DeepSynthBody: the beginning of the end for data deficiency in medicine

Vajira Lasantha Bandara Thambawita

PhD Programme in Engineering Science
Department of Computer Science
Faculty of Technology, Art and Design
OsloMet – Oslo Metropolitan University

Autumn 2021

OsloMet Avhandling 2021 nr 45

ISSN 2535-471X (trykket)
ISSN 2535-5455 (online)

ISBN 978-82-8364-354-1 (trykket)
ISBN 978-82-8364-381-7 (online)

OsloMet – storbyuniversitetet
Universitetsbiblioteket
Skriftserien
St. Olavs plass 4,
0130 Oslo,
Telefon (47) 64 84 90 00

Postadresse:
Postboks 4, St. Olavs plass
0130 Oslo

Trykket hos Byråservice

Trykket på Scandia 2000 white, 80 gram på materiesider/200 gram på coveret

Preface

I wrote this thesis, titled "DeepSynthBody: the beginning of the end for data deficiency in medicine," to fulfill the requirement for completing my Ph.D. for the Ph.D. program in Engineering Science Faculty of Technology, Art and Design, Oslo Metropolitan University, Oslo, Norway. The total time for thesis was around three years. I carried out my work under the supervision of Professor Michael A. Riegler, Professor Pål Halvorsen, and Professor Hugo L. Hammer. I have completed the thesis in the Department of Holistic Systems in Simula Metropolitan Center for Digital Engineering (SimulaMet), which provided the infrastructure and all the financial support to this full research.

This Ph.D. time became a golden period in my life because I have been exploring the real research world which is not limited to a thesis. As a result, I felt my research works and perceived them, which forced me to learn new things every day until I am writing this preface. In addition to the general responsibilities of my life, I was a responsible person for performing quality research works in the medical domain, which is the field no one can argue the importance of it. I was forced to be responsible for this field because the success of our research can save human life and a fault of our research can indirectly cause death.

I hope that you love this thesis reading.

Vajira Thambawita
May, 2021 at Oslo, Norway

Acknowledgments

First, I would like to thank my principal supervisor and two co-supervisors, Michael Alexander Riegler, Pål Halvorsen, and Hugo Lewi Hammer, for their support, motivation, and always behind me. Without them, this would be only a dream. After joining the HOST department as a Ph.D. candidate in 2018, I started experiencing a completely new environment with new people from different countries and cultures.

Michael Riegler became my principal supervisor. I did not know anything about him despite his academic background. However, after few weeks, I realized that he is more than my principal supervisor for my life. Within few months, he became a game-changer in my life. I do not have words to express his qualities and how his advice, encouragement, kindness, and motivations are important to my life. Therefore, I would like to give my most enormous thanks to my primary supervisor, Michael, who is always behind me to support my academic life and get advice for my personal life.

Pål Halvorsen is the department head and one of my co-supervisors, and he is a very kind person supporting us always silently. His advice and encouragements make me better and better every day in my academic life. Then, I would like to thank Pål for his kindness showed me through my Ph.D. journey and the wordless help given to me. Hugo Hammer is my second co-supervisor who supports me in handling advanced statistical problems in my Ph.D. research. Then, I would like to thank Hugo for his friendly help, as always.

The research has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which the Research Council of Norway financially supports under contract 270053. Tore was the person who helped me to use these infrastructures. So, I would like to give my special thanks to him for his outstanding support.

I also would like to thank the closest colleague, Steven Hicks, a good friend met within the department and in my life. Moreover, I thank all the other colleagues, Debesh, Hanna, Pia, Hakon, and Daniel, who work closely in my Ph.D. periods. There are more people

I want to thank, but writing them all here is not possible. However, they all are equally important for me, and therefore I would like to thank them all equally. Master students, other Ph.D. students from different departments and other counties, all the co-authors, and persons who are from the administration are a few of them.

Finally, I would like to thank my wife, Shalike, who sacrificed the freedom of her life for finishing my Ph.D. successfully. Not only that, she always provided me food and other necessary stuff without asking anything from me. Last but not least, I would like to give my thanks to my parents, who sacrificed their whole life to provide us a good life while facing a challenging lifestyle in my country, Sri Lanka.

Abstract-English

Recent advancements in technology have made artificial intelligence (AI) a popular tool in the medical domain, especially machine learning (ML) methods, which is a subset of AI. In this context, a goal is to research and develop generalizable and well-performing ML models to be used as the main component in computer-aided diagnosis (CAD) systems. However, collecting and processing medical data has been identified as a major obstacle to produce AI-based solutions in the medical domain. In addition to the focus on the development of ML models, this thesis also aims at finding a solution to the data deficiency problem caused by, for example, privacy concerns and the tedious medical data annotation process.

To accomplish the goals of the thesis, we investigated case studies from three different medical branches, namely cardiology, gastroenterology, and andrology. Using data from these case studies, we developed ML models. Addressing the scarcity of medical data, we collected, analyzed, and developed medical datasets and performed benchmark analyses. A framework for generating synthetic medical data has been developed using generative adversarial networks (GANs) as a solution to address the data deficiency problem. Our results indicate that our generated synthetic data may be a solution to the data challenge. As an overarching concept, we introduced the DeepSynthBody as a basis for structured and centralized synthetic medical data generation. The studies presented in the thesis, such as generating synthetic electrocardiograms (ECGs), gastrointestinal (GI)-tract images and videos with and without polyps, and sperm samples, showed that DeepSynthBody can help to overcome data privacy concerns, the time-consuming and costly data annotation process, and the data imbalance problem in the medical domain. Our experiments showed that we can generate realistic synthetic data providing comparable results to experiments using real data to tackle the identified problems. The final DeepSynthBody framework is available as an open-source project that allows researchers, industry, and practitioners to use the system and contribute to future developments.

Abstract-Norwegian

Teknologiske fremskritt har gjort kunstig intelligens til et populært verktøy innen medisin. Spesielt metoder innen maskinlæring, en underkategori av kunstig intelligens, er mye brukt. Et mål i denne fobindelse er å utvikle gode, generaliserbare modeller for bruk i systemer for datamaskinassistert-diagnose, men en stor utfordring her er innsamling og behandling av medisinske data på grunn av for eksempel personvernhensyn og kostbare annoteringsprosesser. Denne oppgaven fokuserer derfor både på utvikling av maskinlæringsmodeller og å finne en løsning på problemet med manglende medisinske data.

For å nå oppgavens mål har vi undersøkt tre forskjellige medisinske eksempler, nemlig kardiologi, gastroenterologi og andrologi. Ved hjelp av data fra disse medisinske områdenehar vi utviklet maskinlæringsmodeller. For å løse mangelen på medisinsk data, har vi samlet inn, analysert og utviklet medisinske datasett, og vi har utført referanseanalyser. I tillegg, et rammeverk for generering av syntetiske medisinske data er utviklet ved hjelp av "generative adversarial networks" for å løse problemet med datamangel, hvor resultatene våre indikerer at slike genererte data kan være en mulig løsning. Som et overordnet konsept introduserer vi DeepSynthBody som grunnlag for strukturert og sentralisert generering av syntetisk medisinsk data. Studiene presentert i oppgaven, slik som generering av syntetiske elektrokardiogram, bilder og videoer fra tarmsystemet og sædprøver, viser at DeepSynthBody kan bidra til å overvinne personvernproblemer, redusere tid og ressursbruk innen dataanmerkingsprosessene, og utjevne problemene med data ubalanse innen det medisinske domenet. Våre eksperimenter viser at vi kan generere realistiske syntetiske data som gir sammenlignbare resultater med eksperimenter hvor man bruker reelle data. Det endelige DeepSynthBody-rammeverket er tilgjengelig som et åpent kildekode-prosjekt som gjør det mulig for både forskere og industri å bruke systemet og å bidra til fremtidig utvikling.

Contents

Acronyms

1-D one-dimensional. 16, 28, 46–49, 51

2-D two-dimensional. 16, 28, 29, 46–49, 53, 54

3-D three-dimensional. 16, 28, 29, 46–49, 53, 55

4-D four-dimensional. 16, 28, 100

AI artificial intelligence. vii, 5–7, 9, 10, 12–14, 16, 22, 24, 27, 39, 41, 42, 50, 61, 89

CAD computer-aided diagnosis. vii, 6–8, 13–15, 17–25, 27, 29, 39–41, 57, 58, 87, 91–94, 96–98, 100–103

CNN convolutional neural network. 56–60, 62, 66, 69

DL deep learning. 5, 6, 13, 35, 69, 77, 79

DNN deep neural network. 5, 9, 34

ECG electrocardiogram. vii, 9, 15, 19–22, 24, 28, 43, 46–48, 51–53, 56, 57, 61–65, 78, 83, 84, 88, 93–95, 97, 98, 100, 102

EHR electronic health records. 37, 38, 40

FID Frechet inception distance. 68

GAN generative adversarial network. vii, 14–19, 21–24, 27, 33–41, 47, 50, 53, 54, 61, 62, 64–71, 73, 75, 77–79, 81–84, 88, 89, 91, 92, 94–97, 100–103

GDPR general data protection regulation. 10

GI gastrointestinal. vii, 8, 15, 20–22, 24, 31, 32, 43, 46, 47, 51, 54, 57–59, 61, 62, 65–69, 73, 78, 83–85, 88, 93–96, 100, 101

GMCNN generative multi-column convolutional neural networks. 70

GUI graphical user interface. 61, 79

IOU intersection over union. 33

MAE mean absolute error. 33, 57

MCC Matthews correlation coefficient. 33, 58

ML machine learning. vii, 5–9, 12–15, 17–25, 27, 29–33, 39, 40, 49, 50, 55–60, 66, 79, 84, 87, 91–95, 97, 98, 100–103

MRI magnetic resonance imaging. 16, 29, 43, 48, 81, 100

MSE mean square error. 33

N-D N-dimensional. 47, 48

PyPI Python package index. 24, 77, 97, 102

RMSE root-mean-squared error. 33, 57

VAE variational autoencoder. 33, 34

XAI explainable artificial intelligence. 6, 13, 84, 87

Chapter 1

Introduction

> "*The data-driven world will be always on, always tracking, always monitoring, always listening and always watching – because it will be always learning*" (Rydning [1]).

Artificial intelligence (AI) has become a popular tool in most of the main industries, for example, financial service [2, 3], manufacturing [4, 5], media and entertainment [6, 7], transportation [8, 9] and healthcare [10, 11]. As a result, AI interacts more closely with the day-to-day life of people. While AI has many definitions, the main goal of AI today is to enable faster, more reliable, and more accurate data analysis. Additionally, AI applies to the tasks which cannot be proceeded by humans, such as operations in space, in deep oceans, or deep underground. These AI applications are successful as a result of improvements in machine learning (ML) algorithms [12] used in AI, particularly deep learning (DL) [13], as well as great advances in computational hardware running the compute-heavy ML algorithms, such as deep neural networks (DNNs). Despite such advancements, the algorithms need data to learn. The limited availability of data to train the ML algorithms [14, 15] is a crucial factor in developing successful AI solutions in all domains. The interconnections between the terminology, AI, ML, and DL used in this section are depicted in Figure 1.1.

With the success of applying AI as a tool in the main industries, using AI in the medical domain has received more attention in the recent decade, for example, seen in the news headings[1] and quotes[2] about AI and medicine presented in Figure 1.2. These news

[1] `https://futurism.com/ai-medicine-doctor`
[2] `https://news.harvard.edu/gazette/story/2020/11/`

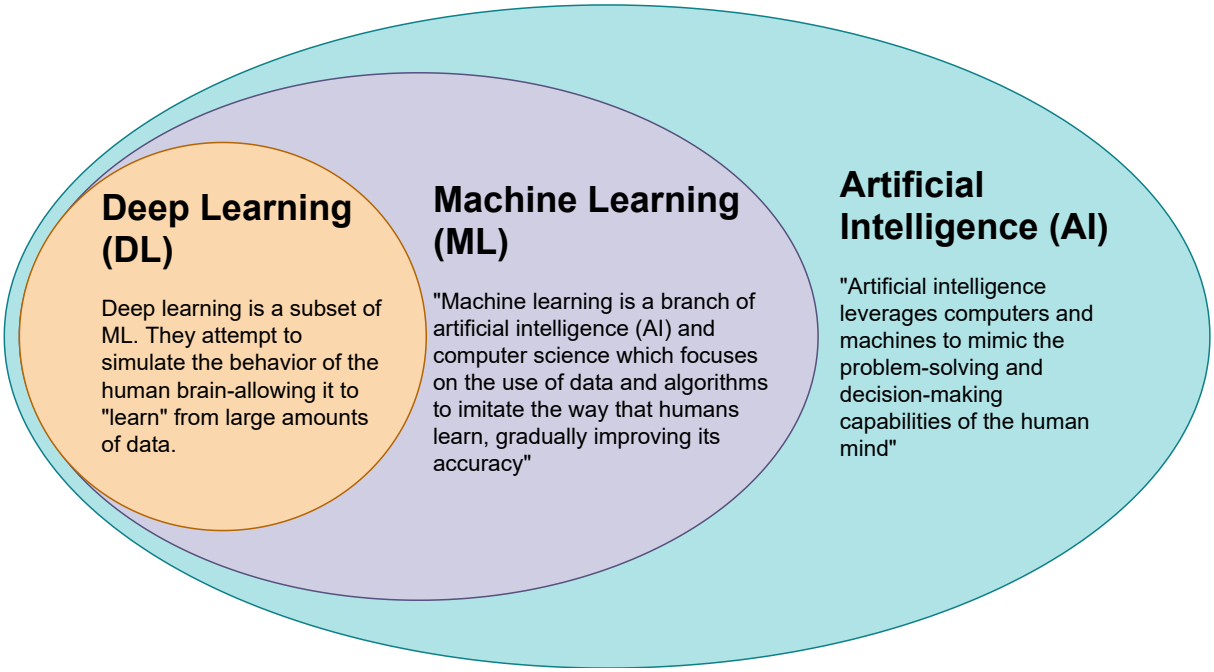


Figure 1.1: Definitions [16] and relations between AI, ML and DL.

shows contradictory ideas about AI in medicine, such as some believe that AI will replace byhuman doctors and others believe that AI will "just" become a supportive tool for human doctors. Nevertheless, it seems like many believe that AI will become more popular in the coming years. Thus, applying AI in medicine is important because it may directly affect humans' personal lives, and successful medical systems are directly correlated with life expectancy and quality. Therefore, producing AI systems with reliability and integrity is important in the medical domain. To understand the process of applying AI in medicine for developing computer-aided diagnosis (CAD) systems, we should understand the full medical AI pipeline. A simplified version of this pipeline is depicted in Figure 1.3 with four steps: (I) collecting data, (II) annotating data using experts, (III) applying ML methods, and (IV) final product and explainable artificial intelligence (XAI). These four steps are discussed further in the next section.

1.1 Background and Motivation

AI-based solutions are used in the medical domain for different purposes, such as to develop treatment protocols, drugs, personalized medicine, patient monitoring systems, robotics, and diagnosis processes [11]. Among these, AI-based diagnosis processes or CAD systems [17] got more attention from AI researchers. CAD systems aid doctors as the

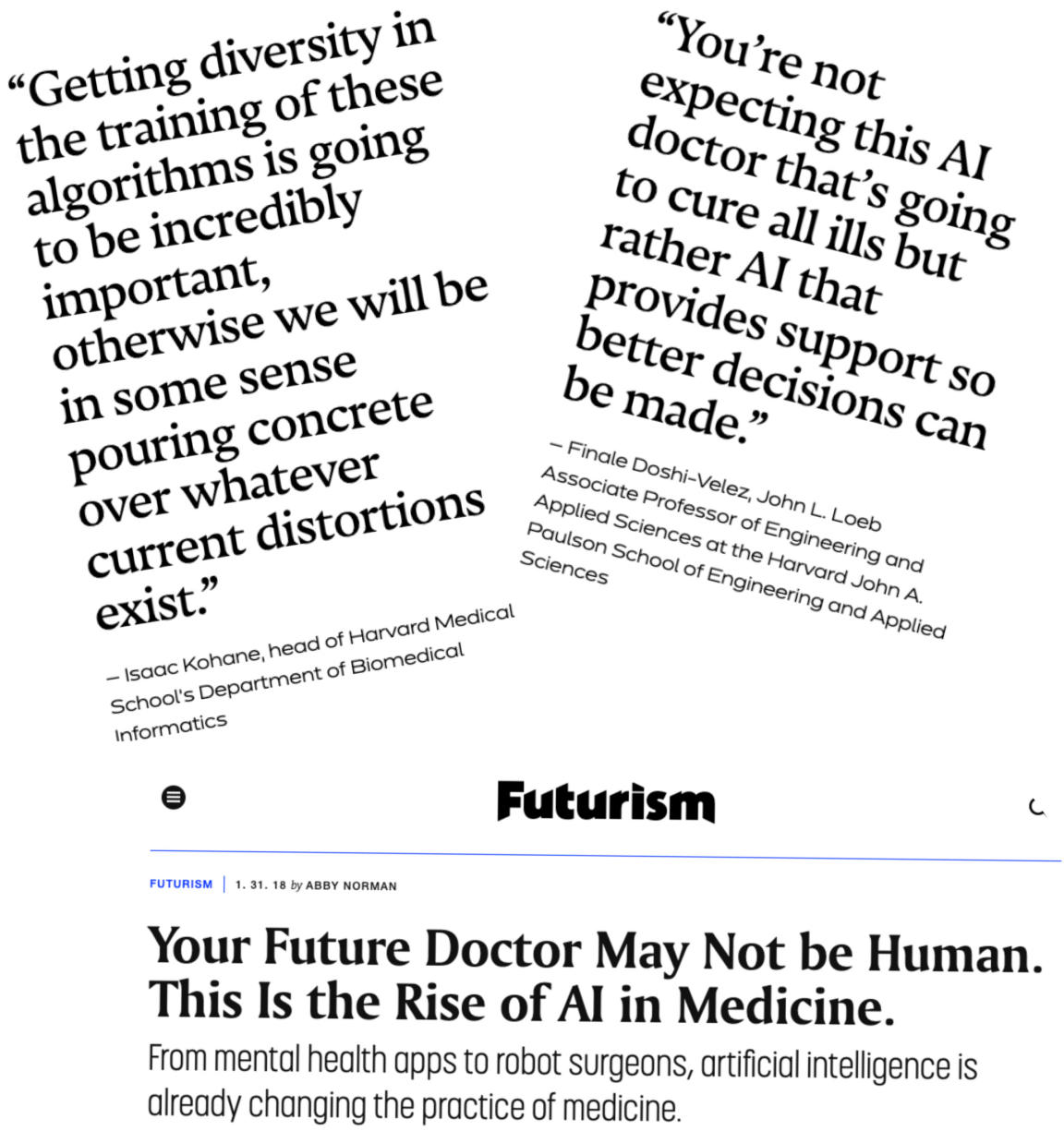


Figure 1.2: Some quotes and headings about AI and medicine in news articles

"second opinion" to finalize decisions.

In this regard, we started to research ML-based solutions for CAD systems by following the above four steps pipeline to help medical experts more correctly and efficiently detect anomalies in medical data from real examinations to save lives ultimately . The goals were to both address large miss-rates [18, 19, 20] and observer variations [21, 22]. The process of researching and developing ML solutions is presented using Step III (Figure 1.3). However, we soon realized a huge lack of medical data to develop good ML models in the domain for various reasons, increasing the importance of the first two steps in Figure 1.3.

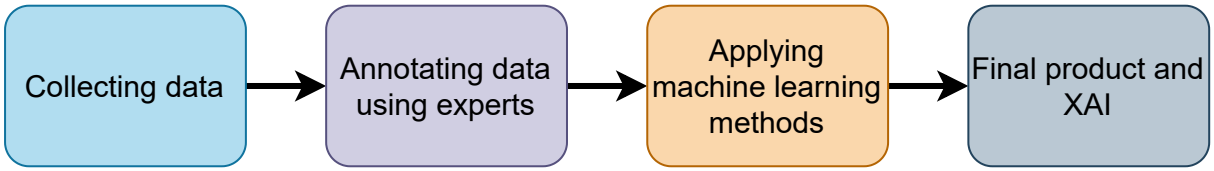


Figure 1.3: The main four steps of applying ML solution in the medical domain.

Therefore, we have studied how datasets should be collected, composed, and published as open datasets. Within the three years of Ph.D. time, a total of seven datasets [23, 24, 25, 26, 27, 28, 29] were successfully collected and published. In these datasets, medical experts labeled or annotated data (Step II), but not all the datasets because the annotation process is costly and time-consuming. For example, our gastrointestinal (GI)-tract dataset [23] has labeled images and pixel-wise annotated polyp images, performed by experienced colonoscopists. However, the biggest part of the GI-tract dataset is still unlabelled data because of the costly and time-consuming data annotation process. We analyzed three branches in medicine, *gastroenterology*, *andrology*, and *cardiology* in parallel to the data collection process. The main motivation for choosing different domains was to show that our methods can work on different problems (are generalizable) and to produce ML-based CAD solutions to help experts by providing more efficient and accurate automated assistance for their tasks.

In the gastroenterology branch, classification models [30, 31, 32, 33, 34] to classify GI-tract findings and segmentation models [35, 36] to segment polyp regions were investigated. When producing these ML solutions, we identified that generalizability is one of the main issues for both classification and segmentation due to the lack of labeled and annotated data to train ML models. The classification models introduced in our studies [30, 31] showed good performance when the validation and testing data are a subset of the same dataset used to prepare the training dataset. However, the performance of the best models showed poor performance for completely new datasets collected from different hospitals. The problem was caused as a result of the over-fitting [37]. In addition to the data bias problem, we also identified that an imbalanced number of images of different classes makes less accurate ML models. Detailed discussion on this issue can be found in [31], where we analyzed and experimented with different datasets. Similar to the classification models, we noticed that polyp segmentation models show poor performance due to small datasets to train segmentation models. We tried to solve the problem by

introducing a novel data augmentation method called PYRA[3] [36] and introducing a novel segmentation model called DivergentNets [35]. However, we had only small datasets to train segmentation models compared to the training datasets used in classification models. Researchers or data providers usually provide only small segmentation datasets for medical image segmentation tasks due to the time-consuming and costly pixel-wise image annotation process. The medical image annotation process is more challenging than the general image annotation process because experts of the specific medical domain should perform these manual segmentations or review them, and these experts are often rare or have not much time.

In addition to providing ML solutions in gastroenterology, we have investigated ML solutions [38, 39, 40] to predict motility and morphology level of sperm samples which are videos recorded using microscopic analysis. These research works are considered under the andrology branch. The proposed models show acceptable performance, but those performance values were insufficient to use the solution practically. By researching ML solutions to predict motility and morphology levels of sperm samples, we noticed that our models could be improved if we can prepare pixel-wise annotated datasets to perform segmentation before predicting morphology and motility levels. However, performing pixel-wise annotations for a sperm-like medical dataset is a complicated problem for experts because of having hundreds of sperms in a single frame of the dataset. A possible solution is annotating sperms using an unsupervised way and processing those annotated sperm samples to find motility and morphology levels.

In cardiology, we built an electrocardiogram (ECG) analysis system [41] using ML models to predict the properties of ECGs. This experiment used a big ECG dataset to train the ML models and showed that the ML models could outperform expert's analyses. Unfortunately, the dataset used to train our models is a private dataset, and publishing them to reproduce our solutions is not possible due to privacy concerns. In this context, we noticed that there should be a way for omitting privacy concerns. In this ECG study, we have presented an explainable AI mechanism called gradcam [42] to find the most important regions for DNNs to predict the properties of ECGs. However, we could use only the explainable methods that do not expose the real dataset to the public because of privacy concerns. Suppose we have a method to omit and work around the privacy

[3] `https://vlbthambawita.github.io/PYRA/`

concerns. In that case, we can use any explainable method which uses the real dataset, for example, to explain using examples [43].

The success of AI solutions in medicine is highly dependent on the data to train the AI algorithms. However, collecting and sharing medical data is harder than for other general data because of reasons like the privacy restrictions attached with the medical data. The collection of medical data (Step I) is presented using the first box in Figure 1.3. If the training data cannot provide useful information to AI algorithms, the algorithms become less accurate and generalizable. Therefore, medical data is essential for developing successful AI solutions. However, medical data collection and preparation are not straightforward. The unrolled cumbersome internal process of Step I is presented in the first seven steps depicted in Figure 1.4, as discussed by Willemink et al. [44]. However, following these steps is a complex task because of privacy concerns such as ethical approval and data de-identification process, in addition to the data preparation process. Medical data need post prepossessing because the raw medical data producing from medical instruments are not designed for sharing. A lot of research discusses the protection of digital data in a learning health system [45], the privacy of big medical data [46, 47, 48], and making a balance between health data access and privacy [49]. These research discussions show the importance of considering privacy rules and regulations with health data. As a result, the privacy restrictions applied with the medical data make the process in Step I harder and slow down the whole pipeline depicted in Figure 1.3.

The rules and regulations for producing open access medical data vary from country to country and region to region according to data protection regulations introduced in the specific regions. For example, Norway should follow the rules given by the Norwegian data protection authority (NDPA) [50] and enforce the personal data act [51] in addition to following general data protection regulation (GDPR) [52], which is the common guideline for European countries. While there is no central level privacy protection guideline in the US like GDPR in Europe, rules and regulations in the US are coming through other US privacy laws, such as Health Insurance Portability and Accountability Act (HIPAA) [53] and California Consumer Privacy Act (CCPA) [54]. In Asian counties, they follow their own set of rules country-wise, such as Japan's Act on Protection of Personal Information [55], South Korea's Personal Information Protection Commission [56], and the Personal Data Protection Bill in India [57]. If researchers can perform research

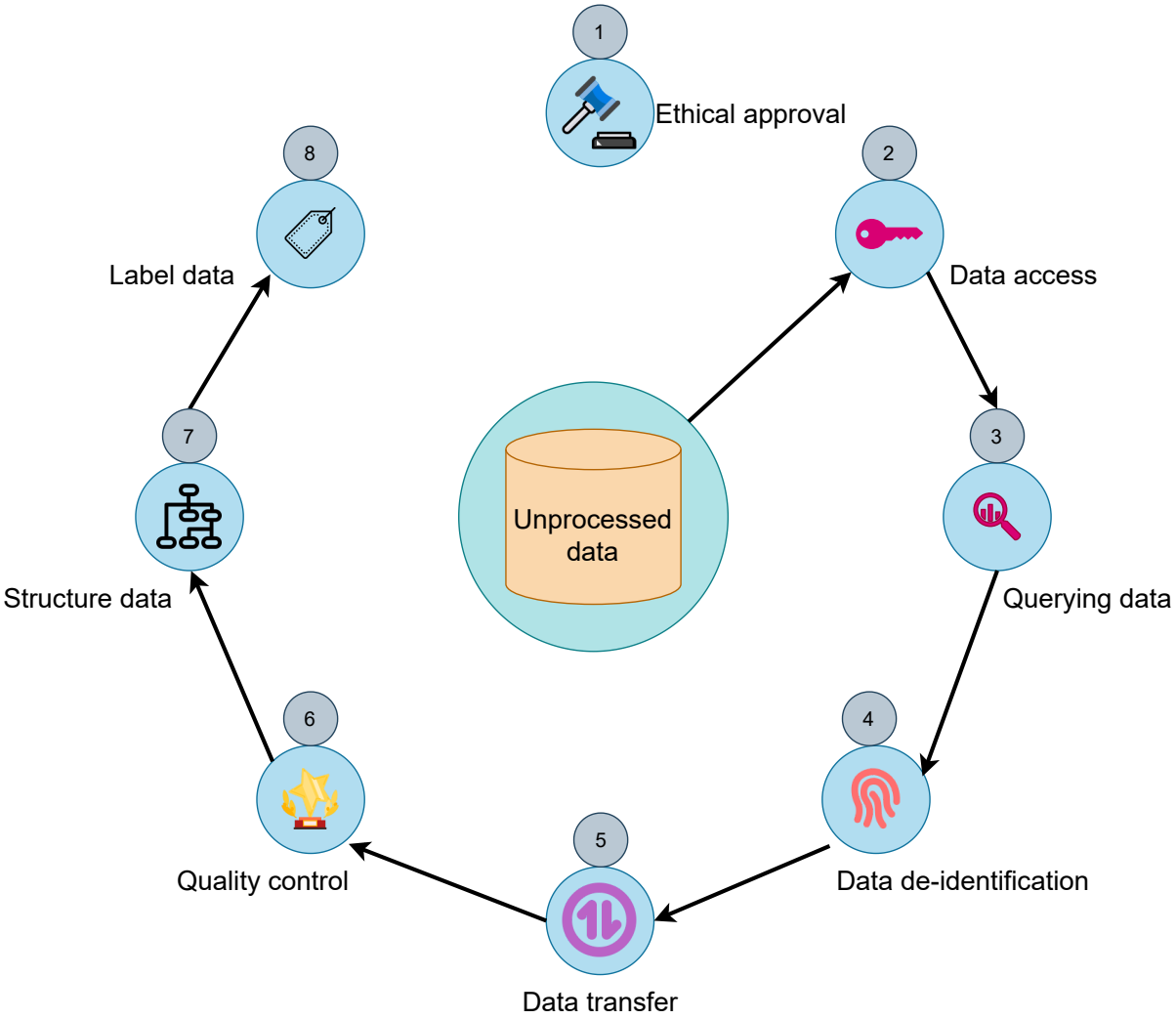


Figure 1.4: Medical data preparation process as discussed in [44]

with these privacy restrictions, the papers published are often theoretical methods only. As a consequence, the results of those studies are not reproducible, and fair and correct comparisons between methods are hard to achieve. All these consequences are due to a lack of available data and sharing restrictions. Furthermore, universities or other research institutes that use medical domain data for teaching purposes use the same medical domain datasets for years, which affects the quality of education. Therefore, data sharing restrictions resulting from privacy protocols are identified as one of the main problems and obstacles, and we have researched to address this challenge in this thesis.

In addition to the privacy concerns, the cost of medical domain experts for extracting useful information from medical data is another obstacle to producing big datasets, which are helpful for AI. This stage is presented as the second box in Figure 1.3 and task number 8 in Figure 1.4. For example, to train the most common supervised ML techniques, ground truth data are needed. In other words, annotated datasets are essential. Because of this necessity of annotated data, new companies and job opportunities are opened to perform data annotations for datasets used to train AI algorithms [58, 59]. For example, the pricing list in Google for annotating datasets is presented in Table 1.1. However, medical data annotation (or producing ground truth) is not easy as making ground truth for general datasets. Medical data annotation is more challenging than other general data annotations because only the experts in the medical domain can perform the annotations fully trustable in terms of correctness. If the data annotation by experts is not possible, the experts should do at least a review process to make the annotations trustable before using them in AI algorithms. The importance of having accurate annotations from experts for medical data is, for example, discussed by Yu et al. [60] using a mandible segmentation dataset of CT images. Because only the medical experts can accurately do the medical data annotation process, the expert annotation process becomes expensive. Additionally, this annotation process takes considerable time to produce ground truth data precisely [44], consuming time that clinicians usually rather spend on treating patients.

The third step in Figure 1.3 represents applying ML methods after collecting medical data and annotating the data using domain experts. However, due to privacy protocols and the aforementioned complex data retrieval and annotation problems, researchers and industry, who apply ML solutions for medical data, do not have access to open-access expert-annotated datasets. Because of this limited data problem, the models become less

Table 1.1: Google labelling cost (to date: 05-05-2021). [61]

Data type	**Objective**	**Unit**	**Tier 1**	**Tier 2**
Image	Classification	Image	$35	$25
	Bounding box	Bounding box	$63	$49
	Segmentation	Segment	$870	$850
	Rotated box	Bounding box	$86	$60
	Polygon/polyline	Polygon/Polyline	$257	$180
Video	Classification	5sec video	$86	$60
	Object tracking	Bounding box	$86	$60
	Event	Event in 30sec video	$214	$150
Text	Classification	50 words	$129	$90
	Entity extraction	Entity	$86	$60

reliable [31] (as a result of poor generalizability) and have fewer functionalities such as limited interpretability [62]. These limitations and our own experience of developing ML models for CAD systems emphasize the requirement of having an alternative fast track to getting medical data into the third step (Step III) of applying ML.

The fourth step in Figure 1.3 represents the final stage of producing products using ML to use in clinical settings. In this stage, explaining the prediction results (XAI) is an important step because it is the only step in which one can convince doctors to accept decisions made by ML solutions. Explanation by example is currently a preferred XAI method by non-experts [63]. Privacy issues can limit these XAI functionalities, such as explaining DL solutions by examples [64], when the example data is restricted to publish.

In summary, the problems related to collecting and processing medical data can be identified as a major bottleneck to produce enough open-access medical data for developing well-performing ML solutions to be used with CAD systems. The privacy concerns with the medical data and the costly and time-consuming medical data annotation process are two reasons for the data deficiency problem. In addition, we identified that a lack of true-positive data compared to true-negative data in the medical domain, giving large class imbalances, is a problem for producing AI-based systems. In this regard, this thesis focus on producing well-performing ML models for CAD systems after finding a way to tackle the data deficiency problem by generating synthetic data using a new concept and the framework named DeepSynthBody.

1.2 Research Question and Objectives

The main overall goal of our research is to investigate and develop accurate, generalizable, and well-performing ML models for CAD systems for biomedical applications assisting doctors in clinical practice. In this thesis, we have a particular focus on the problems and challenges coming from medical data. These challenges of collecting and processing medical data, identifying that the lack of medical data due to, for example, privacy issues, resource-consuming data annotation processes, and data imbalance problems are major obstacles for AI-based medical technology research and development. Therefore, we focus on researching a way to address the data deficiency problem in the medical domain while researching and developing well-performing and generalizable ML models for CAD systems for selected three domains as case studies. The overall research question for this study therefore is:

> **What are the problems that emerge from data in computer-aided diagnosis systems, and how can these problems be tackled?**

After identifying the research question, we have defined the objectives of this thesis as follows:

- **Main objective**: Research and develop ML models which are the main component of CAD systems for different medical applications with a focus on the problems of limited availability of biomedical data.
- **Sub-objective I**: Research and develop ML models for CAD systems to assist doctors.
- **Sub-objective II**: Collect, research and develop datasets to be used for developing ML models for CAD systems for biomedical applications.
- **Sub-objective III**: Research and develop benchmark analysis with the medical datasets to identify the problems for producing well-performing ML solution in medical domain.
- **Sub-objective IV**: Research and develop deep generative adversarial networks (GANs) which can produce synthetic data to address the data deficiency problem which is the major obstacle for developing medical AI-based solutions.

This thesis has used three different medical case studies for Sub-objective I, Sub-objective III, and Sub-objective IV. The medical fields chosen are *cardiology*, *gastroenterology*, and *andrology*. We chose these three domains since they are diverse from each other in terms of data. In Sub-objective II, we have introduced additional datasets in addition to the main three case studies as its main goal is collecting and developing medical datasets.

1.3 Scope and Limitations

This research was started to developing well-performing and generalizable ML models for CAD systems to assist doctors. However, the early identification that the medical data is a major obstacle for developing ML models, solving the data deficiency problem in the medical domain became another objective of this thesis. Therefore, in this thesis, two major development streams can be seen. One is developing ML models for CAD systems, and one is researching and developing GANs to overcome the data deficiency problem. As the main finding of this thesis, we could introduce a novel concept and the framework based on GANs to tackle the data deficiency problem. The framework has been demonstrated with a few selected case studies as a proof of concept. However, the novel concept and the framework are not limited to the presented case studies. All other possible research areas using our concept and framework are discussed in the future work section.

In this thesis, three types of datasets were used. In particular, we have used ECG signals, GI images, and a sperm video dataset as case studies that cover three different medicine branches: gastroenterology, andrology, and cardiology. These three datasets were selected because they were the initial studies used to develop ML models for CAD systems. Additionally, the same datasets were used as proof of concept to demonstrate the potentials of the new concept, and the framework introduces as a solution to the data deficiency problem in the medical domain. It is worth mentioning that the new concept is also developed as a big open-source project planning to have contributions worldwide. Therefore, all the case studies and experiments were performed just to prove the new concept. The ECG dataset covers biomedical signal data in the selected case studies, while the GI image datasets cover biomedical images. The sperm dataset is

related to medical video data as well as medical images. In addition to time restriction, the scope of this study is limited to selected data formats such as one-dimensional (1-D), two-dimensional (2-D), and three-dimensional (3-D) because of limited access to other types of medical data such as magnetic resonance imaging (MRI) which are considered four-dimensional (4-D) with a temporal dimension.

The proposed concept consists of a four-step pipeline. These are collecting real data and analysis, developing generative models, generating synthetic data, and explainable DeepSynth AI and DeepSynth Explainable AI. While the thesis covers the first three, the most important steps, data handling, applying GANs, and producing synthetic data via the end functionalities, the last step of researching explainability is not investigated under this thesis due to time limitations and is regarded as an important future research direction. Additionally, we have published an online platform for the concept. This online platform will be changed in the future as a result of improvements over time.

1.4 Research Methodology

In computer science, it is harder to practice traditional research methodology followed by classic sciences as described by Dodig-Crnkovic [65] because computer science can be identified as a combination of various scientific disciplines. In sciences, we can identify three paradigms, theory, abstraction, and design [66]. Generally, the theory is for mathematical sciences. The abstraction or modeling is for natural sciences. The design or experimentation is for engineering. However, it is not easy to explicitly map computer science for one of these three paradigms. While these three are inseparable from computer science, they are distinct from each other. Therefore, we define this thesis work in each of the above paradigms as follows.

- **Theory:** Major elements of the theory of the concept introduced in this thesis consist of the major theories related to AI introduced in the report [66] produced by the task force of ACM and IEEE. This report has introduced four steps to developing a coherent, valid theory in any science. They are:

    1. Characterize objects of study (definition).
    2. Hypothesize possible relationships among them (theorem).

3. Determine whether the relationships are true (proof).
4. Interpret results.

In this regard, we have introduced our main objective and four sub-objectives to research ML models for CAD systems in the medical domain and a novel concept to overcome the data deficiency problem. We hypothesize that generative models can generate synthetic data to overcome the data deficiency problem of developing ML models in the medical domain. Using three different case studies, we have presented the performance of our ML models. Moreover, using the same case studies, we proved how to use GAN-generated synthetic data to solve the data obstacles in the medical domain.

- **Abstraction (modeling):** is defined based on the experimental scientific methods. In the ACM report, they have described four stages for investigations of phenomenons such as:

  1. Form a hypothesis.
  2. Conduct a model and make a prediction.
  3. Design an experiment and collect data.
  4. Analyze results.

  According to this modeling paradigm, deep generative models can be identified as the main component of modeling our hypothesis. Under different medical data formats, we analyzed generative models and collected synthetic data. To find the best generative models for generating synthetic data, we have studied them qualitatively and quantitatively using experimental prototypes. Not only deep generative models, but we have also experimented with baseline experiments and benchmark experiments, which were performed to develop experimental prototype ML models for CAD systems.

- **Design:** In this paradigm, four stages can also be identified to build a system to solve a specific problem. They are

  1. State requirements.
  2. State specifications.

3. Design and implements the system.
4. Test the system.

The medical data was identified as a key requirement to research and design well-performing ML models for CAD systems. Therefore, real medical datasets and synthetic medical datasets were collected and developed. Then, we designed ML models using the real medical datasets and synthetic medical datasets. Moreover, a complete framework to generate synthetic data in the medical domain was introduced and implemented. We have tested our ML models, and GANs introduced in the framework using three different case studies.

1.5 Contributions

The research in this thesis contributes to the area of medical AI technology aimed to assist clinicians in their daily work, improving the quality of the health care systems. We started to research and develop ML models for CAD systems using small existing datasets and collecting our medical datasets, where the developed models performed very well. However, the major challenge identified was the data deficiency problem, where dataset development was cumbersome due to various reasons. This challenge then becomes the major challenge addressed in this thesis while still developing ML models.

In particular, in this thesis, four sub-objectives were introduced to accomplish the main objective, which aims to develop ML models for CAD systems to assist doctors in improving the efficiency of diagnosis. These four sub-objectives were initiated to develop well-performing ML models and solve the data deficiency problem of the current applied machine learning pipeline used in the medical domain, as depicted in Figure 1.3. We started researching and developing ML models for CAD systems to achieve Sub-objective I. Then, in Sub-objective II, collecting data was initiated after finding that data is an important factor for achieving Sub-objective I. Then, the performing benchmark experiments are mainly used to achieve Sub-objective III to study the medical datasets to understand the related problems to research and address in Sub-objective IV. Sub-objective IV was achieved by experimenting and investigating GANs to generate synthetic data to overcome the data deficiency problem in the medical domain. Figure 1.5 shows all the contributions via these four sub-objectives and the main-objective. Some of the contributions can be

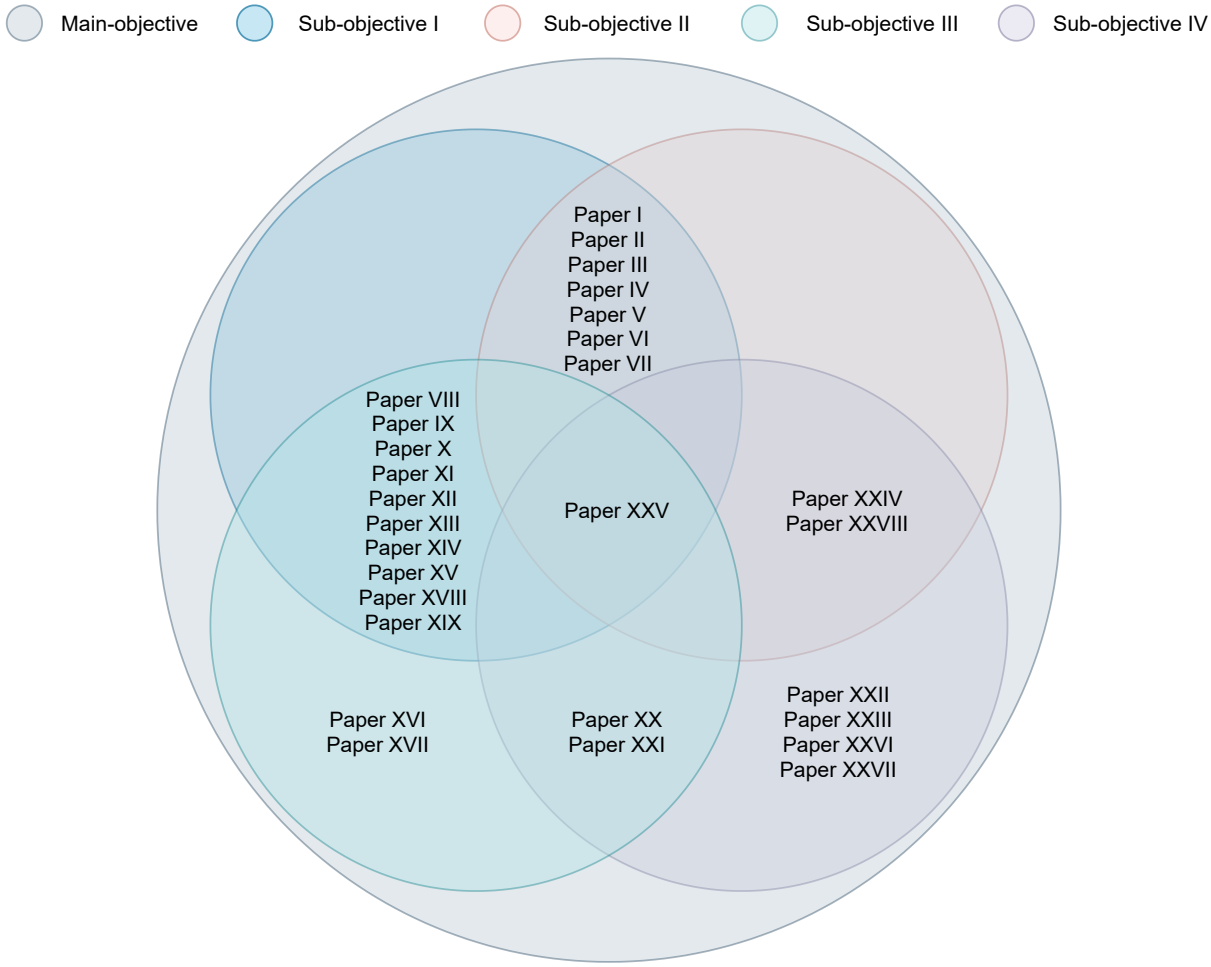


Figure 1.5: Paper-wise contribution to all objectives.

identified through two or more objectives, while all the contributions are directly attached to achieve the main objective.

The following bullet points show all contributions to sub-objectives and the main objective. Within these contributions, dataset papers, ML-based CAD models or benchmark papers, and GAN-related papers can be found. The dataset paper, HyperKvasir [23][4], got much attention from the research community within a short period because of the richness of data diversity. Not only that, the results of most benchmark papers were within the top 5%. For example, we won the 1^{st} place for the EndoCV grand challenge[5] 2021. Similarly, GAN-based experiments also became popular within a short period in the research community because of the competitiveness of the presented qualitative and quantitative results of novel methods used to generate synthetic data. For example, The DeepFake ECG paper was read by many people within a few days after publishing the

[4] `https://www.nature.com/articles/s41597-020-00622-y/metrics`
[5] https://endocv2021.grand-challenge.org/

pre-print, and it became a part of news heading about recent developments of interests in cardiovascular medicine[6]. The following section discusses all the contributions toward the objectives of this thesis. The main objective is discussed at the end of the following list to emphasize how sub-objectives contribute to accomplishing the main objective.

- **Sub-objective I**: The main focus of this sub-objective is to research and develop well-performing ML models for CAD systems to assist doctors. As case studies, we have selected three branches of medicine. These are cardiology, gastroenterology, and andrology. In gastroenterology, images collected from colonoscopies were the main data stream to apply ML algorithms which are the core algorithms in CAD systems. Several classification models [30, 31] and segmentation models [35, 36] were researched and implemented for the gastroenterology branch under this thesis in different timeline stages. In addition to real data, we used synthetic data with segmentation models [67] used to predict polyps in GI-tract data. Similarly, ML-based regression models were investigated and developed for the andrology branch [38, 39, 40, 68]. For the cardiology branch, an ML-based ECG analysis system [41] was researched and implemented. Moreover, all the dataset papers [23, 24, 25, 26, 27, 28, 29] introduced ML models as baseline experiments which can be considered initial models for developing CAD systems.

- **Sub-objective II**: The main task of this sub-objective is to collect and produce medical datasets, which is identified as the main bottleneck for developing ML-based CAD systems. Moreover, these datasets are the main assets for initiating the novel concept and the corresponding framework, DeepSynthBody, introduced in this thesis. Different types of real medical datasets [23, 24, 25, 26, 27, 28, 29] were collected and published to the research community with the baseline experiments under this thesis to accomplish the sub-objective I,. All the datasets contribute to designing ML models for CAD systems (sub-objective I) because of the baseline experiments introduced in every dataset paper.

  In addition to our datasets, two additional datasets were used from outside of the dataset contributions. One is an ECG dataset, which is a private medical signal dataset. The second one is a sperm dataset [69] which represents sperm video data.

[6] `https://www.medpagetoday.com/cardiology/`

The additional datasets were selected to design ML models for CAD systems in completely two different branches: cardiology and andrology. At the end of the thesis, we showed using synthetic datasets to overcome the data deficiency problem. These synthetic datasets, which consist of a synthetic ECG dataset [70], a synthetic GI-tract landmark dataset [71] and, a synthetic polyp dataset [67] generated using the GAN models introduced as a result of our new concept and the corresponding framework.

- **Sub-objective III**: Initially, we focused on designing generalizable ML models, which are the core of CAD systems to achieve Sub-objective I. Later, we identified that the medical data deficiency in training ML models should be tackled. We have performed benchmark analyses with selected three medical datasets to investigate the data-related problems and investigate them. As a result, a set of benchmark articles for the selected datasets as case studies were published to achieve the benchmark analysis objective (Sub-objective III). These benchmark analyses helped to identify the problems of designing ML models. Additionally, these benchmark experiments give preliminary knowledge about medical datasets, which we will use to generate synthetic data to achieve Sub-objective IV. Different types of quality control benchmark analysis with the GI-tract data [32, 33] were performed to support this objective. Moreover, we can consider the ML models [30, 31, 36, 35] introduced in Sub-objective I as benchmark analysis studies for Sub-objective III because they are correlated with each other. Similarly, the ECG analysis [41] and sperm analyses [38, 39, 40, 68] experiments were considered benchmark analyses to identify data-related problems to address using synthetic data. Without having benchmark analysis or baseline experiments, it is not recommended to researching GANs for the new framework because data problems related to a medical dataset cannot be identified without benchmark analysis. We have also performed benchmark analysis with synthetic data [67, 72, 73] to identify the usability of synthetic data instead of real medical data.

- **Sub-objective IV**: Research and developing GANs is the core of the DeepSynthBody concept [71] (`www.deepsynthbody.org`) proposed as a solution to the data deficiency problem identified and investigated in this thesis. We started investigating possibilities of using GANs with GI-tract data such as prepossessing GI tract

images using a GAN [72, 73] to fill blank regions and to predict blurry pill cam video frames using a GAN [74], which can predict the future frames for given input frames to solve the data problems of developing ML models. These experiments gave the basic understanding of how GANs use in the medical domain and how hard it of producing synthetic data in the medical domain. Then, an advanced GAN experiment, namely Pulse2pulse [70], which can generate synthetic 12-leads 10-seconds ECG indistinguishable from real ECGs was introduced to overcome the data sharing problem as a result of privacy issues. Ultimately, we proved that our synthetic ECG dataset shows very close characteristics to the real data distribution [70].

Moreover, to address the costly and time-consuming expert's data annotation process, we experimented and introduced novel pipelines [75] of GAN architectures using GI-tract dataset to generate synthetic polyp data from the clean colon to overcome data imbalance problems in the medical domain, such as having more true-negative samples compared to true positive samples. Furthermore, we researched and presented a new pipeline to generate synthetic polyp data with the corresponding mask from a single polyp image [67], namely SinGAN-Seg, and showed that generated synthetic medical data is a solution to overcome data problems in the medical domain. Additionally, we investigated the usability of GANs to produce synthetic sperm data [76] instead of blurry-looking sperm video samples to have a better quality data stream for training AI-based sperm analysis systems in the future. To get active contributions of performing GAN-related research to produce synthetic data from non-computer science people, we have proposed a tool [77] to run GAN experiments without writing a single line of code.

- **Main-objective**: The final objective was to connect these all together and produce well-performing and more accurate ML models for CAD systems to assist doctors for efficient diagnosis by addressing the data deficiency problem. The initial ML models designed to achieve the Sub-objective I showed the effects of the data deficiency problem in the medical domain. Then, we collected, researched, and developed datasets (real and synthetic) to develop ML models for biomedical applications. In Sub-objective III, benchmark analyses were performed to identify the data problem to be addressed. We proposed the new concept and the corresponding framework, DeepSynthBody, based on GANs as a solution to the data deficiency problem in the

medical domain (Sub-objective IV). Finally, we published our solution as an open-source project for getting more collaborations worldwide at `www.deepsynthbody.org`.

As described above, our research addresses the stated objectives. Then, regarding the overall research question, what problems emerge from data in computer-aided diagnosis systems, and how can these problems be tackled? We first identified the problems and proposed the DeepSynthBody concept to tackle them. As the problems, we could identify that data to train ML models in the medical domain is lacking due to several data preparation problems, such as privacy concerns and the costly and time-consuming data annotation process. Then, this data deficiency problem causes generalisability issues and performance issues for ML models, which are the core algorithms used in CAD systems. To answer the data deficiency problem, we have experimented and developed synthetic data and showed that generated synthetic data could solve the data deficiency problem in the medical domain because synthetic data can address some of the restrictions emerging from privacy issues coming with sensitive data. We also show that synthetic data is an alternative way to prepare data and corresponding segmentation masks for the costly and time-consuming real data annotation process.

In addition to the main contributions aligning to this thesis work, the author contributed as a development team member of the Norwegian "Smittestopp" app, which was developed to trace Covid-19 contacts. Algorithms to find contacted regions of interest using GPS coordinates were investigated under this Covid-19 app development project. Moreover, several master students were supervised, and they successfully completed their master's degrees with good grades and publications [24, 72, 73, 74], which were great contributions to the GAN development stage of DeepSynthBody. Not only these, the author contributed to a research study [78], which was focused on detecting soccer events from video clips, but this study is out of the scope of the thesis.

1.6 Outline

Our initial contributions were focused on designing ML models for CAD systems to aid doctors by achieving the Sub-objectives I and II. However, the data-related problems of the current pipeline of applying ML motivated us to find a new way to overcome the data

deficiency problem in the medical domain. Therefore, this thesis mainly focuses on designing a novel concept, DeepSynthBody, and the corresponding framework introduced to bypass the data-related problems such as privacy-related problems with medical data and resource-consuming medical data annotation process. To discuss, research, and present the DeepSynthBody concept, we organized the thesis as follows:

- Chapter 2: Related Work - gives more required background knowledge to follow this thesis. In this chapter, the basic knowledge about ML concepts and corresponding references used in designing CAD systems are given. Then, deep generative models and the state-of-the-art GANs are discussed with greater details to give enough knowledge to understand the new concept introduced in this thesis. Additionally, similar frameworks to DeepSynthBody and other studies about synthetic medical data generations are discussed.

- Chapter 3: DeepSynthBody - In this chapter, the DeepSynthBody concept, which is the new concept introduced in this thesis to overcome the data deficiency problem, is formalized by developing the corresponding framework. The theoretical behavior of the framework is discussed in this chapter with four main sections, which are collecting real data and analysis, developing GANs, producing DeepSynth data, and explainable DeepSynth AI and DeepSynth explainable AI of this framework. The first three sections are explained using three case studies of ECG data, GI-tract data, and sperm data. These use cases are discussed with the significant findings, which were identified as the most influenced results for the success of DeepSynthBody.

  Under the collecting of real data and analysis, data collection procedures and analysis procedures are discussed. Then, the core of this framework, GAN development, is discussed in developing GANs. In the same section, a novel tool, namely GANEx, used to performing GAN experiments, is introduced. The process of producing Python package index (PyPI) packages is explained using the use case studies in the same section. The website `www.deepsynthbody.org`, which is the online platform of this concept, is introduced in the third section. Finally, the optional step, explainable DeepSynth AI and DeepSynth explainable AI, are discussed theoretically.

- Chapter 4: Discussion and Conclusion - discusses limitations, other advanced func-

tionalities, which can be researched with DeepSynthBody as future directions, and the conclusion about how the DeepSynthBody concept and its formal DeepSynthBody framework help to overcome the data deficiency problem related to the development process of ML models for CAD systems.

- Appendix A: All the papers counted as contributed under this thesis are listed here with the publication details and corresponding contribution statements.

Chapter 2

Related Work

This chapter covers the basic concepts of this thesis and a literature review to discuss similar research directions and their limitations. We give appropriate knowledge to understand the development of ML models forCAD systems with limited medical data. The first section provides an overview of medical datasets. Then, the common ML solutions used in medicine are discussed with the corresponding evaluation criteria because they are the basics for developing CAD systems. Afterward, GANs are introduced with their theoretical background because GAN is the basic model used to generate synthetic data to overcome the data deficiency problem, which is identified as a major problem in the medical domain. Finally, a review and discussion about previous studies, which use a similar direction to DeepSynthBody to address the lack of medical data, is provided.

2.1 Medical Data

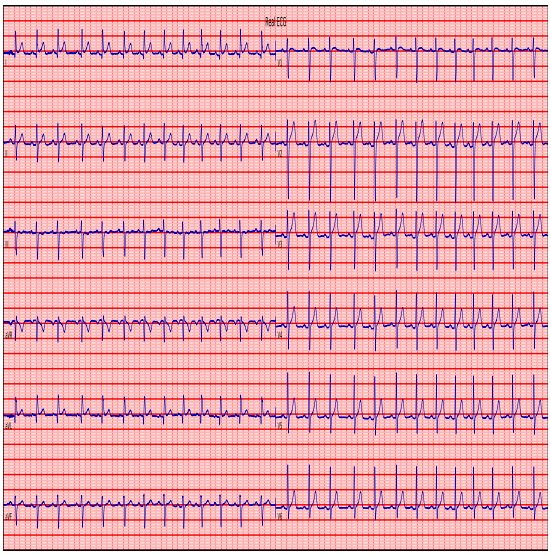
Data is the most important factor for developing AI solutions [79, 80, 81], and it cannot be separated from the field of AI. In this regard, medical datasets are the key to develop successful ML solutions in the medical domain for CAD systems. Therefore, AI researchers try to collect as much as possible medical data from data providers such as hospitals or medical research institutions. As a result, many public repositories are available for medical data, and a few of them are shown in Table 2.1. As we can see in the table, some medical repositories have a specific type of medical data like NITRC, while some collect all types of data, such as the UC Irvine machine learning repository. However, most datasets in these repositories are smaller than general datasets such as Imagenet [82]

Table 2.1: Sample data repositories with various medical data. Some of the data repositories have specific type of data. Some of them have data collections from multiple domains including the medical domain.

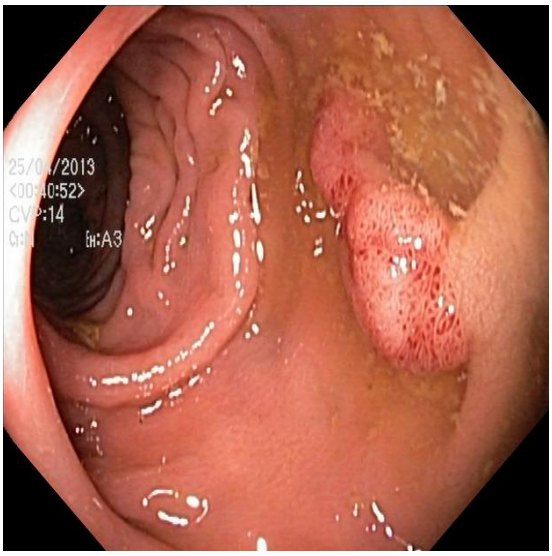
Repository	Link to access	Description
The cancer imaging archive (TCIA)	https://www.cancerimagingarchive.net/	A large archive of medical images of cancers.
NeuroMorpho	NeuroMorpho.Org	Digitally reconstructed neurons from vaiours animal types. Human is included as one type.
NeuroImaging Tools and Resource Collaboratory (NITRC)	https://www.nitrc.org/	Neuroinformatics data, from MR, PET/SPECT, CT, EEG/MEG, optical imaging, clinical neuroimaging.
OpenNEURO	https://openneuro.org/	Sharing MRI, MEG, EEG, iEEG, ECoG, and ASL data.
PhysioNet	https://physionet.org/	A repository for Physiologic Signals.
OSF.io	https://osf.io/	Open datasets from all the domains including the medical domain.
The UC Irvine Machine Learning Repository	https://archive.ics.uci.edu	Open access datasets from many domains including the medical domain.
Registry of Open Data on AWS	https://registry.opendata.aws/	Open access datasets from many domains including the medical domain.
IEEE DataPort	https://ieee-dataport.org/	Datasets from different domains around 25 categories defined by IEEE DataPort such as Biomedical and Health Sciences , Biophysiological Signals, Environmental and more other general categories including health data.

because, for example, collecting medical datasets should follow specific protocols to avoid privacy restrictions, and annotating medical data is costly and time-consuming.

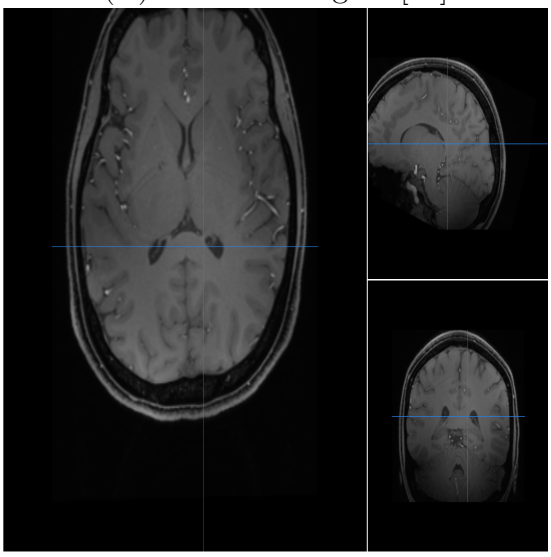
Medical data have different formats, which vary from a simple single value to advanced multi-dimensional data types such as two-dimensional (2-D), three-dimensional (3-D), and four-dimensional (4-D). Multi-dimensional data has more than one value to represent a single data point. Visual representations of sample biomedical data with various data formats are depicted in Figure 2.1. Figure 2.1(A) represents a simple 1-D ECG signal, and Figure 2.1(B) shows an image (2-D) taken from an endoscopy. Some medical data



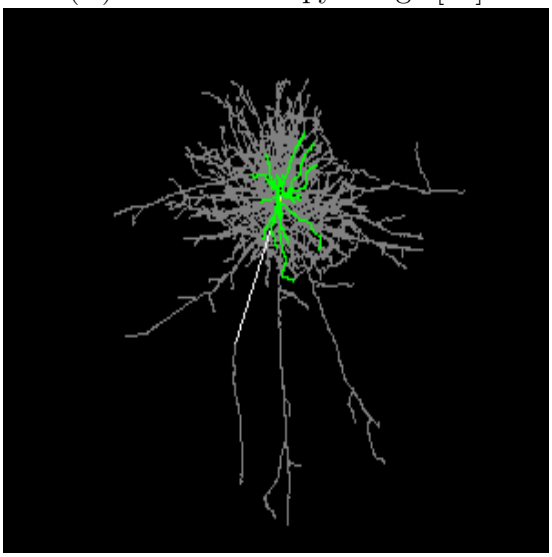
(A) - An ECG signal [41]



(B) - An endoscopy image [23]



(C) - An MRI representation [83]



(D) - A digitally reconstructed neuron [84]

Figure 2.1: Visual representations of different types of biomedical data.

cannot be simply presented in a 2-D plane and need software tools to get actual 3-D visualizations such as an MRI as depicted in Figure 2.1(C), and a digitally reconstructed neuron depicted in Figure 2.1(D). Therefore, considering data formats is important in developing ML solutions, such as deep generative models, which will be discussed in later sections.

Medical datasets, which can be either public or private, are the foundation for developing ML models for CAD systems to assist doctors. Therefore, collecting medical data is identified as a key step for the thesis. As a result, several datasets [23, 24, 25, 26, 27, 28, 29] were published. More details about these datasets are discussed in Section 3.1.1. In DeepSynthBody, which is the novel concept introduced to overcome the data-related

Table 2.2: Sample datasets for the 11 categories of the biological anatomy classification. These datasets were selected randomly using Google search. These datasets are selected from the outside of the dataset contributions introducing under this DeepSynthBody study.

Data class	Sample datasets
Cardiovascular	Cardiac MRI dataset [88], ECG data [89]
Digestive	Endoscopy dataset [90, 91], Capsule endoscopy [92]
Endocrine	Hyperspectral imaging [93], Thyroid ultrasound image [94]
Integumentary	Skin lesions [95], Skin image dataset (melanomas) [96]
Lymphatic	CT lymph nodes [97], Lymphography Data Set [98]
Muscular	MRI of muscles of the hand [99], Full body data with muscle [100]
Nervous	Brain activity fMRI data [101], PET-MR Dataset [102]
Urinary	Kidney dataset [103], CI images kidney stones [104]
Reproductive	Human sperm images [105], Embryo dataset [106]
Respiratory	Chest X-ray data [107], Chest CT datast [108]
Skeletal	Bone X-ray dataset [109], Knee MRIs [110]

problems faced during the development stage of ML solutions, all the medical datasets had to be categorized to make a clear data organization process for the contributors and the end-users of the framework. For this, a biological anatomy classification [85] (11 categories) was used to classify most of the medical datasets (except genome data [86, 87] which is related to the full human body. The genome data will be considered for the DeepSynthBody framework in the future. Table 2.2 presents the 11 classes selected as our classification and corresponding example open-access datasets. These example datasets indicate that most of the data can be classified into these 11 categories.

Even if publicly available, medical datasets can come with other challenges that need to be taken into account. One challenge is the sizes and distributions of medical datasets. If the sizes of these datasets are limited, such as having few data samples, then it directly affects the final performance of ML models. Similarly, if a dataset is imbalanced such as one class has more data and another class is lacking data, then it also affects the performance of the ML models [111, 112, 113, 114]. Despite these problems, privacy concerns of the medical data [115], containing information about patients, is another problem. These privacy concerns directly cause problems for publishing the medical data because medical dataset publishers should follow all the protocols related to publishing medical datasets, as discussed in Section 1.1. In addition to the privacy concerns, making the ground truth data for the medical data is costly and time-consuming. In the medical domain, experts (medical doctors) should perform the data annotation process. Therefore,

one of the goals of this thesis is to overcome the data annotation problem and introduce an efficient way to produce medical datasets with ground truth to train ML solutions, i.e., both reducing the need for medical experts to produce ground truths and bypassing the privacy challenges.

2.2 Machine Learning in Medicine

Different types of ML algorithms are applied to medical data. When researchers and other medical data providers publish datasets to train ML models, they have intended goals to achieve using the datasets. For example, when GI-tract polyp datasets are published with the corresponding ground truth masks [116, 117, 118], the main goal of the datasets is to train ML models to perform polyp segmentation tasks. Therefore, baseline experiments (experimental results coming with dataset papers) and benchmark experiments (experiments performing to achieve the best results compared to the state-of-the-art performance) of a particular dataset are essential to know the capabilities of the ML models trained using the dataset and identify the related practical problems, for example, the generalizability issue of an ML model trained using a single dataset. The baseline and benchmark results coming from ML models can be used to identify the limitations of datasets. For example, suppose every machine learning model shows poor performance for a specific class of a data classification problem. In that case, the problem might be with the data of the particular class. In this regard, this thesis discuss baseline experiments and benchmark experiments. The baseline experiments are discussed with our dataset papers [23, 24, 25, 26, 27, 28, 29], and the benchmark experiments are discussed in our benchmark articles [30, 38, 39, 40, 68, 41, 36, 32, 33, 35, 34].

Most of the ML models trained with medical data can be classified into a regression task [119, 120, 121], classification task [122, 123], detection task [124, 125] or segmentation task [126, 127]. These tasks depend on medical datasets and their intended purposes. ML models trained to solve regression tasks want to predict continuous values (parameters) for a given input data such as numerical input, images, or video inputs. For example, predicting motility or morphology level, which are percentage values, of a sperm sample given as a video is a regression model. In the classification task, ML models need to predict class labels of input data, such as predicting the GI-tract landmark for a given

image captured from an endoscopy. In detection tasks, ML modules focus on predicting bounding boxes for regions of interest on images or videos (normally, videos are processed frame by frame, and this video processing also can be considered as image processing), i.e., predicting polyps in an image of GI-tract. Advanced segmentation tasks perform pixel-wise predictions to mark the region of interest, and this task gives greater details than all other three tasks, for example, predicting the exact regions of polyps using the pixel-wise classification of a GI-tract image. These ML methods have specific evaluation methods based on the objectives.

Evaluating ML models have to be performed properly, which means evaluation processes should reflect the real performance of ML models. For example, data leakage problems [128] should be avoided, the generalizability of ML models should be tested using cross-dataset evaluations, and multiple evaluation metrics should be calculated to show the performance from different perspectives. Otherwise, researchers may produce inefficient solutions which cannot be applied in practical scenarios. According to the type of the ML task, the evaluation methods should be selected. A summary of these evaluation methods is presented in Table 2.3.

One of our studies [31] discusses the importance of evaluating ML models with multiple evaluation metrics and cross datasets for producing better generalizable ML solutions. In addition to the cross dataset evaluations, we have discussed problems of current articles with incomplete evaluation metrics using a literature review of polyp classification as a case study [33]. To overcome this incompleteness of the evaluation results, we have introduced an online tool called MediMetric[1], which can be used to get complete evaluation metrics from the incomplete evaluation metrics for binary classification tasks. The evaluation performance of ML models can be found in baseline experiments, which come with dataset papers, and benchmark papers, which aim to produce state-of-the-art results for a particular dataset. In this thesis, these baseline results and benchmark results are essentials to develop ML solutions to achieve our Sub-objective I and develop DeepSynthBody, which is the main solution introduced in this thesis to achieve Sub-objective IV. Therefore, contributions of ML methods with corresponding evaluations are presented in our series of benchmark articles [30, 38, 39, 40, 68, 41, 36, 32, 33, 35, 34] in addition to the evaluations presented in our dataset publications [23, 24, 25, 26, 27, 28, 29].

[1] `https://medimetrics.no/`

Machine learning (ML) type	Evaluation method
Regression	R Squared (Coefficient of Determination), mean square error (MSE) or root-mean-squared error (RMSE), mean absolute error (MAE)
Classification	Accuracy, F1, Recall (sensitivity), Precision, Matthews correlation coefficient (MCC)
Detection	Intersection over union (IOU), Precision, Recall
Segmentation	IOU(Jaccard index) , F1-score (dice coefficient)

Table 2.3: Example evaluation methods using for the most common ML methods applied with medical data.

2.3 Generative Adversarial Networks

In the above section, regression, classification, detection, and segmentation models known as discriminative models were discussed. As a mathematical definition, the discriminative models capture the conditional probability, for example, $p(Y|X)$, in which X represents data instances and Y represents a set of corresponding labels. In this section, generative models are discussed. These generative models are the most important ML model used in DeepSynthBody, which is introduced as a solution to overcome the data deficiency problem. Generative models learn joint probability distribution compared to the conditional probability of discriminative models. In the formal definition of generative models, they capture the joint probability $p(X, Y)$ if both data instances (X) and labels (Y) exist. Otherwise, the generative models capture only data distribution $p(X)$. There are several types of generative models. Autoregressive models, variational autoencoders (VAEs), Latent Dirichlet Allocation (LDA), Hidden Markov Model, Gaussian Mixture Model, Bayesian Network, VAE, and generative adversarial network (GAN) are a few of them. Among these generative models, two deep generative models, namely VAE [129] and GAN [130], have become popular in the recent research studies [131, 132, 133] of generating synthetic data.

VAE [129] consists of two networks, namely encoder and decoder networks. The basic architecture diagram is illustrated in Figure 2.2 with the basic elements. In the training stage, the encoder converts input data into a latent space represented using mean (μ_x) and standard deviation (σ_x). Then, in the inference stage, only the decoder generates data by sampling the latent vector from the latent space. However, the main disadvantage of using VAEs to generate synthetic data is generating blurry output [134]. In synthetic

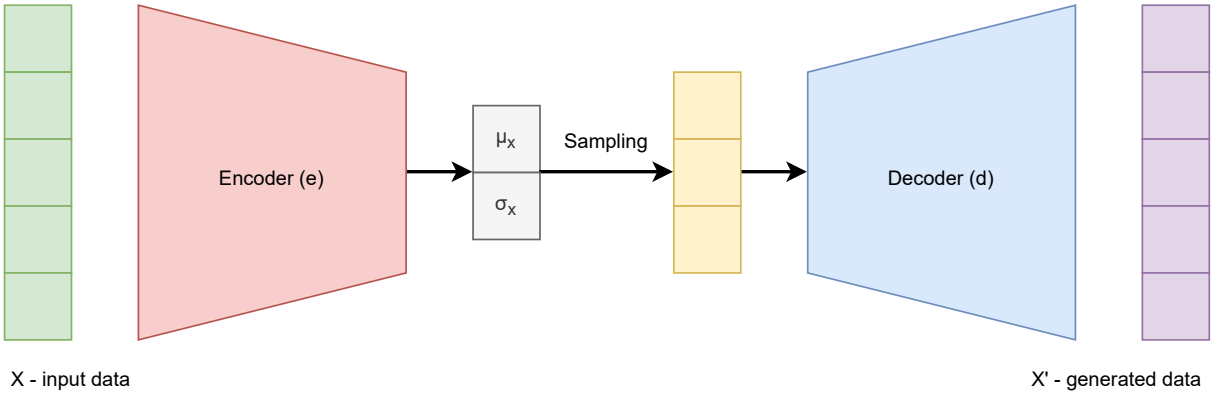


Figure 2.2: Basic architecture of a VAE.

medical data generations, every feature is essential. Therefore, GANs were selected to use as the main generative models to generate synthetic data in this thesis because of high-quality feature-rich generation capabilities. In contrast, GANs are harder to train than VAEs [135].

The basic GAN architecture introduced in 2014 by Ian et al. [130] consists of two DNNs. One is called the generator, and the second one is called the discriminator. The generator's main task is to generate synthetic data by taking a random noise vector as input. The noise vector can be sampled from any statistical distribution, such as normal distribution or Gaussian distribution. Then, the discriminator learns to distinguish generated data from the real data, used to train the GAN architecture. In the training process, the generator and the discriminator are leaning together, which results in a Nash equilibrium [136] problem. If successfully trained, the generator can generate realistic synthetic data samples, which can fool the discriminator. This process is illustrated in Figure 2.3. The objective function (loss function) used in this vanilla GAN architecture is presented in Equation 2.1. However, not every GAN architecture uses the same objective function to optimize the training process. The most common loss functions are summarized in a large study about GAN architectures done by Lucic et al. [137]. Using the most appropriate loss function to generate realistic synthetic data with a stable training process or investigating novel loss functions for a GAN is another important factor in generating realistic synthetic data. Therefore, studying and having comprehensive knowledge about GANs and the corresponding loss functions is essential before developing GANs to generate synthetic data. Otherwise, synthetic data generated from GANs will not cover the real distribution of the training data [138], or the mode collapse behavior [139] of GANs

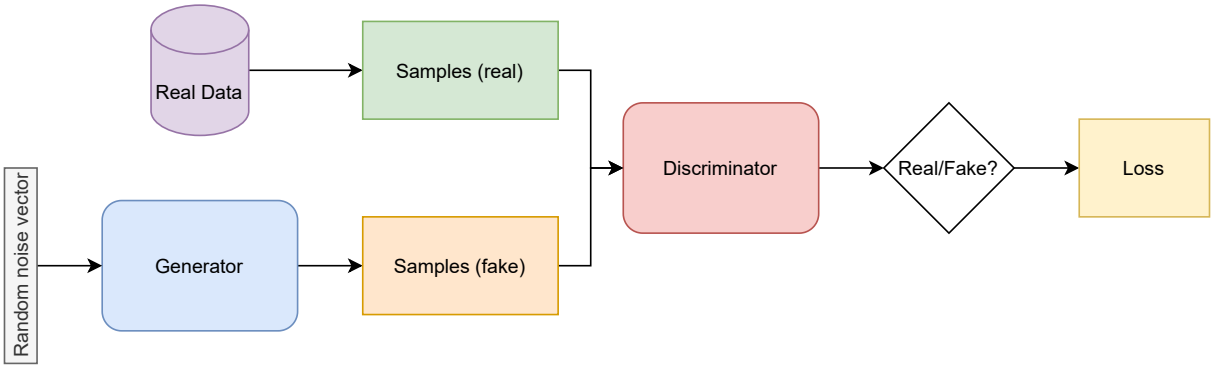


Figure 2.3: A simple representation of the vanila GAN architecture.

may cause.

$$min_G max_D V(D,G) = \mathbb{E}_{x \sim p_d(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))] \quad (2.1)$$

After the vanilla GAN, it became one of the trending fields in DL, and different GAN versions for different purposes were published. A summary of the most popular GAN architectures is shown in Table 2.4. For producing better quality synthetic medical data for the DeepSynthBody, the contributors should use the most appropriate GAN architecture. A literature review or preliminary experiments should be conducted to determine the best fitting GAN architecture for a given problem. Good knowledge about the state-of-the-art GANs methods is important for finding a better GAN model for generating synthetic data. In this thesis, novel GANs [70, 75, 67] and modified versions of different GAN architectures [72, 73, 74, 76] were researched and developed. More details about these GANs are presented in Chapter 3.

Not only the designing and implementation of GANs is essential, but also evaluating them. Evaluation of GANs is an active research area by itself. GAN evaluation is not well-defined in terms of how to measure the quality of the generated synthetic data. Theoretically, GANs should produce synthetic data which looks like real data from the whole distribution of the real data used to train the GANs. To measure the performance of GANs, qualitative and quantitative evaluation metrics were introduced in several research papers. Table 2.5 shows standard GAN evaluation metrics presented in the paper [150]. In the synthetic data generation process, the evaluation process plays a significant role in finding suitable GANs to produce synthetic data to replace the real medical data. For example, evaluation metrics can compare two or more GAN models developed for the

Table 2.4: A little from the most popular GAN architectures and their main functionalities. More about other GAN architectures can be found in [140, 141, 142]

GAN name	Description
Vanila GAN architecture [143]	This is the first GAN architecture introduced in 2014. This is capable of generating low resolution images but they are noisy.
Pix2pix [144]	This is a conditional GAN GAN architecture. This model convert an input image from one domain to an output image in another domain. The training process need paired images from two domain which have one to one mapping.
CycleGAN [145]	This paper present a similar mechanism to the Pix2pix implementation. However, the CycleGAN does not need paired training data, then the model can be train using unpaired two datasets from two different domains. Cycle consistency loss was introduced in this study.
StyleGAN, StyleGANv2 [146]	This GAN architecture is capable of generating realistic high-resolution images and the GAN can be controlled to change high-end features as well as fine features. The major drawback of this GAN is, a large training dataset is required to train the model. However, recent advancements introduced to data augmentation method [147] with GANs shows that limited datasets are enough to train new GAN models.
BigGAN [148]	This is another GAN architecture which can generate high-resolution images with high fidelity. A large dataset is required to train BigGAN also, but the quality of generated samples are high.
SinGAN [149]	This GAN architecture is trained using a single image and then, synthetic data is generated similar to the local and global features of training images but different from the training images. As use cases, generating high-resolution images, image editing , harmonization and making animations are focused.

same purpose. However, evaluating GAN models developed by different developers is not an easy task until a common reference calculates evaluation metrics. Therefore, in this thesis, we recommend using qualitative and quantitative criteria to understand the quality of the generated synthetic data.

Table 2.5: A few of GAN evaluation metrics. The complete list of these evaluation metrics and corresponding details with the original references can be found in [150]

GAN evaluation type	Metircs
Qualitative	Average Log-likelihood Coverage Metric Inception Score (IS) Modified Inception Score (m-IS) Mode Score (MS) AM Score Fréchet Inception Distance (FID) Maximum Mean Discrepancy (MMD) The Wasserstein Critic
Quantitative	Nearest Neighbors Rapid Scene Categorization Preference Judgment Mode Drop and Collapse Network Internals

2.4 Synthetic Data in Medicine

Researchers have experimented with GAN in the medical domain for different purposes. In most cases, GAN models have been used as augmentation techniques to increase the size of the medical datasets [151, 152, 153]. Some of them have focused on improving classification [151], detection [154, 155], or segmentation [156] performance using synthetic data generated by GANs. Besides increasing or augmenting data, special types of GANs can perform medical segmentation tasks [157, 158] and generate super-resolution images to make a precise medical diagnosis [159]. AsynDGAN [160], introduced by Chang et al., is another GAN architecture focusing on solving privacy concerns by distributing discriminator networks among data providers to train a GAN architecture.

To the best of our knowledge, there is no other similar concept proposed like the DeepSynthBody concept, which focuses on producing synthetic medical data for the whole human body to solve the data deficiency problem in the medical domain by addressing, for example, privacy concerns and overcome costly and time-consuming medical data annotation processes. However, few studies developed frameworks to solve privacy concerns of the medical data. The closest framework similar to DeepSynthBody is Synthea[2] [161] which was developed to generate synthetic electronic health records (EHR). Synthea is

[2] `https://synthetichealth.github.io/synthea/#technology-landing`

also running as an open-source project to get contributions from other researchers. This framework focus on generating synthetic EHR to free the medical data from legal, privacy, security, and intellectual property restrictions. Although Synthea focuses its primary goal on producing privacy restriction-free synthetic EHR, which is one of the primary goals of DeepSynthBody, significant differences can be found between Synthea and DeepSynthBody. For example, DeepSynthBody focuses on building a synthetic human body model, while Synthea focuses on making synthetic patient records using synthetic EHRs, which are text-based medical records. Pre-generated records can be downloaded from the Synthea website[3]. The DeepSynthBody concept is not targeting text-based EHR generations like Synthea. Our main focus is on generating realistic medical data similar to the medical data collected from medical instruments used to examine patients, such as biomedical signals and biomedical images.

Moreover, DeepSynthBody provides an advanced well-defined flow from data collection to the end of synthetic data generations focusing on much more advanced additional objectives. These additional objectives provide synthetic data with annotations, define a novel model for the human body, and provide a restriction-free GAN repository for generating synthetic medical data. Additionally, the DeepSynthBody concept publishes GAN models instead of pre-generated synthetic data for the end-users.

Anonymization through data synthesis using generative adversarial networks (ADS-GAN) [162] is another framework to generate synthetic EHR datasets. This framework provides pre-trained GANs to generate synthetic EHR records. Their generation method is based on conditional-GAN, which means to generate synthetic data, real data values should be available. Therefore, they propose to have a trusted intermediate partner to generate synthetic EHR data from real data records. In comparison, DeepSynthBody does not need any intermediate partner because of the in-house GAN training capability introduced in the framework with the corresponding tools. In addition, DeepSynthBody focuses on diverse, complex medical data types compared to normal EHR data considered in the ADS-GAN study.

SynSigGAN [163] was developed by Hazra and Byun to generate privacy restriction-free biomedical signals. However, despite the results in the paper, the GAN is not available in public to generate synthetic data. Similarly, different generative models for dif-

[3] `https://synthea.mitre.org/downloads`

ferent types of medical datasets can be found, such as synthetic embryo images [164] and COVID-19 X-ray images [165]. The study of synthesis of COVID-19 chest X-rays shows improvement for ML models used to detect Covid-19 when this synthetic data is used with real data to train the ML model. They also discuss how GAN is used for anonymization. The improvement achieved for the performance motivated us to make a formal framework for synthetic data in the medical domain. DeepSynthBody provides a framework and infrastructure that can share these anonymized data generators compared to the above solutions.

2.5 Summary

Medical data is the key to apply AI solutions in medicine. Therefore, there are many public repositories, which are collecting medical data and share them with researchers. These medical data have different formats. However, the sizes of the datasets are not enough to train a generalizable and well-performing ML model. The sizes of datasets are limited in the medical domain due to, for example, privacy restrictions and the costly and time-consuming data annotation process. These data deficiency problem motivated us to find a solution to tackle the problems. Identifying the correct organ system, the data source, and the medical data formats are essential for developing ML models for CAD systems, such as deep generative models used in our DeepSynthBody concept.

Applying ML techniques and finding suitable models to get better predictions are the main tasks for developing AI-based CAD systems for medical scenarios. The main ML methods include regression, classification, detection, and segmentation. Different ML methods have implicit evaluation techniques, and following them strictly to evaluate ML models is required to find accurate and generalizable AI solutions. Producing ML solutions for baseline experiments or benchmark analyses can give a first idea about a medical dataset and the quality of dataset's content. Additionally, baseline experiments are necessary for analyzing the quality of synthetic data, which will be be used as alternatives.

To generate synthetic data, we selected GANs as the core generative model in this thesis because of the ability of GANs to generate synthetic data with rich features. However, training GANs is more challenging than training other generative models. Therefore, having a good understanding of GAN types and their evaluation methods are important

factors in implementing good generators that can produce synthetic data for solving the data deficiency problem associated with developing ML models for medical CAD systems.

Our proposed DeepSynthBody is a novel concept and a framework addressing the data deficiency problem identified while developing ML models for CAD systems to assist doctors. In this chapter, existing frameworks were explored with similar directions as DeepSynthBody. Most of the solutions focus on text-based EHRs. Our solution, namely DeepSynthBody is designed to generate all the medical data coming through medical instruments except text-based medical data. While some solutions need a third-party data handler to maintain privacy concerns, the DeepSynthBody concept proposes a mechanism to design GANs in-house of the medical data providers. In the next chapter, the DeepSynthBody concept and the corresponding framework are introduced with three case studies.

Chapter 3

DeepSynthBody

In this section, the flow of the DeepSynthBody concept [71], which is the main solution discussed in this thesis to overcome the data deficiency problem, is introduced. The whole framework is discussed under four major steps: collecting real data and analysis, developing generative models, creating DeepSynth data, and explainable DeepSynth AI and DeepSynth Explainable AI. The first section is further divided into two, collecting real data and analyzing real data to discuss the real data collection process and the process of analyzing them, respectively. Under the second step, namely developing generative models, three sub-section are discussed. These are designing generative models and evaluation, publishing deep generative models, and developing a tool called GANEx to perform GAN experiments. In the third section, creating DeepSynth Data is discussed. At the end of the chapter, explainable DeepSynth AI and DeepSynth explainable AI is presented, followed by a summary.

We have developed this framework to tackle the data deficiency problem identified as a major bottleneck to develop AI-based CAD systems in medicine. The main focus of the DeepSynthBody concept is producing synthetic medical data to overcome barriers attached with medical data, such as privacy concerns, the costly and time-consuming medical data annotation process, and the data imbalance problem in the medical domain. The DeepSynthBody concept is not limited to achieve the primary objectives, but it opens new research directions such as finding a synthetic model to define the human body. Additionally, DeepSynthBody can be considered a modern repository to store medical data without any privacy concerns. It can be used as a medical data compression method to store big datasets in limited spaces.

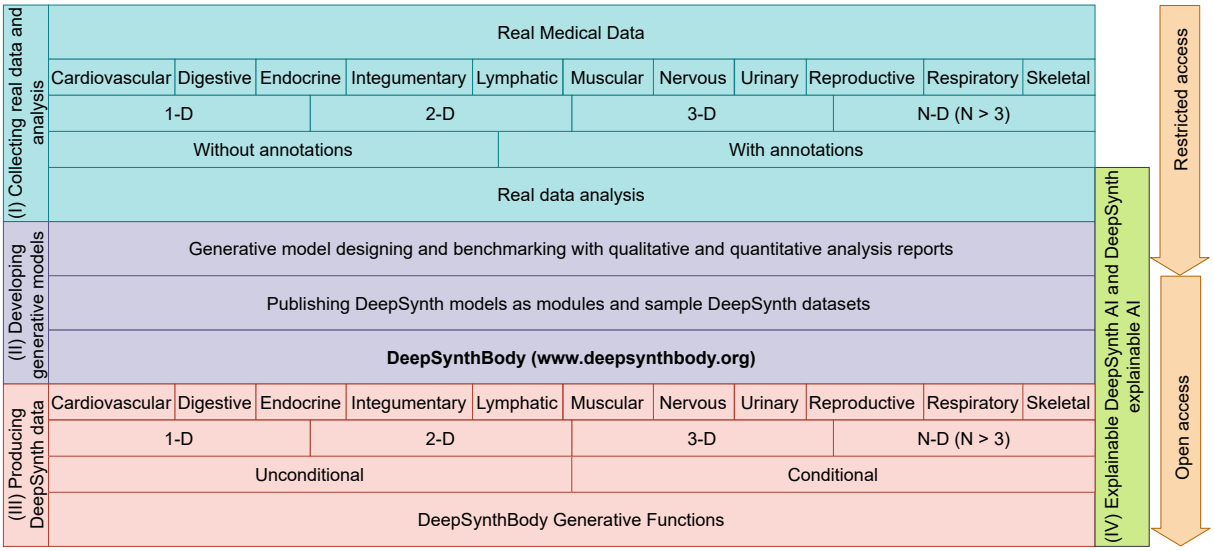


Figure 3.1: Complete framework of DeepSynthBody. Reference for the figure: [71]

An overview of the DeepSynthBody framework is shown in Figure 3.1. There are four major steps namely:

I. collecting real data and analysis.

II. developing generative models.

III. producing deep synthetic data.

IV. explainable DeepSynth AI and DeepSynth explainable AI.

The right-side top arrow in Figure 3.1, *Restricted access*, represents the flow having privacy-related restrictions. The *Open access* arrow represents the open access flow of synthetic data generated to replace real private datasets. These steps are discussed in detail in the following sections.

3.1 Step I: Collecting Real Data and Analysis

Step I in Figure 3.1 is collecting real data and analysis. In this step, real medical data are collected and analyzed for the later steps in DeepSynthBody. Real medical data can be either public or private. If data is private, this Step I should be completed by authorized data providers. If data is public, anyone who wants to contribute to this framework can complete this step. The three sub-processes, data classification, annotation and labeling, and analysis, are discussed separately to simplify the process of the step. The two types

of research contributions can be identified in this step. They are publishing open access medical datasets with baseline experiments and performing benchmark experiments of medical data.

3.1.1 Collecting Real Data

Medical datasets are the key to initiate the DeepSynthBody framework. Hospital and medical research institutions are the sources for collecting real medical data. These medical data come from different sources, such as ECG machines [166], X-ray machines [167], endoscopy machines [168], MRI machines [169], and various other advanced types of machinery collecting human body data. In this thesis, the medical data collection process was performed continuously to achieve Sub-objective II, which also contributes to the data collection process of DeepSynthBody. As a result, seven open datasets were published. The datasets collected in this thesis are tabulated in Table 3.1 with additional two datasets used as case studies. These additional two datasets were not published as dataset papers of this thesis, but we have used them to have different case studies in the later stages of this thesis.

The first three datasets presented using bolded text in Table 3.1 were the selected three datasets. The first dataset is an ECG dataset, but it is restricted for public use because of privacy restrictions. HyperKvasir [23] is the largest public GI-tract dataset consisting of images and videos collected from real endoscopy examinations. This GI-tract dataset consists of polyp images with the corresponding annotations done by experts, unlabelled images, and a set of images belongs to 23 classes. VISEM [69] (sperm video data) was not collected as a part of the thesis, but the dataset is considered as one of the case studies. We have selected this dataset to represent the video data type in our experiments.

The fifth and the sixth in Table 3.1 are two other GI-tract datasets related to Hyper-Kvasir. These are the Kvasir-Capsule [27] and Kvasir-instruments [29] datasets. Kvasir-Capsule consists of images and video data collected from capsule endoscopy. This dataset has $47,238$ labeled images, 43 labeled videos, $4,694,266$ unlabeled images and, 74 unlabeled videos. By comparing the number of labeled and unlabeled images and videos, we understand the capabilities of this dataset for supervised and unsupervised machine learning techniques. However, this dataset does not have any segmented GI-tract images. In contrast to this, Kvasir-instruments is a segmentation dataset with manually annotated

segmentation masks of endoscopic tools. The dataset has 590 images with corresponding mask images of the segmented tools. Providing fewer images with this dataset indicates how hard it is to prepare this type of segmentation dataset with the help of medical experts. So, finding an alternative way to prepare segmentation datasets with medical datasets is important.

The PMData [25] dataset contains general life-logging data and sport activity data. Fitbit versa 2 fitness smartwatch was used to collect sensory data for this dataset. Therefore, the participants of this data collection process were encouraged to wear the watch as much as possible. In addition to this sensor data, all the participants were asked to record their daily activities and fitness levels, such as sleep hours, the mood of the person, etc., in PMSys sports logging app[1]. Furthermore, a Google form was used to collect another set of data: demographic data, food images including drinking, and weights. While this type of data collection is not directly connected with any pure medical data, such as collecting images and signals of the human body using medical instruments, these data are important to know the relationship between daily life and health problems. However, collecting these types of daily activities is challenging, and careful de-identification is needed before publishing data to the public.

PSYKOSE [26] is a motor activity dataset collected from 22 schizophrenia patients and 32 healthy control persons. All the motor activities were collected for an average of 12.7 days using a wrist-worn actigraph device[2]. In addition to the motor activity data, demographic data and the data about medical assessments are given. This kind of detests is essential for predicting the health states and performance outcome of a person. However, motor activity data and the corresponding demographic data are susceptible to privacy. Additionally, collecting health data with multiple sources for the same person is important because finding correlations among health data and other factors such as motor activity can lead researchers to discover hidden behaviors of our human body.

HTAD [28] presents a dataset with wrist-accelerometer data and sound data for the four most common daily activities of human life. These activities are sweeping, brushing teeth, washing hands, and watching TV. Fining the pattern of these kinds of activities can lead to finding new research directions such as assistive technology for older people. Not only as assistive technology, identifying unique patterns of sensor data corresponding

[1]https://forzasys.com/pmsys.html
[2]Actiwatch, Cambridge Neurotechnology Ltd, England, model AW4

to specific health conditions such as mental disorders can lead to treat such patients. However, collecting data about daily routines has a significant impact on privacy concerns. Therefore, these kinds of datasets are scarce, and publishing them needs a careful de-identification process. Otherwise, reaching a way to produce similar synthetic data can lead to share data without privacy concerns.

Toadstool [24] dataset has sensor data collected through an Empatica E4 wristband while a set of people are playing Super Mario Bros. In addition to the sensor data, videos captured during the playtime of the game were included. The data was collected from 10 participants of different ages, sex, and different experience levels. Toadstool looks like a non-medical dataset. However, finding correlations between sensor data and game playing patterns will encourage researchers for new areas like health conditions and game playing. Monitored heart rate and facial expression captured during the playtime can be used to find hidden correlations. While many people can collect this data type, data sharing is not as straightforward as a lack of privacy-preserving data sharing mechanisms. Therefore, we made this dataset to perform research to find suitable data sharing techniques and find a way to produce synthetic data alternatives to replace these advanced data collection processes.

The raw medical data should be classified using data classification methods introduced in DeepSynthBody. First, we have to identify the organ systems which we use as a biological classification method.

Table 3.1: Datasets discussed in this study and corresponding DeepSynthbody Categories. **Bolded** text lists datasets discussing thoroughly in this study as a proof of concepts for the framwork. *Italic* text shows a dataset which was taken from open access datasets and it is not a dataset paper contributed under this study. All other datasets are published as the contributions from this research and they are not analysed.

Dataset	DeepSynthBody category (biological)	DeepSynthBody category (data format)	Availability	Description
ECG data	Cardiovascular	1-D	Restricted	An ECG dataset consists of 15606 samples with eight-leads readings for 10s long duration.
HyperKvasir [23]	Digestive	2-D, 3-D	Public	A dataset with images and videos collected from endoscopy examinations of GI tract.
VISEM (sperm dataset) [69]	Reproductive	3-D	Public	A video dataset consists of 85 video samples collected from microscopic examinations.
Kvasir-capsule [27]	Digestive	2-D, 3-D	Public	A dataset contains 117 videos captured from video capsule endoscopy (VCE) devices.
Kvasir-intruments [29]	Digestive	2-D	Public	An image dataset of 590 annotated frames containing GI-tract procedure tools.
PMData [25]	Cardiovascular, Digestive	1-D, 2-D	Public	The dataset consist of sensor data collected from smart watches and photos taken from smart phones.
PSYKOSE [26]	Muscular, Nervous	1-D	Public	A dataset collected from actigraph devices from 22 patients with schizophrenia and 32 healthy control persons.
HTAD [28]	Muscular	1-D	Public	A dataset consist of home task activities measured using smart wrist-bands and microphones.
Toadstool [24]	Cardiovascular, Muscular	1-D, 3-D	Public	A dataset collected from sensors of Empatica E4 wristband wore to game players and corresponding video frames of the game.

Biological Data Classification

The second row and the third row of Figure 3.1 represent the data classification methods. First, all medical data are classified into 11 categories [85] based on the anatomy of the human body, as presented in the second row of the figure. Then, all data are classified using data formats as represented in the third row. This biological classification was introduced to sort the data in a biological way to identify data using the organ systems of the human body. Then, the data format classification is applied as a supporting classification layer for developers who contribute to developing GANs to generate synthetic data.

The biological categories are cardiovascular, digestive, endocrine, integumentary, lymphatic, muscular, nervous, urinary, reproductive, respiratory, and skeletal. All the input medical data from various sources are considered through one of these categories (see the second column of Table 3.1). For example, ECG data, GI-tract data, and sperm data can be classified under the cardiovascular, digestive, and reproductive categories, respectively (the first three datasets in Table 3.1). If data cannot be considered for only one category, then the data can be classified under several categories. For example, PMData [25], PSYKOSE [26], and Toadstool [24] are classified as multi-classes according to biological categories in Table 3.1. It is essential to identify the correct biological class for data coming from various data sources to find the final categories in DeepSynthBody.

Data Dimension Classification

In addition to the biological classification, the medical data can be further classified using data dimensionality [170, 171]. In this classification, all data formats are classified into four classes, 1-D, 2-D, 3-D, and N-dimensional (N-D), for where $N > 3$. In the DeepSynthBody framework, data dimensionality means data dimensions coming through data sources (medical devices), but not the data dimensions used in data processing techniques. The third column of Table 3.1 presents this classification for our dataset contributions. Considering the dimensionality of real data is important because the dimensions of the real data increase the complexity of generative models (GANs) implementing in later sections (Step II) to generate synthetic representations for the real data. Additionally, data dimensionality decides which GAN architectures to use in Step II: developing generative models.

For the 1-D data format, biosignals (biomedical signals) collected from the human body are considered in this framework. Well-known biosignals are Electroencephalogram (EEG), Electrocardiogram (ECG), Electromyogram (EMG), Mechanomyogram (MMG), Electrooculography (EOG), Galvanic skin response (GSR), and Magnetoencephalogram (MEG). The ECG dataset, PSYKOSE [26] dataset, and HTAD dataset [28] are identified as the datasets with 1-D data format in our dataset contributions in Table 3.1.

On the other hand, medical imaging techniques [172, 173, 174] are commonly used to visualize human body organs, functions, and states for assisted diagnosis and treatment suggestions. Radiography, magnetic resonance imaging, nuclear medicine, ultrasound, elastography, photoacoustic imaging, tomography, functional near-infrared spectroscopy, and magnetic particle imaging are few examples of medical imaging data. Various technologies produce different types of medical images. In DeepSynthBody, medical imaging data is considered under three data format categories: 2-D, 3-D, and N-D, based on the dimensionality of the data obtained. For example, images collected from video cameras can be considered under the 2-D data type. Similarly, videos can be identified as a 3-D data type when the time (represented as consecutive video frames) is considered as the third dimension. However, some data sources produce 3-D data in a spatial domain, e.g., MRI data. However, this type of 3-D data can be classified into 4-D (into N-D because $N > 3$) when the source produces a series of 3-D data points along the time. In addition to 4-D data, some data sources have 5-D data [175], which are considered under the N-D data category. For example, dynamic MRI data with additional information layers such as tracking information has a 5-D data format. Under this definition, all real data sources are identified through 1-D, 2-D, 3-D, or N-D classes.

The data format classifications for the datasets collected under this thesis are presented in the third column of Table 3.1. In this table, multiple data format classifications can be seen for some datasets when the datasets have different types of data. The ECG dataset, which is not public, has the 1-D data format per channel as they received from the data source, and one sample has eight channels in total. While the original data format is 1-D, these ECG samples can be processed as 2-D as well by combining multiple channels together. However, we consider the data format of the original data source as the data format classification to simply this classification. In contrast to this ECG dataset, HyperKvasir [23], Kvasir-Capsule [27] have two different types of data formats. They are

2-D and 3-D. The images collected from endoscopy or capsule endoscopy are considered as 2-D data format. The videos collected from the same instruments are classified as 3-D. These data formats are important to process the data in later steps.

For example, designing image generators are easier than designing video generative models because video generators should consider temporal features compared to considering spatial features of images in the image generators. VISEM [69] dataset has only video data as the main data format, while ground truth data is presented using tabular data. On the other hand, PMData [26] and HTAD [28] data were considered as 1-D data because the main data format coming from the data collection instruments are signals. Toadstool dataset [24] has signals and videos, which means 1-D and 3-D data. Data coming from the Empatica E4 wristband, which was used to collect the players' physiological data streams, is considered 1-D data. The videos recorded from the computer which was used to play the game are considered as the 3-D data format. However, these are the format of raw data. In contrast to raw data formats, one can process these data with a different format; for example, video data can be processed as images if temporal information is unimportant.

The data format classification is done for only the development purpose. This format classification is important only for developers to find proper ML models such as classification, detection, segmentation, and generative models, which are compatible with the dataset.

Data Annotation Classification

After collecting medical data and classifying them according to DeepSynthBody classification, the data can be further categorized into two categories: (i) data without annotations (or labels) and (ii) data with annotations. This classification is represented in the fourth row of Figure 3.1. In this step, whether the data was labeled by experts or not is considered. Generally, most of the data coming from medical systems do not have expert annotations or labels, which are essential to training supervised ML algorithms. Advanced deep generative models such as conditional generative models [175] can be developed if the medical datasets have ground-truth data annotated by medical experts. The conditional generative models take labels (or other kinds of annotations such as pixel-wise classification) as input parameters and produce synthetic data conditioning on the input

annotations. While one of the primary objectives of DeepSynthBody is to reduce annotation cost and time required from experts, conditional GANs should be investigated. Therefore, producing annotated medical data by experts in this stage can help to train deep generative models to overcome the problem of medical data annotations.

Annotations or labels of medical data are different from dataset to dataset. Generally, medical datasets have continuous numerical values, discrete numerical values, class labels, coordinates such as bounding boxes or pixel-wise classifications (e.g., segmented mask). Medical experts can use different kinds of tools for annotating different types of ground truths. These tools may vary from simple image viewers to advanced AI-aided image mask generation tools or expensive medical data analysis tools [176, 177, 178]. However, an expert in the medical domain must operate these tools. While some tools can suggest or predict similar types of annotations, the experts should confirm the final annotations, which will be used as ground truth data for ML algorithms. This expert annotation process needs the medical experts' valuable time, which is costly. Therefore, the DeepSynthBody framework targets handling this problem also.

As explained above, if experts annotations are available, the annotations can be used to train advanced generative models such as conditional GANs. Therefore, experts' annotations were collected for most of the data sets tabulated in Table 3.1. The HyperKvasir dataset [23] consists of image labels and pixels-wise annotations (segmentation masks) for a part of this big dataset. Providing image labels is easier than providing segmentation masks, which represent pixel-wise annotations. Experts' knowledge was used in both annotation processes, but the segmentation annotation process took more time as expected than classifying into the labels. The HyperKvasir dataset consists of unlabeled data, images and videos also. In this context, this dataset can be classified as a dataset with and without data annotations.

The Kvasir-Capsule dataset [27] has labels for the images and the videos. However, in the current version of this dataset, there is not data with pixel-wise annotations. However, classification labels assigned by experts are used to prepare the labeled data. In addition to these labeled images and videos, the rest of the unlabeled images and videos were included without ground truth data because labeling them all is the costly and time-consuming task. If an alternative way to prepare labeled or annotated datasets automatically can be researched, then the expensive and time-consuming medical data annotation process can

be avoided.

In addition to the above GI-tract datasets, the Kvasir-instrument [29] dataset consists of only pixel-wise segmented images, which include instruments used in the colonoscopy examinations and operations. Therefore, this dataset can be identified as a dataset with annotated data. On the other hand, datasets [24, 25, 26, 28] collected through smart watch sensors or special wearable sensors can be considered datasets with annotations because manually identified events were reported in these datasets.

Selecting Case Studies

From the datasets presented in Table 3.1, only three different medical datasets were selected for case studies in this thesis, i.e., representing the various data types supported by our framework. They are an ECG signal dataset, a GI-tract image dataset, and a sperm video dataset. The ECG dataset is not published as a dataset paper. Therefore, this restricted ECG dataset is a perfect example for our Sub-objective IV, which focuses on generating synthetic data instead of the real dataset. On the other hand, the GI-tract [23] dataset is the largest image dataset published under this thesis, and this dataset represents biomedical images. The third dataset is an open-access video dataset [69]. This sperm dataset was selected because of the video data format, and the dataset represents another organ of the human body, while this dataset was not published as a contribution of this thesis. In this section, we discuss the three case studies with comprehensive details.

The ECG dataset is restricted, and only authorized people can access it. As a result, a dataset paper cannot be published. This dataset represents the biomedical signal data format which is considered under cardiovascular class and 1-D data format in DeepSynthBody. In this dataset, each ECG signal consists of readings from eight channels called in the cardiovascular context as channels `I, II, V1, V2, V3, V4, V5, V6` for 10-sec long duration. The eight readings can be converted to 12-leads ECGs mathematically by calculating missing leads `III, aVR, aVL, and aVF` using the following equations 3.1. The sample rate is 500 per ECG sample. Then, there are 5000 data points per lead. A

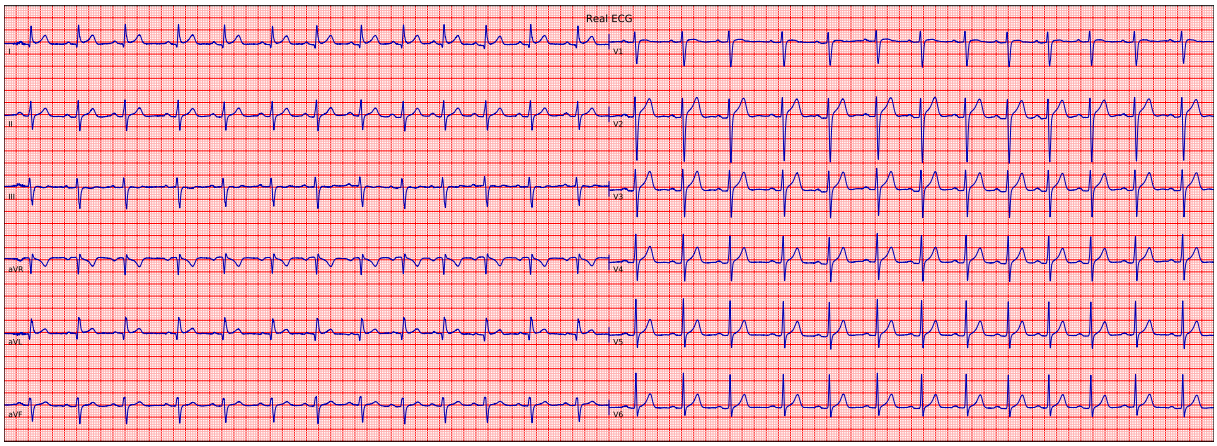


Figure 3.2: A sample of 12-leads 10-sec real ECG. Figure reference: [41]

sample from this dataset is depicted in Figure 3.2.

$$\begin{aligned} III &= II - I \\ aVR &= -0.5 \times (I + II) \\ aVL &= I - 0.5 \times II \\ aVF &= II - 0.5 \times I \end{aligned} \tag{3.1}$$

These ECG signals have been collected from two populations. One population is the Danish General Suburban Population Study (GESUS) [179] which consists of $8,939$ samples, and the other one is the Inter99 study [180] (CT00289237, ClinicalTrials.gov) consists of $6,667$ samples. In total, there are $15,606$ ECG samples. All the collected ECGs were analyzed using a well know ECG analysis system named MUSE [181]. These MUSE reports are used as ground truth for this ECG dataset, and the reports contain important characteristics of ECG signals. The important characteristics of a single ECG pulse are depicted in Figure 3.3. According to the MUSE reports, all the ECGs are classified under four main classes as tabulated in Table 3.2. Other important ECG properties collected from the MUSE system are discussed in the benchmark paper [41].

The HyperKvasir dataset [23] consists of labeled images, segmented polyp images, and unlabelled images and videos. The labeled images consist of $10,662$ images under 23 classes. In the segmented polyp images, there are 1000 polyp images and corresponding ground truth masks annotated by experts. The unlabelled images have $99,417$ images, and there are 374 videos with 30 different classes. This dataset represents the biomedical

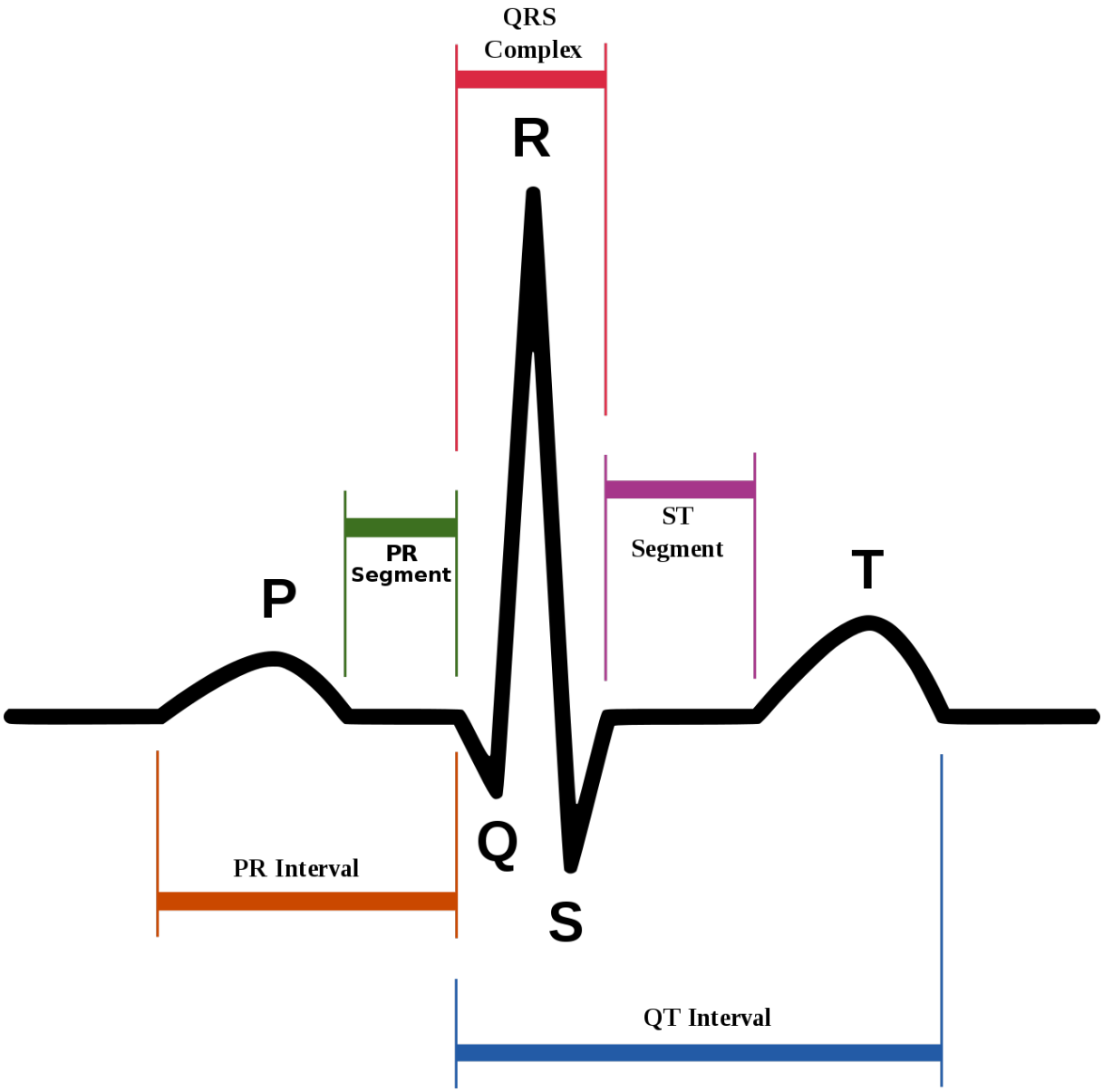


Figure 3.3: The common ECG characteristics. Reference for the image: [182]

Table 3.2: Different classes identified using the MUSE system analysis. Bold numbers represent "Normal" category ECGs which are going to be used as training data for GAN models used in later stages of DeepSynthBody. Reference for the table: [70]

Category	GESUS dataset	int99 dataset	Total
Normal	**3558**	**3675**	**7233**
Otherwise Normal	2370	1536	3906
Abnormal ECG	2118	905	3023
Borderline ECG	893	526	1419
Total	8939	6642	15581

imaging data format considered under digestive class and 2-D and 3-D data formats in DeepSynthBody. However, the labeled images, the segmented images, and the unlabelled

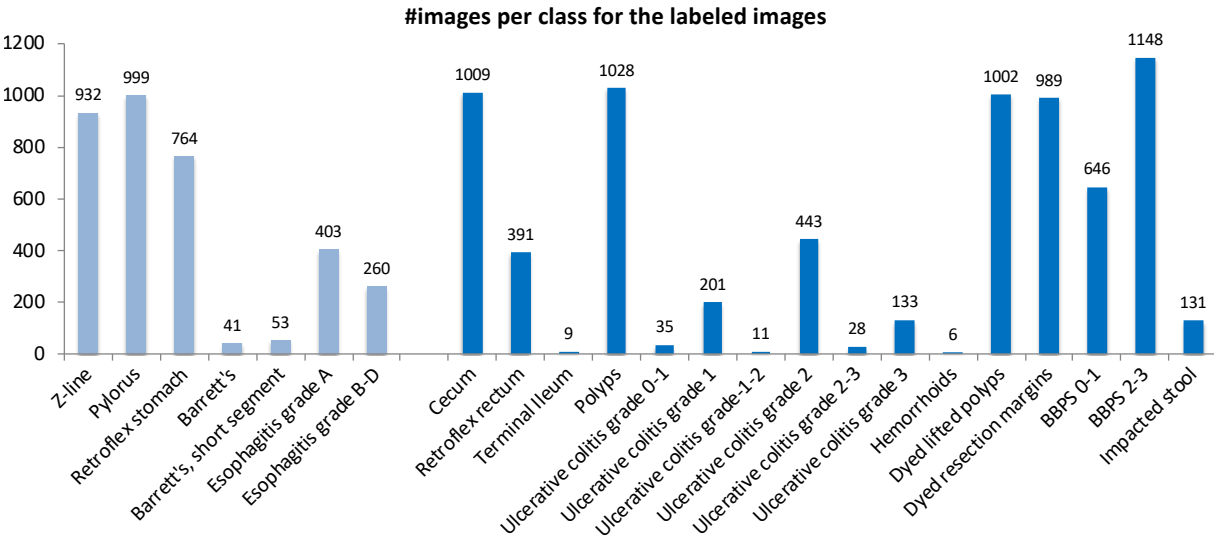


Figure 3.4: The 23 classes of the HyperKvasir dataset and the number of iamges per class. The light blue bars represent classes under upper GI-tract and the dark blue bars represent lower GI-tract images. Reference for the plot: [23]

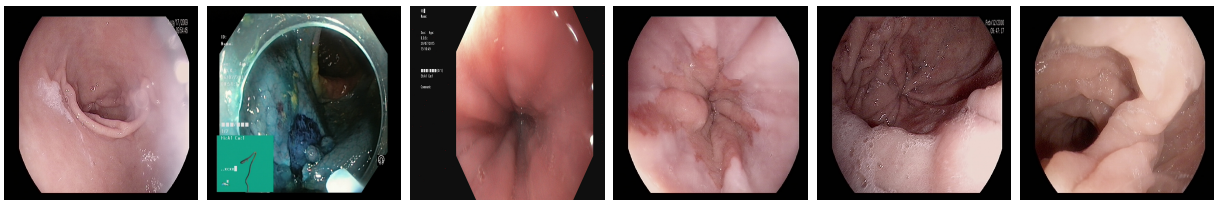


Figure 3.5: Sample images from unlabelled folder from HyperKvasir dataset. Reference for the image: [23].

images are used as case studies in this thesis, and it means, only 2-D data format is considered.

The labeled 23 classes and the number of images per class are illustrated in the graph in Figure 3.4. These images and corresponding class labels were used in baseline experiments performed for the dataset paper [23]. Then, unlabelled GI-tract images of the HyperKvasir dataset, as depicted in Figure 3.5, were used to train a GAN in developing generative models of DeepSynthBody. Polyp images and corresponding masks from the segmentation data folder are illustrated in Figure 3.6. The polyp data was used to train a GAN model, which was developed to show the possibility of using GANs as an alternative method for the costly and time-consuming data annotation process performed by domain experts. More details about the whole HyperKvasir dataset are presented in our dataset paper [23].

The VISEM dataset introduced by Haugen et al. [69] has 85 sperm videos recorded from sperm samples collected from different participants. The sperm video dataset con-

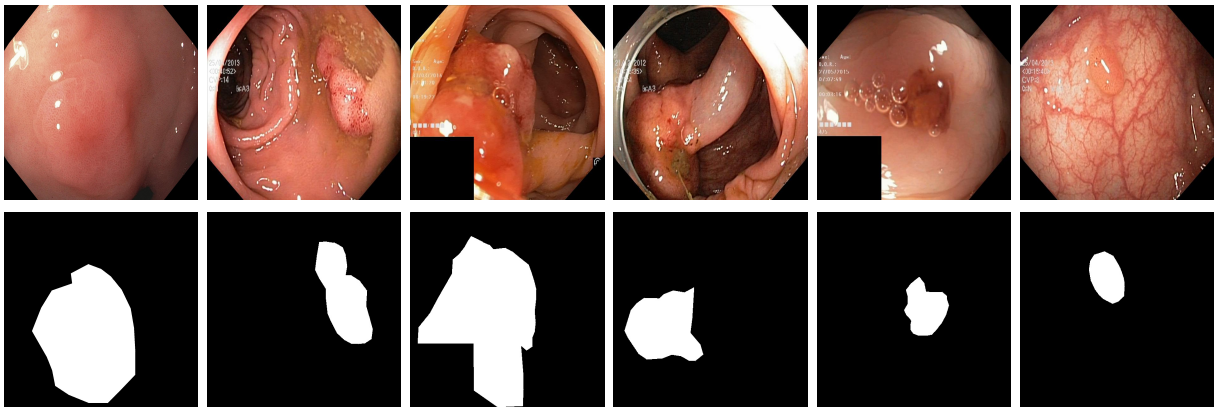


Figure 3.6: Sample images and corresponding masks from HyperKvasir dataset. Reference for the image: [23]

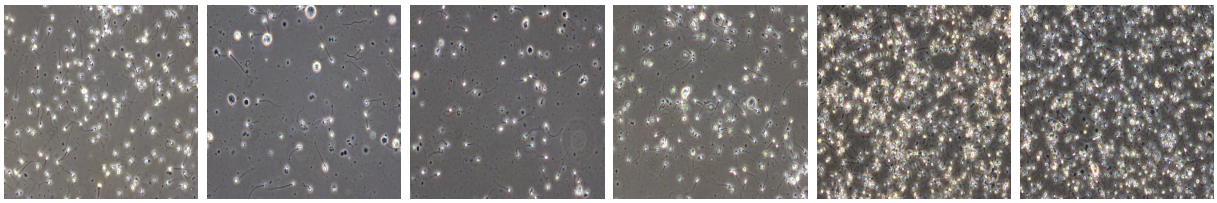


Figure 3.7: Sample frames extracted from different sperm videos from the sperm dataset (VISEM) [69].

sists of analysis data reports produced by experts in the domain of sperm analysis. The sperm dataset is classified under the reproductive system, and it covers the 3-D data format. Example frames extracted from the videos of this dataset are illustrated in Figure 3.7. Different density amounts of sperm counts are shown in this figure from left to right with low-density to high-density, respectively. The collected analysis reports attached with the sperm dataset give information about the morphology and motility level of the 85 sperm samples. Figure 3.8 shows the common quality measurements performing in sperm analysis. They are counting sperms, finding abnormal sperms (sperm morphology level), and finding abnormal movements of sperms (motility level), as illustrated in Figure 3.8 from left to right. Then, the main goal of this dataset is to predict the values in the analysis report automatically using ML techniques. More details about the sperm dataset can be found in the original dataset paper [69].

3.1.2 Analysis of Real Data

Performing benchmark analysis of real medical data is an important step in the DeepSynthBody framework because it gives the initial understanding and inherited challenges about the datasets incoming to use the framework. Generally, baseline experiments and

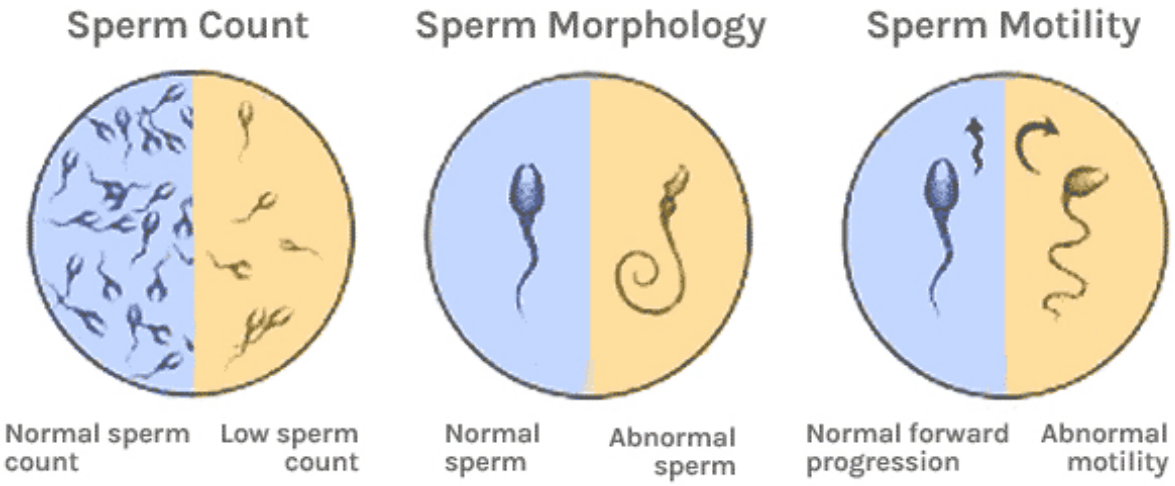


Figure 3.8: An illustration showing important sperm quality measurements. Reference for the figure: [183].

corresponding results are presented through dataset papers. However, benchmark experiments are the only source to know statistics about the private medical datasets when publishing dataset papers are not allowed because of privacy restrictions. Moreover, advanced analyses are performed in benchmark studies that focus on developing ML solutions rather than publishing datasets. Therefore, this section discusses benchmark analysis details of the selected datasets. The selected datasets are the ECG dataset [41], HyperKvasir dataset [23] and sperm dataset [69].

Electrocardiogram (ECG) Signal Analysis

The ECG benchmark analysis study [41] conducted under this thesis has two objectives. One is to predict ECG properties (see Figure 3.3), namely QT-interval, PR-interval, QRS-duration, heart-rate, J-point elevation, T-wave amplitude, and R-peak amplitude using regression ML methods. The second objective is to predict a person's sex (gender) from ECG signals using ML methods used for classification.

Using 12-leads 10-sec format or median ECGs produced from 12-leads ECGs can be used to predict regression values of the ECGs. The median ECG is a normalized single beat version of the long ECGs. Therefore, both types of input formats, 12-leads 10-sec, and the median format were evaluated as inputs to our ML models used to predict the properties of the ECGs. For each property of ECGs, separate ML models were implemented using convolutional neural network (CNN) techniques. On the other hand, to

predict the sex, only the median ECGs were used because we needed to find the correlation between interval-specific features. Medical people are not interested in rhythm-based sex prediction.

All the CNNs were trained and evaluated using five-fold cross-validation to perform a better generalizable evaluation. Quantitative evaluations have been done using MAE and RMSE. In addition to evaluating models' predictions, the GradCAM [42] approach was applied to explain the predictions from CNNs. More details about this ECG analysis and benchmark results can be found in the full article [41]. Referring to this benchmark analysis is the only way to understand this dataset because of the restrictions on sharing the real dataset. However, the methods are not reproducible because the dataset is restricted. The capabilities of DeepSynthBody to solve such privacy issues are discussed in later sections.

Gastrointestinal-tract Image Analysis

For GI-tract benchmark analysis, several experiments were performed for two different types of tasks, classification and segmentation. Under the initial objectives, we performed these experiments to develop ML models for CAD systems to assist doctors. However, under DeepSynthBody, the main goal of these experiments was changed to benchmark analysis. The summary of all the GI-tract analyses performed for the thesis is tabulated in Table 3.3. Some of the GI-tract analyses [30, 36, 35] have been performed as a part of competitions such as MedicoTask [184] and EndoCV-2021[3] grand challenge, which has used similar GI-tract data to HyperKvasir data [23] used in this thesis. Participating in competitions and solving the tasks given by the organizers helps to make benchmarks and analyze them globally with other participants of the competitions. Our initial objective was to produce well-performing ML models for CAD systems to assist doctors. However, the participating competitions and providing well-performing solutions such as the winning solution [35] provided to the EndoCV-2021 make them popular and get researchers' awareness to enhance them.

Moreover, the cross-data evaluations performed in our paper [31] show the data-bias problem occurred due to training ML models using a single training dataset. This generalizability issue occurs due to the lack of diverse medical datasets. This medical data

[3] https://endocv2021.grand-challenge.org/

shortage is identified as the main research question of this thesis. Additionally, in this study, the requirement of fair evaluations using multiple metrics such as accuracy, recall, precision, F1, MCC, and specificity were discussed when the cross dataset evaluations are performed as proof of generalizability.

Not only producing benchmark results, but proper evaluation criteria used to analyze them are essential. Therefore, an online calculator[4] [33] to calculate proper evaluation metrics for binary classification models was implemented with given proper guidelines using the GI-tract images classification as a case study. Using this tool, researchers (or other users of this tool) can calculate evaluation metrics for their studies and calculate missing metrics of other relevant studies that need to be compared. This tool makes a common platform for comparing studies fairly.

The performance of an ML model can depend on the resolution of input images to CNNs. Therefore, another investigation [32] was conducted to find the correlations between input resolutions and output performance using GI-tract images. Two different CNNs models (ResNet-152 [185] and Densenet-161 [186]) to classify 23 classes of the labeled folder in the HyperKvasir dataset [23] were investigated and presented the importance of having high resolutions images for CNNs.

A total of six benchmark analyses have been performed in this thesis to achieve Sub-objective III using the GI-tract datasets. The six models consists of four classifications [30, 31, 36, 32, 33], and two segmentation tasks [36, 35]. However, in this section, we considered all these implementations as benchmark analyses because our Sub-objective III is not producing ML models for CAD systems but investigate the data-related problems. These benchmark evaluations and corresponding results using the GI-tract data can be found in original articles tabulated in Table 3.3.

Sperm Video Analysis

According to the data and ground truth provided in the sperm dataset [69], the intended research work is to predict the morphology and motility level of the sperm samples in the dataset. The prediction of morphology and motility levels were identified as regression problems. The summary of all the studies conducted using this sperm dataset for this thesis is tabulated in Table 3.4.

[4] `https://medimetrics.no/medimetrics/`

Table 3.3: GI-tract analysis done for producing baseline results and benchmark results. Two-type of analysis and different type of ways to produce baseline and benchmark results are tabulated here. These analysis results are relevant to the layer of collecting real data and analysis of DeepSynthBody.

Study	Analysis type	Description
[30]	Classification	Two CNNs are presented in this study to classify 16 classes of GI-tract finding given in the dataset of MediCoTask 2018 [184].
[31]	Classification	This study shows the importance of performing cross-dataset evaluations because training ML models using small datasets shows the data-bias behaviours [187].
[33]	Classification	Importance of fair evaluations of the predictions from ML solutions is discussed in this study. Therefore, an online tool to produce proper evaluation results is presented and published with this paper. The tool can help researchers to evaluate classification models. The study was validated using a review of studies of GI-tract analysis.
[32]	Classification	These studies show the effect of the resolutions of the input images using with CNNs. The importance of having high-resolution medical images is emphasised in these studing using GI-tract images as case study data.
[36]	Segmentation	The data augmentation method (PYRA) introduced in this study discuss how grid-like augmentation can improve the generalizability of polyp segmentation. This the segmentation solution proposed to the benchmark challenge in the Medico task at MediaEval 2020 [188].
[35]	Segmentation	The winning solution of EndoCV2021 is presented in this paper. Participating competitions and producing ML solutions for them help to figure out limitations and challenges of real medical data sources.

We have performed three studies [38, 39, 40] using different pre-processing techniques and various types of CNNs. The main objective was to predict the morphology and motility levels of the sperm videos, which contain recorded videos of microscopic sperm analyses. Dense-optical flow [189] and Lucas-Kanade's algorithm [191] to predict sparse optical flow were investigated as pre-processing techniques for the study [38]. In addition to the optical flow extractions, stacked gray-scale video frames as input were tested. Moreover, video frames were reshaped to vertical frames and stacked to prepare new data structures to compress multiple video frames into one. This new structured data were also investigated in the study [38].

The dense-optical flow extractions with different amounts of frame strides were in-

Table 3.4: Summary of real sperm data analysis. Predicting motility and morphology is the main research problem with this dataset. The analysis type of this dataset is regression.

Study	Analysis type	Description
[38]	Regression	Four type of pre-processing techniques were experimented to predict morphology and motility level of the sperm videos.
[39]	Regression	Using the Dense-optical flow [189] algorithm, the videos were pre-processed before passing them into CNN architectures to predict morphology and motility levels. This implementation was submitted to MedicoTask-2019 [190].
[40]	Regression	Auto-encoders were used to extract temporal features into $2D$ spatial domain and the featured were analysed using CNNs tp predict morphology and motility levels of perm samples. The solution was proposed for MedicoTask-2019 [190].
[68]	Regression	A challenge named BioMedia organised for the ACM Multimedia grand challenge 2020. Participants were asked to develop ML solutions to predict morphology and motility levels automatically.

vestigated in our study [39] for the sperm benchmark analysis problem in MedicoTask 2019 [190]. The stride amount is the gap between video frames extracted to calculate the dense-optical flow. The three-fold cross-validation with ResNet-34 [185] was performed to evaluate the models. For the same task, an auto-encoder-based solution was presented as a new submission [40]. In the second solution, auto-encoders were used to extract temporal information from stacked input video frames. The extracted temporal features act as images to CNNs to predict morphology and motility levels of the sperm videos. The extracted features are not readable to humans. However, CNNs trained using these features could learn to predict the morphology and the motility levels of the sperm samples.

These benchmark results show how difficult to predict the motility and morphology levels only using a small dataset. The results of these experiments reflect the quality of the dataset and also the requirements to improve it. More details about these benchmark analyses performed for the sperm dataset can be accessed from the original papers [38, 39, 40].

3.2 Step II: Developing Generative Models

Step II is the core step of the DeepSynthBody framework. This step is two folds. First one is designing generative models and finding the best models using evaluation processes. The second is publishing the best generative models to the end-users who need synthetic data. Different GAN types and the evaluation methods used to evaluate deep generative models are discussed followed by the methods for publishing GAN models in the DeepSynthBody framework for the end-users.

3.2.1 Generative Model Design and Evaluation

Designing and evaluating GANs for generating synthetic data is the first process in Step II, developing generative models. After collecting and analyzing real medical datasets in Step I, GANs should be investigated to generate synthetic data to achieve the sub-objective III. Sub-objective III focuses on generating synthetic data to overcome the medical data deficiency problem which is the major obstacle for developing AI-based solutions in the medical domain.

The three datasets, analyzed in the data analysis stage, the ECG dataset, the GI-tract dataset, and the sperm dataset, were used as case studies. Comprehensive details of the designing GANs are discussed in this section because the GAN designing and getting state-of-the-art performances are essential for DeepSynthBody as it is the core of this framework. In addition to the GANs design methodology, a novel tool named "GANEx", a graphical user interface (GUI)-based GAN training tool, was introduced. A summary of all GAN-related studies performed for this thesis is summarized in Table 3.5.

Generating Synthetic electrocardiogram Signals

The ECG dataset discussed in our benchmark paper [41] would be a popular dataset for the people doing ECG analysis if it is not a private dataset. Unfortunately, many datasets like this are hidden from researchers as a result of privacy concerns. Therefore, GANs for generating synthetic ECGs were developed in this thesis to generate synthetic ECG data to share public instead of the restricted real dataset.

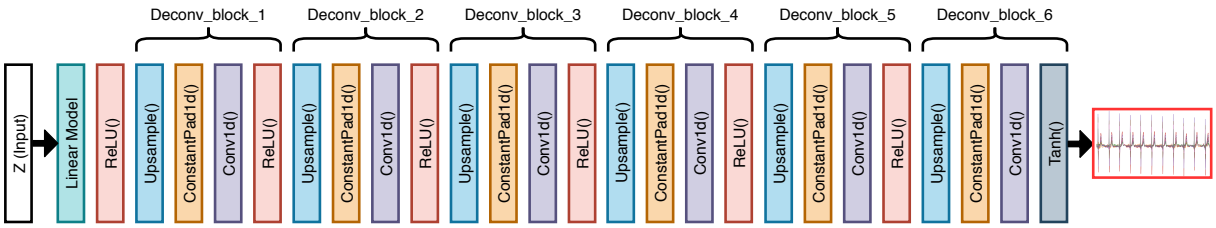
The first GAN architecture to generate synthetic ECG data was inspired by the WaveGAN [192] architecture introduced by Donahue, McAuley, and Puckette. The original

Study	Task of GANs	Description
[70]	Generate synthetic ECG	A novel GAN architecture called Pulse2pulse was introduced to generate synthetic 10s long ECGs with eight-leads to overcome privacy issues of the real dataset.
[72, 73]	Pre-process input data	GAN architectures were experimented to fill a part of GI-tract images, which is the green box appeared at the bottom right corner of the images in HyperKvasir dataset [23].
[74]	Generate synthetic video frames	A GAN architecture named *Vid2pix* with a 3D CNN were investigated to generate synthetic Pilcam video frames [27] for time step $t+1$ conditioning on time steps $t, t-1, t-2$.
[67, 75]	Generate synthetic images with corresponding ground truth mask	GAN architectures were experimented to generate synthetic polyp images and corresponding ground truth mask as proof of concepts to solve privacy issues and medical data annotation cost problem.
[76]	Generate synthetic painting to sperm video frames	A GAN model was experimented to generate a painting like spots instead of sperms in a sperm video frame. This study was focused to generate sperm locations in a synthetic paintings for simple sperm analysis.
[77]	A tool to prefrom GAN experiment	GANEx is a tool with a GUI to perform series of GAN experiments for non-computer science people who want to produce data to DeepSynthBody.

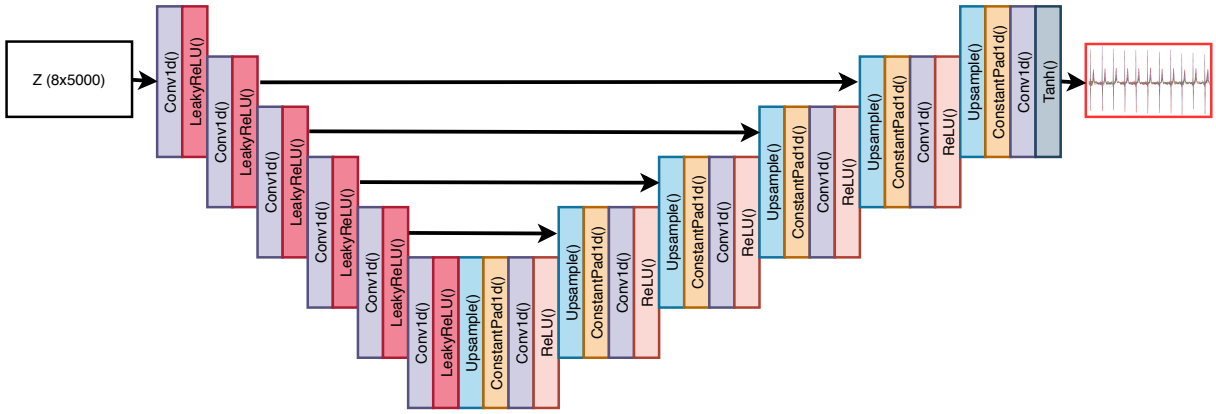
Table 3.5: Summary of GAN-related experiments preformed under this thesis.

WaveGAN was developed to generate synthetic music. Therefore, in the first stage, the WaveGAN architecture was modified to generate ECG signals having a shape of 8×5000, which is the shape of eight-leads $10s$ long ECG samples of the dataset, and it was named WaveGAN*. Then, generated samples from WaveGAN* were analyzed qualitatively and quantitatively. The qualitative analysis was done by inspecting 12-leads plots, and for quantitative analysis, the evaluation reports collected from the MUSE system were used. According to the results, WaveGAN* had to be improved further to get better synthetic ECGs. Therefore, a novel architecture named Pulse2pulse [70], inspired by the UNet architecture [193], was introduced for the DeepSynthBody framework in this thesis.

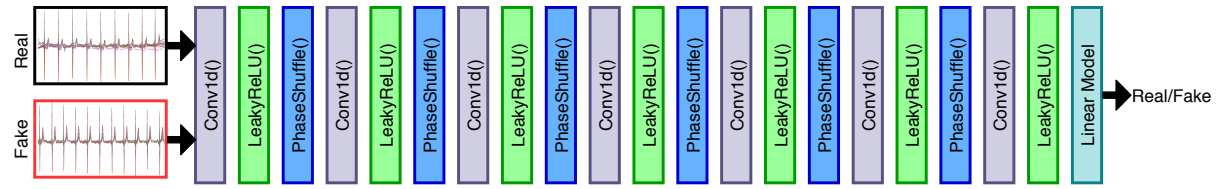
ECGs from the *Normal ECG* category of the dataset were used to train both GAN architectures because the *Normal ECG* category is the biggest population of the dataset (refer the Table 3.2). The discriminator used for both GAN architectures was adapted from the discriminator introduced in WaveGAN [192]. The modified WaveGAN generator, Pulse2pulse generator, and discriminator used for both GAN networks are illustrated in



(a) WaveGAN* generator.



(b) Pulse2pulse generator.



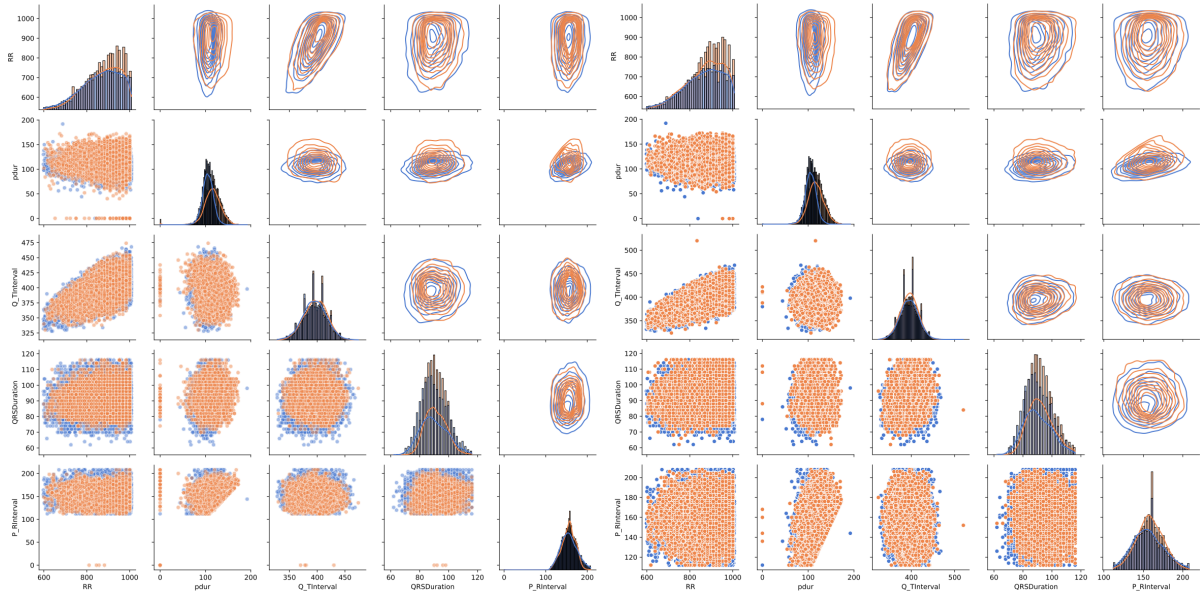
(c) Architecture of the discriminator used for both generators.

Figure 3.9: Model architectures of the generators and the discriminator used to generate synthetic ECGs. WaveGAN* uses a 1D noise vector with 100 points. Pulse2Pulse uses a 2D noise vector with size of 8×5000. Reference for the figure: [70].

Figure 3.9. The complete architecture details are discussed in the full paper [70].

For both models, WaveGAN* and Pulse2Pulse, the best checkpoints were found using MUSE analysis reports collected from generated $10,000$ ECGs per checkpoint from every 500 epochs. Then, the two best checkpoints of WaveGAN* and Pluse2pluse were evaluated further for better understanding before publishing them to the end-users of DeepSynthBody. Five main properties of an ECG, namely RR, P duration, QT interval, QRS duration, and PR interval, were selected to compare the distributions of the selected best checkpoints. The distribution plots are illustrated in Figure 3.10. The blue color dots represent real normal ECG samples, and orange color dots represent generated ECG samples from WaveGAN* and Pulse2pulse.

Comparing distributions of ECG properties, WaveGAN* shows less accurate distribution overlaps with the distributions of real data compared to Pulse2Pulse. This difference



(a) From WaveGAN*. (b) From Pulse2pulse.

Figure 3.10: Comparisons of MUSE predictions using characteristic distribution plots. Blue color plots represent real normal ECG distributions. Orange color plots represents distribution of fake ECGs generated by WaveGAN* and Pulse2pulse respectively in Figure 3.10a and Figure 3.10b. The "nan" values of the selected five features of "Normal ECG"s were converted into 0 to identify predicted "nan" values by the MUSE system.

can easily be noticed from the row presenting correlations between PR interval and other properties. Also, WavGAN* generated faulty synthetic ECG samples making more "nan" values in the MUSE analysis report than Pulse2Pulse. The MUSE algorithms give "nan" values when the algorithm cannot predict the specific property of an ECG. These statistical comparisons are discussed in our full paper [70].

After finding that the novel Pulse2Pulse architecture can generate better quality synthetic ECGs than WaveGAN*, a large synthetic ECG dataset with 150,000 samples was generated using the best checkpoint of Pulse2Pulse. Then, the synthetic dataset was analyzed using the MUSE system to predict the properties of the ECGs. From the MUSE analysis report, the most important nine properties, namely *RR, Ventricular Rate, pdur, QT interval, QRS duration, PR interval STJ V5, RPeakAmp V5, and TPeakAmp V5*, were further analyzed statistically, and the collected results are tabulated in Table 3.6 to compare with the real Normal ECG data statistics.

Table 3.6 presents statistics collected from three datasets for the selected parameters. First, statistics about the real ECG data (filtered "Normal" ECGs), used to train the GAN models are tabulated. Then, statistics about all the generated 150,000 ECGs and

Table 3.6: Comparison of MUSE analysis reports' statistics for selected ECG properties.

	Real				150k				All Normal (121977)			
	Mean	Std	2.5%	97.5%	Mean	Std	2.5%	97.5%	Mean	Std	2.5%	97.5%
RR	866	90	670	1000	870	91	667	1000	870	87	682	1000
VentricularRate	70	8	60	90	70	8	60	90	70	8	60	88
pdur	105	12	82	130	118	17	84	152	117	17	86	152
Q_TInterval	395	21	352	436	395	21	354	436	395	20	354	434
QRSDuration	90	9	74	110	93	10	78	114	92	9	78	112
P_RInterval	156	19	120	198	159	18	124	194	158	17	124	192
STJ_V5	2	27	-44	58	16	36	-54	92	18	33	-44	87
RPeakAmp_V5	1287	402	600	2163	1272	404	561	2114	1276	370	615	2031
TPeakAmp_V5	343	137	126	664	360	141	141	678	364	134	151	668

statistics about filtered "Normal" ECGs (121977) from 150,000 ECGs were tabulated. To achieve sub-objectives II and IV, collecting and developing medical data and developing generative models to generate synthetic data, the synthetic ECGs should have similar characteristics as real ECGs. According to the results presented in Table 3.6, the synthetic ECGs show similar statistical properties to real ECGs, such as equal or very close mean and std values for ventricular rate and QT interval. To present the qualitative properties of synthetic ECGs generated from Pulse2pulse, Figure 3.11 shows two synthetic ECG samples identified as "Normal" according to the MUSE report. Additionally, the 150,000 synthetic ECG dataset and the filtered 121977 "Normal" ECGs can be downloaded with the corresponding MUSE reports from `https://osf.io/6hved/`.

In summary, we could see that our Pulse2Pulse generates realistic synthetic data with very close properties to the real dataset. Then, these generated synthetic ECGs can be used to share to the public instead of the real dataset with privacy concerns.

Generating Synthetic gastrointestinal-tract Images

The HyperKvasir dataset [23] is used as the main case study to experiment with GANs for GI-tract data. Additionally, the Kvasir-Capsule [27] dataset is used. Using these datasets, several GANs were developed to investigate how GANs can generate synthetic medical image data, in this case, GI-tract images.

Several preliminary experiments were performed to use GANs to fill missing parts of GI-tract images [72, 73] and predict future frames of the Pilcam videos of GI-tract videos [74]. The studies [72, 73] focused on removing green boxes that appeared in GI-tract images by generating synthetic filling using GANs. Sample GI-tract images with green boxes are presented in Figure 3.16. The green box removing process is a preprocessing

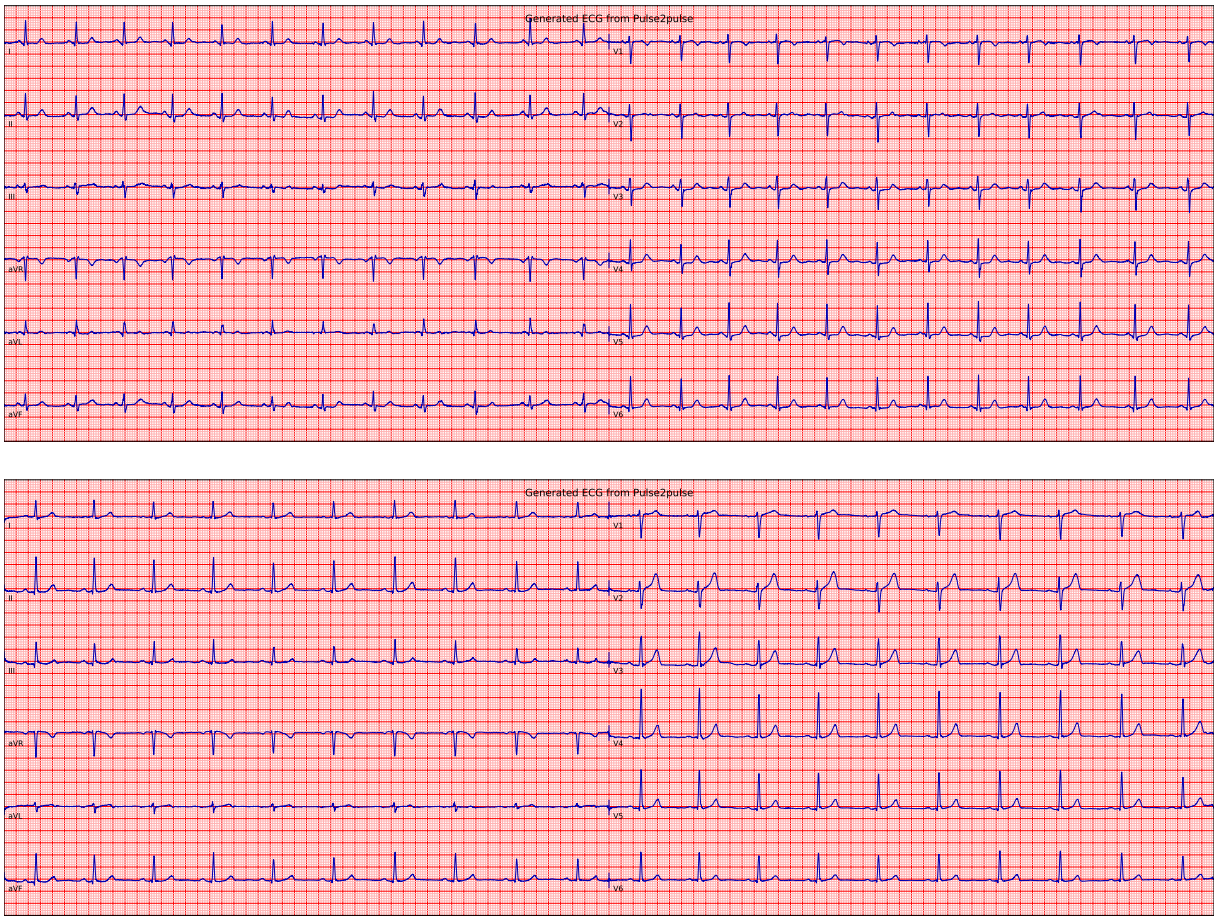


Figure 3.11: 12-leads plots of fake ECG samples from the novel ECG generator introduced in this study: Pulse2pulse.

step to prepare GI-tract images for other ML models. Then, the main goal of this study is to find the effect of removing green boxes that appeared in the GI-tract images on the HyperKvasir dataset by replacing the green box with a generated realistic replacement.

In the preliminary experiment [74], a GAN was researched and developed to generate synthetic video frames for capsule endoscopy (pill cam) videos [27]. The GAN architecture experimented for the video generations process has used $3D$ CNN to predict future frames of the videos to extend the available real dataset to improve the dataset. Then, the goal of improving data is to improve the performance of other machine learning algorithms which use extended synthetic videos.

The generative models discussed with the preliminary experiments have given the foundation to build other GANs discussing in this section. However, quantitative and qualitative analyses show that the performance of these preliminary experiments was not enough for solving Sub-objective II by generating synthetic medical data. Still, exper-

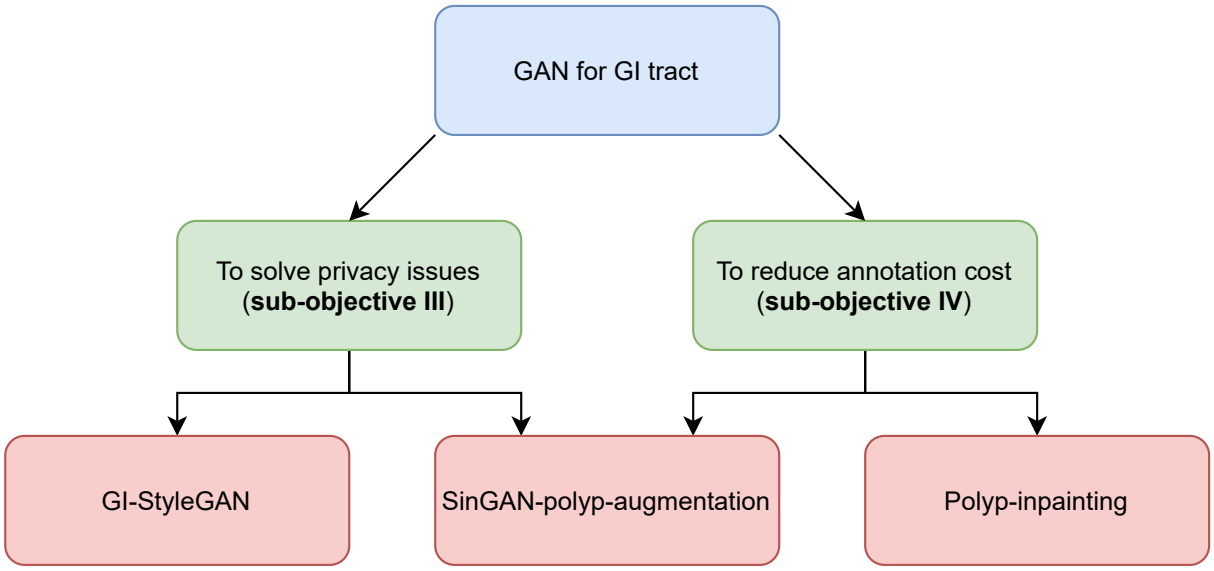


Figure 3.12: Different type of GANs for generating synthetic GI-tract findings for different purpose.

iments discussed in studies [72, 73] are contributed to Sub-objective III of this thesis. Therefore, those GAN architectures were excluded from the final DeepSynthBody platform until improving these using future research works.

Another three advanced GAN architectures were investigated with the HyperKvasir dataset after the foundation analysis from preliminary studies [72, 73, 74]. These three studies, namely GI-StyleGAN [71], SinGAN-polyp-augmentation [67], and Polyp-inpainting [75], were conducted as proof of concepts to mainly address Sub-objective IV, which focus on generating synthetic medical data to solve the data deficiency problem in the medical domain. These three studies and corresponding contributions to the sub-objectives are depicted in Figure 3.12.

The GI-StyleGAN experiment presented in the concept paper of this thesis [71] used StyleGAN-v2 introduced by Karras et al. [194] with the unlabelled data folder of the HyperKvasir dataset to generate synthetic GI-tract images. The main objective of this experiment was to achieve Sub-objective II and IV, which are collecting and developing medical datasets and researching and developing GANs to generate synthetic data. All the unlabelled images (around $100,000$) from HyperKvasir were used to train the StyleGAN-v2 model because the model is prone to a large training dataset. Pytorch implementation of StyleGAN-v2[5] was trained $10,000,000$ steps for more than eight days to get good output. In this training process, checkpoints were saved after every $1000,000$ steps (not using

[5] `https://github.com/lucidrains/stylegan2-pytorch`

Table 3.7: FID scores calculated from different checkpoints of StyleGAN trained for generating GI-tract findings.

chk_point	FID_64	FID_192	FID_768	FID_2048
0	39.1090	189.4938	2.6159	342.0751
100	1.7710	8.3480	0.3030	58.9490
200	1.6616	8.0271	0.2977	59.7215
300	1.6575	7.8310	0.2671	52.6597
400	1.2801	6.1183	0.2429	48.5694
500	1.2262	5.8759	0.2372	49.3512
600	1.5974	7.4586	0.2626	52.9441
700	1.3826	6.5063	0.2363	46.2668
800	1.1938	5.9112	0.2312	46.7931
900	**0.6537**	**3.0260**	0.2017	44.3310
1000	0.8736	4.2926	**0.1980**	**41.2039**

epochs) to check the progress of the quality of generated GI images and Frechet inception distance (FID) values introduced by Heusel et al. [195] were calculated to find the best checkpoint. The calculated FID values from different feature layers, namely 64: first max-pooling features, 192: second max-pooling features, 768: preaux classifier features, and 2048: final average pooling features, are tabulated in Table 3.7. Randomly picked synthetic colon images are presented in Figure 3.13. The presented images show that the StyleGAN implementation is capable of generating realistic synthetic colon images. This colon StyleGAN is not only for generating random images, but it can generate interpolated images between two randomly generated images, as depicted in Figure 3.14. This functionality introduced in the vanilla implementation of StyleGAN [196] can generate synthetic data as needed for end-users. This particular generative model can be used to achieve the **sub-objective II and IV**, aiming to develop medical datasets and solve privacy concerns by generating synthetic medical data.

Generating synthetic data with corresponding ground truth is challenging than generating random synthetic data samples solely. However, generating both synthetic data and ground truth is essential to overcome the data deficiency problem to achieve sub-objectives II and IV. We can use synthetic data to replace the costly and time-consuming medical data annotation process, which is identified as one of the reasons causing the data deficiency problem. We can generate both synthetic data and the corresponding ground truth using GANs to solve the problem.

The polyp inpainting GAN [75], capable of generating synthetic polyps on clean

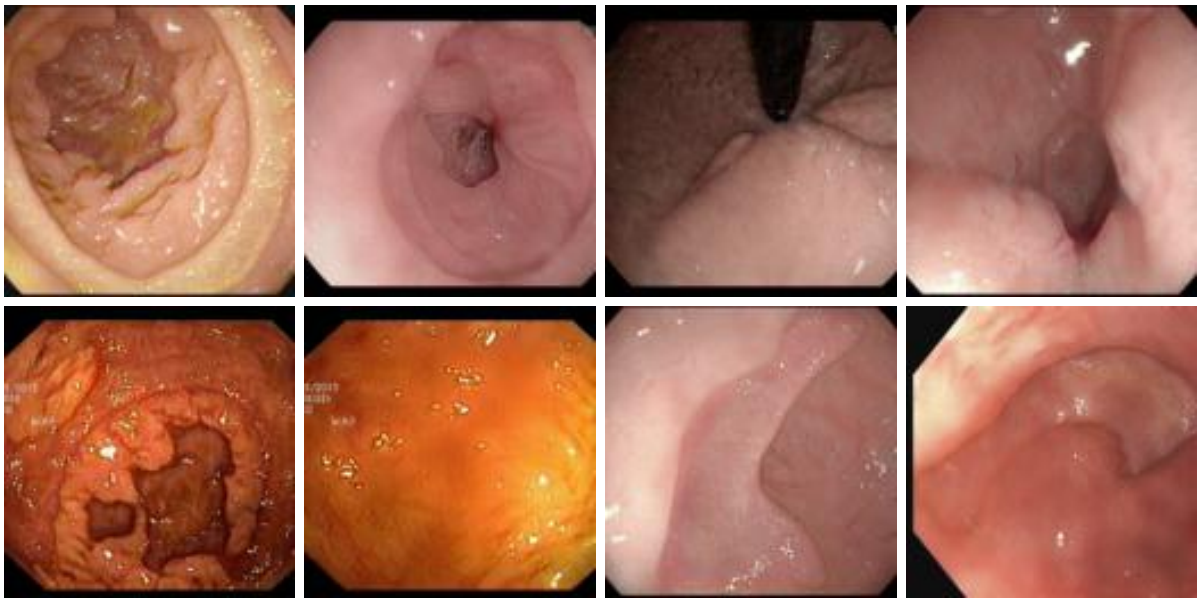


Figure 3.13: Style-GAN generated random gastrointestinal-tract findings.

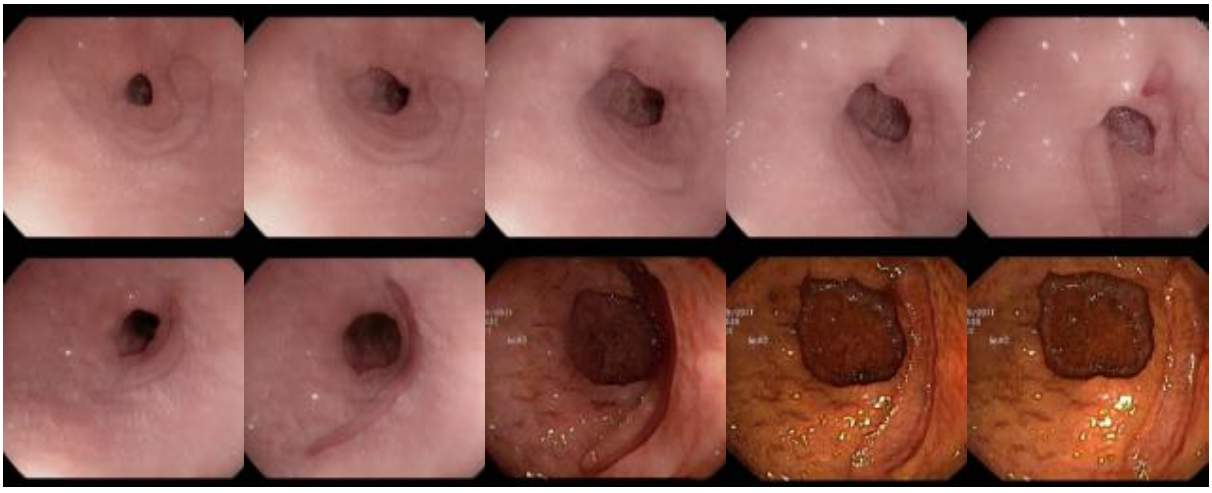


Figure 3.14: First five samples generated with 200 interpolation steps for two different random seeds. First and second row represent the two different random seeds. [71]

colon images, is another study performed with GI data. This gan was researched and developed as the first solution to overcome the data annotation problem, as presented using the third leaf node of Figure 3.12. In this experiment, image inpainting using generative multi-column CNN presented by Wang et al. [197] was studied, researched, and developed to do polyps inpainting for non-polyp images using given masks that represent regions of interest to have polyps. However, the available polyp data in the polyp datasets are not enough to train the GAN from scratch. Therefore, the inpainting GAN model was trained from clean colon image folders as the first step. The clean colon image folders have enough images identified as non-disease images by experts to train a DL model. After training with the clean colon data, the model was retrained using polyp data and

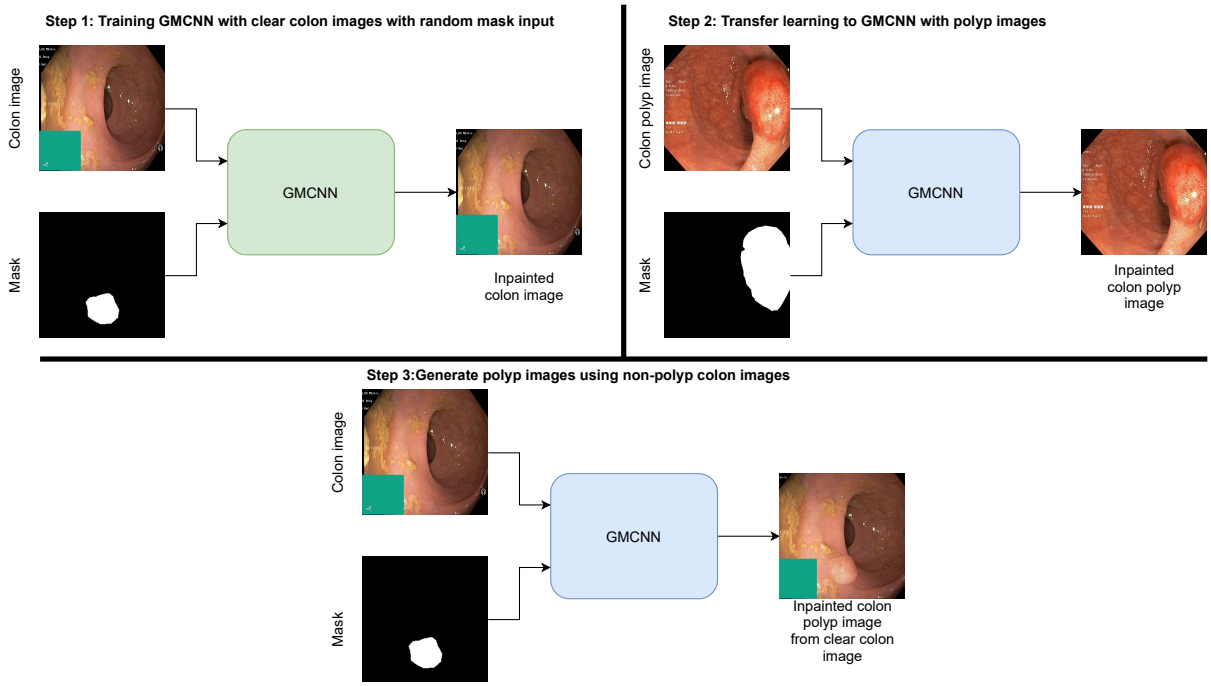


Figure 3.15: Steps of the polyp inpainting training process as discussed in [75]. Generative multi-column convolutional neural networks (GMCNN) [197] is the core network in this process.

corresponding masks using the transfer learning mechanism [198] to generate polyps on clean colon images for given masks. This training process is illustrated in Figure 3.15 according to the steps discussed in the original publication [75].

After the training process, the polyp inpainting GAN can convert clean colon images into corresponding polyp images using given masks. Therefore, this inpainting GAN can generate synthetic polyp datasets with the masks of the polyp regions. Then, the inpainting GAN can be used as a solution to achieve Sub-objective IV by producing synthetic data as alternatives to the resource-consuming medical data annotation process. The inpainting GAN can generate synthetic polyps for given random polyp masks without any aid from experts. Therefore, we can use this type of GANs to generate synthetic true positive data from true negative data, which are common and easy to find. We showed that synthetic polyps show visual properties also indistinguishable from real samples for the domain experts.

A qualitative analysis was done using a survey with medical experts to evaluate the quality of the synthetic polyps generated from the polyp inpainting GAN. Using the polyp inpainting GAN, synthetic polyps were generated and analyzed by domain experts. The experts analyzed five synthetic and five real polyps samples. The samples used for a

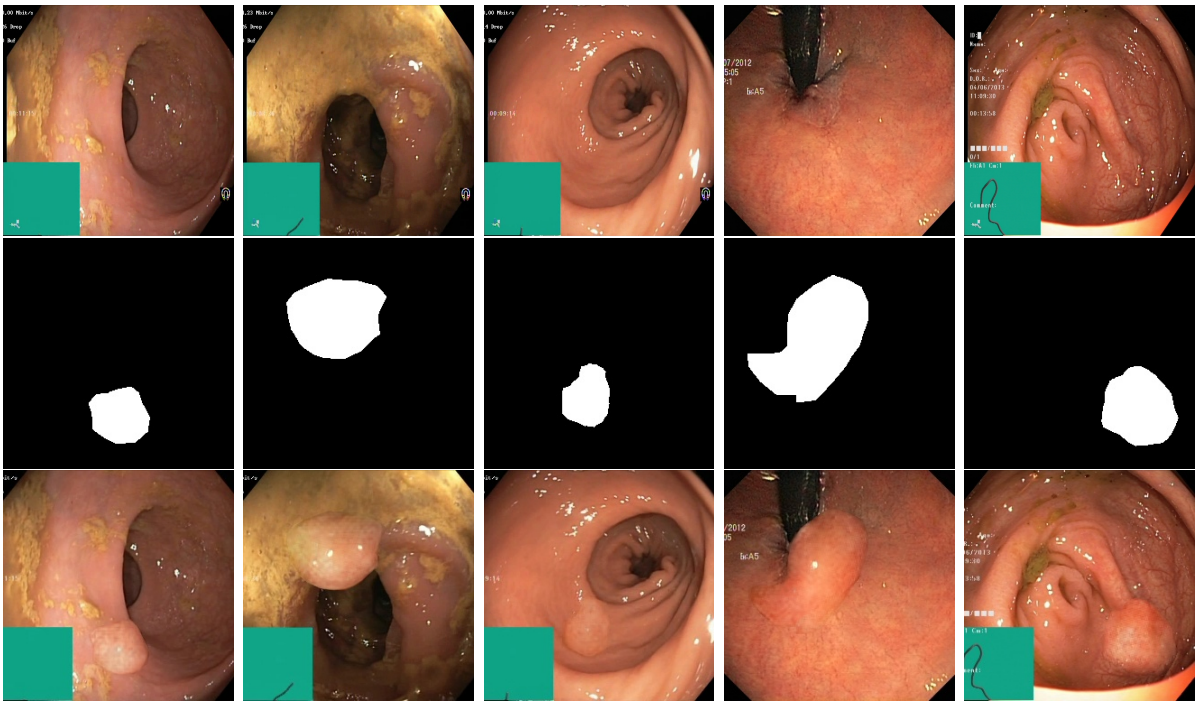


Figure 3.16: Polyp inpainted samples from polyp inpainting gan. The first row illustrates input images. The images in the second row represent input masks used with input images. The third row represents the output images from the polyp inpanting GAN.

questionnaire are presented in Figure 3.16. In this questionnaire, experts were asked to discriminate synthetic polyps from real polyps and give a confidence score for the particular selection. Two experts, three non-experts and three internal medicine residents (total is eight) have participated in this questionnaire. The summary of the results collected from this questionnaire is presented in Table 3.8. Finally, the proposed GAN architecture can generate synthetic polyp image conditioned on a clean colon image and a random mask representing a polyp region. The polyp inpainting GAN shows that modified GAN architectures can generate synthetic data with corresponding masks, usually prepared by experts manually, which is a costly and time-consuming task. More details about this polyp inpainting GAN can be found in our original paper [75]. However, this inpainting GAN is not suitable for a privacy-preserving data sharing technique because the non-polyp regions are identical to the real clean colon images.

SinGAN-Seg [67] was investigated in this thesis to achieve sub-objectives III and IV. The SinGAN-Seg implementation was inspired by the original SinGAN introduced by Rott Shaham, Dekel, and Michaeli [149]. The vanilla SinGAN learns from a single image and generates synthetic samples similar to the pixel distribution of the image used to train it. The original paper presents different applications such as paint to image,

Table 3.8: Overview of obtained results from all 8 readers (2 experts - EE and 3 non-experts - NE, 3 internal medicine residents - IM) for discriminating real and inpainted polyps.

Reader	TP	FN	FP	TN	Accuracy
EE1	3	4	2	1	0.4
EE2	3	5	2	0	0.3
NE1	2	1	3	4	0.6
NE2	3	2	2	3	0.6
NE3	4	2	1	3	0.7
IM1	4	2	1	3	0.7
IM2	4	3	1	2	0.6
IM3	4	1	1	4	0.8

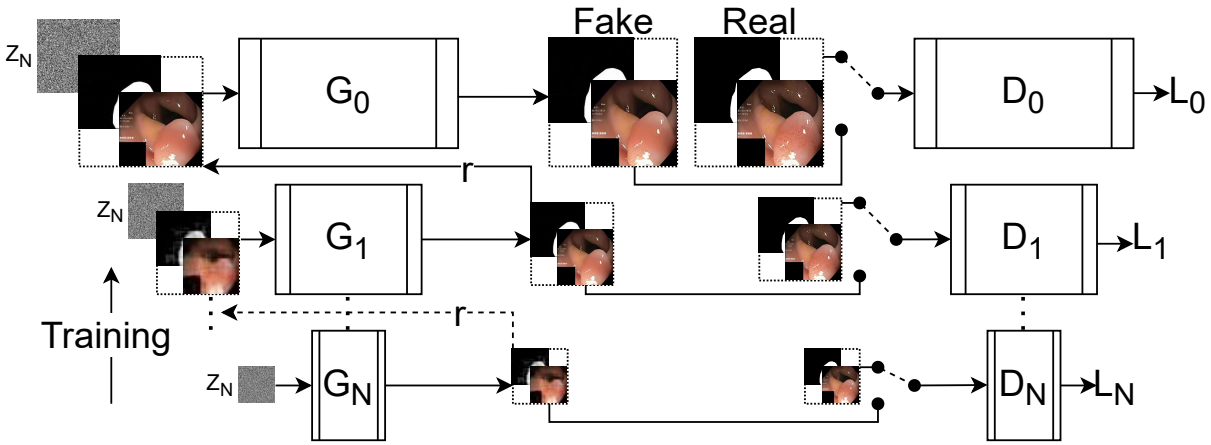


Figure 3.17: A representation of the four-channels SinGAN training step.

super-resolution, editing images, harmonization, and generating animations using a single image. In our SinGAN study [67], the original SinGAN was changed to input four channels containing the input image and its ground truth mask. Then, the modified SinGAN was named SinGAN-Seg because it has a generated synthetic image and its ground truth mask (segmentation mask). So, SinGAN-Seg is a modified version of SinGAN to perform the novel application that generates random images and the corresponding segmentation masks. This SinGAN-Seg was introduced in this thesis to address the sub-objectives III and IV. The complete training process of SinGAN-Seg is depicted in Figure 3.17.

The SinGAN-Seg architectures were trained using the 1000 polyps images of the HyperKvasir dataset. Then, 1000 different checkpoints were generated to replace the 1000 polyp images to demonstrate the capabilities of novel sinGAN-Seg to solve privacy concerns and resource-consuming medical data augmentation process. Synthetic polyp images and corresponding ground truth masks generated automatically using SinGAN-Seg are

depicted in Figure 3.18. The first column of the figure presents real images and corresponding masks of polyp regions, annotated by experts manually. Other columns present generated synthetic polyps and generated masks from SinGAN-Seg learned from the input image of the first column. While the training data consists of only polyp images, SinGAN-Seg can generate non-polyp images as presented in the 3rd and 4th rows in Figure 3.18. This novel SinGAN-Seg implementation contributed to sub-objectives I, II, III, and IV by presenting a well-performing polyp segmentation model, generating realistic synthetic polyps and corresponding ground truth masks to replace private medical data, and tackling costly and time-consuming medical data annotation process.

After generating synthetic polyp images and corresponding maks using our SinGAN-seg, the global features of the synthetic images look awkward because of the unrealistic texture of synthetic images (see Figure 3.18). As a solution to this, the style-transfer algorithm [199] was used to transfer styles from the training image to generated synthetic images. More details about this style-transfer method can be seen in our paper [67].

In summary, we could generate realistic synthetic GI tract images in three ways. The StyleGAN model can generate random synthetic GI-tract landmarks that are indistinguishable from real samples. The polyp inpainting GAN can generate synthetic polyp images by converting a true-negative sample into a true positive sample. The qualitative analysis shows that domain experts also cannot differentiate between real and synthetic samples generated from this polyp inpainting GAN. Synthetic data generated from polyp inpainting GAN addresses data imbalance problems. SinGAN-Seg is another GAN architecture that is capable of generating synthetic polyps and ground truth masks. This GAN can be used to overcome the costly and time-consuming medical image annotation process, which experts usually do.

Generating Synthetic Sperm Video

Synthetic sperm data generation is another area considered as a case study. However, limited data and time constraints were barriers to producing successful GAN architectures that can be plugged in to the DeepSynthBody framework. However, we performed several experiments using SinGAN to generate painting-like sperm video frames [76] to represent the real sperm data because SinGAN learns from a single image it does not need large datasets.

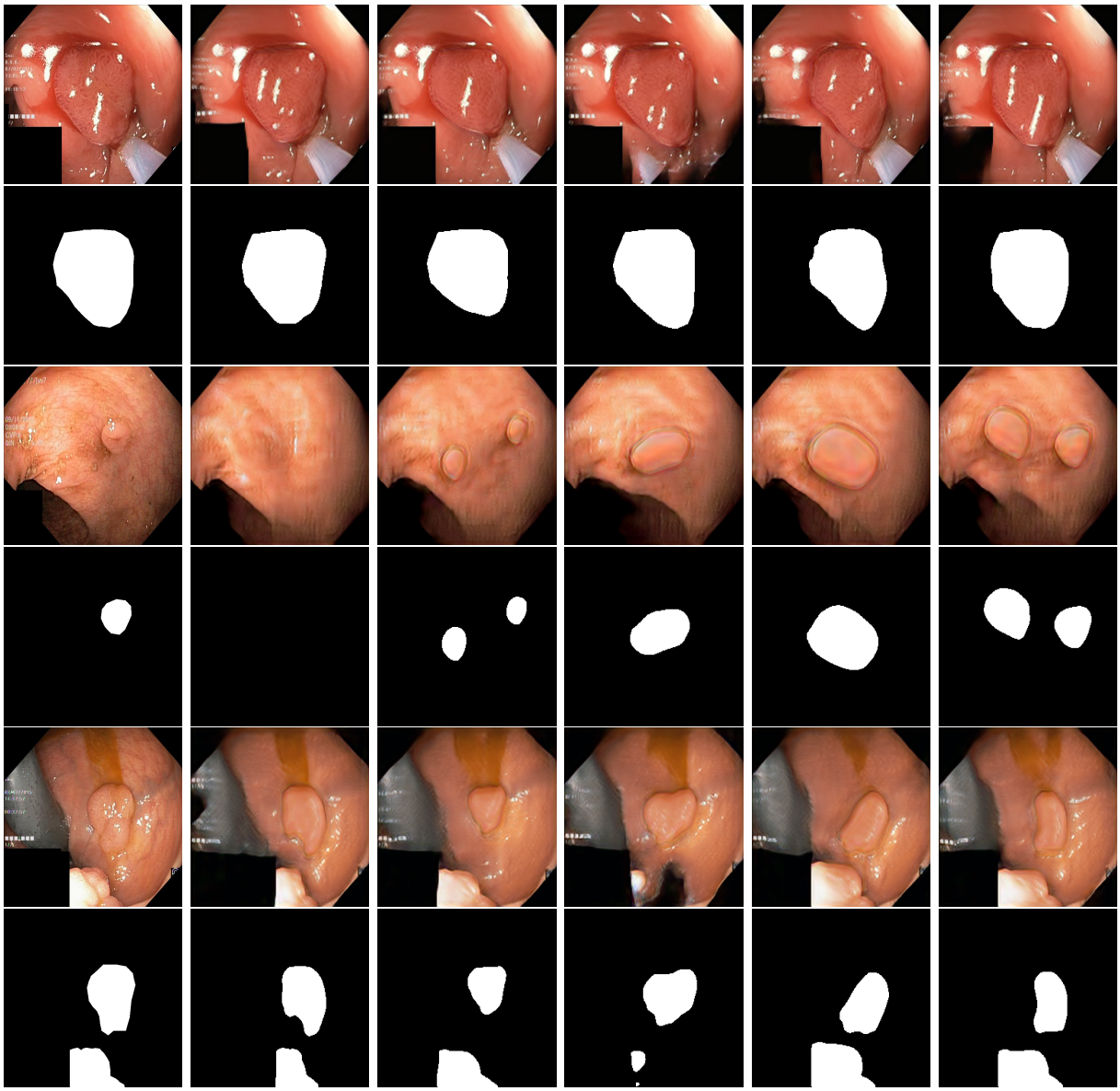


Figure 3.18: Sample real images and corresponding SinGAN generated synthetic GI-tract images with corresponding masks. The first column is illustrated with real images and masks. All other columns represent randomly generated synthetic data from SinGANs which were trained from the image on the first column.

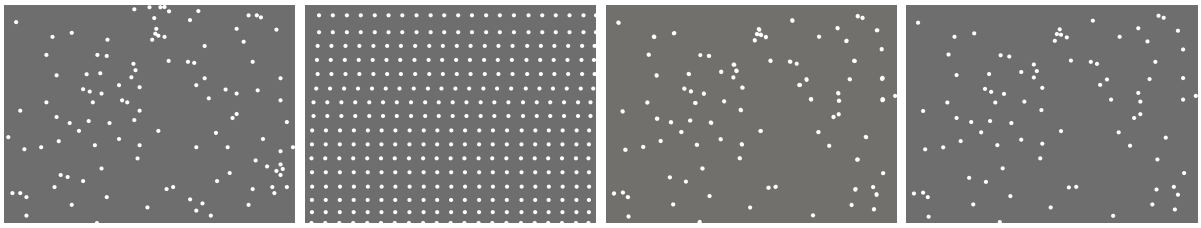


Figure 3.19: Sperm like paintings used to train SinGANs to generate sperm tracking. The last two images have same dot patterns except the background colour.

We used vanilla SinGAN [149] to experiment with the sperm dataset to perform unsupervised sperm segmentation to achieve Sub-objective IV. In this case, the data deficiency problem will be solved by reducing the annotation cost of medical data. In this task, SinGAN was used to track the locations of sperms in an unsupervised way. The complement operation of the paint-to-image operation introduced in the original SinGAN, image-to-paint, was investigated to generate sperm sample-like paintings to represent sperm locations with a clear background. To achieve this, the SinGAN model was trained from a sperm-like picture. Sample training images investigated to train SinGANs are depicted in Figure 3.19. Then, video frames were input into the pre-trained SinGAN using different scale levels as introduced in the original SinGAN implementation. Results were analyzed qualitatively with different input scales. Generated sperm-like paintings from real sperm images can be used to identify sperm locations using this method.

Sample synthetic sperm paintings to represent real sperm sample images are depicted in Figure 3.20. However, the quality of synthetic sperm video frames generated from our SinGAN is not enough for publishing in DeepSynthBody. The results implies that future experiments are required with different GAN architectures and high-quality sperm datasets. A successful GAN architecture to produce sperm like painting can be used to overcome the Sub-objective IV because synthetic sperm like painting can be used for sharing data when privacy concerns are there and, the synthetic sperm like painting is an alternative representation for real sperm video frames which are hard to analyze by experts.

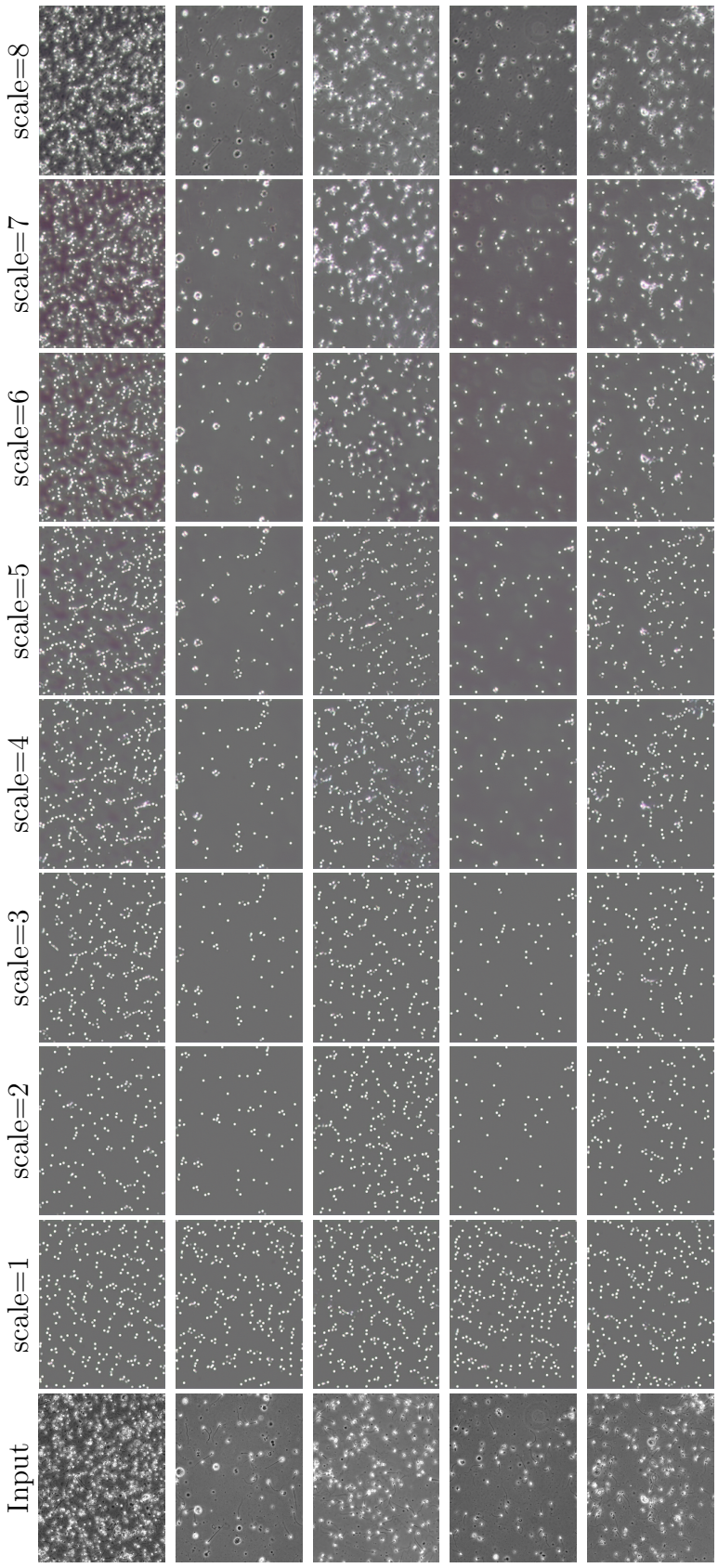


Figure 3.20: Predicted sperm lacations using the SinGAN trained with synthetic sperm samples

In summary, we developed a GAN architecture, based on the SinGAN architecture to generate synthetic sperm data to replace real data. We have generated painting-like sperm images that can measure the quality of the sperm sample. Moreover, this GAN could tract sperm locations using white dots in an unsupervised way. Therefore, usual image processing techniques (without DL) can be used to analyze the sperm samples easily.

3.2.2 Publishing Deep Generative Models

After researching and developing GANs which can generate synthetic data to overcome privacy issues and the costly and time-consuming medical data annotation process, these deep generative models should be published to the end-users. Therefore, the contributors who are developing GANs in DeepSynthBody should have a common platform to share them. As a platform to share the final GAN models with the end-user in this initial stage, the PyPI was selected. Therefore, all the developments were done in the most popular programming language, Python [200], because PyPI is for Python.

> The joy of coding Python should be in seeing short, concise, readable classes that express a lot of action in a small amount of clear code, not in reams of trivial code that bores the reader to death. – Guido van Rossum (creator of Python)

First, the contributors who develop GANs can publish their work as an individual package in PyPI. Then, the PyPI package can be included as a sub-module in the main PyPI called `deepsynthbody`. In cases where PyPI does not work, authors of GAN models, which will be connected with our framework, can share the checkpoints of their deep generative models with corresponding source codes with the main contributors of the framework. If any of these options do not work, researchers can publish only synthetic data in any public data repository, and the corresponding links can be connected to the DeepSynthBody. However, in the latter case, the end-users cannot control the synthetic data generation process.

The flow of PyPI packages is depicted in Figure 3.21. The figure shows how individual PyPI packages are contributing to the main Python package, `deepsynthbody`. First, GAN developers should produce python packages for individual GAN trained for a specific real

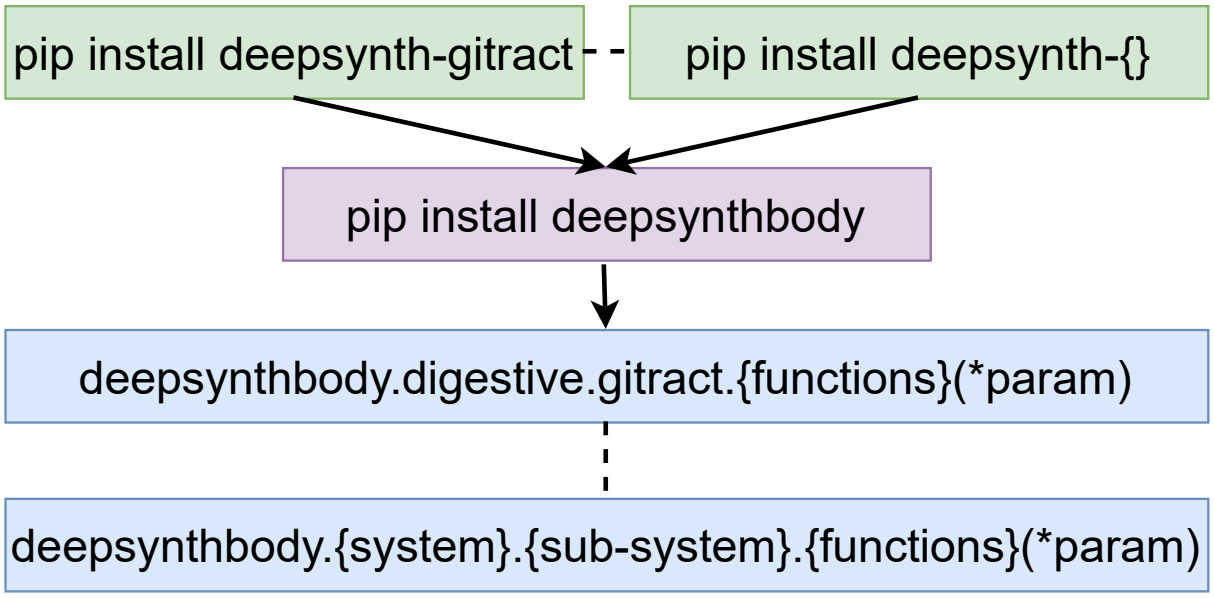


Figure 3.21: The flow of python packages which act as sub-modules of DeepSynthBody framework. The figure reference: [71]

dataset. After training a GAN to produce realistic synthetic data, the GAN can be used as a replacement to the real dataset used to train the GAN. Then, a python package with functionalities to generate synthetic data and the best checkpoints of the GAN model should be packaged into a python package independently. This individual independent python package development process was introduced to reduce the development overhead of the main python package. Finally, these individual packages are connected to the main `deepsynthbody` package according to the human body categorization introduced in Step IV of the framework (see Figure 3.1).

As a proof of concept, two Python packages were developed following the above criteria. First, a python package named deepfake-ecg[6] (`pip install deepfake-ecg`) was published to generate synthetic data from the best checkpoint of the pre-trained Pulse2Pulse [70] ECG GAN. Second, for generating synthetic GI-tract images using the StyleGAN implementation introduced in the paper [71], a python package called `deepsynth-gitract`[7] was published. These packages were developed independently from the `deepsynthbody` package. After publishing the individual packages, they have been connected to the `deepsynthbody`[8] main package.

[6] `https://pypi.org/project/deepfake-ecg/`
[7] `https://pypi.org/project/deepsynth-gitract/`
[8] `https://pypi.org/project/deepsynthbody/`

3.2.3 A Tool to Experiment with Generative Adversarial Networks: GANEx

The DeepSynthBody framework should interact with medical data providers to collect deep generative models,. However, the main challenge is all medical data providers do not have ML programmers who can perform GAN experiments to produce generative models. Additionally, data providers may not have the authority to share the data with intermediate partners to develop GANs. In this context, GANEx (GAN Experimenter) [77] is a tool introduced in this thesis to overcome the barrier of performing GAN experiments by one who does not have a deep understanding of ML or DL. This tool makes a bridge between DeepSyntBody and multi-disciplinary medical data providers.

GANEx consists of two main components: a FastGAN library and a GUI. The FastGAN library is a high-level GAN library, which provides functionalities to create predefined GANs, train GANs and analyze them through a high-end abstract layer called FastGAN Runner, as depicted in Figure 3.22. Using this FastGAN library as the backend, the GUI has been developed to interact with the backend. The GUI of GANEx provides functionalities to create GAN projects, experiments using a predefined collection of GANs provided from the FastGAN library. Then, using the same GUI, users can run and analyze series of GANs using their datasets without writing a single line of code. The whole process of the GUI is illustrated as a flow diagram in Figure 3.23. After completing the GAN training process, the users have GAN checkpoints, which can be shared to generate synthetic data without any privacy concerns.

The sample screenshots of the tool are presented in Figure 3.24. Training progress, a setting page of hyperparameters, and generated sample synthetic data from the CelebA dataset [201] are given in the figure. The given screenshots show how the tool manages every GAN training step without programming (coding). GANEx was developed as a supporting tool to achieve the main objective, which focusing on combing all sub-objectives together to make the functional full framework DeepSynthBody. In addition to this GAN tool, `deepsynthbody.org` is hosted as the main website to achieve the main objective. The website is for both contributors and end-users of the DeepSynthBody framework.

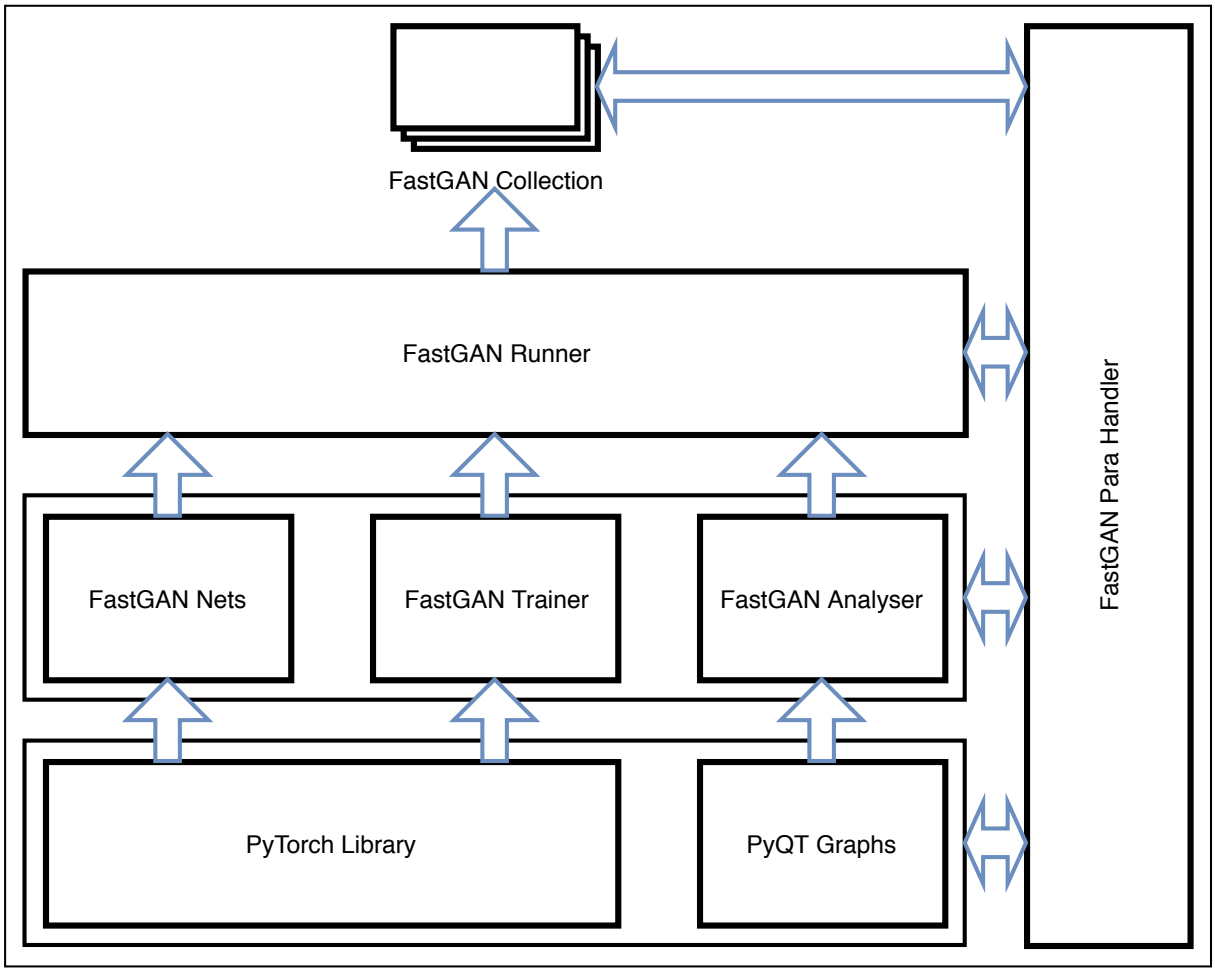


Figure 3.22: The FastGAN library [77] introduced to connect multi-disciplinary user to DeepSynthBody framework.

3.3 Step III: Producing DeepSynth Data

Producing DeepSynth data in Step III is presented in the big picture of DeepSynthBody in Figure 3.1. In other words, this is the layer for the end-users who want to generate synthetic data. This Step III has a flow similar to Step I, but the objectives are slightly different. In Step I, the categorization is used to classify input data, while Step III uses the same categorization to generate synthetic data. The data annotation layer presented in Step I was replaced with two new data generation processes: unconditional and conditional. As the final layer of Step III, synthetic data generation functionalities are used instead of the real data analysis in Step I.

We use the same 11 categories in Step III as used in Step I to generate synthetic data for the end-users. Step III is the output layer of the DeepSynthBody pipeline. We further split the 11 categories into four categories based on the data dimensionality (1-D, 2-D,

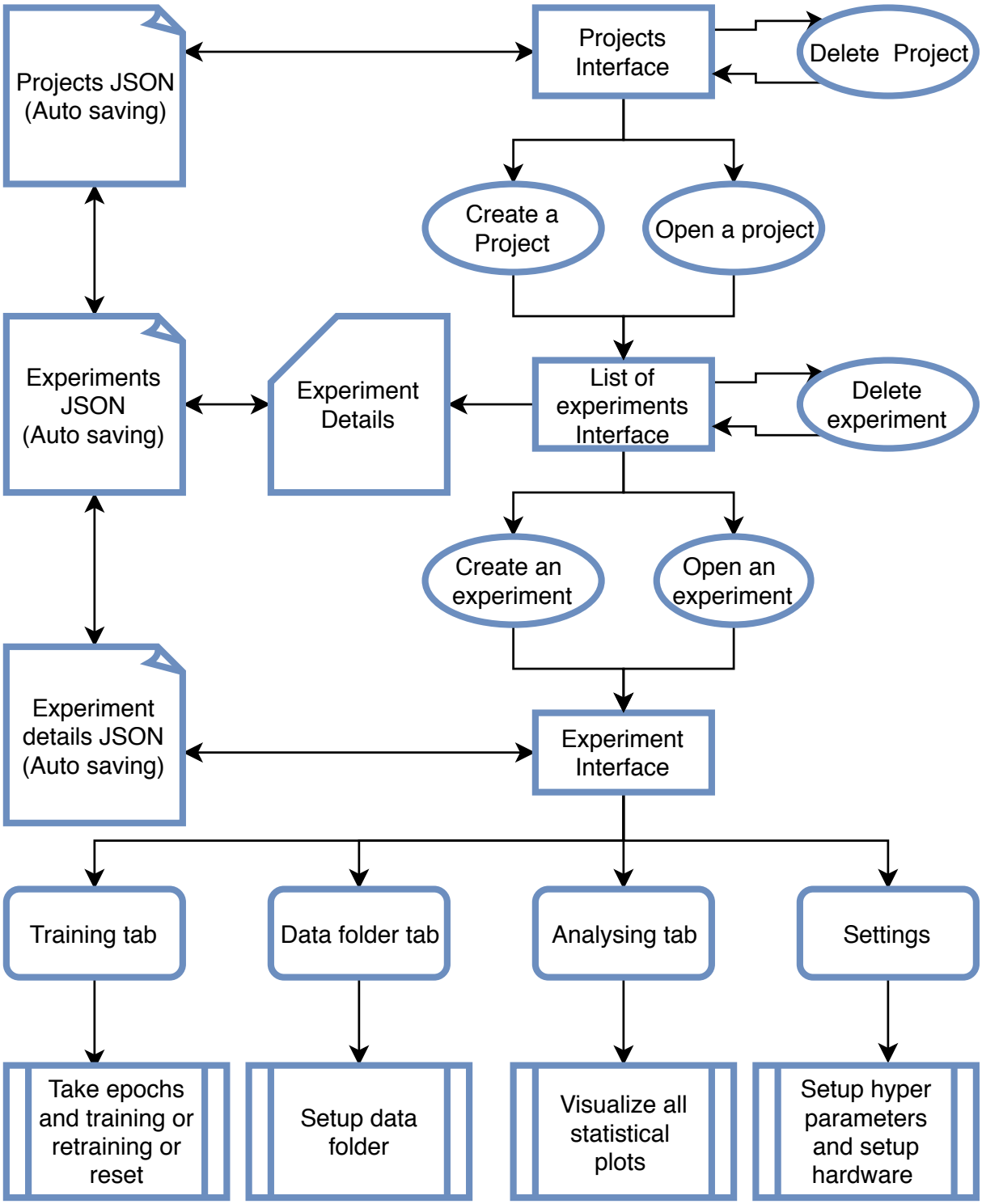


Figure 3.23: GUI flow of GANEx which is a tool to handle GAN experiments for non computer science users of DeepSynthBody.

3-D, and N-D) as discussed in Section 3.1.1. The data dimension layer decides the data output format when there are multiple data formats to generate synthetic data. For example, MRI data can be generated as images (2-D) or volume data (3-D) if both formats are available at DeepSynthBody. In addition, the end-users can decide that the genera-

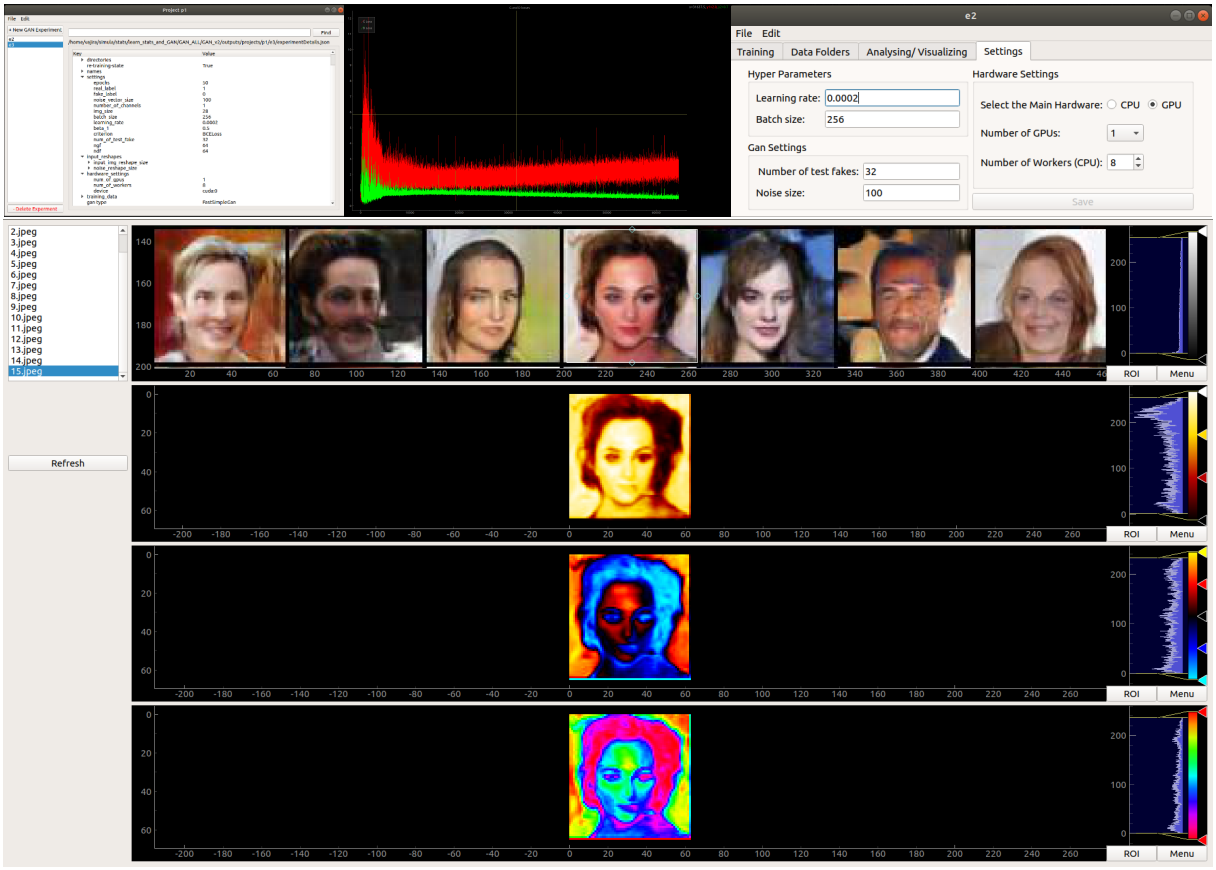


Figure 3.24: Sample screenshots of the GANEx GUI showing user friendly GUI design which can be handled by non computer science multi-disciplinary people. **Top-left**: is showing a screenshot of GAN project management window which shows the summary of all experiments saved in GANEx. **Top-middle**: is showing a screen shot taken from real-time analysis of a GAN experiment using generator loss and discriminator loss. **Top-right**: is showing the window of GANEx which gives functionalities to users to change configuraions of GANs. **Bottom**: is showing GAN generated sample analyser which has functionalities to produce histogram and heat maps of images.

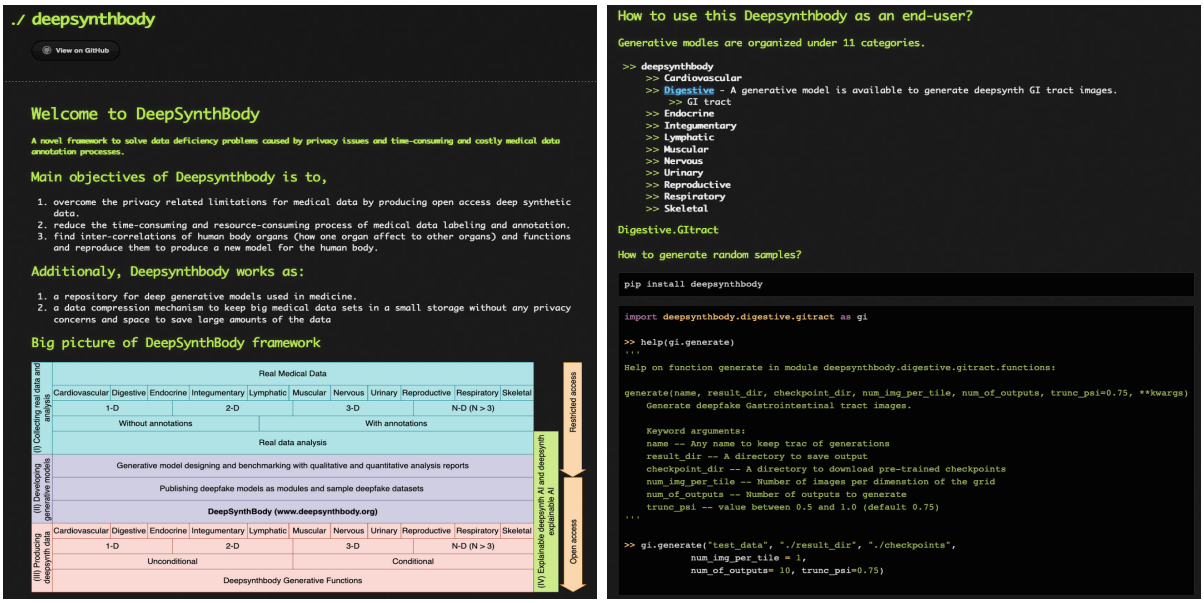


Figure 3.25: Sample screenshots of **deepsynthbody.org**.

tion process is either unconditional or conditional if both options are available. Several generative models can exist in this framework for a specific generative task (e.g., two different conditional GAN models to generate synthetic ECGs). If more than one model exists, the end-users can choose one for their specific application based on the benchmark reports or using their own qualitative and quantitative comparisons. Similarly, multiple GANs can be used together to generate diverse data distributions because different GAN models may have different data distributions based on the training data used to train them. The website named `deepyynthbody.org` is an online platform for providing all the information about functionalities and their usage to the end-users of the DeepSynthBody concept [71].

The website `deepsynthbody.org` links the researchers and the end-users. The main purpose of this online platform is to connect everything to achieve the main objective. Sample screenshots of the current website are given in Figure 3.25. This site provides the necessary information to contributing to DeepSynthBody and the end-users of this concept. However, the content of this site is subject to change based on new contributions and user experiences. Like the contents, the functional flow of the site is also subject to change to give better user experiences in the future. At the moment, two functionalities to generate synthetic data are presented on the website. One is for generating synthetic ECGs, and others for generating synthetic GI-tract data.

Abstract functions to generate synthetic ECGs were implemented as presented in

```
import deepsynthbody
deepsynthbody.cardiovascular.ecg.generate
            ("number of ECG to generate",
             "Path to generate",
             "start file ids from this number",
             "device to run")
```

Listing 1: The generative function to generate synthetic ECGs that are 10s long and having 8-leads.

Listing 1. Using this generation function, the end-users of DeepSynthBody can generate an unlimited number of 8-leads 10-sec long ECGs, which are convertible to 12-leads ECGs. However, this ECG generative model does not generate ground truth properties such as *PR interval, QT interval, heart rate*, and other properties discussed in the ECG analysis paper [41]. Suppose the end-users are interested in pre-analyzed ECGs. In that case, the generated ECGs can be analyzed using the MUSE system or the pre-generated dataset from the best checkpoint of Pulse2Pulse, and the corresponding MUSE analysis report can be downloaded here: `https://osf.io/6hved/` as presented in our ECG GAN paper [70].

Similarly, the end-users can generate an unlimited number of GI-tract images using the function presented in Listing 2. In addition to the main generation function, an additional generation function, originally discussed in the vanilla implementation of StyleGANv2 [194], was presented to generate intermediate generations using interpolations between two random points of synthetic generations. This function is presented in Listing 3.

3.4 Step IV: Explainable DeepSynth AI and DeepSynth Explainable AI

Step IV in the framework, namely explainable DeepSynth AI and DeepSynth XAI, is introduced to embed explainability and transparency into all other layers. This layer covers an essential concept to explain our deep generative models to increase trust and enable deeper failure analysis. Additionally, it allows another way to explain other ML methods using synthetic examples when the data restrictions are applied with real medical data.

If additional explanations are available to explain the synthetic data generation pro-

```
import deepsynthbody.digestive.gitract as gi

>> help(gi.generate)
'''
Help on function generate in module
    deepsynthbody.digestive.gitract.functions:

generate(name, result_dir, checkpoint_dir, num_img_per_tile,
        num_of_outputs, trunc_psi=0.75, **kwargs)
    Generate deepfake Gastrointestinal tract images.

    Keyword arguments:
    name -- Any name to keep trac of generations
    result_dir -- A directory to save output
    checkpoint_dir -- A directory to download pre-trained checkpoints
    num_img_per_tile -- Number of images per dimenstion of the grid
    num_of_outputs -- Number of outputs to generate
    trunc_psi -- value between 0.5 and 1.0 (default 0.75)
'''

>> gi.generate("test_data", "./result_dir", "./checkpoints",
            num_img_per_tile = 1,
            num_of_outputs= 10, trunc_psi=0.75)
```

Listing 2: Random GI-tract image generation function using StyleGAN.

```
import deepsynthbody.digestive.gitract as gi
>> help(gi.generate_interpolation)
'''
Help on function generate_interpolation in module
    deepsynthbody.digestive.gitract.functions:

generate_interpolation(name, result_dir, checkpoint_dir,
    num_img_per_tile, num_of_outputs, num_of_steps_to_interpolate,
    save_frames, trunc_psi=0.75, **kwargs)

    Generate deepfake Gastrointestinal tract images.

    Keyword arguments:
    name -- Any name to keep trac of generations
    result_dir -- A directory to save output
    checkpoint_dir -- A directory to download pre-trained checkpoints
    num_img_per_tile -- Number of images per dimenstion of the grid
    num_of_outputs -- Number of outputs to generate
    num_of_steps_to_interpolate -- Number of step between
                    two random points
    save_frames -- True if you want frame by frame,
                    otherwise .gif will be generated
    trunc_psi -- value between 0.5 and 1.0 (default 0.75)
'''

>> gi.generate_interpolation("test_data", "./result_dir",
                          "./checkpoints",
                          num_img_per_tile=1,
                          num_of_outputs=1,
                          save_frames=True,
                          num_of_steps_to_interpolate=100,seed=100)
```

Listing 3: The interpolation function to generate random images between two points of generation.

cess before using the synthetic data to replace real medical data, the trust of the end-users to use synthetic data can be improved. Therefore, DeepSynthBody introduces eXplainable DeepSynth Artificial Intelligence (XSAI). XSAI's primary goal is to explain deep generative models [202, 203] to increase understanding of the generative process and the quality of the generated data.

On the other hand, in the medical domain, XAI should be applied in increasing trust to accept solutions from ML models that generally perform classification, detection, and segmentation. While XSAI discusses the explainability of generative models, deep synthetic data can be used to support the XAI of other ML models. This functionality is discussed under DeepSynth XAI (SXAI). In this context, the main goal is not to explain the deep generative models but rather to explain other ML models used to classify, detect and segment medical data using synthetic data as examples. This DeepSynth XAI can overcome the privacy issues occurring when real data is used to explain ML models. For example, when researchers cannot explain their ML models by examples because the real data is restricted to share, they can use synthetic examples to explain their models with less concern about the privacy restrictions.

Both XSAI and SXAI concepts are discussed in the theoretical model. However, this explainable layer is a value-added layer to the DeepSynthBody framework. Therefore, Step IV is an optional step, and as a result, the DeepSynthBody framework functions without these XSAI and SXAI implementations. In this regard, we keep these options for future research works.

3.5 Summary

The DeepSynthBody concept was introduced in this thesis as the main solution to the data deficiency problem, which was identified during researching and developing ML models for CAD systems to assist doctors (Sub-objective I). The concept and the corresponding framework were discussed in four steps. These are, collecting real data and analysis, developing generative models, producing deep synthetic data, and explainable DeepSynth AI and DeepSynth explainable AI. In this chapter, these four steps were discussed one by one with the corresponding contributions.

Medical data is the core of any ML solution. Therefore, we successfully collected and

published seven dataset papers [23, 24, 25, 26, 27, 28, 29] to achieve Sub-objective II. Additionally, these datasets are required data to initiate DeepSynthBody. The datasets were classified according to the novel classification protocol introduced using the biological organ classification and the data dimension classification. Since analyzing the four steps with all types of medical data is impractical, an ECG signal dataset, a GI-tract image dataset, and a sperm video dataset were analyzed as case studies.

The ECG dataset is private, and it is not shareable. Therefore, this dataset was used as a case study to show how synthetic data is shared instead of a real dataset, which has privacy restrictions to share. A benchmark experiment was performed to understand the ECG dataset and implemented a novel GAN architecture, Pulse2Pluse, to generate realistic synthetic data. The Pulse2Pulse can generate synthetic 12-leads, 10-sec ECGs as alternative data to represent the restricted ECG data. The results show that synthetic ECGs generated from Pulse2Pulse are preserving the quality of the real dataset.

The GI-tract dataset was used as case studies to implement synthetic image generators to demonstrate synthetic medical image data sharing to avoid privacy concerns and present the capabilities of using synthetic data to solve the costly and time-consuming medical data annotation process. The deepsynth-gi generator using StyleGAN-v2 was implemented to generate synthetic GI-tract data. Additionally, the image inpainting GAN and SinGAN-Seg were demonstrated as solutions to the resource-consuming medical data annotation process.

The sperm dataset was analyzed, and SinGAN was investigated to perform an unsupervised medical video annotation process. The SinGAN functionality of converting paint-to-image was reversed and used as image-to-paint to accomplish this unsupervised sperm localization mechanism to use as another implementation to prove the capability of DeepSynthBody to use as an alternative data provider for the costly and time-consuming medical data annotation process. These sperm analysis experiments are in the early stage. Therefore final version of synthetic sperm generations will not be available at `deepsynthbody.org` until the GAN can produce quality output that can be published for the end-users of DeepSynthBody. Other than this synthetic sperm generator, the end-users of the DeepSynthBody can access both the synthetic ECG generator and the GI-tract image generator via `deepsynthbody.org`.

Overall, we could generate synthetic data using three different case studies represent-

ing other data formats, such as signals, images, and videos. In most cases, we have generated realistic-looking synthetic data that can be replaced for real data with privacy restrictions. Moreover, we showed that GANs could generate synthetic data with ground truths to overcome the costly and time-consuming data annotation process. Furthermore, we presented how to convert true negative data into true positive data using GANs to address the data imbalance problem. Presented qualitative and quantitative analyses imply that synthetic data can overcome the data deficiency problem in the medical domain.

Explainable DeepSynth AI and DeepSynth explainable AI were introduced as an optional step in this framework, and therefore, contributors can decide that they are following this step or not. This functionality was kept for future research. However, adding explainability to generative models used in this framework can improve the trust of the end-users to use the synthetic data.

Chapter 4

Discussion and Conclusion

The main objective of this thesis is to research and develop generalizable, accurate and well-performing ML models which can be used in CAD systems to aid doctors by detecting more anomalies to save lives ultimately. However, we identified that the lack of medical data is a major problem in the current pipeline of applying ML methods in the medical domain. Therefore, we have defined several objectives to find a way to overcome the data deficiency problem in applying ML solutions in the medical domain. As a result, we introduced a novel concept and the corresponding framework, DeepSynthBody, to bypass the data deficiency problem.

In this thesis, the main research question stated was "**What are the problems that emerge from data in computer-aided diagnosis systems, and how can these problems be tackled?**". To address the research question, we have researched and analyzed ML models used in CAD systems. To support it, we collected and investigated the real medical datasets, researched and developed benchmark analysis to identify the data problems to be addressed. We could identify that data deficiency is the main problem in the medical domain. This problem has occurred due to privacy concerns, the time-consuming and costly data annotation, and the data imbalance problem in the medical domain. To overcome these problems, we researched and developed a GAN-based concept and a framework to tackle the data deficiency problem in the medical domain, namely DeepSynthBody. In the DeepSynthBody solution, the main focus is to overcome the data deficiency problems using synthetic medical data. We show that synthetic data can overcome the data deficiency problem by omitting privacy concerns, generating synthetic data with ground truth and generating synthetic data to overcome data imbalance problems

by converting true negatives to true positives.

To achieve the main objective, **seven datasets** [23, 24, 25, 26, 27, 28, 29], 12 **benchmark analysis studies and ML models to use with CAD systems** [30, 31, 38, 39, 40, 68, 41, 36, 32, 33, 35, 34] and **eight GAN studies** [72, 73, 77, 74, 70, 67, 75, 76, 71] were published to cover all the sub-objectives and finally achieve the main objective and answer the research question. Some of these papers contribute to multiple objectives, while others contribute to only a single objective. These contribution overlaps are illustrated in Figure 1.5 in Section 1.5.

4.1 Contributions and Discussions

The main focus of our research, in general, is to find generalizable and well-performing ML models, which are the main component of CAD systems to assist doctors, and this thesis address several of the challenges arising in this context. In particular, we have focused on researching ML models for CAD systems with special attention to the challenges medical data scarcity introduces. To accomplish this, Sub-objective I was introduced. However, the data deficiency problem was identified as a significant barrier to achieve the sub-objective I. Therefore, this thesis also introduced sub-objectives II, III, and IV to research and develop medical datasets, research and establish benchmarks to identify the data problems, and research and develop GAN-based frameworks to generate synthetic data as the solution. Sub-objective I and Sub-objective III are overlapped greatly because designing ML models for CAD systems consists of implicit benchmark analysis and vice-versa. Finally, we achieved Sub-objective IV by introducing the novel DeepSynthBody concept and the corresponding framework. Three different medical branches, *gastroenterology*, *andrology*, and *cardiology*, were used as the case studies for sub-objectives I, II, III, and IV:

- **Sub-objective I**: The main focus of this sub-objective is to research and develop well-performing ML models for CAD systems to assist doctors. As case studies, we have selected three branches of medicine. In gastroenterology, images collected from colonoscopies were the main data stream to apply ML algorithms which are the core algorithms in CAD systems. Several classification models [30, 31] and segmentation models [35, 36] were researched and implemented for the gastroenterology branch

under this thesis in different stages of the timeline. Not only using real data, but also synthetic data was used with segmentation models [67] used to predict polyps in GI-tract data. Similarly, ML-based regression models were investigated and developed for the andrology branch [38, 39, 40, 68]. For the cardiology branch, an ML-based ECG analysis system [41] was researched and implemented. Moreover, all the dataset papers [23, 24, 25, 26, 27, 28, 29] introduced ML models as baseline experiments considered as initial models for developing CAD systems.

- **Sub-objective II**: The main task of this sub-objective is to collect and produce medical datasets. Collecting medical data and producing baseline results to understand the data is the first step of developing CAD systems. Therefore, different types of medical datasets [23, 24, 25, 26, 27, 28, 29] representing different types of human body organs were collected and published with the baseline experiments. While all the datasets contribute to the main objective, the GI-tract dataset [23] was selected to use as one of the case studies for other sub-objectives because of the data diversity and a large amount of data. Despite our dataset contributions, two additional datasets were used as the case studies. They are an ECG dataset, which is a private medical dataset representing biomedical signal, and a sperm dataset representing video data. The additional datasets were selected to research and develop ML models for CAD systems in the initial stage. Later, these additional dataset were used to maintain the diversity of the case studies used as proof of concepts.

  From the perspective of DeepSynthBody, which is the solution introduced in this thesis to overcome the data deficiency problem, this data processing step is an in-house step if the datasets are private. In this thesis, one private dataset and two public datasets were used to prove the concept of DeepSynthBody. For further investigating the concept's possibilities, experimenting with new medical data types can be started with public datasets with other data types, which were not covered in this thesis. At the end of the successful implementation of DeepSynthBody, we could introduce synthetic datasets, such as synthetic ECGs, synthetic polyps, and the corresponding ground truth masks, and randomly generated synthetic GI-tract landmarks to support the main objective.

- **Sub-objective III**: The selected datasets were used to design generalizable and

well-performing ML models for CAD systems in our Sub-objective I. However, after identifying the data problems of the ML-based CAD system designing process, we re-analyzed the process of designing ML models as benchmark analysis to investigate the data deficiency problem to be addressed in Sub-objective IV. Under Sub-objective I, different types of ML solutions for CAD systems were investigated under the three different selected medical branches, gastroenterology [30, 31, 36, 32, 33, 35], andrology [38, 39, 40, 68], and cardiology [41]. However, all findings were considered as benchmark articles under new Sub-objective III as well because these studies reflect the real problems associated with the medical data.

A set of benchmark articles for the selected datasets as case studies were published to achieve the benchmark analysis objective (Sub-objective III). While all the datasets should have benchmark analysis results, we chose the same three datasets selected in Sub-objective I, as case studies to achieving Sub-objective III. They are the ECG data, the GI-tract dataset, and the sperm dataset. Then, different types of benchmark analysis experiments done for developing ML models for CAD systems in Sub-objective I with the GI-tract data [30, 31, 36, 32, 33, 35] were re-considered to support this objective. Similarly, the ECG analysis [41] and sperm analysis [38, 39, 40, 68] experiments were investigated again as benchmark analyses for identifying the data-related problems to address experimenting GANs. Without having benchmark analysis, it is not recommended to research GANs under this DeepSynthBody framework because the end-user of the DeepSynthBody framework will not have results to compare the quality of synthetic data coming from this framework in addition to understanding the data-related problems. In these benchmark analyses, we contributed to organizing a competition, namely BioMedia 2020 [68], and participated in a competition, namely EndoCV 2021 [35], to maintain higher standards for the benchmark results. A detailed analysis of GI tract landmark classification was performed within the benchmark analyses to introduce proper generalizable analyses using cross datasets of GI data [31]. As a result of the cross dataset evaluation, we further discussed proper evaluation mechanisms and guidelines for binary classification of medical data in our Medimetrics[1] [33], an open-access tool for fair evaluations among different research findings. Furthermore, the effect of image resolution [32]

[1] https://medimetrics.no/medimetrics/

was investigated to show that high-resolution data can improve the performance of ML models using the GI-tract data as the case study.

- **Sub-objective IV**: In this sub-objective, the main purpose is to generate synthetic medical data to overcome privacy-related problems, the time-consuming and costly medical data annotation process, the data bias problem in the medical domain, and the medical data imbalance problem. Before studying synthetic data generation experiments, we investigated possible use cases of GANs with GI-tract data. One study has investigated to preprocessing GI tract images using a GAN [72, 73] to fill green regions of endoscopic images. Another study was performed to predicting blurry pill cam video frames using a GAN [74]. The later GAN experiment shows that the GAN can predict the fifth frame for the given four input frames of a pill cam endoscopic video. These experiments helped us to get a basic understanding of GANs in the medical domain.

  Then, advanced GAN experiments to overcome the privacy issues were researched and developed. The privacy concerns were identified as one of the major issues that caused the data deficiency problem in the medical domain. The private ECG dataset was investigated and successfully published a novel GAN architecture called Pulse2Pulse [70], which can generate synthetic 12-leads 10-seconds long ECGs indistinguishable from real ECGs. Not only this ECG generation GAN, we investigated a GI-tract image generation in the concept paper [71], which introduced the DeepSynthBody concept. The synthetic GI-tract data generator introduced in the concept paper showed how to generate controllable synthetic data as an alternative to real medical image data if the real datasets have privacy concerns.

  Not only privacy concerns, but the results collected from the ECG generation and GI-tract image generation experiments show that synthetic data can represent the real data distributions. Remarkably, the synthetic ECGs clearly show the exact distribution of the properties of the real dataset used to train the Pulse2Pulse GAN. Besides generating the synthetic samples within the distribution, the generated synthetic data can cover untouched regions of the real data distribution. For more information about the distribution overlap between the real and the synthetic data, refer to the original article of Pulse2Pulse [70]. Similar to the synthetic ECGs, the synthetic GI-tract images shows realistic GI-tract landmark within the generated

images. Then, these GANs are an indication that synthetic data can be used to generate uniform data distributions or missing data.

Furthermore, under this objective, to introduce an alternative method for the costly and time-consuming expert's data annotation process, we researched and developed novel pipelines of GAN architectures using two case studies, the GI-tract dataset [23] and the sperm dataset [69]. In one study, the GI-tract dataset was used to train a GAN to generate synthetic polyp data from clean colon images [75]. This study also contributed to the data imbalance problem in the medical domain because the pipeline introduced in this study converts a real clean colon image (true-negative sample) into a synthetic polyp image (true-positive sample). In another study, SinGAN-Seg, synthetic polyp data were generated with the corresponding mask from a single polyp image [67]. In the SinGAN-Seg study, an unlimited number of synthetic samples can be generated with the corresponding segmentation masks of polyps. This GAN can solve the time-consuming and costly data annotation process by generating synthetic data and the corresponding segmentation masks automatically. Moreover, we show that generated synthetic samples can improve the performance of polyp segmentation algorithms used in CAD systems when the manually annotated dataset is small. Additionally, we have investigated the usability of GANs to produce synthetic sperm data [76] instead of blurry-looking sperm video samples to have better quality assessments.

In this thesis, we researched an unsupervised way to segment sperms using a GAN-based model. The results showed promising directions of converting real sperm video frames into synthetic clear video frames with sperm locations, which can be used to analyze the sperm samples in future studies. This sperm study also a proof for using GANs to overcome the time-consuming and costly data annotation process in the medical domain.

In addition to generating synthetic data with segmentation masks representing the most advanced ground truth type, which is pixel-wise classification, all other ground truth generations, such as continuous values, class labels, and bounding boxes, can be explored and considered to overcome the data deficiency problem using GANs as explored under Sub-objective IV. For example, conditional GANs generating synthetic medical data using simple numerical values as input conditions can make

synthetic datasets with numerical ground truth data. Similarly, using class labels as input to GANs can produce synthetic datasets with the corresponding class labels. Moreover, bound box ground truth, one of the famous medical image analysis techniques, can be made using similar conditional GANs.

Finally, we formalized the GAN development process using the novel concept and the framework called DeepSynthBody to overcome the data deficiency problem. In this framework, we pipeline the synthetic data generation process in the medical domain using four steps. Developers who are researching GANs and end-users who need synthetic data can use our framework via `www.deepsynthbody.org`. In this framework, we encouraged to publish generative models as PyPI package instead of publishing pre-generated billions of synthetic data samples. This encouragement is a trade-off because it has advantages and disadvantages. Pre-trained GAN models need less space than publishing pre-generated data is an advantage. If pre-trained GANs are conditional GANs, then the end-users can generate synthetic data as they needed. This custom data generation is another advantage. The main disadvantage of using pre-trained models instead of pre-generated synthetic data is the reproducibility of research works performed using privately generate synthetic datasets. However, publishing the synthetic datasets used to perform the research in other public data repositories can solve this problem. Therefore, overall we recommend publishing pre-generated GAN models instead of pre-generated datasets.

- **Main-objective**: The final objective was to connect all sub-objectives to produce well-performing and more accurate ML models for CAD systems to assist doctors in efficient diagnoses by addressing the data deficiency problem. The initial ML models designed to achieve the Sub-objective I showed the effects of data deficiency problems in the medical domain. Then, we collected, researched, and developed datasets (real and synthetic) for developing ML models for CAD systems for biomedical applications. In Sub-objective III, benchmark analyses were performed to identify the data problem to be addressed using GANs. Then, we proposed DeepSynthBody, which is based on GANs to address the data deficiency problem in the medical domain (Sub-objective IV). Finally, we published our solution as an open-source project for getting more collaborations worldwide at `www.deepsynthbody.org`.

  Generated synthetic ECG data show that our concept can avoid the privacy concerns

in the medical domain. We proved the usability of synthetic ECG data qualitatively and quantitatively in our DeepFake ECG paper [70]. Moreover, synthetic polyp generation studies [67, 75] showed that the data imbalance problem and the time-consuming and costly data annotation problem can be solved using synthetic data. Additionally, SinGAN-Seg [67] showed performance improvements when synthetic datasets are used instead of small real datasets. Ultimately, we could show that the main-objective is achievable using the novel concept and the corresponding framework, namely DeepSynthBody, introduced in Sub-objective IV and achieving other three sub-objectives I, II, and III.

By achieving the four sub-objectives, we reached our main objective: research and develop ML models for CAD systems for different medical applications focusing on the problems of limited availability of biomedical data. Finally, we showed that the research question, "**What are the problems that emerge from data in computer-aided diagnosis systems, and how can these problems be tackled?**" could be answered using our novel concept called DeepSynthBody. Now the concept is public. All the necessary infrastructure of the DeepSynthBody framework is ready for contributions from researchers who can provide deep generative models to this framework to make a fully functional open-source DeepSynthBody. This concept will open a new era for open science in the medical domain. For contributions, researchers can visit our online platform: `www.deepsynthbody.org`.

4.2 Ethical Consideration

Medical data collected from one patient, one hospital, one region, or one human race to train and develop ML models used in CAD systems can lead to ethical problems because the models based on this data can make biased predictions. Therefore, researchers should pay more attention to this problem in their research. For example, when an ML model is trained from a patient's data, then the model should consider the patient's anonymity and confidentiality. In this regard, we have maintained all the participants' anonymity and privacy for our data collections by de-identifying data samples, thus, making it impossible to connect the data to real persons. Furthermore, we combined data collected from several hospitals to avoid patient bias problems and hospital bias

problems. However, to avoid the race and country bias problems, a more extensive data collection should be performed. Collecting data in the medical domain is challenging due to privacy concerns, the costly and timely medical data annotation process, and data bias problems. The DeepSynthBody concept can address these problems by omitting anonymization and confidentiality concerns (privacy concerns) by generating synthetic data with ground truths, and generating synthetic data by converting true negatives to true positives.

Although DeepSynthBody is created to solve the data problems in the medical domain, the ethics of synthetic data, which is the core of the concept, is a critical topic. Deep Fakes [204], a popular topic in synthetic data generation, can fool people by generating realistic-looking face images and videos. In this context, Deep Fakes are sometimes used to make fake news about famous people. While some of these Deep Fakes are used to entertain society, others are purposely harming both people and the society.

The same problems may happen with synthetic data in the medical domain. Some possibilities are that someone can generate fake medical reports with generated realistic-looking medical images and videos, etc. People may use these fake reports to cheat their companies to get social benefits such as additional money. This kind of circumstance cannot be avoided, and making a fully secure link with hospitals to get approval can be a solution. Another ethical issue arises with converting true negatives into true positives. Somebody can argue that this is not an ethical procedure because one converts healthy medical data to unhealthy data. However, if true positives are not identified using a real name, we believe that this conversion is ethical.

In sum, we believe that the possible negative effects of synthetic data in the medical domain are outweighed by the positive aspects. We presented in this thesis how to use synthetic data to share private datasets in order to avoid privacy concerns. Furthermore, we showed that synthetic data is a possible solution to overcome the data bias problems. For example, we converted non-polyp images into polyp images. In other words, we converted true negative samples into true positives. A similar mechanism can overcome the data imbalance problems by converting data from one racial background to another racial background to avoid ethical issues related to imbalanced data. Moreover, ground truth data in the medical domain can raise ethical issues due to differences from an expert to another expert who performs the ground truth preparation process. These differences

affect the final performance of the ML models trained from the data. The ML models, in some way, reflect the skills of the person who prepared the ground truth data. In this context, we have proposed a way to generate synthetic data with the corresponding ground truth. Therefore, experts' knowledge can be used to verify the ground truth rather than preparing ground truths which have differences from one expert to another. Overall, we can see that synthetic data in medicine can rather help to solve ethical issues than producing new ones. Nevertheless, like for all research where humans are involved, one needs to be very careful and sensitive in addressing ethical questions for each specific medical application area where synthetic data might be used.

4.3 Future Works

Our solution, DeepSynthBody, which was introduced to tackle the data deficiency problem for developing ML models for CAD systems in the medical domain, can be improved in different ways from Step I to Step IV. In Step I, many datasets from different organ systems have been collected. However, in this thesis, only one dataset from the data collection was investigated with additional two datasets from the outside of our data collection. Therefore, benchmark studies and GAN experiments should be performed with the rest [24, 25, 26, 27, 28, 29] of our data collection. In addition to the collected datasets, other open-access datasets can be used as case studies, such as, MRI datasets representing 4-D datatype, which was not considered in this thesis. In Step II, the evaluation process of benchmark results can be improved by introducing a common guideline to measure the performance of detection and segmentation ML models such as MediMetrics [33], which was introduced to improve the quality of evaluations used with binary classifications.

The GAN models used for the three case studies, which used ECG dataset, GI-tract dataset, and sperm dataset as main data sources, can be further improved. For example, Pulse2Pulse [70] can be enhanced by adding conditional input such as ECG properties. Additionally, continuous ECG pulse generation can be researched with a modified version of the Pulse2Pulse generator by conditioning on the first half of ECGs as input to the generator. Moreover, GI-tract style GAN [71] can be improved using conditional-GANs of GI-tract images to generate specific landmarks of GI-tract. However, the main challenge for training conditional GANs for generating synthetic GI-tract images is a lack of

labeled GI-tract images. In this case, researchers can experiment with transfer learning mechanisms for GAN training [205, 206]. Further investigations with SinGAN-Seg and polyp inpainting GANs can improve the quality of the synthetic data generated from these GANs, i.e., adding super-resolution GAN [207] to the pipeline of synthetic polyp image generation. In the end, we have considered three branches of medicine, cardiology, gastroenterology, and andrology. Other branches of medicine should be considered in future research to build the complete DeepSynthBody, and ML models for CAD systems to assist doctors.

The GANEx [77] tool can be further improved with several functionalities. For GANEx, we can introduce functionalities to publish checkpoints of trained GAN architectures directly into the DeepSynthBody framework. In this functionality, the submitted checkpoint can be reviewed by computer science experts of the future community of the framework before merging them into the final `deepsynthbody` package. Adding these kinds of functionalities can help non-computer science people to publish their GAN modules without any coding burdens. Additionally, integrating interaction between GANEx and the online platform can introduce real-time performance comparisons, such as qualitative comparisons for synthetic images, if there are two or more models for the same purposes. Not only that, Federated learning techniques [208, 209] can be investigated for GANEx to enable distributed GAN learning to input bigger training data distribution to get better realistic synthetic data. However, some re-engineering, such as added web services for distributed computing, can be researched for online interactions and distributed computing.

We believe that DeepSynthBody will open new research directions and overcome the data deficiency problem in medicine. For example, DeepSynthBody can produce a new model for representing the human body and its intra-correlations of functionalities of the organs. These functionalities can be achieved by collecting multi-model datasets consisting of various types of medical data correlated with each other. Suppose we can investigate GAN models, which can condition on one datatype and generate synthetic data on another data type. In that case, those models can be used to find correlations among different medical data types. Finally, GANs can be trained to generate synthetic data conditioned on one organ's data and generate data for another organ system. Successful findings of these correlations can lead to finding correlations about organs' functions

because data coming from organ systems is inherited from their functions. Additionally, this platform will act as a large medical data repository without any privacy concerns and data storage shortages because successful GANs can act as a data compression method. For example, the size of the training dataset used in the Pusle2Pulse [70] implementation is around 3GB for around $15,000$ ECGs. However, if we use the `deepfake-ecg` PyPI package, it takes around 50MB to store in cloud platforms. Still, it can generate an unlimited number of realistic synthetic ECGs from a similar distribution of the real data. In this thesis, proper evaluations were not done focusing on this data compression because it was not our main goal. Thus, future studies can be focused on evaluating this privacy-preserving data compression and storage.

4.4 Conclusion

In conclusion, ML-based CAD systems are a great value addition to medicine because these systems have the capabilities to assist doctors by performing automated diagnosis processes. However, we showed that a lack of medical data to train ML models causes generalizability and performance issues. Collecting and processing medical domain data is a basic solution to overcome this problem. However, collecting and processing data is not easy in the medical domain because of privacy restrictions and the costly and time-consuming data annotation process. Generating synthetic medical data to train ML models is an alternative solution to overcome this data deficiency problem.

Well-performing GAN architectures can generate realistic synthetic data. These synthetic data can represent real medical data when the real datasets are not permitted to share. Moreover, conditional GAN architectures can generate synthetic datasets with the corresponding ground truth data, which domain experts normally do. For example, we showed that how to generate synthetic polyps and the corresponding ground truth masks. Furthermore, GANs can generate synthetic medical data by converting true negative data samples into true positive data samples. Data conversion, such as true negatives to true positives, can solve the data imbalance problem in the medical domain.

DeepSynthBody framework, which was introduced as the main solution in this thesis to overcome the data deficiency problem, provides a complete framework to generate synthetic data and develop generative models. We published this concept and the framework

as an open-source project to get contributions worldwide. Getting more contributions, we hope to produce the largest synthetic data repository in the world. Ultimately, this DeepSynthBody concept can be improved to use as a model to represent the human body. Furthermore, the data compression ability of GANs is a solution for storing medical data in a limited space avoiding privacy concerns.

4.5 Final Remarks

In this thesis, we researched and developed ML-based components for CAD systems in three different branches, gastroenterology, andrology, and cardiology. All the data collected under these three branches were collected from hospitals in Norway and Denmark. In most of the cases, datasets were analyzed by experts in the domains. In the cases where we generated synthetic data, domain experts helped us to perform a qualitative analysis with their expertise. Furthermore, our solution proposed in the thesis, namely DeepSynthBody, shows a high potential to be an important part of the future of developing well-performing ML models for developing CAD systems. However, the success of the future directions of DeepSynthBody depends on the contributions from the research community of ML and the medical data providers. Therefore, the framework is available as an open-source project at `deepsynthbody.org` to get more contributions and to the end-users who want to generate synthetic medical data. Moreover, we showed advanced future directions of our DeepSynthBody, such as using the framework as a novel model to the human body and a novel way to store medical data.

Bibliography

[1] David Reinsel–John Gantz–John Rydning. "The digitization of the world from edge to core". In: *Framingham: International Data Corporation* (2018).

[2] Daniel E O'Leary. "AI in accounting, finance and management". In: *Intelligent Systems in Accounting, Finance and Management* 4.3 (1995), pp. 149–153.

[3] Henri Arslanian and Fabrice Fischer. *The Future of Finance: The Impact of FinTech, AI, and Crypto on Financial Services*. Springer, 2019.

[4] Bo-hu Li, Bao-cun Hou, Wen-tao Yu, Xiao-bing Lu, and Chun-wei Yang. "Applications of artificial intelligence in intelligent manufacturing: a review". In: *Frontiers of Information Technology & Electronic Engineering* 18.1 (2017), pp. 86–96.

[5] Cihan H Dagli. *Artificial neural networks for intelligent manufacturing*. Springer Science & Business Media, 2012.

[6] Richard Lachman and Michael Joffe. "Applications of Artificial Intelligence in Media and Entertainment". In: *Analyzing Future Applications of AI, Sensors, and Robotics in Society*. IGI Global, 2021, pp. 201–220.

[7] Sylvia M Chan-Olmsted. "A Review of Artificial Intelligence Adoptions in the Media Industry". In: *International Journal on Media Management* 21.3-4 (2019), pp. 193–215.

[8] Fei-Yue Wang. "Toward a revolution in transportation operations: AI for complex systems". In: *IEEE Intelligent Systems* 23.6 (2008), pp. 8–13.

[9] Adel W Sadek. "Artificial intelligence applications in transportation". In: *Transportation Research Circular* (2007), pp. 1–7.

[10] Fei Wang and Anita Preininger. "AI in health: state of the art, challenges, and future directions". In: *Yearbook of medical informatics* 28.1 (2019), p. 16.

[11] Thomas Davenport and Ravi Kalakota. “The potential for artificial intelligence in healthcare”. In: *Future healthcare journal* 6.2 (2019), p. 94.

[12] Susmita Ray. “A quick review of machine learning algorithms”. In: *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE. 2019, pp. 35–39.

[13] Rocio Vargas, Amir Mosavi, and Ramon Ruiz. “Deep learning: a review”. In: *Advances in intelligent systems and computing* (2017).

[14] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. “Overview and Importance of Data Quality for Machine Learning Tasks”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 3561–3562.

[15] Xue-Wen Chen and Xiaotong Lin. “Big data deep learning: challenges and perspectives”. In: *IEEE access* 2 (2014), pp. 514–525.

[16] IBM. *IBM Cloud Learn Hub*. Accessed: 2021-04-25. URL: `https://www.ibm.com/cloud/learn/artificial-intelligence`.

[17] Kunio Doi. “Computer-aided diagnosis in medical imaging: historical review, current status and future potential”. In: *Computerized medical imaging and graphics* 31.4-5 (2007), pp. 198–211.

[18] David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. “Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the ”Big Three””. In: *Diagnosis* 1.ahead-of-print (2020).

[19] Eric J Chin, Andrew Bloom, and Andrew Thompson. “A comparison of perceived acceptable missed diagnosis rates for high-risk emergency medicine diagnoses: A brief report”. In: *The American journal of emergency medicine* 35.12 (2017), pp. 1973–1977.

[20] Nam Hee Kim, Yoon Suk Jung, Woo Shin Jeong, Hyo-Joon Yang, Soo-Kyung Park, Kyuyong Choi, and Dong Il Park. "Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies". In: *Intestinal research* 15.3 (2017), p. 411.

[21] LM Blendis, WJ McNeilly, Louise Sheppard, Roger Williams, and JW Laws. "Observer variation in the clinical and radiological assessment of hepatosplenomegaly". In: *Br Med J* 1.5698 (1970), pp. 727–730.

[22] Paul Brennan and Alan Silman. "Statistical methods for assessing observer variability in clinical measures." In: *BMJ: British Medical Journal* 304.6840 (1992), p. 1491.

[23] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy". In: *Scientific Data* 7.1 (2020), pp. 1–14.

[24] Henrik Svoren, Vajira Thambawita, Pål Halvorsen, Petter Jakobsen, Enrique Garcia-Ceja, Farzan Majeed Noori, Hugo L. Hammer, Mathias Lux, Michael Alexander Riegler, and Steven Alexander Hicks. "Toadstool: A Dataset for Training Emotional Intelligent Machines Playing Super Mario Bros". In: *Proceedings of the 11th ACM Multimedia Systems Conference*. MMSys '20. Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 309–314. ISBN: 9781450368452. DOI: 10.1145/3339825.3394939. URL: https://doi.org/10.1145/3339825.3394939.

[25] Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, Simon Nordvang, Sigurd Pedersen, Anders Gjerdrum, Tor-Morten Grønli, Per Morten Fredriksen, Ragnhild Eg, Kjeld Hansen, Siri Fagernes, Christine Claudi, Andreas Biørn-Hansen, Duc Tien Dang Nguyen, Tomas Kupka, Hugo Lewi Hammer, Ramesh Jain, Michael Alexander Riegler, and Pål Halvorsen. "PMData: A Sports Logging Dataset". In: *Proceedings of the 11th ACM Multimedia Systems Conference*. MMSys '20. Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 231–236. ISBN:

9781450368452. DOI: 10.1145/3339825.3394926. URL: https://doi.org/10.1145/3339825.3394926.

[26] P. Jakobsen, E. Garcia-Ceja, L. A. Stabell, K. J. Oedegaard, J. O. Berle, V. Thambawita, S. A. Hicks, P. Halvorsen, O. B. Fasmer, and M. A. Riegler. "PSYKOSE: A Motor Activity Database of Patients with Schizophrenia". In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 2020, pp. 303–308. DOI: 10.1109/CBMS49503.2020.00064.

[27] Pia H. Smedsrud, Vajira Thambawita, Steven A. Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc Tien Dang Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T. Schmidt, Ervin Toth, Hugo L. Hammer, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. "Kvasir-Capsule, a video capsule endoscopy dataset". In: *Scientific Data* 8.1 (2021), p. 142. DOI: 10.1038/s41597-021-00920-z. URL: https://doi.org/10.1038/s41597-021-00920-z.

[28] Enrique Garcia-Ceja, Vajira Thambawita, Steven A. Hicks, Debesh Jha, Petter Jakobsen, Hugo L. Hammer, Pål Halvorsen, and Michael A. Riegler. "HTAD: A Home-Tasks Activities Dataset with Wrist-Accelerometer and Audio Features". In: *MultiMedia Modeling*. Ed. by Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras. Cham: Springer International Publishing, 2021, pp. 196–205. ISBN: 9783030678357.

[29] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael Riegler, Thomas de Lange, Peter T Schmidt, Håvard Johansen, et al. "Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy". In: ().

[30] Vajira Thambawita, Debesh Jha, Michael Riegler, Pål Halvorsen, Hugo Lewi Hammer, Håvard D Johansen, and Dag Johansen. "The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning". In: *Proc. of MediaEval* (2018).

[31] Vajira Thambawita, Debesh Jha, Hugo Lewi Hammer, Håvard D. Johansen, Dag Johansen, Pål Halvorsen, and Michael A. Riegler. "An Extensive Study on Cross-

Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification". In: *ACM Trans. Comput. Healthcare* 1.3 (June 2020). ISSN: 2691-1957. DOI: 10.1145/3386295. URL: https://doi.org/10.1145/3386295.

[32] Vajira Thambawita, Steven Hicks, Strümke Inga, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. "Impact of image resolution on convolutional neural networks performance in gastrointestinal endoscopy". In: *Gastrointestinal Endoscopy* (2021). DDW 2021 AGA Program and Abstracts.

[33] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. "On evaluation metrics for medical applications of artificial intelligence". In: *medRxiv* (2021). DOI: 10.1101/2021.04.07.21254975. eprint: https://www.medrxiv.org/content/early/2021/04/09/2021.04.07.21254975.full.pdf. URL: https://www.medrxiv.org/content/early/2021/04/09/2021.04.07.21254975.

[34] Henrik L. Gjestang, Steven A. Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A. Riegler. "A self-learning teacher-student framework for gastrointestinal image classification". In: *Proceedings of International Symposium on Computer-Based Medical Systems (CBMS)*. 2021.

[35] Vajira Thambawita, Steven A. Hicks, Pål Halvorsen, and Michael A. Riegler. "DivergentNets: Medical Image Segmentation by Network Ensemble". In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) co-located with with the 17th IEEE International Symposium on Biomedical Imaging (ISBI 2021)*. 2021.

[36] Vajira Thambawita, Steven Hicks, Pål Halvorsen, and Michael A Riegler. "Pyramid-Focus-Augmentation: Medical Image Segmentation with Step-Wise Focus". In: *arXiv preprint arXiv:2012.07430* (2020).

[37] B. Mac Namee, P. Cunningham, S. Byrne, and O.I. Corrigan. "The problem of bias in training data in regression problems in medical decision support". In: *Artificial Intelligence in Medicine* 24.1 (2002), pp. 51–70. ISSN: 0933-3657. DOI: https://doi.org/10.1016/S0933-3657(01)00092-6. URL: https://www.sciencedirect.com/science/article/pii/S0933365701000926.

[38] Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Vajira Thambawita, Pål Halvorsen, Hugo L. Hammer, Trine B. Haugen, and Michael A. Riegler. "Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction". In: *Scientific Reports* 9.1 (2019). DOI: 10.1038/s41598-019-53217-y.

[39] Vajira Thambawita, Pål Halvorsen, Hugo Hammer, Michael Riegler, and Trine B Haugen. "Stacked Dense Optical Flows and Dropout Layers to Predict Sperm Motility and Morphology". In: *Proc. of MediaEval* (2019).

[40] Vajira Thambawita, Pål Halvorsen, Hugo Hammer, Michael Riegler, and Trine B Haugen. "Extracting Temporal Features into a Spatial Domain Using Autoencoders for Sperm Video Analysis". In: *Proceedings of MediaEval* (2019).

[41] Steven A. Hicks, Jonas L. Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary M. Maleckar, Michael A. Riegler, and Jørgen K. Kanters. "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis". In: *Scientific Reports* 11.1 (2021), p. 10949. DOI: 10.1038/s41598-021-90285-5. URL: https://doi.org/10.1038/s41598-021-90285-5.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[43] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey". In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.

[44] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. "Preparing medical imaging data for machine learning". In: *Radiology* 295.1 (2020), pp. 4–15.

[45] Deven McGraw and Kenneth D. Mandl. "Privacy protections to encourage use of health-relevant digital data in a learning health system". In: *npj Digital Medicine* 4.1 (2021), p. 2. DOI: 10.1038/s41746-020-00362-8. URL: https://doi.org/10.1038/s41746-020-00362-8.

[46] 2nd Price W Nicholson and I Glenn Cohen. "Privacy in the age of medical big data". In: *Nature medicine* 1 (Jan.), pp. 37–43. DOI: 10.1038/s41591-018-0272-7.

[47] Marcello Ienca, Agata Ferretti, Samia Hurst, Milo Puhan, Christian Lovis, and Effy Vayena. "Considerations for ethics review of big data health research: A scoping review". In: *PloS one* 13.10 (2018), e0204937.

[48] Bartha Maria Knoppers and Adrian Mark Thorogood. "Ethics and big data in health". In: *Current Opinion in Systems Biology* 4 (2017), pp. 53–57.

[49] Julia Lane and Claudia Schur. "Balancing access to health data and privacy: a review of the issues and approaches for the future". In: *Health services research* 45.5p2 (2010), pp. 1456–1467.

[50] *The Norwegian Data Protection Authority*. Accessed: 2021-04-25. URL: https://www.datatilsynet.no/en/.

[51] *The Personal Data Act*. Accessed: 2021-04-25. URL: https://www.forskningsetikk.no/en/resources/the-research-ethics-library/legal-statutes-and-guidelines/the-personal-data-act/.

[52] Paul Voigt and Axel Von dem Bussche. "The eu general data protection regulation (gdpr)". In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10 (2017), p. 3152676.

[53] Peter Edemekong, Pavan Annamaraju, and Micelle Haydel. "Health Insurance Portability and Accountability Act". In: *StatPearls* (2020).

[54] *California Consumer Privacy Act*. 2018. URL: https://oag.ca.gov/privacy/ccpa.

[55] *Act on the Protection of Personal Information*. 2003. URL: https://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf.

[56] *Personal Information Protection Commission*. 2011. URL: http://www.pipc.go.kr/cmt/main/english.do.

[57] *THE PERSONAL DATA PROTECTION BILL*. 2018. URL: https://www.meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf.

[58] *Anolytics - Making AI Possible with Right Data and Image Annotation Services*. Accessed: 2021-04-25. URL: https://www.anolytics.ai/solutions/healthcare/.

[59] *Mindy Support*. Accessed: 2021-04-25. URL: https://mindy-support.com/industries-posts/healthcare/.

[60] Shaode Yu, Mingli Chen, Erlei Zhang, Junjie Wu, Hang Yu, Zi Yang, Lin Ma, Xuejun Gu, and Weiguo Lu. "Robustness study of noisy annotation in deep learning based medical image segmentation". In: *Physics in Medicine & Biology* 65.17 (2020), p. 175007. DOI: 10.1088/1361-6560/ab99e5. URL: https://doi.org/10.1088/1361-6560/ab99e5.

[61] Google. *Labeling costs*. https://cloud.google.com/ai-platform/data-labeling/pricing. Apr. 2021.

[62] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC medicine* 17.1 (2019), pp. 1–9.

[63] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. "How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods". In: *Advances in Neural Information Processing Systems* 33 (2020).

[64] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. "Evaluating XAI: A comparison of rule-based and example-based explanations". In: *Artificial Intelligence* 291 (2021), p. 103404. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2020.103404. URL: http://www.sciencedirect.com/science/article/pii/S0004370220301533.

[65] Gordana Dodig-Crnkovic. "Scientific methods in computer science". In: *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia*. 2002, pp. 126–130.

[66] D. E. Comer, David Gries, Michael C. Mulder, Allen Tucker, A. Joe Turner, Paul R. Young, and Peter J. Denning. "Computing as a Discipline". In: *Commun. ACM* 32.1 (Jan. 1989), pp. 9–23. ISSN: 0001-0782. DOI: 10.1145/63238.63239. URL: https://doi.org/10.1145/63238.63239.

[67] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven Hicks, Hugo L. Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. "SinGAN-Seg: Synthetic Training Data Generation for Medical Image Segmentation". In: *arXiv preprint* (2021).

[68] Steven A. Hicks, Vajira Thambawita, Hugo L. Hammer, Trine B. Haugen, Jorunn M. Andersen, Oliwia Witczak, Pål Halvorsen, and Michael A. Riegler. "ACM Multimedia BioMedia 2020 Grand Challenge Overview". In: New York, NY, USA: Association for Computing Machinery, 2020, pp. 4655–4658. ISBN: 9781450379885. URL: https://doi.org/10.1145/3394171.3416287.

[69] Trine B. Haugen, Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Hugo L. Hammer, Rune Borgli, Pål Halvorsen, and Michael A. Riegler. "VISEM: A Multimodal Video Dataset of Human Spermatozoa". In: *Proceedings of the 10th ACM on Multimedia Systems Conference*. MMSys'19. Amherst, MA, USA: ACM, 2019. DOI: 10.1145/3304109.3325814. URL: http://doi.acm.org/10.1145/3304109.3325814.

[70] Vajira Thambawita, Jonas L Isaksen, Steven Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Inga Strümke, Hugo L. Hammer, Mary M Maleckar, Pål Halvorsen, Michael A. Riegler, and Jørgen K. Kanters. "DeepFake electrocardiograms: the beginning of the end for privacy issues in medicine". In: *medRxiv* (2021). DOI: 10.1101/2021.04.27.21256189. eprint: https://www.medrxiv.org/content/early/2021/04/30/2021.04.27.21256189.full.pdf. URL: https://www.medrxiv.org/content/early/2021/04/30/2021.04.27.21256189.

[71] Vajira Thambawita, Steven A. Hicks, Jonas Isaksen, Mette Haug Stensen, Trine B. Haugen, JØrgen Kanters, Sravanthi Parasa, Thomas de Lange, Håvard D. Johansen, Dag Johansen, Hugo L. Hammer, Pål Halvorsen, and Michael A. Riegler.

"DeepSynthBody: the beginning of the end for data deficiency in medicine". In: *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. 2021, pp. 1–8. DOI: 10.1109/ICAPAI49758.2021.9462062.

[72] Mathias Kirkerød, Vajira Thambawita, Michael Riegler, and Pål Halvorsen. "Using Preprocessing as a Tool in Medical Image Detection." In: *Proceedings of MediaEval*. 2018.

[73] Mathias Kirkerød, Rune Johan Borgli, Vajira Thambawita, Steven Hicks, Michael Alexander Riegler, and Pål Halvorsen. "Unsupervised preprocessing to improve generalisation for medical image classification". In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019, pp. 1–6. DOI: 10.1109/ISMICT.2019.8743979.

[74] Oda O. Nedrejord, Vajira Thambawita, Steven A. Hicks, Pål Halvorsen, and Michael A. Riegler. "Vid2Pix - A Framework for Generating High-Quality Synthetic Videos". In: *2020 IEEE International Symposium on Multimedia (ISM)*. 2020, pp. 25–26. DOI: 10.1109/ISM.2020.00010.

[75] Vajira Thambawita, Steven Hicks, Strümke Inga, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. "Generative Adversarial Networks For Creating Realistic Artificial Colon Polyp Images". In: *Gastrointestinal Endoscopy* (2021). DDW 2021 ASGE Program and Abstracts.

[76] Vajira Thambawita, Trine B. Haugen, Mette Haug Stensen, Oliwia Witczak, Hugo L. Hammer, Pål Halvorsen, and Michael A. Riegler. "Identification of spermatozoa by unsupervised learning from video data". In: *Proceedings of ESHRE*. 2021.

[77] Vajira Thambawita, Hugo Lewi Hammer, Michael Riegler, and Pål Halvorsen. "GANEx: A complete pipeline of training, inference and benchmarking GAN experiments". In: *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2019, pp. 1–4.

[78] Olav A. Norgård Rongved, Steven A. Hicks, Vajira Thambawita, Håkon K. Stensland, Evi Zouganeli, Dag Johansen, Michael A. Riegler, and Pål Halvorsen. "Real-Time Detection of Events in Soccer Videos using 3D Convolutional Neural Networks". In: *2020 IEEE International Symposium on Multimedia (ISM)*. 2020, pp. 135–144. DOI: 10.1109/ISM.2020.00030.

[79] M. C. Elish and danah boyd. “Situating methods in the magic of Big Data and AI”. In: *Communication Monographs* 85.1 (2018), pp. 57–80. DOI: 10.1080/03637751.2017.1375130. eprint: https://doi.org/10.1080/03637751.2017.1375130. URL: https://doi.org/10.1080/03637751.2017.1375130.

[80] Kristian Kersting and Ulrich Meyer. “From Big Data to Big Artificial Intelligence?” In: *KI - Künstliche Intelligenz* 32.1 (2018), pp. 3–8. DOI: 10.1007/s13218-017-0523-7. URL: https://doi.org/10.1007/s13218-017-0523-7.

[81] Kurt Benke and Geza Benke. “Artificial Intelligence and Big Data in Public Health”. In: *International journal of environmental research and public health* 15.12 (Dec. 2018), p. 2796. DOI: 10.3390/ijerph15122796. URL: https://pubmed.ncbi.nlm.nih.gov/30544648.

[82] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[83] *A sample MRI - 7-Tesla MRI scanner*. Accessed: 2021-04-25. URL: https://openneuro.org/datasets/ds003642/versions/1.0.0/file-display/sub-075:ses-001:anat:sub-075_ses-001_INV2.nii.gz.

[84] *A sample MRI - 7-Tesla MRI scanner*. Accessed: 2021-04-25. URL: http://neuromorpho.org/.

[85] *Organ System Definition*. https://biologydictionary.net/organ-system/. Accessed: 2021-01-25.

[86] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. “NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy”. In: *Nucleic acids research* 40.D1 (2012), pp. D130–D135.

[87] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[88] Alexander Andreopoulos and John K Tsotsos. “Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI”. In: *Medical Image Analysis* 12.3 (2008), pp. 335–357.

[89] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients". In: *Scientific Data* 7.1 (2020), p. 48. DOI: 10.1038/s41597-020-0386-x. URL: https://doi.org/10.1038/s41597-020-0386-x.

[90] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection". In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. MMSys'17. Taipei, Taiwan: ACM, 2017, pp. 164–169. ISBN: 9781450350020. DOI: 10.1145/3083187.3083212. URL: http://doi.acm.org/10.1145/3083187.3083212.

[91] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. "Nerthus: A Bowel Preparation Quality Video Dataset". In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. MMSys'17. Taipei, Taiwan: ACM, 2017, pp. 170–174. ISBN: 9781450350020. DOI: 10.1145/3083187.3083216. URL: http://doi.acm.org/10.1145/3083187.3083216.

[92] Romain Leenhardt, Cynthia Li, Jean-Philippe Le Mouel, Gabriel Rahmi, Jean Christophe Saurin, Franck Cholet, Arnaud Boureille, Xavier Amiot, Michel Delvaux, Clotilde Duburque, Chloé Leandri, Romain Gérard, Stéphane Lecleire, Farida Mesli, Isabelle Nion-Larmurier, Olivier Romain, Sylvie Sacher-Huvelin, Camille Simon-Shane, Geoffroy Vanbiervliet, Philippe Marteau, Aymeric Histace, and Xavier Dray. "CAD-CAP: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy". In: *Endoscopy international open* 8.3 (Mar. 2020), E415–E420. DOI: 10.1055/a-1035-9088. URL: https://pubmed.ncbi.nlm.nih.gov/32118115.

[93] Manuel Barberio, Marianne Maktabi, Ines Gockel, Nada Rayes, Boris Jansen-Winkeln, Hannes Köhler, Sebastian M. Rabe, Lena Seidemann, Jonathan P. Takoh, Michele Diana, Thomas Neumuth, and Claire Chalopin. "Hyperspectral based discrimination of thyroid and parathyroid during surgery". In: *Current Directions in*

Biomedical Engineering 4.1 (2018), pp. 399–402. DOI: doi:10.1515/cdbme-2018-0095. URL: https://doi.org/10.1515/cdbme-2018-0095.

[94] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. "An open access thyroid ultrasound image database". In: *10th International Symposium on Medical Information Processing and Analysis*. Vol. 9287. International Society for Optics and Photonics. 2015, 92870W.

[95] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5.1 (2018), p. 180161. DOI: 10.1038/sdata.2018.161. URL: https://doi.org/10.1038/sdata.2018.161.

[96] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". In: *Scientific data* 8.1 (2021), pp. 1–8.

[97] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. "A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2014, pp. 520–527.

[98] Kemal Polat and Salih Güneş. "Automatic determination of diseases related to lymph system from lymphography data using principles component analysis (PCA), fuzzy weighting pre-processing and ANFIS". In: *Expert Systems with Applications* 33.3 (2007), pp. 636–641. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2006.06.004. URL: https://www.sciencedirect.com/science/article/pii/S0957417406001898.

[99] G. Andreisek, Benedikt Kislinger, R. Dessouky, and A. Chhabra. "MRI of the Intrinsic Muscles of the Hand". In: *Seminars in Musculoskeletal Radiology* 21 (2017), pp. 392–402.

[100] F. S. Gayzik, D. P. Moreno, C. P. Geer, S. D. Wuertzer, R. S. Martin, and J. D. Stitzel. "Development of a Full Body CAD Dataset for Computational Modeling: A Multi-modality Approach". In: *Annals of Biomedical Engineering* 39.10 (2011), p. 2568. DOI: 10.1007/s10439-011-0359-5. URL: https://doi.org/10.1007/s10439-011-0359-5.

[101] Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. "BOLD5000, a public fMRI dataset while viewing 5000 visual images". In: *Scientific Data* 6.1 (2019), p. 49. DOI: 10.1038/s41597-019-0052-3. URL: https://doi.org/10.1038/s41597-019-0052-3.

[102] Florian Knoll, Martin Holler, Thomas Koesters, Ricardo Otazo, Kristian Bredies, and Daniel K Sodickson. "Joint MR-PET Reconstruction Using a Multi-Channel Image Regularizer". In: *IEEE Transactions on Medical Imaging* 36.1 (2017), pp. 1–16. DOI: 10.1109/TMI.2016.2564989.

[103] Konrad S Famulski, Declan G de Freitas, Chatchai Kreepala, Jessica Chang, Joana Sellares, Banu Sis, Gunilla Einecke, Michael Mengel, Jeff Reeve, and Philip F Halloran. "Molecular phenotypes of acute kidney injury in kidney transplants". In: *Journal of the American Society of Nephrology* 23.5 (2012), pp. 948–958.

[104] Mehmet Sarier, Ibrahim Duman, Mehmet Callioglu, Ahmet Soylu, Sabri Tekin, Emrah Turan, Hasan Celep, Asuman Havva Yavuz, Alper Demirbas, and Erdal Kukul. "Outcomes of conservative management of asymptomatic live donor kidney stones". In: *Urology* 118 (2018), pp. 43–46.

[105] Soroush Javadi and Seyed Abolghasem Mirroshandel. "A novel deep learning method for automatic assessment of human sperm images". In: *Computers in Biology and Medicine* 109 (2019), pp. 182–194. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.04.030. URL: https://www.sciencedirect.com/science/article/pii/S0010482519301386.

[106] Fernando H Biase, Xiaoyi Cao, and Sheng Zhong. "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing". In: *Genome research* 24.11 (2014), pp. 1787–1796.

[107] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[108] SP Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. "Mosmeddata: Chest ct scans with covid-19 related findings dataset". In: *arXiv preprint arXiv:2005.06465* (2020).

[109] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. "Mura: Large dataset for abnormality detection in musculoskeletal radiographs". In: *arXiv preprint arXiv:1712.06957* (2017).

[110] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet". In: *PLoS medicine* 15.11 (2018), e1002699.

[111] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain". In: *Applied Sciences* 11.2 (2021), p. 796.

[112] Sandeep Dutta and Eric Gros. "Evaluation of the impact of deep learning architectural components selection and dataset size on a medical imaging task". In: *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 10579. International Society for Optics and Photonics. 2018, p. 1057911.

[113] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance". In: *Neural networks* 21.2-3 (2008), pp. 427–436.

[114] Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232.

[115] Edward S Dove and Mark Phillips. "Privacy law, data sharing policies, and medical data: a comparative perspective". In: *Medical data privacy handbook*. Springer, 2015, pp. 639–678.

[116] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. "Kvasir-SEG: A Segmented Polyp Dataset". In: *MultiMedia Modeling*. Ed. by Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve. Cham: Springer International Publishing, 2020, pp. 451–462. ISBN: 9783030377342.

[117] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodrıguez, and Fernando Vilariño. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized Medical Imaging and Graphics* 43 (2015), pp. 99–111.

[118] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer". In: *International journal of computer assisted radiology and surgery* 9.2 (2014), pp. 283–293.

[119] D.F. Specht. "A general regression neural network". In: *IEEE Transactions on Neural Networks* 2.6 (1991), pp. 568–576. DOI: 10.1109/72.97934.

[120] Xueheng Qiu, Le Zhang, Ye Ren, P. N. Suganthan, and Gehan Amaratunga. "Ensemble deep learning for regression and time series forecasting". In: *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*. 2014, pp. 1–6. DOI: 10.1109/CIEL.2014.7015739.

[121] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. "Accurate Uncertainties for Deep Learning Using Calibrated Regression". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 2796–2804. URL: http://proceedings.mlr.press/v80/kuleshov18a.html.

[122] Heba Mohsen, El-Sayed A El-Dahshan, El-Sayed M El-Horbaty, and Abdel-Badeeh M Salem. "Classification using deep learning neural networks for brain tumors". In: *Future Computing and Informatics Journal* 3.1 (2018), pp. 68–71.

[123] Waseem Rawat and Zenghui Wang. "Deep convolutional neural networks for image classification: A comprehensive review". In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[124] Rachel Huang, Jonathan Pedoeem, and Cuixian Chen. "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 2503–2510.

[125] Zhuoling Li, Minghui Dong, Shiping Wen, Xiang Hu, Pan Zhou, and Zhigang Zeng. "CLU-CNNs: Object detection for medical images". In: *Neurocomputing* 350 (2019), pp. 53–59.

[126] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges". In: *Journal of Digital Imaging* 32.4 (2019), pp. 582–596. DOI: 10.1007/s10278-019-00227-x. URL: https://doi.org/10.1007/s10278-019-00227-x.

[127] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation". In: *Medical Image Analysis* 63 (2020), p. 101693.

[128] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, and Sathvik Koneru. "Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks". In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol. 11314. International Society for Optics and Photonics. 2020, p. 1131416.

[129] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[130] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[131] Lu Mi, Macheng Shen, and Jingzhao Zhang. "A probe towards understanding gan and vae models". In: *arXiv preprint arXiv:1812.05676* (2018).

[132] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. "Variational autoencoder for semi-supervised text classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[133] Xin Yi, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review". In: *Medical image analysis* 58 (2019), p. 101552.

[134] Shengjia Zhao, Jiaming Song, and Stefano Ermon. "Towards deeper understanding of variational autoencoding models". In: *arXiv preprint arXiv:1702.08658* (2017).

[135] T. Jaydeep. "Comparative Study of GAN and VAE". In: *International Journal of Computer Applications* 182 (2018), pp. 1–5.

[136] Farzan Farnia and Asuman Ozdaglar. "Do GANs always have Nash equilibria?" In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 3029–3039. URL: http://proceedings.mlr.press/v119/farnia20a.html.

[137] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. "Are GANs Created Equal? A Large-Scale Study". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[138] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. "Seeing what a gan cannot generate". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4502–4511.

[139] Zhaoyu Zhang, Mengyan Li, and Jun Yu. "On the convergence and mode collapse of gan". In: *SIGGRAPH Asia 2018 Technical Briefs*. 2018, pp. 1–4.

[140] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. "Generative adversarial network: An overview of theory and applications". In: *International Journal of Information Management Data Insights* (2021), p. 100004.

[141] Lan Lan, Lei You, Zeyang Zhang, Zhiwei Fan, Weiling Zhao, Nianyin Zeng, Yidong Chen, and Xiaobo Zhou. "Generative Adversarial Networks and Its Applications in Biomedical Informatics". In: *Frontiers in Public Health* 8 (2020), p. 164.

[142] Zhengwei Wang, Qi She, and Tomas E Ward. "Generative adversarial networks in computer vision: A survey and taxonomy". In: *arXiv preprint arXiv:1906.01529* (2019).

[143] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[144] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR* (2017).

[145] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017.

[146] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: *CoRR* abs/1912.04958 (2019).

[147] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training generative adversarial networks with limited data". In: *arXiv preprint arXiv:2006.06676* (2020).

[148] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=B1xsqj09Fm.

[149] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. "SinGAN: Learning a Generative Model from a Single Natural Image". In: *Computer Vision (ICCV), IEEE International Conference on*. 2019.

[150] Ali Borji. "Pros and cons of gan evaluation measures". In: *Computer Vision and Image Understanding* 179 (2019), pp. 41–65.

[151] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. "Synthetic data augmentation using GAN for improved liver lesion classification". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 289–293.

[152] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Plácido Rogerio Pinheiro. "Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection". In: *Ieee Access* 8 (2020), pp. 91916–91923.

[153] Talha Iqbal and Hazrat Ali. "Generative adversarial network for medical images (MI-GAN)". In: *Journal of medical systems* 42.11 (2018), pp. 1–11.

[154] Hitesh Tekchandani, Shrish Verma, and Narendra Londhe. "Performance improvement of mediastinal lymph node severity detection using GAN and Inception network". In: *Computer Methods and Programs in Biomedicine* 194 (2020), p. 105478.

[155] Avi Ben-Cohen, Eyal Klang, Stephen P Raskin, Shelly Soffer, Simona Ben-Haim, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. "Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection". In: *Engineering Applications of Artificial Intelligence* 78 (2019), pp. 186–194.

[156] Shuangting Liu, Jiaqi Zhang, Yuxin Chen, Yifan Liu, Zengchang Qin, and Tao Wan. "Pixel level data augmentation for semantic image segmentation using generative adversarial networks". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 1902–1906.

[157] Marco Domenico Cirillo, David Abramian, and Anders Eklund. "Vox2Vox: 3D-GAN for brain tumour segmentation". In: *arXiv preprint arXiv:2003.13653* (2020).

[158] Zhongyi Han, Benzheng Wei, Ashley Mercado, Stephanie Leung, and Shuo Li. "Spine-GAN: Semantic segmentation of multiple spinal structures". In: *Medical image analysis* 50 (2018), pp. 23–35.

[159] Jin Zhu, Guang Yang, and Pietro Lio. "How can we make gan perform better in single medical image super-resolution? A lesion focused multi-scale approach". In:

2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE. 2019, pp. 1669–1673.

[160] Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, and Dimitris N Metaxas. "Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13856–13866.

[161] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: *Journal of the American Medical Informatics Association* 25.3 (2018), pp. 230–238.

[162] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. "Anonymization through data synthesis using generative adversarial networks (ads-gan)". In: *IEEE journal of biomedical and health informatics* 24.8 (2020), pp. 2378–2388.

[163] Debapriya Hazra and Yung-Cheol Byun. "SynSigGAN: Generative Adversarial Networks for Synthetic Biomedical Signal Generation". In: *Biology* 9.12 (2020), p. 441.

[164] Darius Dirvanauskas, Rytis Maskeliūnas, Vidas Raudonis, Robertas Damaševičius, and Rafal Scherer. "Hemigen: human embryo image generator based on generative adversarial networks". In: *Sensors* 19.16 (2019), p. 3578.

[165] Hasib Zunair and A Ben Hamza. "Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation". In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–12.

[166] Anthony Dupre, Sarah Vincent, and Paul A Iaizzo. "Basic ECG theory, recordings, and interpretation". In: *Handbook of cardiac anatomy, physiology, and devices*. Springer, 2005, pp. 191–201.

[167] Clive Whiston and Florence Elizabeth Prichard. "X-ray Methods". In: (1987).

[168] Jyoti Shah. "Endoscopy through the ages". In: *BJU international* 89.7 (2002), pp. 645–652.

[169] AM Blamire. “The technology of MRI—the next 10 years?” In: *The British journal of radiology* 81.968 (2008), pp. 601–617.

[170] Michele Larobina and Loredana Murino. “Medical image file formats”. In: *Journal of digital imaging* 27.2 (2014), pp. 200–206.

[171] Alois Schlögl. “An overview on data formats for biomedical signals”. In: *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*. Springer. 2009, pp. 1557–1560.

[172] FJ Murphy. “The paradox of imaging technology: a review of the literature”. In: *Radiography* 12.2 (2006), pp. 169–174.

[173] Zachary Munn and Zoe Jordan. “The patient experience of high technology medical imaging: a systematic review of the qualitative evidence”. In: *Radiography* 17.4 (2011), pp. 323–331.

[174] Jacob Beutel, Harold L Kundel, and Richard L Van Metter. *Handbook of medical imaging*. Vol. 1. Spie Press, 2000.

[175] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014).

[176] *Get to production AI faster*. `https://labelbox.com/`. Accessed: 2021-04-17.

[177] *Basic AI*. `https://www.basic.ai/`. Accessed: 2021-04-17.

[178] *Awesome data annotation*. https://github.com/taivop/awesome-data-annotation. Accessed: 2021-04-17.

[179] CR Juhl, IM Miller, GB Jemec, JK Kanters, and C Ellervik. “Hidradenitis suppurativa and electrocardiographic changes: a cross-sectional population study”. In: *British Journal of Dermatology* 178.1 (2018), pp. 222–228.

[180] Jonas Ghouse, Christian Theil Have, Peter Weeke, Jonas Bille Nielsen, Gustav Ahlberg, Marie Balslev-Harder, Emil Vincent Appel, Tea Skaaby, Søren-Peter Olesen, Niels Grarup, et al. “Rare genetic variants previously associated with congenital forms of long QT syndrome have little or no effect on the QT interval”. In: *European heart journal* 36.37 (2015), pp. 2523–2529.

[181] *MUSE v9 Cardiology Information System*. Accessed: 2021-03-26. URL: https://www.gehealthcare.com.au/products/diagnostic-ecg/cardio-data-management/muse-v9.

[182] *Electrocardiography-Wikipedia*. Accessed: 2021-03-26. URL: https://en.wikipedia.org/wiki/Electrocardiography.

[183] *Sperm and Semen Testing and Evaluation*. Accessed: 2021-03-26. URL: https://www.fertility-docs.com/programs-and-services/sperm-evaluation/sperm-and-semen-testing.php.

[184] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Hicks, Kristin Ranheim Randel, Duc Tien Dang Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange. "Medico multimedia task at mediaeval 2018". In: *CEUR Workshop Proceedings*. Vol. 2283. Technical University of Aachen. 2018, pp. 1–4.

[185] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[186] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[187] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. "A Deeper Look at Dataset Bias". In: *Domain Adaptation in Computer Vision Applications*. Ed. by Gabriela Csurka. Cham: Springer International Publishing, 2017, pp. 37–55. ISBN: 9783319583471. DOI: 10.1007/978-3-319-58347-1_2. URL: https://doi.org/10.1007/978-3-319-58347-1_2.

[188] Debesh Jha, Steven A Hicks, Krister Emanuelsen, Håvard Johansen, Dag Johansen, Thomas de Lange, Michael A Riegler, and Pål Halvorsen. "Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation". In: *arXiv preprint arXiv:2012.15244* (2020).

[189] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: *Scandinavian conference on Image analysis*. Springer. 2003, pp. 363–370.

[190] Steven Hicks, Pål Halvorsen, Trine B Haugen, Jorunn M Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L Hammer, Duc-Tien Dang-Nguyen, Mathias Lux, and Michael Riegler. "Medico Multimedia Task at MediaEval 2019". In: *CEUR Workshop Proceedings-Multimedia Benchmark Workshop (MediaEval)*. 2019.

[191] Bruce D Lucas, Takeo Kanade, et al. "An iterative image registration technique with an application to stereo vision". In: Vancouver, British Columbia. 1981.

[192] Chris Donahue, Julian McAuley, and Miller Puckette. "Adversarial Audio Synthesis". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=ByMVTsR5KQ.

[193] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[194] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8110–8119.

[195] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6629–6640. ISBN: 9781510860964.

[196] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410.

[197] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. "Image Inpainting via Generative Multi-column Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 331–340.

[198] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. "A survey on deep transfer learning". In: *International conference on artificial neural networks*. Springer. 2018, pp. 270–279.

[199] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. "A neural algorithm of artistic style". In: *arXiv preprint arXiv:1508.06576* (2015).

[200] KR Srinath. "Python–the fastest growing programming language". In: *International Research Journal of Engineering and Technology* 4.12 (2017), pp. 354–357.

[201] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[202] Vineel Nagisetty, Laura Graves, Joseph Scott, and Vijay Ganesh. "xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems". In: *arXiv preprint arXiv:2002.10438* (2020).

[203] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2019.

[204] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and beyond: A survey of face manipulation and fake detection". In: *Information Fusion* 64 (2020), pp. 131–148.

[205] Yaël Frégier and Jean-Baptiste Gouray. "Mind2Mind: transfer learning for GANs". In: *arXiv preprint arXiv:1906.11613* (2019).

[206] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. "Transferring gans: generating images from limited data". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 218–234.

[207] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. "To learn image super-resolution, use a GAN to learn how to do image degradation first". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[208] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. "Federated learning: Strategies for improving communication efficiency". In: *arXiv preprint arXiv:1610.05492* (2016).

[209] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H Brendan McMahan, et al. "Towards federated learning at scale: System design". In: *arXiv preprint arXiv:1902.01046* (2019).

Appendix A

Published Articles

Note: Due to the original size of this file, all of the originally attached papers/articles in this dissertation have been removed to make the file more manageable. All links to papers/articles on the web remain. Another number of articles have been removed due to copyright.

A.1 Paper I - HyperKvasir, a Comprehensive Multi-class Image and Video Dataset for Gastrointestinal Endoscopy

Authors: Hanna Borgli, **Vajira Thambawita**, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K. Stensland, Enrique Garcia-Ceja, Peter T. Schmidt, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, Thomas de Lange

Abstract: Artificial intelligence is currently a hot topic in medicine. However, medical data is often sparse and hard to obtain due to legal restrictions and lack of medical personnel for the cumbersome and tedious process to manually label training data. These constraints make it difficult to develop systems for automatic analysis, like detecting disease or other lesions. In this respect, this article presents HyperKvasir, the largest image and video dataset of the gastrointestinal tract available today. The data is collected during real gastro- and colonoscopy examinations at Bærum Hospital in Norway and partly labeled by experienced gastrointestinal endoscopists. The dataset contains 110,079 images and 374 videos, and represents anatomical landmarks as well as pathological and normal findings. The total number of images and video frames together is around 1 million. Initial experiments demonstrate the potential benefits of artificial intelligence-based computer-assisted diagnosis systems. The HyperKvasir dataset can play a valuable role in developing better algorithms and computer-assisted examination systems not only for gastro- and colonoscopy, but also for other fields in medicine.

Published: Nature scientific data, 2020. DOI: https://doi.org/10.1038/s41597-020-00622-y

Candidate contributions: Vajira contributed (as one of the main authors) to the conception and design of the paper and doing the deep learning baseline experiments in this manuscript for the classification task by critically analyzing the data. He developed and analyzed four different deep learning methods (ResNeet-152, DenseNet-161, averaged ResNet-152, DenseNet-161, and combined ResNet-152 and DenseNet-161 through an MLP) using two folds cross-validation and Pytorch deep learning

framework. These deep learning experiments show the best baseline performance of this paper. He contributed to drafting the manuscript and revising it.

Thesis objectives: Sub-objective I, Sub-objective II

A.2 Paper II - Toadstool: A Dataset for Training Emotional Intelligent Machines Playing Super Mario Bros

Authors: Henrik Svoren, **Vajira Thambawita**, Pål Halvorsen, Petter Jakobsen, Enrique Garcia-Ceja, Farzan Majeed Noori, Hugo L. Hammer, Mathias Lux, Michael Alexander Riegler, Steven Alexander Hicks

Abstract: Games are often defined as engines of experience, and they are heavily relying on emotions, they arouse in players. In this paper, we present a dataset called Toadstool as well as a reproducible methodology to extend on the dataset. The dataset consists of video, sensor, and demographic data collected from ten participants playing Super Mario Bros, an iconic and famous video game. The sensor data is collected through an Empatica E4 wristband, which provides high-quality measurements and is graded as a medical device. In addition to the dataset and the methodology for data collection, we present a set of baseline experiments which show that we can use video game frames together with the facial expressions to predict the blood volume pulse of the person playing Super Mario Bros. With the dataset and the collection methodology we aim to contribute to research on emotionally aware machine learning algorithms, focusing on reinforcement learning and multimodal data fusion. We believe that the presented dataset can be interesting for a manifold of researchers to explore exciting new interdisciplinary questions.

Published: The ACM Multimedia Systems Conference (MMSys) - 2020. DOI: https://doi.org/10.1145/3339825.3394939

Candidate contributions: Vajira contributed to the conception and designing of the theoretical models. He contributed to collecting data as a participant also. Vajira contributed to publishing data in osf.io and organizing it. He contributed to drafting the paper and revising it.

Thesis objectives: Sub-objective I, Sub-objective II

A.3 Paper III - PMData: A Sports Logging Dataset

Authors: Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, Simon Nordvang, Sigurd Pedersen, Anders Gjerdrum, Tor-Morten Grønli, Per Morten Fredriksen, Ragnhild Eg, Kjeld Hansen, Siri Fagernes, Christine Claudi, Andreas Biørn-Hansen, Duc Tien Dang Nguyen, Tomas Kupka, Hugo Lewi Hammer, Ramesh Jain, Michael Alexander Riegler, Pål Halvorsen

Abstract: In this paper, we present PMData: a dataset that combines traditional lifelogging data with sports-activity data. Our dataset enables the development of novel data analysis and machine-learning applications where, for instance, additional sports data is used to predict and analyze everyday developments, like a person's weight and sleep patterns; and applications where traditional lifelog data is used in a sports context to predict athletes' performance. PMData combines input from Fitbit Versa 2 smartwatch wristbands, the PMSys sports logging smartphone application, and Google forms. Logging data has been collected from 16 persons for five months. Our initial experiments show that novel analyses are possible, but there is still room for improvement.

Published: The ACM Multimedia Systems Conference (MMSys) -2020. DOI: https://doi.org/10.1145/3339825.3394926

Candidate contributions: Vajira contributed to the analysis and interpretation of data. He contributed to drafting the article, hosting the data in an open access data hosting location (osf.io), and revising the manuscript. He presented the paper at MMSys 2020.

Thesis objectives: Sub-objective I, Sub-objective II

A.4 Paper IV - PSYKOSE: A Motor Activity Database of Patients with Schizophrenia

Authors: Petter Jakobsen, Enrique Garcia-Ceja, Lena Antonsen Stabell, Ketil Joachim Oedegaard, Jan Oystein Berle, **Vajira Thambawita**, Steven Alexander Hicks, Pål Halvorsen, Ole Bernt Fasmer, Michael Alexander Riegler

Abstract: Using sensor data from devices such as smart-watches or mobile phones is very popular in both computer science and medical research. Such movement data can predict certain health states or performance outcomes. However, in order to increase reliability and replication of the research it is important to share data and results openly. In medicine, this is often difficult due to legal restrictions or to the fact that data collected from clinical trials is seen as very valuable and something that should be kept "in-house". In this paper, we therefore present PSYKOSE, a publicly shared dataset consisting of motor activity data collected from body sensors. The dataset contains data collected from patients with schizophrenia. Schizophrenia is a severe mental disorder characterized by psychotic symptoms like hallucinations and delusions, as well as symptoms of cognitive dysfunction and diminished motivation. In total, we have data from 22 patients with schizophrenia and 32 healthy control persons. For each person in the dataset, we provide sensor data collected over several days in a row. In addition to the sensor data, we also provide some demographic data and medical assessments during the observation period. The patients were assessed by medical experts from Haukeland University hospital. In addition to the data, we provide a baseline analysis and possible use-cases of the dataset.

Published: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). DOI: https://doi.org/10.1109/CBMS49503.2020.00064

Candidate contributions: Vajira contributed to hosting the data on a public data repository (osf.io) and preparing the corresponding wiki pages as a manual for users who will use the dataset. He contributed to revising the manuscript based on the reviews.

Thesis objectives: Sub-objective I, Sub-objective II

[Article not attached due to copyright]

A.5 Paper V - Kvasir-Capsule, a Video Capsule Endoscopy Dataset

Authors: Pia H. Smedsrud, **Vajira Thambawita**, Steven A. Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc Tien Dang Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T. Schmidt, Ervin Toth, Hugo L. Hammer, Thomas de Lange, Michael A. Riegler, Pål Halvorsen

Abstract: Artificial intelligence (AI) is predicted to have profound effects on the future of video capsule endoscopy (VCE) technology. The potential lies in improving anomaly detection while reducing manual labour. Existing work demonstrates the promising benefits of AI-based computer-assisted diagnosis systems for VCE. They also show great potential for improvements to achieve even better results. Also, medical data is often sparse and unavailable to the research community, and qualified medical personnel rarely have time for the tedious labelling work. We present Kvasir-Capsule, a large VCE dataset collected from examinations at a Norwegian Hospital. Kvasir-Capsule consists of 117 videos which can be used to extract a total of 4,741,504 image frames. We have labelled and medically verified 47,238 frames with a bounding box around findings from 14 different classes. In addition to these labelled images, there are 4,694,266 unlabelled frames included in the dataset. The Kvasir-Capsule dataset can play a valuable role in developing better algorithms in order to reach true potential of VCE technology.

Published: Nature Scientific Data, 2021. DOI: https://doi.org/10.1038/s41597-021-00920-z

Candidate contributions: Vajira contributed to the main baseline experiments discussed in the paper. He performed the baseline experiments using two different deep learning methods (DenseNet-161 and ResNet-152) using Pytorch deep learning framework. In these baseline experiments, he performed deep analysis using two different loss functions such as Normal Cross-Entropy Loss, Weighted Cross-Entropy Loss, and using a weighted sampling method, and his experiments showed the best results in this paper (Refer to Table 3 in the paper). Vajira contributed to drafting and revising the paper. He especially focused on revising the technical

part of the paper.

Thesis objectives: Sub-objective I, Sub-objective II

A.6 Paper VI - HTAD: A Home-Tasks Activities Dataset with Wrist-Accelerometer and Audio Features

Authors: Enrique Garcia-Ceja, **Vajira Thambawita**, Steven A. Hicks, Debesh Jha, Petter Jakobsen, Hugo L. Hammer, Pål Halvorsen, Michael A. Riegler

Abstract: In this paper, we present HTAD: A Home Tasks Activities Dataset. The dataset contains wrist-accelerometer and audio data from people performing at-home tasks such as sweeping, brushing teeth, washing hands, or watching TV. These activities represent a subset of activities that are needed to be able to live independently. Being able to detect activities with wearable devices in real-time is important for the realization of assistive technologies with applications in different domains such as elderly care and mental health monitoring. Preliminary results show that using machine learning with the presented dataset leads to promising results, but also there is still improvement potential. By making this dataset public, researchers can test different machine learning algorithms for activity recognition, especially, sensor data fusion methods.

Published: MultiMedia Modeling (MMM), 2021. DOI: https://doi.org/10.1007/978-3-030-67835-7_17

Candidate contributions: Vajira contributed to organizing data and publishing the dataset in the public data repository called osf.io. He created the wiki page of this dataset and published it to users as a reference manual to the dataset. He contributed to drafting and revising the paper.

Thesis objectives: Sub-objective I, Sub-objective II

A.7 Paper VII - Kvasir-Instrument: Diagnostic and Therapeutic tool Segmentation Dataset in Gastrointestinal Endoscopy

Authors: Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A. Hicks, **Vajira Thambawita**, Enrique Garcia-Ceja, Michael A. Riegler, Thomas de Lange, Peter T. Schmidt, Håvard D. Johansen, Dag Johansen, Pål Halvorsen

Abstract: Gastrointestinal (GI) pathologies are periodically screened, biopsied, and resected using surgical tools. Usually, the procedures and the treated or resected areas are not specifically tracked or analysed during or after colonoscopies. Information regarding disease borders, development, amount, and size of the resected area get lost. This can lead to poor follow-up and bothersome reassessment difficulties post-treatment. To improve the current standard and also to foster more research on the topic, we have released the "Kvasir-Instrument" dataset, which consists of 590 annotated frames containing GI procedure tools such as snares, balloons, and biopsy forceps, etc. Besides the images, the dataset includes ground truth masks and bounding boxes and has been verified by two expert GI endoscopists. Additionally, we provide a baseline for the segmentation of the GI tools to promote research and algorithm development. We obtained a dice coefficient score of 0.9158 and a Jaccard index of 0.8578 using a classical U-Net architecture. A similar dice coefficient score was observed for DoubleUNet. The qualitative results showed that the model did not work for the images with specularity and the frames with multiple tools, while the best result for both methods was observed on all other types of images. Both qualitative and quantitative results show that the model performs reasonably good, but there is potential for further improvements. Benchmarking using the dataset provides an opportunity for researchers to contribute to the field of automatic endoscopic diagnostic and therapeutic tool segmentation for GI endoscopy.

Published: MultiMedia Modeling (MMM), 2021. DOI: https://doi.org/10.1007/978-3-030-67835-7_19

Candidate contributions: Vajira contributed to drafting and revising the paper.

Thesis objectives: Sub-objective I, Sub-objective II

A.8 Paper VIII - The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning

Authors: Vajira Thambawita, Debesh Jha, Michael Riegler, Pål Halvorsen, Hugo Lewi Hammer, Håvard D. Johansen, Dag Johansen

Abstract: In this paper, we present our approach for the 2018 Medico Task classifying diseases in the gastrointestinal tract. We have proposed a system based on global features and deep neural networks. The best approach combines two neural networks, and the reproducible experimental results signify the efficiency of the proposed model with an accuracy rate of 95.80%, a precision of 95.87%, and an F1-score of 95.80%.

Published: In the Proceedings of MediaEval 2018. URL: http://ceur-ws.org/Vol-2283/

Candidate contributions: In this working notepaper, Vajira is the first author and the corresponding author. He contributed to the main conception and design of three experiments (out of five) using deep learning approaches which use Resnet-152, Densenet-161, and a combination of these. Vajira's experiments achieved the best performance of this paper. He developed and analyzed the results of the three experiments. Vajira contributed to draft the article and revise it.

Thesis objectives: Sub-objective I, Sub-objective III

A.9 Paper IX - An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification

Authors: **Vajira Thambawita**, Debesh Jha, Hugo Lewi Hammer, Håvard D. Johansen, Dag Johansen, Pål Halvorsen, and Michael A. Riegler

Abstract: Precise and efficient automated identification of gastrointestinal (GI) tract diseases can help doctors treat more patients and improve the rate of disease detection and identification. Currently, automatic analysis of diseases in the GI tract is a hot topic in both computer science and medical-related journals. Nevertheless, the evaluation of such an automatic analysis is often incomplete or simply wrong. Algorithms are often only tested on small and biased datasets, and cross-dataset evaluations are rarely performed. A clear understanding of evaluation metrics and machine learning models with cross datasets is crucial to bring research in the field to a new quality level. Toward this goal, we present comprehensive evaluations of five distinct machine learning models using global features and deep neural networks that can classify 16 different key types of GI tract conditions, including pathological findings, anatomical landmarks, polyp removal conditions, and normal findings from images captured by common GI tract examination instruments. In our evaluation, we introduce performance hexagons using six performance metrics, such as recall, precision, specificity, accuracy, F1-score, and the Matthews correlation coefficient to demonstrate how to determine the real capabilities of models rather than evaluating them shallowly. Furthermore, we perform cross-dataset evaluations using different datasets for training and testing. With these cross-dataset evaluations, we demonstrate the challenge of actually building a generalizable model that could be used across different hospitals. Our experiments clearly show that more sophisticated performance metrics and evaluation methods need to be applied to get reliable models rather than depending on evaluations of the splits of the same dataset—that is, the performance metrics should always be interpreted together rather than relying on a single metric.

Appendix A. Published Articles

Published: ACM Transactions on Computing for Healthcare, 2020-2021. DOI: https://doi.org/10.1145/3386295

Candidate contributions: Vajira is the first author and the corresponding author of this journal paper. He contributed to the main conception and design of the experiments in this manuscript. Vajira developed and analyzed the three different deep neural networks critically in this study. Additionally, he analyzed several Gastrointestinal (GI) tract datasets to use in the experiments and evaluated his models using those cross datasets to measure the generalizability of the deep learning solutions in real-world applications. He contributed to drafting the manuscript and revising it. This journal paper is the extended version of "The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning".

Thesis objectives: Sub-objective I, Sub-objective III

A.10 Paper X - Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction

Authors: Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, **Vajira Thambawita**, Pål Halvorsen, Hugo L. Hammer, Trine B. Haugen, Michael A. Riegler

Abstract: Methods for automatic analysis of clinical data are usually targeted towards a specific modality and do not make use of all relevant data available. In the field of male human reproduction, clinical and biological data are not used to its fullest potential. Manual evaluation of a semen sample using a microscope is time-consuming and requires extensive training. Furthermore, the validity of manual semen analysis has been questioned due to limited reproducibility, and often high inter-personnel variation. The existing computer-aided sperm analyzer systems are not recommended for routine clinical use due to methodological challenges caused by the consistency of the semen sample. Thus, there is a need for an improved methodology. We use modern and classical machine learning techniques together with a dataset consisting of 85 videos of human semen samples and related participant data to automatically predict sperm motility. Used techniques include simple linear regression and more sophisticated methods using convolutional neural networks. Our results indicate that sperm motility prediction based on deep learning using sperm motility videos is rapid to perform and consistent. Adding participant data did not improve the algorithms performance. In conclusion, machine learning-based automatic analysis may become a valuable tool in male infertility investigation and research.

Published: Nature scientific reports, 2019. DOI: https://doi.org/10.1038/s41598-019-53217-y

Candidate contributions: Vajira contributed to the conception and design of this ar-ticle. He experimented with two different deep learning methods (out of four) which are based on dense optical flow and a novel preprocessing technique called “vertical frame matrix” to predict motility values of sperm samples. He performed his experi-ments with different input types such as pre-processed video frames and participant

data. He contributed to analyzing the results of his methods, drafting the article, and revising it.

Thesis objectives: Sub-objective I, Sub-objective III

A.11 Paper XI - Stacked Dense Optical Flows and Dropout Layers to Predict Sperm Motility and Morphology

Authors: Vajira Thambawita, Pål Halvorsen, Hugo Hammer, Michael Riegler, and Trine B. Haugen

Abstract: In this paper, we analyse two deep learning methods to predict sperm motility and sperm morphology from sperm videos. We use two different inputs: stacked pure frames of videos and dense optical flows of video frames. To solve this regression task of predicting motility and morphology, stacked dense optical flows and extracted original frames from sperm videos were used with the modified state of the art convolution neural networks. For modifications of the selected models, we have introduced an additional multi-layer perceptron to overcome the problem of over-fitting. The method which had an additional multi-layer perceptron with dropout layers, shows the best results when the inputs consist of both dense optical flows and an original frame of videos.

Published: In the Proceedings of MediaEval 2019. URL: http://ceur-ws.org/Vol-2670/

Candidate contributions: Vajira contributed to the conception and design of this working-note paper. He conducted all the experiments of this paper using two different deep learning approaches to predict motility and morphology of the given videos of sperm samples by organizers of MediaEval 2019-MedicoTask. He analyzed the results collected from his methods using three-folds cross-validation and presented the results at MedicaEval-2019 and they were the best results from all the participants of Medicotask-2019. Vajira contributed to drafting the paper and revising it.

Thesis objectives: Sub-objective I, Sub-objective III

A.12 Paper XII - Extracting Temporal Features into a Spatial Domain Using Autoencoders for Sperm Video Analysis

Authors: **Vajira Thambawita**, Pål Halvorsen, Hugo Hammer, Michael Riegler, Trine B. Haugen

Abstract: In this paper, we present a two-step deep learning method that is used to predict sperm motility and morphology-based on video recordings of human spermatozoa. First, we use an autoencoder to extract temporal features from a given semen video and plot these into image-space, which we call feature-images. Second, these feature-images are used to perform transfer learning to predict the motility and morphology values of human sperm. The presented method shows it's capability to extract temporal information into spatial domain feature-images which can be used with traditional convolutional neural networks. Furthermore, the accuracy of the predicted motility of a given semen sample shows that a deep learning-based model can capture the temporal information of microscopic recordings of human semen.

Published: In the Proceedings of MediaEval 2019. URL: http://ceur-ws.org/Vol-2670/

Candidate contributions: Vajira contributed to the conception and design of this paper. He introduced a novel architecture to extract temporal and spatial features of sperm video using auto-encoder-based architecture. Using the extracted features, Vajra predicted motility and morphology levels of a given sperm sample video. Ad-ditionally, he critically evaluated results using two different baseline experiments and two different input shapes. He contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective I, Sub-objective III

A.13 Paper XIII - ACM Multimedia BioMedia 2020 Grand Challenge Overview

Authors: Steven A. Hicks, **Vajira Thambawita**, Hugo L. Hammer, Trine B. Haugen, Jorunn M. Andersen, Oliwia Witczak, Pål Halvorsen, and Michael A. Riegler.

Abstract: The BioMedia 2020 ACM Multimedia Grand Challenge is the second in a series of competitions focusing on the use of multimedia for different medical use-cases. In this year's challenge, participants are asked to develop algorithms that automatically predict the quality of a given human semen sample using a combination of visual, patient-related, and laboratory-analysis-related data. Compared to last year's challenge, participants are provided with a fully multimodal dataset (videos, analysis data, study participant data) from the field of assisted human reproduction. The tasks encourage the use of the different modalities contained within the dataset and finding smart ways of how they may be combined to further improve prediction accuracy. For example, using only video data or combining video data and patient-related data. The ground truth was developed through a preliminary analysis done by medical experts following the World Health Organization's standard for semen quality assessment. The task lays the basis for automatic, real-time support systems for artificial reproduction. We hope that this challenge motivates multimedia researchers to explore more medical-related applications and use their vast knowledge to make a real impact on people's lives.

Published: Proceedings of the 28th ACM International Conference on Multimedia.
DOI: https://doi.org/10.1145/3394171.3416287
Candidate contributions: Vajira contributed to revising and drafting the manuscript.

Thesis objectives: Sub-objective I, Sub-objective III

A.14 Paper XIV - Explaining Deep Neural Networks for Knowledge Discovery in Electrocardiogram Analysis

Authors: Steven A. Hicks, Jonas L. Isaksen, **Vajira Thambawita**, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary M. Maleckar, Michael A. Riegler, Jørgen K. Kanters

Abstract: Deep learning-based tools may annotate and interpret medical data more quickly, consistently, and accurately than medical doctors. However, as medical doctors are ultimately responsible for clinical decision-making, any deep learning-based prediction should be accompanied by an explanation that a human can understand. We present an approach called electrocardiogram gradient class activation map (ECGradCAM), which is used to generate attention maps and explain the reasoning behind deep learning-based decision-making in ECG analysis. Attention maps may be used in the clinic to aid diagnosis, discover new medical knowledge, and identify novel features and characteristics of medical tests. In this paper, we showcase how ECGradCAM attention maps can unmask how a novel deep learning model measures both amplitudes and intervals in 12-lead electrocardiograms, and we show an example of how attention maps may be used to develop novel ECG features.

Published: Nature Scientific Reports, 2021. DOI: https://doi.org/10.1038/s41598-021-90285-5

Candidate contributions: Vajira contributed to the conception and design of this paper. He contributed to analyzing the results collected from the deep learning experiments discussed in this manuscript. He contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective I, Sub-objective III

A.15 Paper XV - Pyramid-Focus-Augmentation: Medical Image Segmentation with Step-Wise Focus

Authors: Vajira Thambawita, Steven Hicks, Pål Halvorsen, Michael A. Riegler

Abstract: Segmentation of findings in the gastrointestinal tract is a challenging but also an important task which is an important building stone for sufficient automatic decision support systems. In this work, we present our solution for the Medico 2020 task, which focused on the problem of colon polyp segmentation. We present our simple but efficient idea of using an augmentation method that uses grids in a pyramid-like manner (large to small) for segmentation. Our results show that the proposed methods work as indented and can also lead to comparable results when competing with other methods.

Published: In the Proceedings of MediaEval 2020. DOI: https://doi.org/10.48550/arXiv.2012.07430

Candidate contributions: Vajira contributed to the conception and design of the pyramid-focus-augmentation study. He conducted all the experiments for this manuscript and analyzed the results with baseline experiments. Vajira published the finding of this study as a python package index (`https://pypi.org/project/pyra-pytorch/`) and GitHub repository (`https://vlbthambawita.github.io/PYRA/`) which can be used by other researchers. He contributed to drafting the manuscript and revising it.

Thesis objectives: Sub-objective I, Sub-objective III

A.16 Paper XVI - Impact of Image Resolution on Convolutional Neural Networks Performance in Gastrointestinal Endoscopy

Authors: Vajira Thambawita, Steven Hicks, Inga Strümke, Michael Riegler, Pål Halvorsen, Sravanthi Parasa

Abstract: Convolutional neural networks (CNNs) are increasingly used to improve and automate processes in gastroenterology, like the detection of polyps during a colonoscopy. An important input to these methods is images and videos. Up until now, no well-defined, common understanding or standard regarding the resolution of the images and video frames has been defined, and to reduce processing time and resource requirements, images are today almost always down-sampled. However, how such down-sampling and the image resolution influence the performance in context with medical data is unknown. In this work, we investigate how the resolution relates to the performance of convolutional neural networks. This can help set standards for image or video characteristics for future CNN based models in gastrointestinal endoscopy.

Published: AGA, DDW Abstract Issue, 2021. DOI: https://doi.org/10.1016/S0016-5085(21)01616-4.

Candidate contributions: Vajira contributed to the conception and design of this abstract. He conducted all the experiments presenting in this study and he tested the effect of image resolution for deep neural networks using two different well-known neural networks, namely ResNet-151 and DenseNet-161. Vajira contributed to analyzing the results collected from these experiments. He contributed to drafting and revising the abstract.

Thesis objectives: Sub-objective III

A.17 Paper XVII - On Evaluation Metrics for Medical Applications of Artificial Intelligence

Authors: Steven A. Hicks, Inga Strümke, **Vajira Thambawita**, Malek Hammou, Michael A. Riegler, Pål Halvorsen, Sravanthi Parasa

Abstract: Clinicians and model developers need to understand how proposed machine learning (ML) models could improve patient care. In fact, no single metric captures all the desirable properties of a model and several metrics are typically reported to summarize a model's performance. Unfortunately, these measures are not easily understandable by many clinicians. Moreover, comparison of models across studies in an objective manner is challenging, and no tool exists to compare models using the same performance metrics. This paper looks at previous ML studies done in gastroenterology, provides an explanation of what different metrics mean in the context of the presented studies, and gives a thorough explanation of how different metrics should be interpreted. We also release an open source web-based tool that may be used to aid in calculating the most relevant metrics presented in this paper so that other researchers and clinicians may easily incorporate them into their research.

Published: Submitted for publication, Preprint is available at medRxiv. DOI: https://doi.org/10.1101/2021.04.07.21254975

Candidate contributions: Vajira contributed to designing and developing the concept of this paper. He also contributed to the main analysis of the results collected from the literature reviews. Also, Vajira contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective III

A.18 Paper XVIII - DivergentNets: Medical Image Segmentation by Network Ensemble

Authors: Vajira Thambawita, Steven A. Hicks, Pål Halvorsen, Michael A. Riegler

Abstract: Detection of colon polyps has become a trending topic in the intersecting fields of machine learningand gastrointestinal endoscopy. The focus has mainly been on per-frame classification. More recently,polyp segmentation has gained attention in the medical community. Segmentation has the advantageof being more accurate than per-frame classification or object detection as it can show the affected areain greater detail. For our contribution to the EndoCV 2021 segmentation challenge, we propose twoseparate approaches. First, a segmentation model namedTriUNet-composed of three separate UNetmodels. Second, we combine TriUNet with an ensemble of well-known segmentation models, namelyUNet++, FPN, DeepLabv3, and DeepLabv3+, into a model calledDivergentNetsto produce more generalizable medical image segmentation masks. In addition, we propose a modified Dice loss that calculatesloss only for a single class when performing multi-class segmentation, forcing the model to focus onwhat is most important. Overall, the proposed methods achieved the best average scores for each re-spective round in the challenge, with TriUNet being the winning model in Round I and DivergentNetsbeing the winning model in Round II of the segmentation generalization challenge at EndoCV 2021. Theimplementation of our approach is made publicly available on GitHub.

Published: In proceedings of EndoCV 2021. DOI: https://doi.org/10.48550/arXiv.2107.00283

Candidate contributions: Vajira contributed to the conception and design of this study. He introduced two new deep neural network architectures named as TriUNet and DivergentNets to perform the segmentation task of EndoCV grand challenge 2021. Vajira contributed to developing these two architectures and performed the experiments. He collected results from the two networks and submitted them to the cloud platform of the challenge. According to his submissions, his team won first place in the polyp segmentation task. Vajira contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective I, Sub-objective III

A.19 Paper XIX - A Self-learning Teacher-student Framework for Gastrointestinal Image Classification

Authors: Henrik L. Gjestang, Steven A. Hicks, **Vajira Thambawita**, Pål Halvorsen, Michael A. Riegler

Abstract: We present a semi-supervised teacher-student framework to improve classification performance on gastrointestinal image data. As labeled data is scarce in medical settings, this framework is built specifically to take advantage of vast amounts of unlabeled data. It consists of three main steps: (1) train a teacher model with labeled data, (2) use the teacher model to infer pseudo labels with unlabeled data, and (3) train a new and larger student model with a combination of labeled images and inferred pseudo labels. These three steps are repeated several times by treating the student as a teacher to relabel the unlabeled data and consequently train a new student. We demonstrate that our framework can classify both video capsule endoscopy (VCE) and standard endoscopy images. Our results indicate that our teacher-student framework can significantly increase the performance compared to traditional supervised-learning-based models, i.e., an overall increase in the F_1-score of 4.7% for the Kvasir-Capsule VCE dataset and 3.2% for the HyperKvasir colonoscopy dataset. We believe that our framework can use more of the data collected at hospitals without the need for expert labels, contributing to overall better models for medical multimedia systems for automatic disease detection.

Published: In the Proceedings of International Symposium on Computer-Based Medical Systems (CBMS). DOI: https://doi.org/10.1109/CBMS52027.2021.00087

Candidate contributions: Vajira contributed to the conception and designing of the study presented in this paper. He contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective I, Sub-objective III

[Article not attached due to copyright]

A.20 Paper XX - Using Preprocessing as a Tool in Medical Image Detection

Authors: Mathias Kirkerød, **Vajira Thambawita**, Michael Riegler, Pål Halvorsen

Abstract: In this paper, we describe our approach to gastrointestinal disease classification for the medico task at MediaEval 2018. We propose multiple ways to inpaint problematic areas in the test and training set to help with classification. We discuss the effect that preprocessing does to the input data with respect to removing regions with sparse information. We also discuss how preprocessing affects the training and evaluation of a dataset that is limited in size. We will also compare the different inpainting methods with transfer learning using a convolutional neural network.

Published: In the Proceedings of MediaEval 2020. DOI: http://ceur-ws.org/Vol-2283/

Candidate contributions: Vajira contributed to the conception and design of the study discussed in this manuscript. He guided the first author (master student) of this manuscript and contributed to analyzing the results of this study. Vajira contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective III, Sub-objective IV

A.21 Paper XXI - Unsupervised Preprocessing to Improve Generalisation for Medical Image Classification

Authors: Mathias Kirkerød, Rune Johan Borgli, **Vajira Thambawita**, Steven Hicks, Michael Alexander Riegler, Pål Halvorsen

Abstract: Automated disease detection in videos and images from the gastrointestinal (GI) tract has received much attention in the last years. However, the quality of image data is often reduced due to overlays of text and positional data. In this paper, we present different methods of preprocessing such images and we describe our approach to GI disease classification for the Kvasir v2 dataset. We propose multiple approaches to inpaint problematic areas in the images to improve the anomaly classification, and we discuss the effect that such preprocessing does to the input data. In short, our experiments show that the proposed methods improve the Matthews correlation coefficient by approximately 7% in terms of better classification of GI anomalies.

Published: In proceedings of 13th International Symposium on Medical Information and Communication Technology (ISMICT), 2019. DOI: https://doi.org/10.1109/ISMICT.2019.8743979

Candidate contributions: Vajira contributed to the conception, design of this study, and analysis of the results of this manuscript. Vajira contributed to drafting and revising this extended version of the working notepaper: "Using preprocessing as a tool in medical image detection".

Thesis objectives: Sub-objective III, Sub-objective IV

[Article not attached due to copyright]

A.22 Paper XXII - GANEx: A Complete Pipeline of Training, Inference and Benchmarking GAN Experiments

Authors: Vajira Thambawita, Hugo Lewi Hammer, Michael Riegler, Pål Halvorsen

Abstract: Deep learning (DL) is one of the standard methods in the field of multimedia research to perform data classification, detection, segmentation and generation. Within DL, generative adversarial networks (GANs) represents a new and highly popular branch of methods. GANs have the capability to generate, from random noise or conditional input, new data realizations within the dataset population. While generation is popular and highly useful in itself, GANs can also be useful to improve supervised DL. GAN-based approaches can, for example, perform segmentation or create synthetic data for training other DL models. The latter one is especially interesting in domains where not much training data exists such as medical multimedia. In this respect, performing a series of experiments involving GANs can be very time consuming due to the lack of tools that support the whole pipeline such as structured training, testing and tracking of different architectures and configurations. Moreover, the success of generative models is highly dependent on hyper-parameter optimization and statistical analysis in the design and fine-tuning stages. In this paper, we present a new tool called GANEx for making the whole pipeline of training, inference and benchmarking GANs faster, more efficient and more structured. The tool consists of a special library called FastGAN which allows designing generative models very fast. Moreover, GANEx has a graphical user interface to support structured experimenting, quick hyper-parameter configurations and output analysis. The presented tool is not limited to a specific DL framework and can be therefore even used to compare the performance of cross frameworks.

Published: In proceedings of International Conference on Content-Based Multimedia Indexing (CBMI), 2019. DOI: https://doi.org/10.1109/CBMI.2019.8877387

Candidate contributions: Vajira contributed to the conception and the design of this manuscript. He implemented a new GUI-based tool named GANEx to train, do

inference, and evaluate Generative Adversarial Networks (GANs) using pre-defined GAN implementations. Vajira has published this tool in a Github repository (https://github.com/v to use in the research community who need to train GANs without coding. He contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective IV

[Article not attached due to copyright]

A.23 Paper XXIII - Vid2Pix - A Framework for Generating High-Quality Synthetic Videos

Authors: Oda O. Nedrejord, **Vajira Thambawita**, Steven A. Hicks, Pål Halvorsen, Michael A. Riegler

Abstract: Data is arguably the most important resource today as it fuels the algorithms powering services we use every day. However, in fields like medicine, publicly available datasets are few, and labeling medical datasets require tedious efforts from trained specialists. Generated synthetic data can be to future successful healthcare clinical intelligence. Here, we present a GAN-based video generator demonstrating promising results.

Published: In proceedings of IEEE International Symposium on Multimedia (ISM), 2020. DOI: https://doi.org/10.1109/ISM.2020.00010

Candidate contributions: Vajira contributed to designing and implementing the theoretical models discussed in this paper. He contributed to evaluating the results (generated synthetic data) critically using dense optical flow calculations which can be used to identify temporal feature differences between frames in a video. Vajira also contributed to drafting and revising this manuscript.

Thesis objectives: Sub-objective IV

[Article not attached due to copyright]

A.24 Paper XXIV - DeepFake Electrocardiograms: the Beginning of the End for Privacy Issues in Medicine

Authors: **Vajira Thambawita**, Jonas L. Isaksen, Steven A. Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Inga Strümke, Hugo L. Hammer, Molly Maleckar, Pål Halvorsen, Michael A. Riegler, Jørgen K. Kanters

Abstract: Recent global developments underscore the prominent role big data have in modern medical science. Privacy issues are a prevalent problem for collecting and sharing data between researchers. Synthetic data generated to represent real data carrying similar information and distribution may alleviate the privacy issue. In this study, we present generative adversarial networks (GANs) capable of generating realistic synthetic DeepFake 12-lead 10-sec electrocardiograms (ECGs). We have developed and compare two methods, WaveGAN* and Pulse2Pulse GAN. We trained the GANs with 7,233 real normal ECG to produce 121,977 DeepFake normal ECGs. By verifying the ECGs using a commercial ECG interpretation program (MUSE 12SL, GE Healthcare), we demonstrate that the Pulse2Pulse GAN was superior to the WaveGAN to produce realistic ECGs. ECG intervals and amplitudes were similar between the DeepFake and real ECGs. These synthetic ECGs are fully anonymous and cannot be referred to any individual, hence they may be used freely. The synthetic dataset will be available as open access for researchers at OSF.io and the DeepFake generator available at the Python Package Index (PyPI) for generating synthetic ECGs. In conclusion, we were able to generate realistic synthetic ECGs using adversarial neural networks on normal ECGs from two population studies, i.e., there by addressing the relevant privacy issues in medical datasets.

Published: Submitted for publication, Preprint is available at medRxiv.
DOI: https://doi.org/10.1101/2021.04.27.21256189

Candidate contributions: Vajira contributed to the conception and design of the deepfake ECG generation study. He implemented a novel GAN architecture named Pulse2pulse that can generate realistic synthetic ECGs with the properties of real

"Normal" ECGs. Vajira conducted all GAN experiments and evaluated using MUSE reports (ECG evaluation reports generated from a real system using in hospitals). Vajira published his work on GitHub to make it reproducible for other ECG datasets. He generated and published the largest synthetic ECG dataset (around 120,000 ECGs) as a replacement to a restricted real ECG dataset. He contributed to drafting the manuscript and revising it.

Thesis objectives: Sub-objective II, Sub-objective IV

A.25 Paper XXV - SinGAN-Seg: Synthetic Training Data Generation for Medical Image Segmentation

Authors: **Vajira Thambawita**, Pegah Salehi, Sajad Amouei Sheshkal, Steven A. Hicks, Hugo L. Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen,Michael A. Riegler

Abstract: Processing medical data to find abnormalities is a time-consuming and costly task, requiring tremendous efforts from medical experts. Therefore, artificial intelligence (AI) has become a popular tool for the automatic processing of medical data, acting as a supportive tool for doctors. AI tools highly depend on data for training the models. However, there are several constraints to access to large amounts of medical data to train machine learning algorithms in the medical domain, e.g., due to privacy concerns and the costly, time-consuming medical data annotation process.

To address this, in this paper we present a novel synthetic data generation pipeline called SinGAN-Seg to produce synthetic medical data with the corresponding annotated ground truth masks. We show that these synthetic data generation pipelines can be used as an alternative to bypass privacy concerns and as an alternative way to produce artificial segmentation datasets with corresponding ground truth masks to avoid the tedious medical data annotation process. As a proof of concept, we used an open polyp segmentation dataset. By training UNet++ using both real polyp segmentation dataset and the corresponding synthetic dataset generated from the SinGAN-Seg pipeline, we show that the synthetic data can achieve a very close performance to the real data when the real segmentation datasets are large enough. In addition, we show that synthetic data generated from the SinGAN-Seg pipeline improving the performance of segmentation algorithms when the training dataset is very small. Since our SinGAN-Seg pipeline is applicable for any medical dataset, this pipeline can be used with any other segmentation datasets.

Published: Submitted for publication, Preprint is available at arxiv.
DOI: https://doi.org/10.48550/arXiv.2107.00471

Candidate contributions: Vajira contributed to the conception and designing of this study. He developed the whole source code and tested the initial experiments. Moreover, he evaluated the performance of the model introduced in this paper critically by conducting several experiments. He has created and published the synthetic dataset, the corresponding generative models as a PyPI package, and the GitHub repository. Vajira also contributed to the drafting and revising of the article.

Thesis objectives: Sub-objectives I, Sub-objective II, Sub-objective III, Sub-objective IV

A.26 Paper XXVI - Generative Adversarial Networks For Creating Realistic Artificial Colon Polyp Images

Authors: Vajira Thambawita, Inga Strümke, Steven Hicks, Michael A. Riegler, Pål Halvorsen, Sravanthi Parasa

Abstract: Artificial intelligence is increasingly used to detect and classify colon polyps. However, small datasets are a major obstacle, especially for supervised machine learning. Data collection is challenging, and synthetic data generation, using models such as generative adversarial networks (GANs), may help overcome this hurdle. To determine the clinical utility of synthesized images, we generate images containing colon polyps, and eight endoscopists assess their anatomical correctness.

Published: GIE, DDW Abstract Issue, 2021. DOI: https://doi.org/10.1016/j.gie.2021.03.431

Candidate contributions: Vajira contributed to the conception and design of this study. He conducted all the experiments of this research and introduced a novel method to generate synthetic polyp images using a real clean colon image. Vajira evaluated the study critically with experts (doctors) of the domain using a questionnaire. He contributed to drafting the manuscript and revising it.

Thesis objectives: Sub-objective I, Sub-objective II, Sub-objective III, Sub-objective IV

A.27 Paper XXVII - Identification of Spermatozoa by Unsupervised Learning from Video Data

Authors: Michael A. Riegler, Trine B. Haugen, Mette Haug Stensen, Oliwia Witczak, Hugo L. Hammer, Pål Halvorsen, Michael A. Riegler

Abstract: Identification of individual sperm is essential to assess a given sperm sample's motility behaviour. Existing computer-aided systems need training data based on annotations by professionals, which is resource demanding. On the other hand, data analysed by unsupervised machine learning algorithms can improve supervised algorithms that are more stable for clinical applications. Therefore, unsupervised sperm identification can improve computer-aided sperm analysis systems predicting different aspects of sperm samples. Other possible applications are assessing kinematics and counting of spermatozoa. Generative adversarial networks (GANs) have become common AI methods to process data in an unsupervised way. Based on single image frames extracted from videos, a GAN (SinGAN) can be trained to determine and track locations of sperms by translating the real images into localization paintings. The resulting model showed the potential of identifying the presence of sperms without any prior knowledge about data. Visual comparisons of localization paintings to real sperm images show that inverse training of SinGANs can track sperms. Converting colour frames into grayscale frames and using grayscale synthetic sperm-like frames showed the best visual quality of generated localization paintings of sperm frames.

Published: Oxford Academic Press, European Society of Human Reproduction and Embryology (Eshre), 2021. DOI: https://doi.org/10.1093/humrep/deab130.028

Candidate contributions: Vajira contributed to the design of this concept and he developed all the models and the corresponding experiments for tracking sperms using a modified SinGAN generative model. He evaluated results collected from the experiments and published them for public use (https://vlbthambawita.github.io/singan-sperm/). Vajira contributed to drafting and revising the manuscript.

Thesis objectives: Sub-objective IV

Citation: V Thambawita, T B Haugen, M H Stensen, O Witczak, H L Hammer, P Halvorsen, M A Riegler, P–029 Identification of spermatozoa by unsupervised learning from video data, Human Reproduction, Volume 36, Issue Supplement_1, July 2021, deab130.028

A.28 Paper XXVIII - DeepSynthBody: the Beginning of the End for Data Deficiency in Medicine

Authors: Vajira Thambawita, Steven A. Hicks, Jonas Isaksen, Mette Haug Stensen, Trine B. Haugen, Jørgen Kanters, Sravanthi Parasa, Thomas de Lange, Håvard D. Johansen, Dag Johansen, Hugo L. Hammer, Pål Halvorsen, Michael A. Riegler

Abstract: Limited access to medical data is a barrier on developing new and efficient machine learning solutions in medicine such as computer-aided diagnosis, risk assessments, predicting optimal treatments and home-based personal healthcare systems. This paper presents DeepSynthBody: a novel framework that overcomes some of the inherent restrictions and limitations of medical data by using deep generative adversarial networks to produce synthetic data with characteristics similar to the real data, so-called DeepSynth (deep synthetic) data. We show that DeepSynthBody can address two key issues commonly associated with medical data, namely privacy concerns (as a result of data protection rules and regulations) and the high costs of annotations. To demonstrate the full pipeline of applying DeepSynthBody concepts and user-friendly functionalities, we also describe a synthetic medical dataset generated and published using our framework. DeepSynthBody opens a new era of machine learning applications in medicine with a synthetic model of the human body.

Published: In proceedings of the International Conference on Applied Artificial Intelligence(ICAPAI), 2021. DOI: https://doi.org/10.1109/ICAPAI49758.2021.9462062

Candidate contributions: Vajira came with this idea and he contributed to the conception and design. He implemented the whole pipeline and did all the implementations. Vajira has developed this DeepSynthBody as a framework and this is the core of his thesis storyline. He performed several experiments to generate gastrointestinal tract images using GANs as a proof of concept of this framework. Final outcome of this study is available online to end-users of this framework at https://deepsynthbody.org/ which was developed by him. Vajira contributed to drafting the manuscript and revising it.

Thesis objectives: Sub-objective II, Sub-objective IV

[Article not attached due to copyright]