,” *International Journal of Applied and Basic Medical Research*, vol. 9, no. 4. Medknow, p. 226, 2019.

[13] S. Santo et al., “Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines,” *Acta Obstetricia et Gynecologica Scandinavica*, vol. 96, no. 2. Wiley, pp. 166–175, Jan. 06, 2017.

[14] A. Petrozziello, C. W. G. Redman, A. T. Papageorghiou, I. Jordanov, and A. Georgieva, “Multimodal Convolutional Neural Networks to Detect Fetal Compromise During Labor and

Delivery,” IEEE Access, vol. 7. Institute of Electrical and Electronics Engineers (IEEE), pp. 112026–112036, 2019.

[15] Z. Zhao, Y. Zhang, Z. Comert, and Y. Deng, “Computer-Aided Diagnosis System of Fetal Hypoxia Incorporating Recurrence Plot With Convolutional Neural Network,” *Frontiers in Physiology*, vol. 10. Frontiers Media SA, Mar. 12, 2019.

[16] J. Ogasawara et al., “Deep neural network-based classification of cardiocograms outperformed conventional algorithms,” *Scientific Reports*, vol. 11, no. 1. Springer Science and Business Media LLC, Jun. 28, 2021.

[17] V. Chudáček et al., “The CTU-UHB Intrapartum Cardiotocography Database.” physionet.org, 2014.

[18] I. Linardos, “Signal Quality Assessment for Reliable Fetal Heart Rate Detection Using Deep Learning,” Unpublished, 2020.

[19] R. D. I. Puspitasari, M. A. Ma’sum, M. R. Alhamidi, Kurnianingsih, and W. Jatmiko, “Generative adversarial networks for unbalanced fetal heart rate signal classification,” *ICT Express*. Elsevier BV, Jul. 2021.

[20] P. Meena, M. Meena, and M. Gunawat, “Correlation of APGAR score and cord blood pH with severity of birth asphyxia and short-term outcome,” *International Journal of Contemporary Pediatrics*, vol. 4, no. 4. Medip Academy, p. 1325, Jun. 21, 2017.

[21] G. C. Martin, R. S. Green, and I. R. Holzman, “Acidosis in Newborns with Nuchal Cords and Normal Apgar Scores,” *Journal of Perinatology*, vol. 25, no. 3. Springer Science and Business Media LLC, pp. 162–165, Dec. 16, 2004.

[22] J. Esteban-Escañó et al., “Machine Learning Algorithm to Predict Acidemia Using Electronic Fetal Monitoring Recording Parameters,” *Entropy*, vol. 24, no. 1. MDPI AG, p. 68, Dec. 30, 2021.

- [23] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery.” arXiv, 2017.
- [24] M. E. O’Sullivan, E. C. Considine, M. O’Riordan, W. P. Marnane, J. M. Rennie, and G. B. Boylan, “Challenges of Developing Robust AI for Intrapartum Fetal Heart Rate Monitoring,” *Frontiers in Artificial Intelligence*, vol. 4. Frontiers Media SA, Oct. 26, 2021.
- [25] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” arXiv, 2015.
- [26] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN.” arXiv, 2017.
- [27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved Training of Wasserstein GANs.” arXiv, 2017.
- [28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” arXiv, 2017
- [29] J. T. Parer, “Standardization of fetal heart rate pattern management: Is international consensus possible?,” *Hypertension Research in Pregnancy*, vol. 2, no. 2. Japan Society for the Study of Hypertension in Pregnancy, pp. 51–58, 2014.
- [30] S. Santo et al., “Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines,” *Acta Obstetrica et Gynecologica Scandinavica*, vol. 96, no. 2. Wiley, pp. 166–175, Jan. 06, 2017.
- [31] M. Hayashi, A. Nakai, A. Sekiguchi, and T. Takeshita, “Fetal Heart Rate Classification Proposed by the Perinatology Committee of the Japan Society of Obstetrics and Gynecology: Reproducibility and Clinical Usefulness,” *Journal of Nippon Medical School*, vol. 79, no. 1. Medical Association of Nippon Medical School, pp. 60–68, 2012.
- [32] E. Wiberg-Itzel et al., “Determination of pH or lactate in fetal scalp blood in management of intrapartum fetal distress: randomised controlled multicentre trial,” *BMJ*, vol. 336, no. 7656. *BMJ*, pp. 1284–1287, May 25, 2008.

- [33] M. Persson, N. Razaz, K. Tedroff, K. S. Joseph, and S. Cnattingius, “Five and 10 minute Apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in Sweden,” *BMJ*. *BMJ*, p. k207, Feb. 07, 2018.
- [34] “Committee Opinion No. 644,” *Obstetrics & Gynecology*, vol. 126, no. 4. Ovid Technologies (Wolters Kluwer Health), pp. e52–e55, Oct. 2015.
- [35] V. Ehrenstein, L. Pedersen, M. Grijota, G. L. Nielsen, K. J. Rothman, and H. T. Sørensen, “Association of Apgar score at five minutes with long-term neurologic disability and cognitive function in a prevalence study of Danish conscripts,” *BMC Pregnancy and Childbirth*, vol. 9, no. 1. Springer Science and Business Media LLC, Apr. 02, 2009.
- [36] E. M. Assunção Salustiano, J. A. DuarteBonini Campos, S. M. Ibidi, R. Ruano, and M. Zugaib, “Low Apgar scores at 5 minutes in a low risk population: Maternal and obstetrical factors and postnatal outcome,” *Revista da Associação Médica Brasileira*, vol. 58, no. 5. Elsevier BV, pp. 587–593, Sep. 2012.
- [37] A. Modabbernia et al., “Apgar score and risk of autism,” *European Journal of Epidemiology*, vol. 34, no. 2. Springer Science and Business Media LLC, pp. 105–114, Oct. 05, 2018.
- [38] B. M. Casey, D. D. McIntire, and K. J. Leveno, “The Continuing Value of the Apgar Score for the Assessment of Newborn Infants,” *New England Journal of Medicine*, vol. 344, no. 7. Massachusetts Medical Society, pp. 467–471, Feb. 15, 2001.
- [39] F. Li, T. Wu, X. Lei, H. Zhang, M. Mao, and J. Zhang, “The Apgar Score and Infant Mortality,” *PLoS ONE*, vol. 8, no. 7. Public Library of Science (PLoS), p. e69072, Jul. 29, 2013.
- [40] S. G. Walker, “Smith’s Anesthesia for Infants and Children, 9th ed,” *Anesthesia & Analgesia*, vol. 128, no. 2. Ovid Technologies (Wolters Kluwer Health), p. e19, Feb. 2019.
- [41] M. A. Kacho, N. Asnafi, M. Javadian, M. Hajiahmadi, and N. Taleghani. “Correlation between umbilical cord ph and apgar score in high-risk pregnancy”, *Iranian journal of pediatrics*, vol. 20, no. 4. p. 401, Dec. 2010.

- [42] Royal college of obstetricians and gynaecologists. Each Baby Counts: 2019 Progress Report, London: RCOG; 2020.
- [43] M. G. Frasch, G. B. Boylan, H. Wu, and D. Devane, “Commentary: Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial,” *Frontiers in Physiology*, vol. 8. Frontiers Media SA, Sep. 28, 2017.
- [44] I. Nunes et al., “Central Fetal Monitoring With and Without Computer Analysis,” *Obstetrics & Gynecology*, vol. 129, no. 1. Ovid Technologies (Wolters Kluwer Health), pp. 83–90, Jan. 2017.
- [45] S. Berglund, H. Pettersson, S. Cnattingius, and C. Grunewald, “How often is a low Apgar score the result of substandard care during labour?,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 117, no. 8. Wiley, pp. 968–978, Apr. 20, 2010.
- [46] S. Banu, “Relationship between Abnormal Cardiotocography and Fetal Outcome,” *Nepal Journal of Obstetrics and Gynaecology*, vol. 10, no. 2. Nepal Journals Online (JOL), pp. 36–39, Jan. 15, 2016.
- [47] L. Ali, R. Mushtaq, and N. Ahmed. “Frequency of Pathological CTG in Low Risk Women and its Outcomes,” *Pak J Surg*, vol. 30, no. 4, pp. 340–345, 2014.
- [48] M. Podda, D. Bacciu, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, “A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor,” *Scientific Reports*, vol. 8, no. 1. Springer Science and Business Media LLC, Sep. 13, 2018.
- [49] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *arXiv*, 2019.
- [50] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2009.

[51] A. P. Drogtop, R. Ubels, and J. G. Nijhuis, “The association between fetal body movements, eye movements and heart rate patterns in pregnancies between 25 and 30 weeks of gestation,” *Early Human Development*, vol. 23, no. 1. Elsevier BV, pp. 67–73, Jun. 1990.

[52] S. Cnattingius, M. Norman, F. Granath, G. Petersson, O. Stephansson, and T. Frisell, “Apgar Score Components at 5 Minutes: Risks and Prediction of Neonatal Mortality,” *Paediatric and Perinatal Epidemiology*, vol. 31, no. 4. Wiley, pp. 328–337, May 11, 2017.

[53] G. C. Martin, R. S. Green, and I. R. Holzman, “Acidosis in Newborns with Nuchal Cords and Normal Apgar Scores,” *Journal of Perinatology*, vol. 25, no. 3. Springer Science and Business Media LLC, pp. 162–165, Dec. 16, 2004.

[54] B. Weinberger, M. Anwar, T. Hegyi, M. Hiatt, A. Koons, and N. Paneth, “Antecedents and Neonatal Consequences of Low Apgar Scores in Preterm Newborns,” *Archives of Pediatrics & Adolescent Medicine*, vol. 154, no. 3. American Medical Association (AMA), p. 294, Mar. 01, 2000.

[55] H. C. Lee, M. Subeh, and J. B. Gould, “Low Apgar score and mortality in extremely preterm neonates born in the United States,” *Acta Paediatrica*, vol. 99, no. 12. Wiley, pp. 1785–1789, Jul. 07, 2010.

[56] V. E. Torbenson et al., “Intrapartum factors associated with neonatal hypoxic ischemic encephalopathy: a case-controlled study,” *BMC Pregnancy and Childbirth*, vol. 17, no. 1. Springer Science and Business Media LLC, Dec. 2017.

[57] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639. Springer Science and Business Media LLC, pp. 115–118, Jan. 25, 2017.

[58] J. De Fauw et al., “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nature Medicine*, vol. 24, no. 9. Springer Science and Business Media LLC, pp. 1342–1350, Aug. 13, 2018.

[59] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *arXiv*, 2017

- [60] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1. Springer Science and Business Media LLC, Mar. 19, 2019.
- [61] D. Banik and D. Bhattacharjee, "Mitigating Data Imbalance Issues in Medical Image Analysis," *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance*. IGI Global, pp. 66–89, 2021.
- [62] A. Gupta, N. Tatbul, R. Marcus, S. Zhou, I. Lee, and J. Gottschlich, "Class-Weighted Evaluation Metrics for Imbalanced Data Classification." arXiv, 2020.
- [63] I. J. Goodfellow et al., "Generative Adversarial Networks." arXiv, 2014.
- [64] V. Sampath, I. Mourtua, J. J. Aguilar Martín, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of Big Data*, vol. 8, no. 1. Springer Science and Business Media LLC, Jan. 29, 2021.
- [65] Y.-W. Lu, K.-L. Liu, and C.-Y. Hsu, "Conditional Generative Adversarial Network for Defect Classification with Class Imbalance," 2019 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE). IEEE, Apr. 2019.
- [66] S. Guan, "Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks," *Journal of Medical Imaging*, vol. 6, no. 03. SPIE-Intl Soc Optical Eng, p. 1, Mar. 23, 2019.
- [67] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection," arXiv, 2021.
- [68] R. D. I. Puspitasari, M. A. Ma'sum, M. R. Alhamidi, Kurnianingsih, and W. Jatmiko, "Generative adversarial networks for unbalanced fetal heart rate signal classification," *ICT Express*. Elsevier BV, Jul. 2021.
- [69] Q. Wen et al., "Time Series Data Augmentation for Deep Learning: A Survey," arXiv, 2020

- [70] T. Golany, G. Lavee, S. Tejman Yarden, and K. Radinsky, “Improving ECG Classification Using Generative Adversarial Networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 08. Association for the Advancement of Artificial Intelligence (AAAI), pp. 13280–13285, Apr. 03, 2020.
- [71] A. M. Delaney, E. Brophy, and T. E. Ward, “Synthesis of Realistic ECG using Generative Adversarial Networks.” arXiv, 2019.
- [72] A. M. Shaker, M. Tantawi, H. A. Shedeed, and M. F. Tolba, “Generalization of Convolutional Neural Networks for ECG Classification Using Generative Adversarial Networks,” IEEE Access, vol. 8. Institute of Electrical and Electronics Engineers (IEEE), pp. 35592–35605, 2020.
- [73] K. F. Hossain et al., “ECG-Adv-GAN: Detecting ECG Adversarial Examples with Conditional Generative Adversarial Networks,” arXiv, 2021.
- [74] A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante, and K. Veeramachaneni, “TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks.” arXiv, 2020.
- [75] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training.” arXiv, 2018.
- [76] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs.” arXiv, 2016.
- [77] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic Image Inpainting with Deep Generative Models.” arXiv, 2016.
- [78] L. Weng, “From GAN to WGAN.” arXiv, 2019.
- [79] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, “Are GANs Created Equal? A Large-Scale Study.” arXiv, 2017.
- [80] A. Borji, “Pros and Cons of GAN Evaluation Measures: New Developments.” arXiv, 2021.

- [81] S. Zhou, M. L. Gordon, R. Krishna, A. Narcomey, L. Fei-Fei, and M. S. Bernstein, “HYPER: A Benchmark for Human eYe Perceptual Evaluation of Generative Models.” arXiv, 2019.
- [82] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, “Assessing Generative Models via Precision and Recall.” arXiv, 2018.
- [83] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved Precision and Recall Metric for Assessing Generative Models.” arXiv, 2019.
- [84] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable Fidelity and Diversity Metrics for Generative Models.” arXiv, 2020.
- [85] A. G. Cahill, A. B. Caughey, K. A. Roehl, A. O. Odibo, and G. A. Macones, “Terminal Fetal Heart Decelerations and Neonatal Outcomes,” *Obstetrics & Gynecology*, vol. 122, no. 5. Ovid Technologies (Wolters Kluwer Health), pp. 1070–1076, Nov. 2013.

To Whom It May Concern:

Here, I certify that this research entitled “*A Study on Classifying Fetal Distress from Large-Scale Cardiotocographic (CTG) Data Using Different Machine Learning Approaches*” had conducted by Mohannad Alkanan, and corroborated with our research group. This research result indicates the prospect of establishing strategies that are able to predict fetal distress by cardiotocography in the future. Since the beginning of the study, we were deeply involved in all experiments by addressing the issues we face when dealing with CTG and analyzing the achieved results of each study. We actively communicated with Mohannad through emails and physical/online meetings. Each objective in the thesis tries to answer real-world issues in the field of OBGYN. The choices of features, the processing of the dataset, and the denoising and feature extracting algorithms were applied under medical doctors’ supervision. Algorithms that extract medical diagnosis cases such as tachycardia or prolong deceleration were only implemented after medical doctors’ affirmation. We see Mohannad as a potential contributor to future studies and his work can be used by other researchers to further push research on this topic in the right direction.

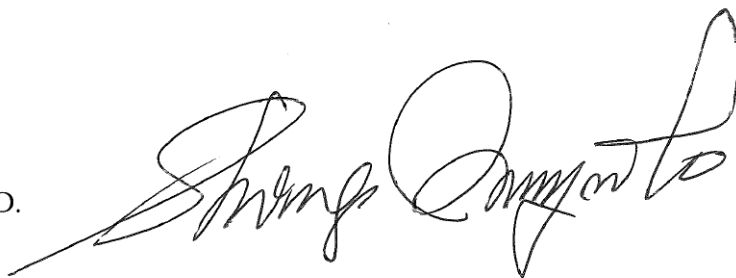
The impact of this thesis helped us achieve an interesting thing and that is higher diagnostic accuracy was acquired with AI compared to humans which used to refer to as about AUC 0.7 clinically. Promotion of mechanization has been undertaken in several fields, also in Medicine, this research surely contributes to the evolution of OBGYN research as well as improvement of fetal prognosis.

Sincerely,

Shingo Miyamoto, M.D., Ph.D.

Chief Professor

Department of Obstetrics and Gynecology, Faculty of Medicine, Fukuoka University

A handwritten signature in black ink, appearing to read 'Shingo Miyamoto', written in a cursive style.