Transparency in Medical Artificial Intelligence Systems

Steven Hicks

October 29, 2021

Preface

This Ph.D. thesis was written as part of the Engineering Science Faculty of Technology, Art and Design Ph.D. program at Oslo Metropolitan University under the supervision of Michael A. Riegler, Pål Halvorsen, and Hugo L. Hammer. The presented research represents three years of work, half of which was done from home office due to the COVID-19 pandemic. For these three years, I was employed at the Department of Holistic Systems in Simula Metropolitan Center for Digital Engineering (SimulaMet), which provided a thriving research environment and the financial support needed to fuel this research. Starting a Ph.D. was never my plan. But now that it is done, I am happy I did.

Steven Hicks
Oslo, Norway

Abstract

Applying machine learning to problems in medicine is a rapidly growing trend in nearly all areas of healthcare. The immense performance attained by using deep learning on tasks like image and time series analysis can profoundly impact how computers are used in hospitals or clinics. There is a lot to gain in developing these systems, both monetary and societal, where deep neural network-based models may someday be in charge of monitoring our health. However, despite the massive responsibility that we give these models, the approach of developing and evaluating these methods is often not clear. Medical artificial intelligence (AI) research usually has imprecise method descriptions, private data, closed-source implementations, and incomplete evaluations. This thesis studies at how AI can be used in different areas within medicine, where a primary focus is to look at the current state of transparency within medical AI systems research and aims to contribute to a more open and public research community. To achieve this, we collected and published several medical datasets, developed several AI models in various medical domains, performed an assortment of different experiments to validate the collected datasets, organized many competitions on medical AI applications, and examined adequate model evaluation methods. The work was done across four fields of medicine to get a thorough understanding of how transparent AI can be applied to different medical domains, which includes cardiology, assisted reproductive technology, gastroenterology, and mental health.

Abstract - Norwegian

Å bruke maskinlæring på problemer innen medisin er en raskt voksende trend i nesten alle områder av helsevesenet. Den enorme ytelsen som oppnås ved å bruke dyp læring på oppgaver som bilde- og tidsserieanalyse kan ha stor innvirkning på hvordan datamaskiner brukes på sykehus eller klinikker. Det er mye å vinne på å utvikle disse systemene, både monetære og samfunnsmessige, der dype nevrale nettverksbaserte modeller en dag kan ha ansvaret for å overvåke helsen vår. Til tross for det enorme ansvaret vi gir disse modellene, er tilnærmingen for å utvikle og evaluere disse metodene ofte ikke klar. Medisinsk kunstig intelligens (AI)-forskning har vanligvis upresise metodebeskrivelser, private data, implementeringer med lukket kilde og ufullstendige evalueringer. Denne oppgaven studerer hvordan AI kan brukes på ulike områder innen medisin, hvor et primært fokus er å se på dagens åpenhet innen medisinsk AI-systemforskning og har som mål å bidra til et mer åpent og offentlig forskningsmiljø. For å oppnå dette, samlet og publiserte vi flere medisinske datasett, utviklet flere AI-modeller i ulike medisinske domener, utførte et utvalg av forskjellige eksperimenter for å validere de innsamlede datasettene, organiserte mange konkurranser om medisinske AI-applikasjoner og undersøkte adekvate modellevalueringsmetoder. Arbeidet ble utført på tvers av fire felt av medisin for å få en grundig forståelse av hvordan transparent AI kan brukes på ulike medisinske domener, som inkluderer kardiologi, assistert reproduktiv teknologi, gastroenterologi og mental helse.

Acknowledgements

First, I want to thank my three supervisors Michael Riegler, Pål Halvorsen, and Hugo Hammer. Thank you for always having my back and pushing me to achieve more than I thought I could. Michael told me that the relationship between a supervisor and student was more liken to a parent and child than a teacher and student. I did not understand this at the time, but now I can genuinely say that he was right. It is difficult to express how grateful I am in such a small area of text, but this work would truly not be possible without your help.

I would also like to thank my fellow HOST Ph.D. students Vajira, Hanna, Pia, Debesh, and Daniel, who made my work life a lot more fun, even though I may have distracted them with my off-topic discussions sometimes. I will never forget our trips to San Diego, Dublin, and France, which were a lot of fun and would not be the same without great company. I hope to someday visit you, Vajira, in Sri Lanka, where I too can try all these types of bananas you were talking about. I would also like to extend this to my newer colleagues that started during the pandemic, namely Inga, Andrea, Thu, Zohaib, Steffen, Cise, Pegah, Sajad, and Malek. I may not have met all of you in person yet, but our video calls have always been nice. I also need to thank my master's students, who have all been great and made me very proud to be their supervisor.

Lastly, I would like to thank my mom for allowing me to escape home during the pandemic and my dad for always being there to talk when needed. Also, I need to thank my cat and dog for being the best fluffy animals in existence.

Contents

Chapter 1

Introduction

Artificial intelligence (AI) has in the last few years shown immense progress in advancing the state-of-the-art in areas such as computer vision [158], robotics [82], and natural language processing (NLP) [155]. As a research field, AI has been around since the mid-1950s [156, 124]. Still, it is not until recently that we have obtained the amount of data and the computational power required to truly see the potential in the algorithms developed during its early years [139]. These days, when we refer to AI, we often mean the subfield machine learning. Figure 1.1 shows the Google search trend for the terms *artificial intelligence* and *machine learning*, where we see that both terms have exploded in popularity over the last ten years. Machine learning encompasses deep learning, which is currently the most popular family of algorithms in AI research and practice. Deep learning uses *deep* neural networks to automatically extract features from data to perform tasks like regression or classification. These methods have shown an extreme range in the problems they can be applied to, stretching from self-driving cars to automatically detecting and diagnosing different types of diseases.

1.1 Motivation and Background

AI-based systems, or AI systems, are slowly making their way into clinical practice [118]. At the same time, the research that goes into developing these systems is often clouded by closed-source implementations, private data, lackluster evaluations, and non-reproducible results [140]. We are on the verge of living in a world where machines will determine what medicine we should take, perform surgeries on us, or diagnose us with a specific illness or

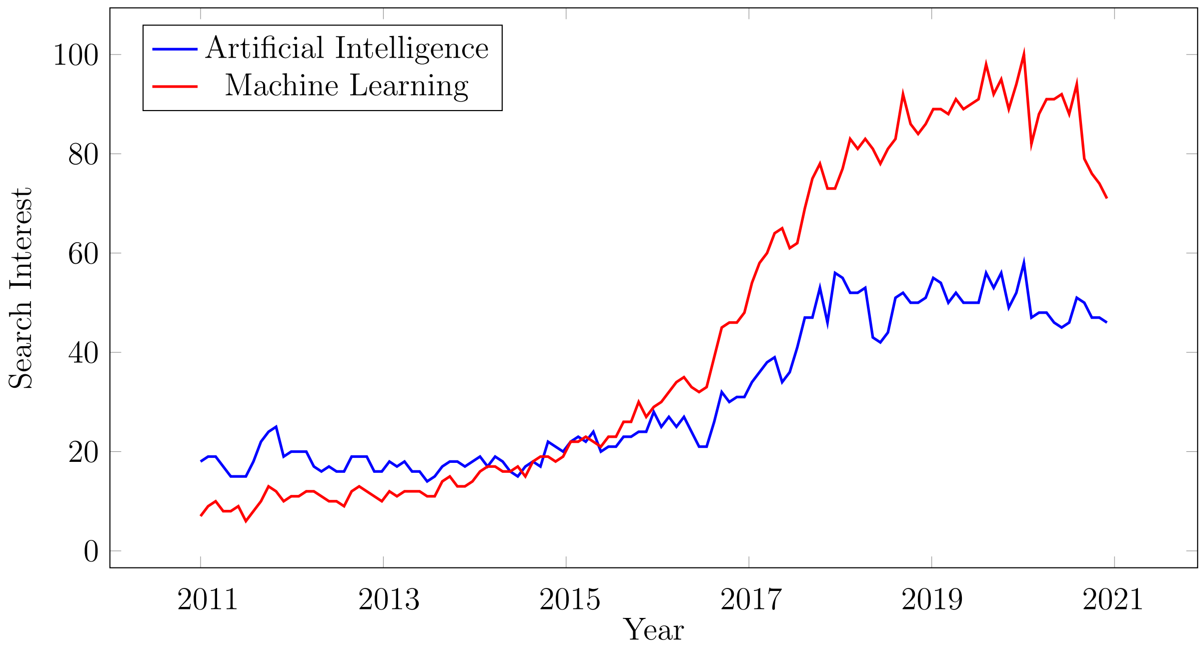


Figure 1.1: The Google search trend over the last ten years for the terms *artificial intelligence* and *machine learning*.

disease. As these algorithms obtain more authority over our everyday lives, transparency in how they were built and how they work becomes essential. In this context, transparency can be seen as an important principle of ethics, where Dr. David Leslie defines it as a combination of two meanings [85]; (*i*) transparency in the sense of a clear, see-through object and (*ii*) transparency as in a justified and explained process open for inspection and free from secrets. The first statement relates to explaining the internal processes and output of the complex models currently being used today. The second statement is tied to open and publicly available data, implementations, and evaluations used to develop and deploy AI systems in practice. Modern AI systems are complex and made up of multiple parts, where each can have a substantial effect on the final prediction. Therefore, it is crucial that each part of the system, from data collection to final evaluation, is transparent for data scientists, engineers, health care professionals, and patients.

1.2 A Definition of Terms

This thesis contains several terms and concepts that are important to understand in order to fully comprehend its content. The following gives a brief description of some terms that will be continuously mentioned throughout this thesis.

1.2.1 Basic Artificial Intelligence Terms

The following describes some common terms relating to basic AI concepts.

Artificial Intelligence Artificial intelligence (AI) refers to the simulated human intelligence demonstrated by machines. Applications include computer vision, natural language processing, robotics, and reasoning systems to name a few.

Machine Learning Machine learning is an application of AI that uses algorithms to automatically learn specific tasks using data. This can either be done using labeled data (supervised learning) or unlabeled data (unsupervised learning). Machine learning is currently the most popular application of AI both in research and in practice.

Deep Learning Deep learning is class of algorithms within machine learning that is based on deep neural networks, meaning neural networks that consists of several hidden layers.

1.2.2 Transparency Terms

Terms such as explanations, transparency, and interpretability are often used interchangeably, and although the terms are quite similar, they have different meanings. An interpretation of a prediction does not necessarily lead to more transparency, neither does it make the model more explainable. For full transparency, we need to know the whole pipeline used to develop an AI system. This aspect of AI has started to get some attention, and we expect it to become more important as these algorithms begin making their way into production. In this section, we discuss the meaning of the words interpretability, explainability, and transparency in the context of medical AI.

Interpretability Interpretation of a model refers to the ability to infer how changes to the input will affect the output. For example, if we train a model to predict whether or not an image contains a polyp, we expect the output to be *polyp* when an image of a polyp is passed through the model. If we modify the image by replacing the polyp with a black box, we expect the output to change. By this definition, one does not necessarily know why a model is making these predictions. Still, we can determine the relationship between the polyp in the image and the output of the model.

Explainability Model explainability concerns the ability to understand why a model is making a specific prediction. Using the example of the polyp prediction model again, if the model predicts that a image contains a polyp, explainability would also tell us why this image contains a polyp. This could be visual description, like the presence of certain features visualized through feature maps. It could also be through similarity scores, like showing similar images that also contains polyps. Model explanations should not only be inheritable through cause and effect, but should also explain why this prediction was made.

Transparency Model transparency not only looks at the model itself but also takes the entire system into account, from the data used for training to final evaluation. For an AI system to be truly transparent, we need to know everything about it. This includes information about the data used to train it, how the model was implemented (which frameworks and libraries), which methods were used for evaluation, where the models fail (failure analysis), which parts of the model contribute to the prediction (ablation study), and a thorough description of how the model was trained.

1.3 Problem Statement

The aim of this thesis is to research how medical AI systems can be developed and be more transparent across data use, model development, and model evaluation. With this in mind, we present the primary research research question:

> *Can medical AI-based systems be made more transparent?*

This research question is rooted in a general lack of transparency found in medical AI research [140], and will be the foundation for the work presented in this thesis. To make the main question more tangible, we break it down into three objectives that, in the end, will give us enough information to provide a conclusion on our research question.

Objective 1 Obtain a better understanding of the role and challenges of data in medical AI systems by collecting, preparing, and publishing datasets from different medical domains in close collaboration with experts. Each dataset should be collected with the purpose of solving real medical problems, and all related materials should be made public for other researchers to extend and improve. Public data is an essential aspect of any transparent AI system as it will determine the strengths and weaknesses of the underlying machine learning model.

Objective 2 Research and develop efficient methods for medical data analysis that aim to solve real medical problems defined by experts. The methods should cover a wide variety of different approaches, including those based on classical machine learning and more modern approaches like deep learning. Furthermore, explore different methods surrounding the training of the model, such as data preprocessing, data augmentation, and multimodal analysis. Transparency in the methods used for analysis is important for reproducibility and comparability.

Objective 3 Evaluate the results using different methods to assess an AI system before being deployed to the real world, such as using AI explanation methods to interpret model predictions, measuring the quantitative performance using standard evaluation metrics, and benchmarking methods for reproducibility and comparability. Transparency in the evaluation and interpretation of the results is crucial for understanding how the model may perform in a production environment.

1.4 Scope and Limitations

The work for this Ph.D. project started with the development of efficient ML models for medical applications with the question of how explainable AI could help medical doctors gain more trust and confidence in black-box models such as deep neural networks. After some time researching this problem and working closely with the doctors, we realized that faith and trust do not come from explainability alone and that there is a much larger problem in medical AI research, namely transparency. Therefore, we switched the main focus of this thesis from only explainable AI to look at the broader picture about how the process of developing and evaluating an AI-based system can be more transparent.

We believe our findings are relevant for various medical areas, but the research is primarily limited to the following four medical fields; gastroenterology, assisted reproductive technology, cardiology, and mental health. This limitation is rooted mainly in the collaboration and relationships we have with hospitals and clinics. Although some of the methods presented in this thesis are specialized in their application, the underlying principles and theory still apply to other areas as well. Furthermore, the data collected with our collaborating parties is limited to the equipment available at the point of collection. This includes the microscopes used to capture video footage of human semen, the endoscopes used to collect image data from colonoscopies, the wristbands used to collect activity data, and the various other equipment used for data collection.

Clinical trials are a time-consuming and costly process, with many unknowns when it comes to the evaluation of AI-based systems. None of the systems described in this thesis have been deployed or are currently being used in clinics or hospitals. Having these systems deployed in the real world requires legal approvals making sure the systems follow all regulations, such as meeting the requirements for data collection, using approved equipment, and several other factors. We hope to one day integrate the research into the existing workflow of medical doctors, but for the purpose of this Ph.D. project, we focus purely on the development and evaluation of transparent AI systems from a research point of view.

1.5 Research Methods

Research can be performed in a many different of ways. For this Ph.D. project, the research mainly consists of building prototypes and implementing AI-based systems that perform a given task. In a more formal context, we generally follow the Association for Computing Machinerys (ACMs) research methodology. In 1989, the ACM Education Board assigned a task force to compile the core fundamentals of computer science and computer engineering into a detailed report [25]. The report describes the discipline of computing as being split between three paradigms; (*i*) theory, (*ii*) abstraction, and (*iii*) design. The research done for this thesis covers all three of the explained paradigms. In the case of this work, the theory paradigm relates to defining a medical problem together with domain experts and collecting data that can be used to solve it. The abstraction paradigm relates to developing the algorithms devised by the previous paradigm and interpreting the results together with the domain experts. The design paradigm relates to the development of the full system, which includes setting the requirements, stating the specifications, implement the system, and having medical doctors test it. In the following, we explain in more detail how our research falls under each paradigm.

1.5.1 Theory

The ”theory” paradigm is rooted in mathematics and relates to developing a coherent and valid theory. The report describes this phase as being made up of four steps, which are described as follows; (*i*) characterize the objects of study (definition), (*ii*) hypothesize the possible relationships among them (theorem), (*iii*) determine whether the relationships are true (proof), and (*iv*) interpret the results.

This paradigm is reflected in collaboration with medical doctors in collecting and developing medical datasets aimed at solving clinical relevant problems in medicine. For example, we hypothesized that we could use videos of human semen to automatically predict the quality of a given semen sample. We collected a dataset and created machine learning models to predict sperm quality in terms of the motility and morphology of the sperm. The results were analyzed and compared against simple baselines to verify that our hypothesis was correct. These steps were repeated several times as each published dataset was accompanied by a set of baseline experiments as a means of technical validation.

1.5.2 Abstraction

The ”abstraction” paradigm is rooted in the experimental scientific method and relates the investigation of a phenomenon, e.g., hypothesis. The report describes this phase as a process consisting of four steps, which are described as follows; (*i*) form a hypothesis, (*ii*) construct a model and make a prediction, (*iii*) design an experiment and collect data, and (*iv*) analyze results.

This paradigm is supported by the numerous experiments performed on solving different problems in medicine across multiple different fields. These experiments started with a hypothesis on how a problem may be solved, for which we developed a machine learning-based model to validate the hypothesis and evaluated the results. For example, we hypothesized that we could generate synthetic images of colorectal polyps that could replace real data and still achieve desirable results. From this hypothesis, we constructed a generative adversarial network (GAN) and generated a large set of synthetic polyp images and performed several experiments that showed that the synthetic data could be used in place of actual polyp data.

1.5.3 Design

The ”design” paradigm is closely related to engineering and relates to the construction of a system, e.g., software, hardware, etc. The report describes this phase as a process consisting of four steps, which are described as follows; (*i*) state requirements, (*ii*) state specifications, (*iii*) design and implement the system, and (*iv*) test the system.

This paradigm is supported by the AI systems developed to perform the experiments for this thesis. The development of each system consists of several steps, from data preparation to performance evaluation, which was implemented and tested using different configurations.

1.6 Main Contributions

The general lack of transparency in medical AI applications motivated us to research this area in more detail, where this thesis is a culmination of our work. Through the three years of this Ph.D. project, we released ten open datasets [17, 39, 50, 58, 67, 67, 69, 137, 143, 151], organized seven medical machine learning challenges [111, 52, 59, 59, 54, 71, 55] (one more ongoing), published many papers on applying AI to different medical problems [60, 53, 32, 154, 120, 161], and looked at how more transparency, or lack thereof, can impact medical AI research [61, 62, 57, 121]. An overview of the papers and how they relate to each part of a transparent AI system is shown in Figure 1.2. The work was primarily performed over four branches of medicine, namely gastroenterology, cardiology, assisted reproductive technology, and mental health. The reason for spreading the research across several medical fields was to get a better understanding of how transparent AI systems can be deployed in different environments. The main contributions of this thesis are supported by publications in top-tier conferences or journals. In the following, we detail the contributions in relation to the research question and research objectives defined in Section 1.3.

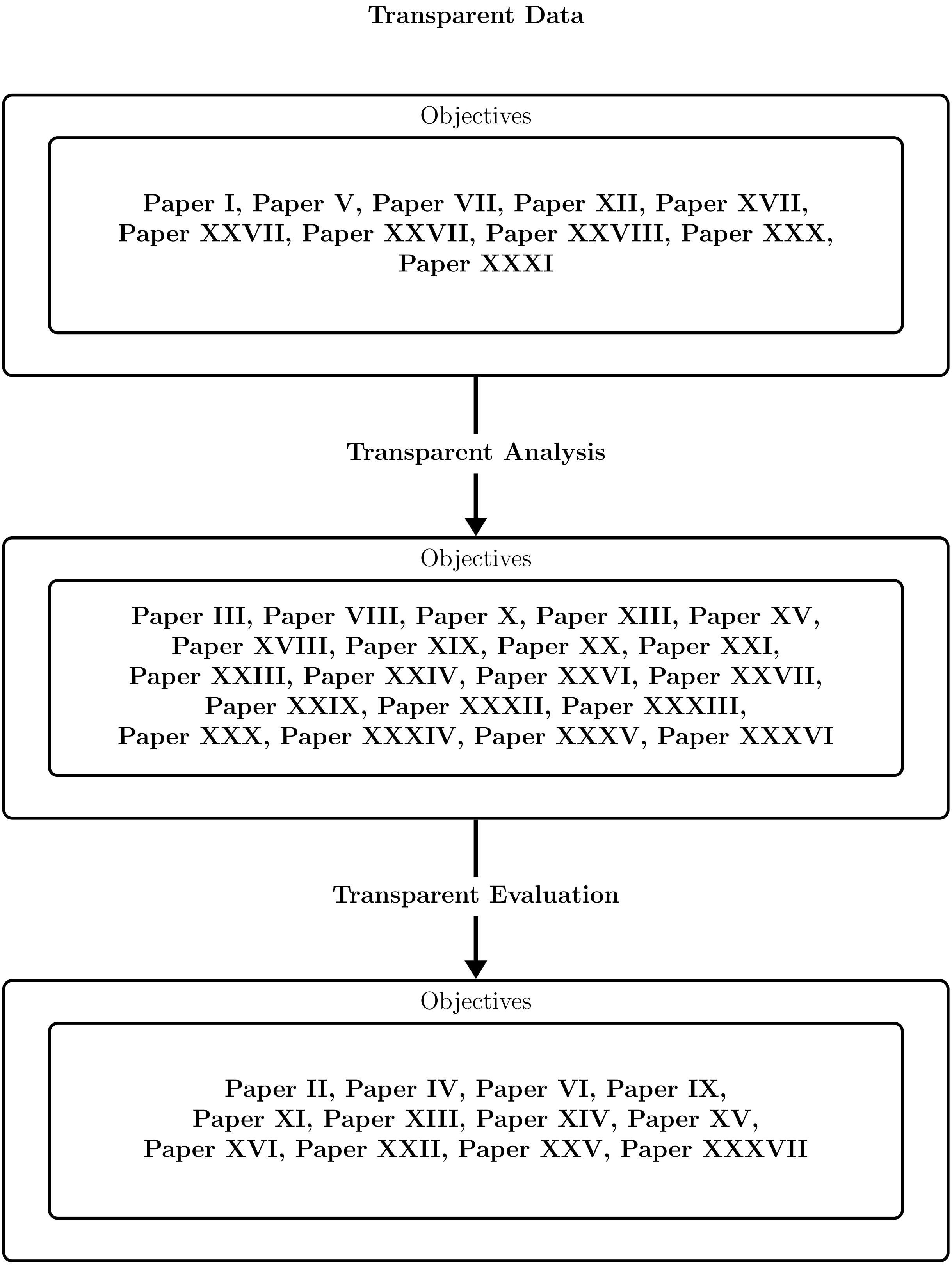

Transparent Data
Objectives
Paper I, Paper V, Paper VII, Paper XII, Paper XVII,
Paper XXVII, Paper XXVII, Paper XXVIII, Paper XXX,
Paper XXXI
Transparent Analysis
Objectives
Paper III, Paper VIII, Paper X, Paper XIII, Paper XV,
Paper XVIII, Paper XIX, Paper XX, Paper XXI,
Paper XXIII, Paper XXIV, Paper XXVI, Paper XXVII,
Paper XXIX, Paper XXXII, Paper XXXIII,
Paper XXX, Paper XXXIV, Paper XXXV, Paper XXXVI
Transparent Evaluation
Objectives
Paper II, Paper IV, Paper VI, Paper IX,
Paper XI, Paper XIII, Paper XIV, Paper XV,
Paper XVI, Paper XXII, Paper XXV, Paper XXXVII

Figure 1.2: An overview of all published papers and how they relate to each part of a transparent AI system.

Contributions to Objective 1 Objective one is supported by the collection and publication of several medical datasets in the field of gastroenterology [17, 137], assisted reproductive technology [50], mental health [67, 58], and cardiology [152]. Each dataset was developed in collaboration with health experts within each respective field and is made publicly available and free to use for research and educational purposes. The datasets were used as a basis for most of the research presented in this thesis and have been used to organize several challenges and benchmarks [56, 52, 59, 54, 71, 55]. Each dataset is accompanied by a research article published in a top-tier conference or journal, where we also include a set of benchmark experiments and possible research directions. The contribution to transparency in medical AI systems comes from the open and publicly available datasets published under non-restrictive licenses that can be used to develop and benchmark machine learning models. We also include information on how the data was collected and verified by the medical experts, including details on the equipment used to collect the data.

Contributions to Objective 2 Objective two is supported by our work on developing and training machine learning models to support medical doctors and clinicians in performing different tasks within medicine. This includes methods for automatically determining the quality of a given semen sample [60], predicting the sex, waves, and intervals of a standard 12-lead ECG [61], detecting disease and other findings in the gastrointestinal (GI) tract [98, 75], and recognizing disorders such as attention deficit hyperactivity disorder (ADHD) or schizophrenia from activity data [58, 67, 37]. The contribution to transparency comes from the implementation descriptions, where we detail how the models are trained, which hyperparameters are used, and usually include an open implementation that is freely available online.

Contributions to Objective 3 Objective three is supported by our work on through evaluation and post validation of machine learning models after training. This work includes using explainable AI to discover new features related to the relationship between sex and electrocardiogram (ECG) signals [61], exploring how evaluation metrics may be used to give an incomplete view of a models predictive performance [62, 121], and the organization of several machine learning challenges for benchmarking and comparability purposes [56, 55, 52, 59, 54, 71, 111]. Openness and accurate reporting of evaluation methods is a key attribute of transparency within the evaluation of a AI system. Therefore, we aimed to show both the negative and positive sides of our developed systems, reporting all relevant metrics and measuring performance against several benchmarks.

Additional Contributions We also contributed to areas that fall outside the aforementioned research question and objectives. Still, these contributions follow the principles of transparency and helped us better understand the requirements for the medical-based case studies. We researched and developed systems for automatically detecting and clipping events from soccer games [101, 123]. These systems were built on open datasets and gave us experience with analyzing video data using deep learning. We also used model interpretation methods to generate explanations over the time-dimension to explain the predictions of the deep neural network. We looked at how machine learning can be used to predict latency in mobile networks, where we used both traditional machine learning methods and deep learning. Through this work we gained experience in analyzing large time-series data. We collected and published a dataset for performing sentiment analysis on disaster-related images and are currently organizing a challenge using this dataset [49, 168]. We collected and published a dataset on developing emotional intelligence machines using the video game Super Mario Bros. as an initial use case for building emotional intelligent machines [143]. This work spawned from the idea of incorporating human ethics into machine learning algorithms. We collected and published a dataset for analyzing sports activity in relation to other personal attributes like sleep and weight [151]. We collected and published a dataset containing activity data from everyday tasks like brushing teeth or watching television together with audio recordings [39]. We collected and published a dataset for predicting cloud fractional cover using satellite observations.

The aforementioned objectives were completed to progress towards our overarching research question, which was as follows.

Can medical AI-based systems be made more transparent?

Through our work developing medical AI systems, we gained a better understanding of the problems medical doctors face in their everyday work. Collecting and publishing datasets for objective 1 gave us a better understanding of the intricacies of making medical data public and the potential obstacles that make this difficult. Developing AI systems for different tasks within medicine for objective 2 helped us to recognize the issues and requirements that medical doctors and clinicians face every day. Researching the evaluation of AI systems in medicine and organizing several machine learning challenges for objective 3 gave us the experience of realizing that there is still a lot of research to be conducted in medical AI research, and a good way to accelerate this process is through public events that encourage collaboration, reproducibility, and comparability. The lessons learned from the fulfillment of these objectives is that a medical AI system can be more transparent if both medical doctors and computer scientists have a solid understanding of why transparency in AI research is important.

1.7 Thesis Structure

The structure of this thesis is organized as a paper collection and is primarily split between two distinct parts. The first part is an introduction to our research and aims to provide the reader with an overview of the research area itself and tie the various papers together in one complete story. The second part contains the published papers, where the reader may obtain more details regarding the specific research areas. In the following, we give a short summary of the following chapters.

Chapter 2: Medical Artificial Intelligence Systems This chapter covers the related work of developing AI systems in medicine. This includes a description on the different stages of developing an AI system, how AI systems are currently being used in medicine today, some of the problems of bringing AI research into the clinic or hospital, and a discussion on safe and ethical use of AI.

Chapter 3: Transparent Artificial Intelligence Systems in Medicine This chapter covers our research on developing transparent AI systems within four medical case studies; a GI tract case study, assisted reproductive technology case study, electrocardiogram case study, and mental health case study. The chapter is organized into three sections, each targeting a different aspect of transparent research in AI systems development.

Chapter 4: Conclusion This chapter summarizes the contributions of this thesis and discuss future works in applying transparent machine learning to different areas of medicine.

Chapter 5: Papers and Author's Contributions In this final chapter, we present all core research papers included and discussed in this thesis. We describe the author's contributions for each paper and recount how it contributes to the thesis' overall mission.

Chapter 2

Medical Artificial Intelligence Systems

Ever since the introduction of AlexNet [78] in 2012, AI has been applied to nearly all facets of society [116, 65], medicine being no exception [20]. AlexNet showed that using graphics processing units (GPUs) could greatly accelerate the training and processing time of deep neural networks, making them a good alternative to what we now refer to as traditional machine learning. Now, research is published on how AI systems are being used to aid in diagnoses or automate specific tasks to alleviate part of the workload of medical doctors. Many works show highly promising results, where some systems claim to perform certain tasks better than the trained professionals [45]. However, one should be careful about comparing the results of an AI system against human experts. The predictions of an AI system can contain unforeseen biases, and the metrics used for measuring performance may be skewed in favor of the algorithm. AI systems are not very different from *standard* automation or software systems. Both involve several parts that work together in order to provide the user with some value or perform a given task. What differentiates an AI system from its counterparts is that it aims to mimic the actions of a person by learning to do so automatically from data. This chapter looks at the current state of AI systems used in medicine. This includes a discussion on how AI systems are developed, how AI systems are used in medicine, the process of deploying a medical AI system in practice, and some of the ethical dilemmas that keep most AI systems pure research. The subjects covered in this chapter will establish the groundwork for the subsequent chapters and support the papers published throughout this Ph.D. project.

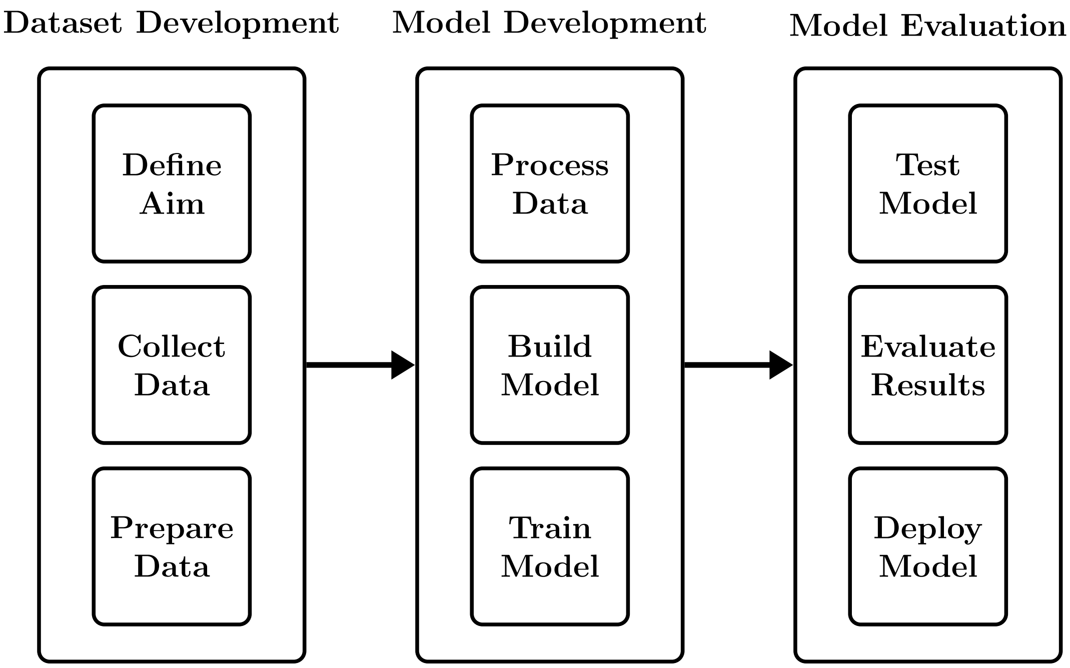


Figure 2.1: This diagram describes a basic machine learning pipeline from data collection to model deployment.

2.1 Artificial Intelligence Systems Development

Developing an AI system can largely be broken into three distinct steps; data collection and preparation, model development and training, and model evaluation and testing. Each step builds on the last and are equally important when developing a safe and high-performing system. In the following, we give a more detailed description of the three steps, which are also showed in Figure 2.1.

Data Collection and Preparation The first step of developing a AI system is defining its purpose. Data should be collected with a specific goal in mind [121]. This could be, for example, collecting magnetic resonance imaging (MRI) images of the human brain to diagnose a specific type of brain disease. Problems should be developed in close collaboration with domain experts, like a neurologist in the aforementioned example. Before a dataset can be used for anything useful, it must first be cleaned and prepared for analysis. This usually consists of removing corrupted samples and organizing the data in a format that makes it easy to use [119]. Depending on the method used for analysis, features may be extracted from the data. This is very common for complex data types like images [108], where complexity is reduced by extracting visual features that represent specific certain characteristics like color distribution [106], coarseness [10], directional [104], and roughness [68].

Model Development and Training After collecting and preparing data for a specific task, the next step is to develop an algorithm that can harness the information contained within and solve the target task. For traditional algorithms like linear regression or decision trees, this can be a simple as applying the respective algorithms using standard hyper-parameters. However, we usually want to tailor the configuration to the task at hand by either manually turning the hyper-parameters or optimizing them using hyperparameter optimization algorithms [165, 36, 90] like random search [14] or Bayesian optimization [131].

Model Evaluation and Testing The last part of the machine learning pipeline is testing and evaluating that the model works as expected. This is usually done by using a set of quantitative evaluation metrics that measure the performance related to the task at hand. This could be, for example, measuring the precision and recall for a classification [62] task or using mean squared error to estimate the error for a regression task [34]. However, basing an evaluation purely on a set of evaluation metrics could yield ill results in a production environment [93]. These metrics are highly dependant on the test dataset, meaning that any scenarios that are not present in the testing dataset will not be reflected in the evaluation metrics [153]. This is why it is important to test the model under certain conditions and use AI explanation methods to ensure that the model works as intended.

2.2 The Role of Artificial Intelligence in Medicine

Data generated by the healthcare industry is astronomical and growing steadily year by year [27]. The amount of data collected far exceeds the ability of any human to accurately analyze and interpret, making a lot of information go unused. At the same time, medical doctors are overwhelmed with work and are losing time used for direct patient care to analysis and administrative tasks. As previously explained, recent machine learning algorithms based on deep neural networks, also called deep learning, have found much success in automatically analyzing large amounts of data. Thus, many see the potential of using deep learning to support doctors in analysis and automate simple mundane administrative tasks [63]. Some AI systems have already made their way into clinics and are currently helping doctors perform surgeries [48] and diagnose different types of disease [81]. There are also several systems that are in the process of being or have been approved by the Food and Drug Administration (FDA) [13]. However, most systems do not make it to this stage and remain pure research. There are several reasons why AI systems fail in production [93, 11]. For example, AI systems can have a hard time adapting to new and unforeseen scenarios [74]. This problem usually stems from a lack of quantity and varied training samples, which translates into the model not understanding how to handle certain situations. This is especially common in medicine where AI have a document track record of not working consistently across different hospitals and clinics [12]. Another issue that AI systems face is that they may learn unintended biases between specific inputs and outputs. An example from the real world is a case where a machine learning model used to predict patient risk determined that black patients were sicker than white patients [102], thus providing black patients with less medical support.

Different medical fields have different needs and requirements regarding the problems they are trying to solve and the data collected. To get a broad overview of how AI systems are integrated into healthcare, we used four medical case studies to focus our research, namely a GI tract case study, a ECG case study, an assisted human reproduction case study, and a mental health case study. There are several reasons why these case studies were selected. First, as our goal was to obtain a broad overview of how transparent AI systems could work within different areas of medicine, selecting case studies that primarily depend on different types of data was paramount. The GI tract case study mostly focuses on analyzing video frames collected from colonoscopies or gastroscopies. The findings generally do not require any temporal information to detect, meaning that analyzing a single video frame at a time is often enough. On the contrary, analysis for the assisted reproductive technology case study is often based on video data and highly dependant on temporal information. For example, if we wish to predict the motility of a given sperm, one needs to know the direction and speed, which a single frame can not determine. The ECG case study uses time-series data to analyze specific intervals and waves that make up a standard ECG. This time-series data is made up of multiple channels that represent the different leads of the ECG, which require algorithms that can utilize this information and make connections across the different channels. For the mental health case study,

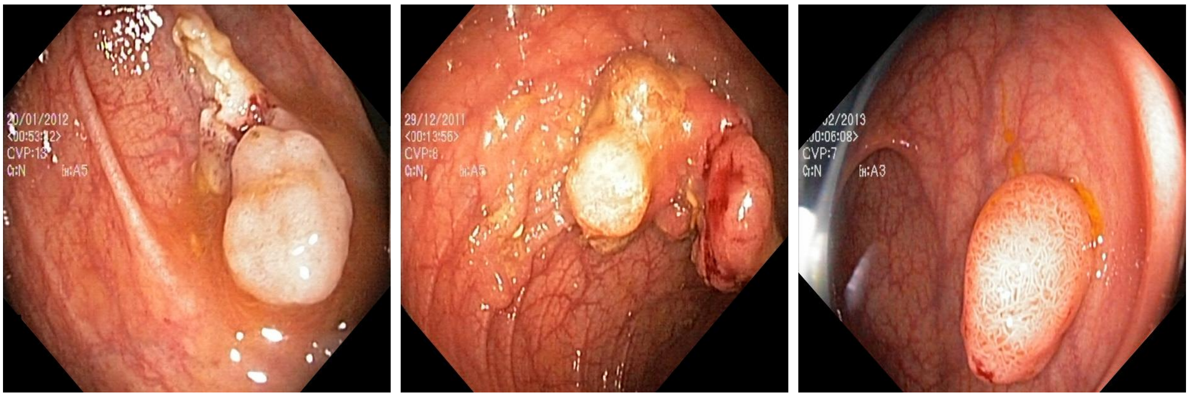
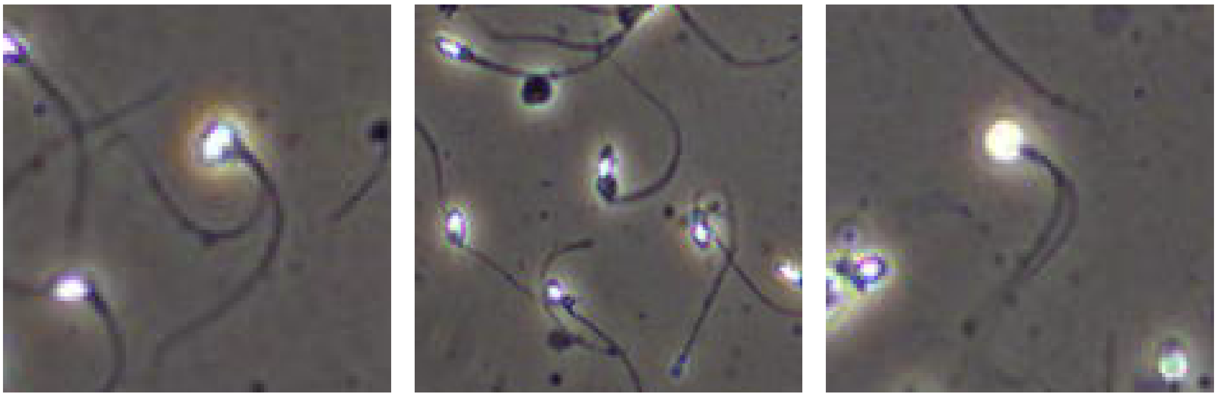


Figure 2.2: Images collected from colonoscopies that contain colon polyps.

we mainly focused on analyzing mental health in terms of how it affects the patient's activity. This data is also time-series but contains a single channel and must be analyzed for a much larger period of time, usually consisting of several days.

2.2.1 Gastrointestinal Computer-Aided Diagnosis

Gastroenterology is the branch of medicine that deals with the digestive system and the various diseases that afflict it. This includes the GI tract and the different organs that support it, like the liver and pancreas. The research for this thesis is mainly focused on automatically detecting different lesions and findings found in the GI tract. This is a rapidly growing area of research, where arguably the most common task is automatic colon polyp detection [146, 144, 147]. Polyps are small growths that grow on the mucosal wall of the colon (shown in Figure 2.2) and are a precursor to colon cancer, one of the most deadly types of cancer according to [142]. Despite being among the most deadly, colon cancer is also one of the most treatable if detected early enough. By removing a polyp, it eliminates the chance of it becoming cancerous. The current gold standard for detecting these lesions is through a procedure called endoscopy, where a long flexible tube attached to a tiny camera, called an endoscope, is inserted into either the mouth or anus. However, one major limitation of this procedure is that it is highly dependant on the skill and experience of the person handling the endoscope [117]. The consequence is polyp miss-rates that varies between 6% and 27% [4], which is a high range when dealing with something as life-threatening as colon cancer. Researchers see the potential of using machine learning to act a digital *third-eye* to catch the lesions that go overlooked during the endoscopy procedure.



(a) Two tails.

(b) Large midpiece.

(c) Large head.

Figure 2.3: Examples of abnormal sperm taken from the VISEM dataset [50].

Several works and research directions for developing computer-aided diagnosis systems to analyze the video feed from endoscopies have been explored [73, 42, 31]. Like previously mentioned, automatic polyp detection is especially sought-after. Several approaches have been applied with different levels of detection granularity. The simplest approach is classifying whether a video frame contains a polyp or not [105], without giving any information regarding the lesion's spatial properties. Other works focus on locating the polyp using bounding boxes or segmentation masks [18]. Beside polyp detection, there are other important findings that can be supported by machine learning. Finding and documenting anatomical landmarks is an important aspect of any endoscopy procedure and a recommendation by the European Society for Gastrointestinal Endoscopy (ESGE) [19]. There are also other lesions besides polyps that may be difficult to spot and diagnose during mucosal inspection. Machine learning may also be used to aid in surgery, such as verifying that the entire polyp has been removed after resection.

2.2.2 Assisted Reproductive Technology

Fertility rates have been dropping steadily in most parts of the developed world [135, 121]. This is partially explained by socioeconomic factors but also due to a rising trend of human infertility. Infertility affects about 12% of all couples worldwide, with about 40% being due to male infertility factors [79]. Concurrent with the drop in fertility rates, several studies have indicated that the sperm count has declined globally during the last few decades [21, 86, 72]. Semen quality is a key component of determining male fertility, but results are not consistent with regards to which parameters are best suited to predict this metric [16, 44, 28, 136]. To measure the quality of a semen sample, laboratory personnel look at certain key attributes of spermatozoa including sperm motility, sperm morphology, sperm vitality, and the concentration of spermatozoa per volume of semen (million/mL). These attributes are measured up against a manual for semen analysis published by World Health Organization (WHO) [164], which lists a set of reference values for semen attributes (parameters) based on semen quality of fertile men, whose partners had a time to pregnancy of 12 months or less. However, a common issue is that manual semen analysis requires trained laboratory personnel. If not performed in agreement with the WHOs guidelines, it might be subjective and prone to intra- and interlaboratory variability. Machine learning algorithms have shown immense power in analyzing visual data, making the visual aspects of semen analysis especially interesting from a machine learning perspective.

Sperm motility and morphology are two aspects of semen analysis that are highly visual and ripe for automation. Motility is measured by the number of progressive, non-progressive, and immotile sperm in a given semen sample. Progressive sperm are forward moving at a consistent pace. Progressive sperm are sometimes split into slow-progressive and rapid-progressive categories to differentiate the sperm that move fast and slow, but this dataset does not make this distinction. Non-progressive sperm is sperm that move but do not have any forward progression. This includes sperm that move in a circle or generally do not swim in a straight line. Immotile sperm are non-moving sperm and can be considered dead. The motility of a given semen sample is often measured as the percentage of sperm that fall into each respective category. Sperm morphology looks at the different parts that make up the sperm, namely the tail, midpiece, and head. Like motility, the sperm morphology of a given semen sample is measured by the number of sperm with tail defects, midpiece defects, and head defects. Common defects include multiple tails or heads, abnormally large or small heads, very short tails, or very long midpieces. Some images of abnormal sperm are shown in Figure 2.3.

2.2.3 Automatic Electrocardiogram Interpretation

Cardiology is the field of medicine that focuses on the heart and the treatment of the diseases that may affect it. The heart is the most important muscle in the human body and is exposed to a wide range of disorders, where one of the cheapest and most popular

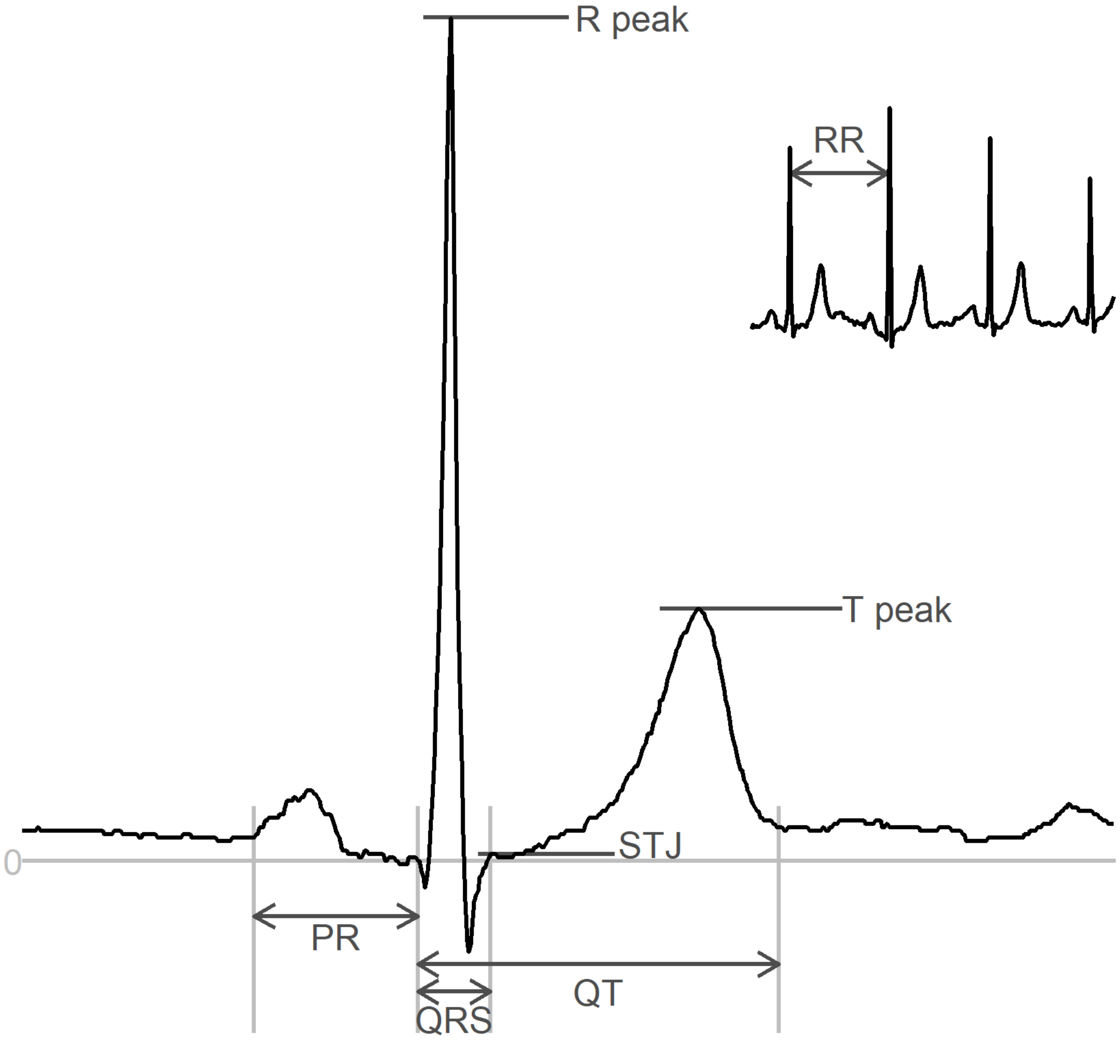


Figure 2.4: An illustration of a standard ECG showing waves and complexes.

procedures used for diagnosis are ECGs. ECGs is essential for a doctor's assessment of a patient's health and well-being, and may uncover a wide range of different cardiovascular diseases and dysfunctions. The examination is done by measuring the electrical signals (voltage) and conduction through the heart in each heartbeat. Every wave or spike depicts the depolarization or repolarization of the cardiac cells in a specific part of the heart. In order to evaluate the heart's electrical conduction system, analysis of certain intervals such as the PR interval, QT interval, and QRS duration is important (shown in Figure 2.4). The timing and the amplitude of these waves contain essential information about morbidity and mortality [26, 43, 100, 99] and automatic analysis of ECGs has been a topic of research since the early 1960s [138]. Still, the recent emergence of deep neural networks has led to more advanced approaches, including feature extraction [92, 113], noise reduction [112, 103], and heart rhythm classification [2, 167, 46].

2.2.4 Mental Health Detection and Aid

Mental health is currently on the decline in many countries, where rising suicide rates and substance abuse are just some of the consequences of mental illness becoming more common [157]. This is especially true since the start of the COVID-19 pandemic, where many have become more isolated and less social [107]. Furthermore, the continuing rise of social media platforms among the younger generation has shown to be a source of anxiety and depression [163]. Early detection and treatment have proven to significantly prevent mental health problems from developing [83]. However, many do not seek professional help before it starts becoming a real problem [51]. This is partially due to social factors such as the negative stigma around mental illness [133], but also due to many not realizing they have a problem in the first place.

Activity data has shown to be effective applied to studies to psychiatric diagnosis like bipolar disorder [127], ADHD [33, 96], and Schizophrenia [145]. For example, a recent systematic review summarized several motor activity studies of schizophrenia [160], which showed that patients with schizophrenia are associated with lower motor activity levels and repetitious and rigid patterns of behavior compared to healthy controls.

2.3 Ethics and Safety

The need for transparency in medical AI is rooted in ethics and safety [85]. Although AI systems have the capability of providing tremendous benefits to both health care providers and patients, several risks accompany the integration of AI into the current health care system [122]. First, there are no guarantees that the integrated AI system will always provide a accurate answer. Of course, one would not expect a medical doctor to be correct 100% of the time either, but in the case of AI, the reason behind a faulty prediction may be difficult to determine. Studies have shown that certain models do not provide the same level of fairness across race, gender, or socioeconomic status [22]. This may be justified in some cases, but it could also be a problem stemming from the development data used to build the AI system. Since these systems may be unreliable, explanation methods and through transparency provided with any system meant to be involved with patient care.

2.4 Summary

AI systems making their way into hospitals are inevitable. However, the point in time in which they will replace existing systems is still far away. Despite there being many useful applications that could potentially help save lives, there are still too many unknowns when it comes to deploying complex models like deep neural networks in the real world. This chapter presented some background on how AI systems are used in the four medical case studies that are the primary focus of this thesis, namely a GI tract case study, assisted reproductive technology case study, ECG case study, and mental health case study. First, we gave a brief introduction on how AI systems are developed by breaking it down into three primary stages; data collection, model development, and model evaluation. Then, we presented how AI systems are currently being used in each of the aforementioned case studies. Lastly, we looked at some of the risks of using AI in medicine, mainly focused on ethics and safety. In the next chapter, we present our work developing efficient and transparent AI systems within our four primary case studies.

Chapter 3

Transparent Artificial Intelligence Systems in Medicine

The primary research objective of this thesis was to understand how medical AI systems can be made more transparent. To answer this question, we need a solid understanding of how AI systems are developed and implemented for medical use cases, and recognize what problems medical doctors and clinicians want to be solved. With this in mind, we aimed to develop efficient transparent AI systems that solve medical problems faced by health care professionals. These systems were developed using four medical case studies, namely a case study on the GI tract, assisted reproductive technology, ECG analysis, and a mental health case study. Each system was built on the concept of transparent AI, which we define by the following three principles.

Data Transparency The principle of data transparency is built on open and publicly available data. This includes being transparency about how the data was collected and how it was prepared for analysis. The data used to develop and test an AI system should be publicly available so that other researchers may reproduce the work, inspect the dataset for any potentially missed biases or other drawbacks, and test their methods to ensure generalizability. Furthermore, public data can incentivize a common benchmark to measure future research and motivate new researchers to contribute to the field.

Analysis Transparency The principle of data transparency is built on open and publicly available data. The data used to develop and test an AI system should be publicly available so that other researchers may reproduce the work, inspect the dataset for any potentially missed biases or other drawbacks, and test their methods to ensure generalizability. Furthermore, public data can incentivize a common benchmark to measure future research and motivate new researchers to contribute to the field.

Evaluation Transparency The principle of analysis transparency is built on public and reproducible implementations of methods used for data analysis. The algorithms and methods used for analysis should be easily accessible, the configuration in terms of the hyperparameters and the architectural design used should be clear, and different approaches should be used and compared to related works to better understand how a system fits into the current research landscape.

This chapter is primarily split into three sections, one for each of the aforementioned principles of transparency in medical AI systems. First, we look at our contributions to making medical data more open and transparent through publishing multiple datasets in our four aforementioned medical case studies. Next, we show how we leveraged these datasets to implement AI systems to solve different problems in medicine. Last, we describe how these models were evaluated and tested to measure the performance, and how we may interpret the results to ensure that they worked as intended.

3.1 Transparent Data

Data is the lifeblood of any AI system. The quality of the data used for training and testing will determine the success and longevity of a model deployed in production. Gathering high-quality data is often easier said than done, as even if the number of data samples is high, it could still be skewed towards or against a particular population. Machine learning algorithms trained on data that does not properly reflect the target distribution could produce models that do not work as intended and become biased towards certain demographics such as sex, race, age, or specific symptoms or conditions. Therefore, knowing the data used to train and evaluate a model is essential when determining whether a model is ready for production. In high-risk fields like medicine, one could argue that this is even more essential because the models may be used as a basis for making life-critical decisions. However, data collected at hospitals is often buried behind laws and regulations in order to protect the privacy of the patient, which makes it difficult to make this data available to the public. Making data private as a means to protect patient privacy is not necessarily a bad thing but seriously impacts the accessibility of conducting research in the area. The following dives a bit deeper into some of the obstacles that make collecting and publishing medical data difficult.

1. The number of data samples for specific problems may be scarce and far below what is usually needed to train a complex model. For example, if we are trying to build a model that detects the presence of a rare disease, the number of real-world data points for this disease would naturally be low. A potential solution to this issue could be using artificial data or transfer learning, but a sufficient number of data points is still needed.

2. The collected data should be annotated by domain experts, meaning specialists within the field that the data is collected. For example, for data collected from colonoscopies, each frame should be viewed and annotated by an experienced gastroenterologist. This is an issue in medicine as medical doctors are often under extreme time constraints, making data annotations expensive and difficult to fit into an already busy schedule.

3. As with any private or sensitive information, there are rules and regulations in place that could make using the data a challenge. These vary between countries and can be quite restrictive. However, if the data is properly anonymized by removing any associations between the data sample and patient, it can usually be published without any possible harm to the provider. There are a few exceptions to this. For example, if the collected data is in and of itself identifiable information, like a recognizable mole or lesions that require affect the patient's face. These types of data would require additional measures before being made public.

Motivated by these obstacles, we collected and published several medical datasets across different areas of medicine. This section focuses on the importance of data transparency, where we present the datasets that were developed and published across three medical case studies; GI tract case study, assisted reproductive technology case study, and mental health case study. Subsequently, we present the synthetic datasets we developed as a response to not being able to publish certain data due to legal constraints. Then, we discuss why public medical data is important and how it can contribute to an overall better research community. Last, we discuss the ethical ramifications of making medical data public and consider potential methods of circumventing the regulations that prevent it from going public.

3.1.1 Medical Datasets

Dataset development is a complicated process that involves tasks like deciding what data should be collected and how and how it should be annotated. For this to be successful, several parties need to be involved. First, data scientists need to be involved as they understand how the data should be prepared and what labels should be included for it to be viable for model training and evaluation. Then, there are the domain experts who know the data itself and understand the value of the information contained within. Without close collaboration between domain experts and data scientists, a dataset may be positioned to solve the wrong problems or contain faulty samples, thereby making it practically useless. Each dataset published during this thesis was developed together with domain experts, both in terms of collecting and annotating the data, but also in building the potential prospects for the dataset. Furthermore, the datasets are published together with an article explaining the primary use case, baseline experiments, and discussing future directions. Baseline experiments are important as they demonstrate the technical validity of the dataset and act as a benchmark for future researchers. In the following, we briefly describe all published datasets and an overview of the currently available open datasets, organized under each case study.

3.1.1.1 Gastrointestinal Tract

For the GI case study, we developed three datasets [17, 69, 137] consisting of data collected from standard endoscopy and capsule endoscopy procedures. Findings such as anatomical landmarks, surgical interventions, and lesions are essential to document for an endoscopy to be considered complete [19]. Most currently open datasets (see Table 3.1) focus on automatic polyp detection, thereby overlooking many other findings. The main aim of developing the GI datasets was to compile a variety of findings, which could be used to build models for different tasks like classification or segmentation.

Table 3.1: An overview of the currently available open datasets containing images from the GI tract. Please note that the size of the datasets may have changed since the publication of this thesis.

Dataset	Target	Ground Truth	# Images	# Videos
[134]	Polyps	Classification	196	-
[15]	Polyps	Classification	612	-
[77]	Capsule endoscopy	Classification	$2,371$	47
[7]	GI findings	Classification	386	-
[97]	GI findings	Classification	-	$5,138$
[110]	GI findings	Classification	$8,000$	-
[6]	Polyps	Segmentation	$3,446$	-
[6]	Polyps	Segmentation	$1,000$	-
[109]	Bowel cleanliness	Classification	-	21
Our Datasets				
[17]	GI findings	Various	$110,079$	374
[137]	Capsule endoscopy	Various	$47,238$	117
[69]	Instruments	Segmentation	560	-

The first developed dataset was HyperKvasir [17], which is currently the largest publicly available GI dataset consisting of videos and images collected from colonoscopies (lower GI tract) and gastroscopies (upper GI tract) using standard endoscopy equipment from Olympus and Pentax at Bærum Hospital in Norway. The dataset contains both labeled and unlabeled data, where the labeled data is made up of $10,662$ images split between 23 different classes (examples shown in Figure 3.1) and 374 labeled videos. The unlabeled data contains $99,417$ image frames that have not yet been annotated. HyperKvasir also contains $1,000$ images of colon polyps that have corresponding hand-made segmentation masks. What differentiates HyperKvasir from most other GI datasets is that it also contains less-common findings, like dyed polyps and resection margins, which are important to document during an endoscopy procedure. The main goal of creating HyperKvasir was to compile a large dataset that could be used for a variety of different use cases, including classification, detection, and segmentation. This required a close collaboration with several medical doctors and export gastroenterologists in order to categorize and segment the different classes contained within the dataset. The number of classes, which classes were included, and the annotation protocol was decided upon by both the medical doctors and computer scientists. Both were necessary to ensure that the dataset was both usable from a medical and machine learning perspective.

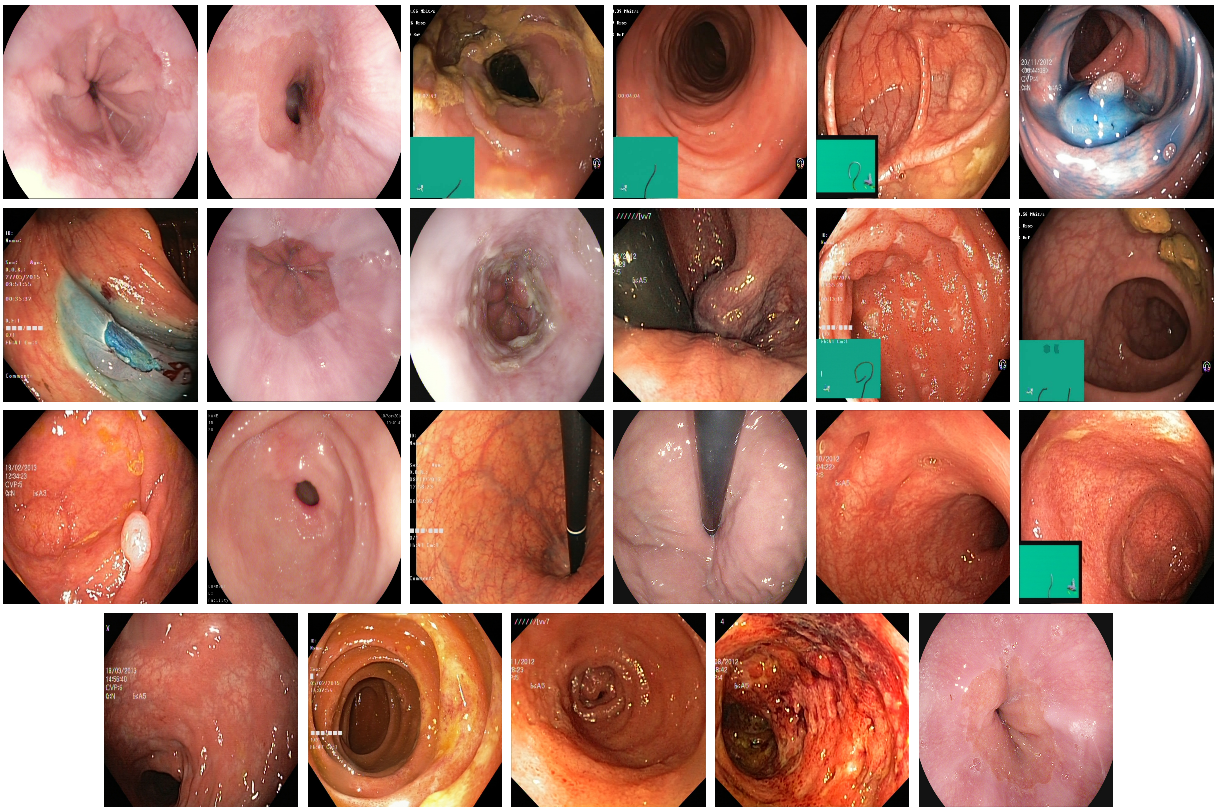


Figure 3.1: Example images taken from each class contained within the labeled part of HyperKvasir [17].

The second published dataset was Kvasir Capsule [137], which is a dataset containing images and videos collected from capsule endoscopy procedures. The data was captured from a Olympus EC-S10 endocapsule using an Olympus RE-10 endocapsule recorder. Like HyperKvasir, this dataset includes both labeled and unlabeled data. The labeled data is made up of $44,228$ images spread between 13 different classes and 44 labeled videos, where the class labels cover standard lesions like polyps and anatomical landmarks. An image sampled from each class is shown in Figure 3.2. The unlabeled data contains 74 videos, which is approximately 25 hours of raw video footage and $2,785,829$ frames. The aim of this dataset was to target capsule endoscopy which is gaining increased attention for analyzing the small bowel. Similar to HyperKvasir, this dataset was developed in close collaboration between both expert gastroenterologists and computer scientists.

Lastly, the final GI dataset we published was Kvasir Instrument [69]. Kvasir Instrument is a GI instrument segmentation dataset containing images and masks of instruments such as snares, balloons, and biopsy forceps. The dataset can be used to develop tool segmentation models that may assist medical doctors in performing surgery during the

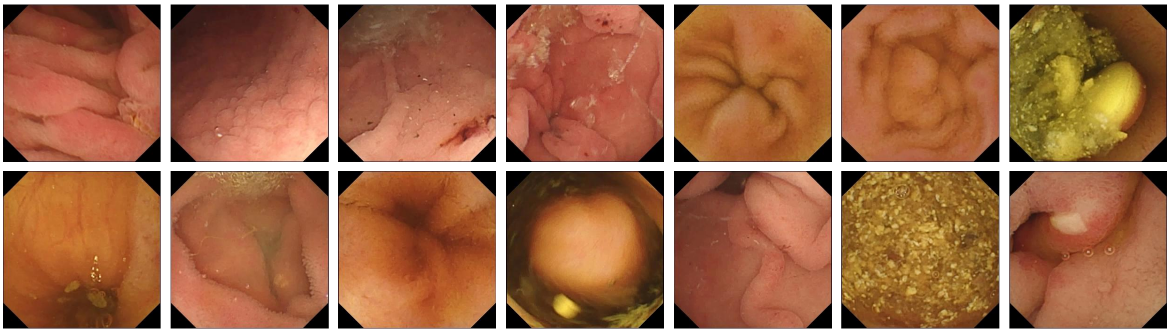


Figure 3.2: Example images taken from each class contained within the labeled part of Kvasir Capsule [137].

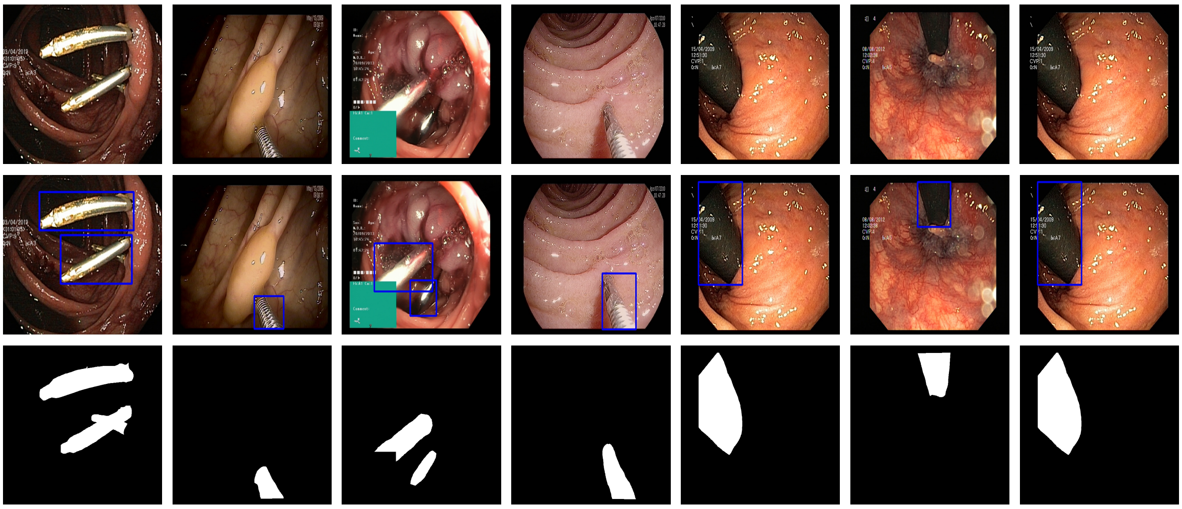


Figure 3.3: Examples taken from the development part of the instrument segmentation dataset Kvasir-Instrument [69].

endoscopy. The ground truth segmentation masks were drawn by a computer science student with much experience in analyzing GI-related data. Still, each segmentation mask was manually verified by an expert gastroenterologist to ensure correctness. Figure 3.3 shows a few examples of the images contained within the dataset together with the corresponding bounding box and segmentation mask.

3.1.1.2 Assisted Reproductive Technology

Machine learning is slowly being adopted by the assisted reproduction community [121], so the number of open datasets is very few. A list of the currently available datasets for both semen and embryo analysis is shown in Table 3.2. For the area of assisted reproductive technology, we published a dataset containing video recordings of human semen with associated analysis and participant-related data [50]. Our dataset, VISEM, contains data

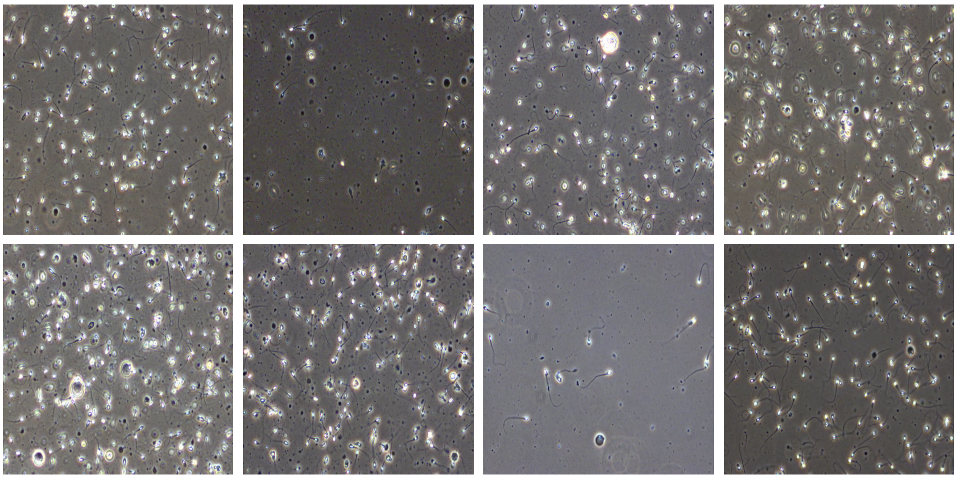


Figure 3.4: Example frames collected from the VISEM dataset [50].

Table 3.2: An overview of the currently available open datasets containing visual data related to assisted reproduction (embryo and semen). Please note that the size of the datasets may have changed since the publication of this thesis.

Dataset	Target	Ground Truth	# Images	# Videos
[8]	Sperm	Classification	1,540	-
[132]	Sperm	Classification	725	-
[66]	Sperm	Classification	200	-
[94]	Sperm	Classification	1,064	-
[126]	Embryo	Segmentation	235	-
Our Datasets				
[50]	Sperm	Regression	-	75

from 85 participants aged 18 years or older that were collected in association with a study on how body mass index (BMI) affects male fertility [9]. The videos were recorded using an Olympus CX31 microscope at 400× magnification at 50 frames per second and had a resolution of 640×480. Figure 3.4 shows a few frames taken from videos included in the dataset with different levels of sperm density. In addition to the video data, each semen sample also comes with a set of meta-data, which includes participant-related information (age, days of abstinence, and BMI), a fatty acid profile of the sperm and serum phospholipids, sex hormone levels, and a preliminary quality analysis done by an expert clinician following the WHO guidelines.

The primary aim of developing this dataset was to develop AI systems that can automatically evaluate the quality of a given semen sample. With this goal, we tried to only include high-quality semen samples that contain minimal drift (moving serum) and videos

Table 3.3: An overview of the currently available open datasets containing activity-related mental health data. Please note that the size of the datasets may have changed since the publication of this thesis.

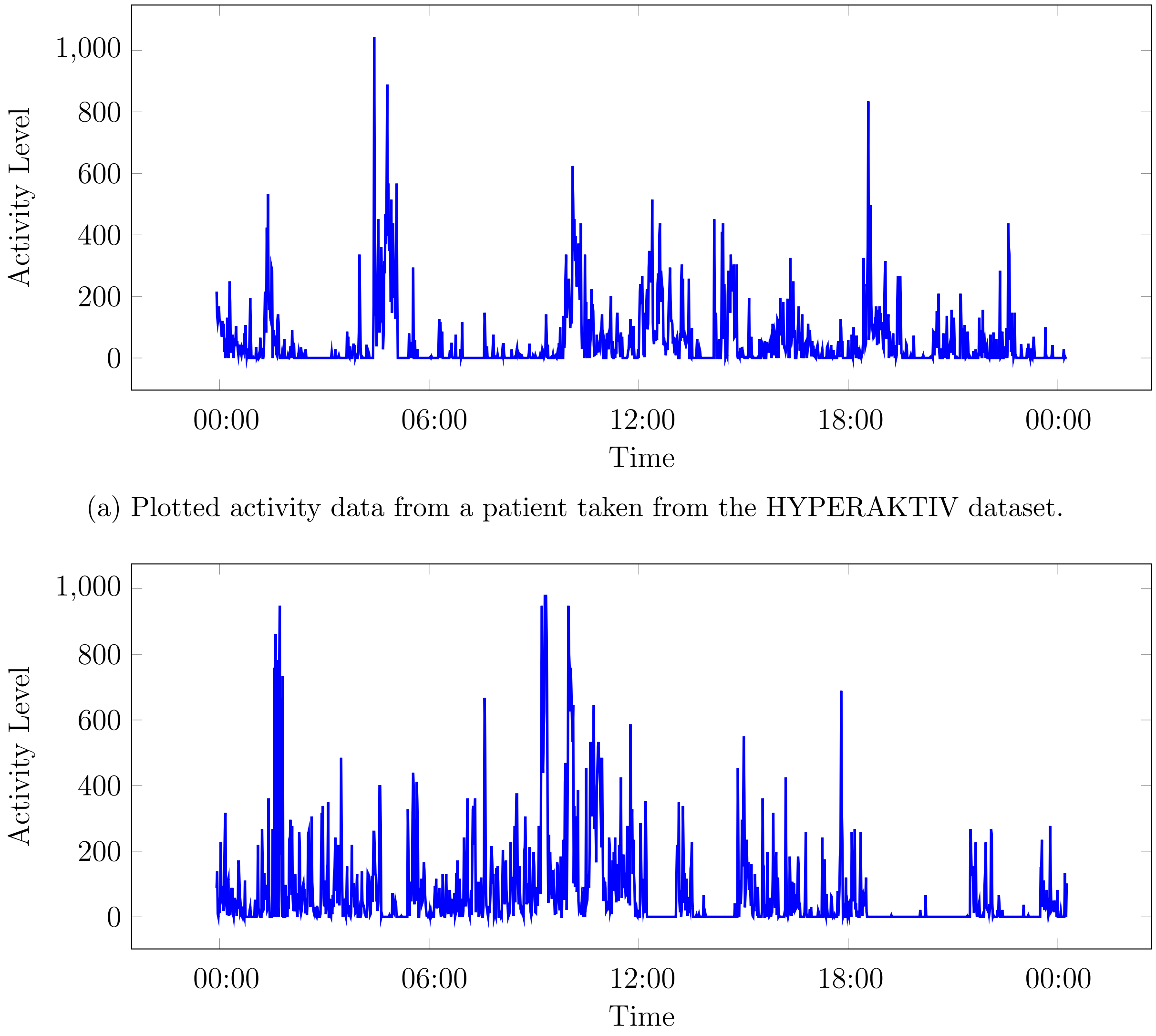
Dataset	Target	Ground Truth	# Samples
[125]	Healthy	Classification	22
[38]	Depression	Classification	55
Our Datasets			
[39]	Schizophrenia	Classification	32
[58]	ADHD	Classification	103

where the microscope is correctly focused. Looking back at that decision now, it may have been better to include the videos with poor quality for a more realistic varied dataset. This would increase the size of the dataset and introduce samples that the clinicians need the most help with analyzing. Furthermore, adding annotations for sperm tracking would make the dataset immensely more interesting from a machine learning perspective as one could then analyze the individual sperm rather than just the sample as a whole. We started on an annotation tool for assisting clinicians in adding tracking annotations to the data, but this is still a work in progress.

3.1.1.3 Mental Health

The world is becoming more *smart*, where seemingly every accessory has some feature to measure specific biometrics about the wearer. Data measured by these devices include information like activity measurements, quality of sleep, heart measurements, and several metrics regarding personal health and fitness. A lot of research is being done analyzing this data as it can discover and support people with disorders such as anxiety or depression [80, 115]. Table 3.3 shows the currently open dataset that primarily target a mental health use case. We developed and published two datasets containing activity data collected from patients hospitalized at a long-term open psychiatric ward at Haukeland University hospital.

The first dataset, Psykose [67], consists of activity data collected from schizophrenia patients, which was obtained through a wrist-worn actigraph device (Actiwatch 4 accelerometer) sampled at $32Hz$ and movements over $0.05g$. In total, data was collected from 22 schizophrenia patients and 32 healthy controls. In addition to the activity data, the dataset also contains information about the patient like their age, sex, type of



(a) Plotted activity data from a patient taken from the HYPERAKTIV dataset.

(b) Plotted activity data from a patient taken from the PSYKOSE dataset.

Figure 3.5: Activity measurements from two patients taken from the HYPERKVASIR and PSYOKSE dataset respectively.

schizophrenia, Brief Psychiatric Rating Scale (BPRS) sum score, or if the patient used some certain types of medication during study period. Figure 3.5b shows an example of the activity data collected from one of the patients plotted over a 24 hour period.

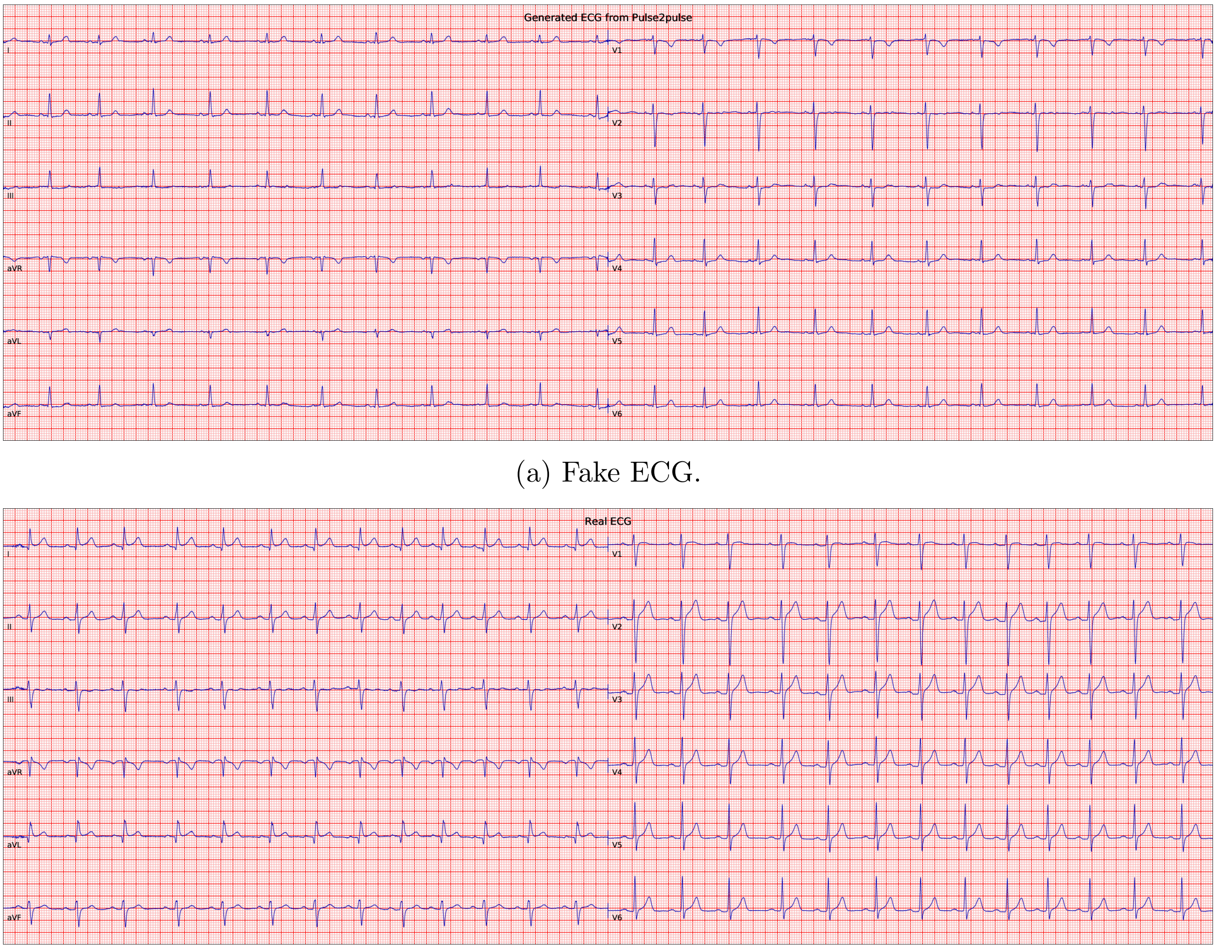
The second dataset, Hyperaktiv [58], contains activity and heart rhythm data collected from patients diagnosed with ADHD. Like the Psykose dataset, the data was collected using a wrist-worn actigraphy device (Actiwatch 4 accelerometer) that registers acceleration in the three-dimensional space. Overall, we collected data from 51 patients with ADHD and 52 clinical controls. We also include a series of patient attributes such as their age, sex, and information about their mental state and output data from a computerized neuropsychological test. The primary purpose of this dataset is to analyze the activity and heart rhythm data to detect whether the patient has ADHD or not. Figure 3.5a shows an example of the activity data collected from one of the patients plotted over 24 hours.

3.1.2 Synthetic Data

As mentioned earlier, medical data is highly personal and is almost always subject to laws and regulations preventing it from being made public. In some cases, making a dataset open is just not feasible due to the legal ramifications or simply because the owners do not want to release it to the public. An alternative could be to release a synthetic version of the dataset that represents the distribution found in the real data. This removes the privacy barriers as the data is fake and not associated with any single patient. Furthermore, Synthetic data could also be useful if the number of data points is relatively small, where more variations could lead to a more general model. Motivated by the privacy barriers that prevented us from publishing datasets and the general lack of data samples in medical datasets, we developed a framework to replace real medical data with synthetic data of equal quality. The idea behind this framework is to select specific parts of the human body and have synthetic data generated from this part. The project is fully open-source and is access online[1].

Currently, the framework supports two types of medical data, synthetic ECGs and synthetic colon polyps. The synthetic ECG data is generated using a GAN, which was trained to represent real ECGs collected from Denmark [47, 40]. Using the GAN, we produced $121,977$ synthetic ECGs and published them online for other researchers to use. We also published the GAN architecture together with the trained weights so other researchers can generate synthetic ECGs by themselves. An example taken from the synthetic ECG dataset is shown in Figure 3.6, which also includes a real ECG for comparison. The synthetic polyp data was generated similarly, using a GAN but with a different architecture. As with the synthetic ECGs, we generated $10,000$ synthetic polyp images with corresponding segmentation masks and made them publicly available through DeepSynthBody.

[1] https://deepsynthbody.org



(a) Fake ECG.

(b) Real ECG.

Figure 3.6: An example ECG generated by our GAN (top) compared to a real ECG (bottom).

3.1.3 Ethical Considerations

Collecting and publishing sensitive information comes with serious ethical considerations. First of all, any information collected from a patient should be fully anonymized before being made public. Data that may seem harmless to some could potentially be devastating for others if leaked depending on a person's situation. Besides being an important ethical point, it is also part of several legal and regulatory requirements of publishing data collected from humans. Therefore, all datasets presented and used in this thesis have been fully anonymized without the possibility of tying a data sample back to the original paper.

There are some ethical dilemmas that can be considered open research questions. For example, as mentioned in Section 3.1.2, we circumvented the privacy-related issues of public medical data by using synthetic data generated by a GAN trained on real samples from a private dataset. The open question here is that assuming that the trained GAN has some recollection of the real data within its weights, would it be possible to reverse engineer the weights to come back to the original data sample? Currently, there is no method for doing this, but one could imagine a model that could learn the association between the fake and real data. Another open question could be if one could identify a patient's data samples in an anonymized dataset using data from that patient at another point in time. Imagine a time series dataset that contains activity data from several different persons. Suppose one were to collect new data from one of the participants in the dataset. Would it be possible to use this new data sample to identify which samples in the dataset belong to that participant? These are some questions that would be interesting future research topics that could have an impact on how we publish data in the future.

3.1.4 Lessons Learned

Through collecting and publishing the aforementioned datasets, we gained a better understanding of the importance of close collaboration between medical doctors and computer scientists in each step of the dataset development process. None of the aforementioned datasets would be possible if neither part were there to make sure the requirements from both sides were met. We as computer scientists can often find problems that seem important but are actually not an issue when speaking directly with the medical doctors we are trying to help. On the other side, medical doctors often misunderstand the requirements of what type of data is needed to solve a specific problem and have either too high or too low expectations of what AI-based solutions can bring. Furthermore, we identified several factors that make publishing medical data especially important.

A Common Benchmark State-of-the-art machine learning models are often determined based on their performance on a common benchmarking dataset. For example, ImageNet [29] is a benchmark for image classification and detection. A common benchmark ensures that all methods are trained and tested on the same data, making the results a cause of the methods rather than the data. A common issue in medicine is that methods are often trained and tested on data that is private, without means of comparing against other works. This is not an oversight from the authors' side but a general lack of public benchmarking datasets in medicine. However, benchmarking datasets are starting to appear in medicine as well, but there are so many different application areas that this is still a problem. By making more medical data public, we have the opportunity to develop benchmarks for popular medical problems, but also in areas that often go overlooked by the majority of medical research.

Increased Awareness Most researchers do not have the luxury of having an established relationship with a hospital or clinic to supply them with medical data for research. Furthermore, those who do have access usually only receive data for a few problems. This lack of data accessibility severely hinders the number of people that are able to work on any given medical problem. Previous works have shown that with the introduction of large open datasets it accelerates the amount of research produced in that field by a lot [114]. By making data public, we open the opportunity for other researchers to contribute to a field that would have otherwise gone unnoticed by the majority. We know that there is interest by both independent and institutional researchers based on the number of contributions to publicly hosted challenges available on sites like Kaggle[2]. Moreover, we have hosted several challenges that confirm the interest in medical multimedia, which will be further discussed in Section 3.3.2.

Transparent Research As we touched upon earlier, knowing what data a machine learning model is trained and evaluated on is an important aspect of understanding its potential limitations. Without knowing the data, it is difficult to understand whether a model is good or bad based on a set of evaluation metrics alone.

3.2 Transparent Analysis

If data is the lifeblood of machine learning, the methods used for analysis are the brain. The parameters and architectural design of a machine learning model determine how well it is able to *learn* from the provided data. This data comes in all shapes and sizes, of which different methods should be used to exploit the nuances of each modality in order to fully use the potential of the information contained within. As we described in Section 2.1, collecting and preparing data is the first step of building an AI system. The next step is developing and implementing the methods to analyze this data. Throughout this work, we implemented and experimented with several different types of methods. Publications

[2] https://www.kaggle.com

were associated with several of these experiments, where we aimed to follow the following principles of transparency.

Open and Reproducible Implementation The implementation of the analysis methods should be fully reproducible. Ideally, the code should be made publicly available through a service like GitHub or GitLab. If not made open source, the methods paper should at the very least contain enough information to implement the method from scratch. This includes the architecture and implementation type of the methods used and the hyperparameters used to train it.

Developed on Open Datasets The methods should be trained and developed on open datasets. This ties in with the previous principle of reproducibility as methods developed as one can not reproduce the results of a method developed on closed data.

In this section, we describe how we leveraged the developed datasets to develop systems within the four main case studies of this thesis. This is organized by case study, where we first look at the systems developed for analysis of data collected from the GI tract. Then, we present our work on automatically assessing the quality of a given semen sample under the assisted reproduction case study. Last, we present our work on automatically analyzing ECG using deep neural networks.

3.2.1 Gastrointestinal Track

The primary aim of developing AI systems for the GI case study was to automatically detect different findings in the GI tract. Previously, we developed a system meant to support medical doctors by automatically generating endoscopy reports from a supplied video [1]. The system scanned through a provided video and presented the user with the findings detected by an underlying convolutional neural network (CNN). The detected findings could be further scrutinized by using a model visualization feature that allowed the doctors to gain more insight into how and why the model categorized a specific image to particular class. Building from this work, we looked at enhancing the AI system by improving the underlying CNN by expanding the scope to cover more GI findings and also include analysis of data collected from capsule endoscopies. The work here can largely be split into two groups; findings classification and findings detection. Both groups aim to automatically find notable findings in the GI tract, where findings are objects that are of particular interest to the gastroenterologist like lesions, anatomical landmarks, or surgical landmarks. In the following, we describe the methods and experiments developed for the GI case study.

3.2.1.1 Endoscopy Image Classification

Image classification is the task of predicting what class or category an image belongs to. This could be, for example, categorizing images of different animals or automatically annotating frames of a video that contain cancer. In the context of GI endoscopy analysis, image classification can help us detect important findings during or after an endoscopy procedure. We experimented with several approaches to endoscopy image classification.

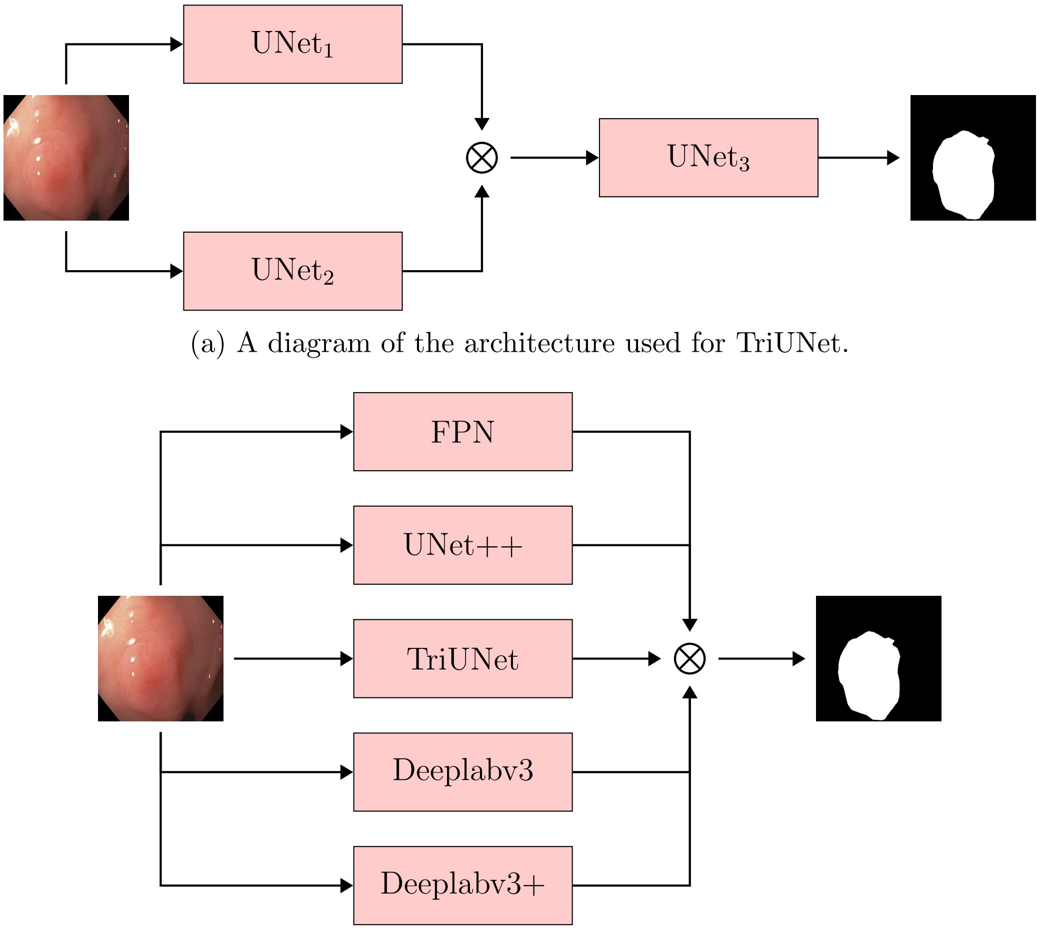
Transfer learning is a common technique where we use the weights of one model to initialize the weights of another. The advantage is that we start the model with some sense of the real world, making the new task easier to learn. This is especially convenient if our target task has a low number of training samples. Usually, we want the transferred weights to be as close to the target domain as possible so that the model does not have to relearn several new concepts. For medical imaging, studies have found that transferring the weights from natural images (like the images found in ImageNet) works very well and leads to faster convergence and a more general model [159]. To get a better understanding of how the domain relevance of the transferred weights affects a model during transfer learning, we trained several models for automatic GI findings classification using two transfer learning domain sources [64]. The first domain was natural images, where we used ImageNet [29], which a huge database of natural images ranging from inanimate objects to different types of animals. The second domain was medical images collected from surgeries, where we used a combination of different medical datasets collected from procedures such as laparoscopy [84]. The results showed that the weights trained on the larger and more diverse dataset, ImageNet, performed better than weights trained on lower quantity but more domain-relevant data.

Annotating every piece of data collected at hospitals is expensive, takes too much time, and requires excessive work from the medical doctors. The consequence is that unlabeled data is substantially more difficult to use for training machine learning models.

We developed a system using unlabeled data for training classification models in standard and capsule endoscopies using a teacher-student framework [41]. The framework consists of a teacher and student, thus the name. The teacher is trained in a standard supervised manner on labeled data and is used to assign pseudo-labels to the unlabeled data that the student model then uses for training. The labels assigned to the unlabeled data are called pseudo-labels because they do not adhere to any actual ground truth and merely reflect what the teacher model has learned from the labeled data. The experiments were run using HyperKvasir [17] and Kvasir-Capsule [137], where we show that using the unlabeled data produce better results than the standard classification paradigm. This showed us that if we have data that experts have not labeled, we can still use this to improve the generalization and predictive performance using pseudo labels. The framework and the source code used to implement the experiments were open-sourced on GitHub[3].

Medical data is sometimes collected with certain artifacts or overlays that may interfere with a model's predictions during training. We experienced this ourselves as some of the data collected in HyperKvasir contain artifacts such as a green navigation box, text overlays, black borders, and some icons placed on certain parts of the frame [63]. Using GradCAM [128] to visualize the predictions, we confirmed that the models had learned to associate the green navigation box with colon polyps, meaning that the model had incorrectly learned that the navigation box is an attribute of colon polyps. Based on these findings, we looked at different methods to replace these artifacts with what would be a natural extension of the image. This task is commonly referred to as image inpainting, where we trained an autoencoder and GAN to replace green navigation boxes and black corners with a colon background. The results showed an improvement over using the raw images which contained the artifacts [76]. These experiments reinforce the principle of collecting high-quality data when the primary purpose is for training a machine learning model. Had we collected the data without these artifacts, this step would be unnecessary, and we may have seen a general increase in performance among the models trained on the dataset. Still, with medical data, we do not always have the privilege to choose what state we receive data and that sometimes overcoming these challenges will be part of the solution.

[3] https://github.com/henriklg/teacher-student-framework



(a) A diagram of the architecture used for TriUNet.

(b) A diagram of the architecture used for DivergentNets.

Figure 3.7: A diagram of the two neural network architectures used to segment polyps for the EndoCV 2021 challenge.

3.2.1.2 Polyp Detection and Segmentation

Image classification signifies the presence of a specific finding somewhere in the given image. However, doctors often want a more granular prediction that can show what regions of the image or frame contain the predicted finding. This can be done through either object detection or object segmentation, where object detection is the process of locating the object encasing it in a bounding box, and object segmentation is locating the object with pixel-level precision. To satisfy this request for more precise prediction, we performed additional experiments that automatically segment colon polyps in a given endoscopy video frame.

Segmentation models are often complex and have had several new contributions in terms of neural network architectures in the last few years [130]. While participating in the 2021 edition of EndoCV [5], we developed multiple ensemble-based segmentation models for automatic polyp segmentation [150]. The main contributions of this work were two architectures, TriUNet and DivergentNets, both of which use an assortment of popular segmentation models. TriUNet consists of three UNet architectures organized in a triangular-like shape. The input is passed through two separate UNet models, whose output is concatenated and passed through a final UNet model. The three models are interconnected, and their weights are updated based on the loss calculated from the output of the final model. The second model, DivergentNets, is a standard ensemble model consisting of five separately trained segmentation models, whose output is produced through majority voting on the pixels. The five model architectures used were UNet++ [166], FPN [87], DeepLabv3 [23], DeepLabv3+ [24], and TriUNet [150]. The architecture of both models is shown in Figure 3.7. Overall, the results showed that the combination of the five different models performed better than any single alone, where some models seemed to tackle certain situations better than others. Still, the added predictive performance gain came at the cost of a substantially larger model and much slower processing speeds, making it questionable whether this model is appropriate depending on the requirements. The proposed solution achieved the best scores among all other participants [150], and the implementation was open-sourced on GitHub[4].

Data augmentation is often used to make models generalize to samples with slight variations and is also used to increase the overall training size of a dataset. We developed a novel method for augmenting the masks of a segmentation pair (image and mask), which aims at making the model learn the features of a class at several levels of granularity. The augmentation framework is called Pyramid-Focus-Augmentation (PYRA) [149], and augments the masks by dividing the region of interest into grids. The number of grids is a hyperparameter, but the remainder should be 0 when dividing the resolution by the grid size. Augmenting the training data resulted in overall better performance. The augmentation framework was open-sourced as a Python library[5].

3.2.2 Assisted Reproductive Technology

In the area of assisted reproductive technology, we developed methods for automatically analyzing the quality of human reproductive data, which includes data from semen and

[4] https://github.com/vlbthambawita/divergent-nets

[5] https://github.com/vlbthambawita/pyra-pytorch

embryos. Although the main focus of our work was aimed at semen, we also performed some experiments on analyzing time-lapse videos of human embryos [53]. The motivation behind automatic semen analysis is that current methods are time-consuming and could be significantly accelerated through automation. Although automatic systems like computer-aided sperm analysis (CASA) exist [95], they are not considered good enough to be recommended for in-clinic use [148]. Using the VISEM dataset described in Section 3.1.1.2, we developed and compared several machine learning methods to predict the morphology and motility of human sperm using videos and the associated meta-data. For the embryo scenario, we used a private dataset to build models that predict human embryo viability. In the future, we hope to make this dataset public but this still requires some work together with the clinicians. This section describes the methods used for analyzing human semen samples and human embryos in detail and shows how we utilized the video data in different ways to make the most of the spatial and temporal information contained within.

3.2.2.1 Semen Analysis using Traditional Machine Learning

As an initial benchmark, we experimented with using traditional machine learning methods and handcrafted features to analyze the sperm videos for motility and morphology prediction [60]. Even though deep learning is the most popular approach today, deep learning does not always outperform traditional methods using handcrafted features. Furthermore, traditional methods are usually more explainable when compared to those based on deep neural networks, making them a viable alternative even though they may perform slightly worse than more modern methods. For these experiments, features were extracted from the first and middle frame of the first 60 seconds of the semen videos using the open-source library Lucene Image Retrieval (LIRE) [91], a Java-based image retrieval library that contains several feature extraction algorithms. We experimented with over 30 different visual features such as Tamura, auto color correlogram, and pyramid histogram of oriented gradients, to name a few. The extracted features were used to train a series of differed algorithms implemented in the Waikato Environment for Knowledge Analysis (WEKA) [162] machine learning software. For predicting semen quality, our findings showed that, in general, Tamura features seemed to capture the information within the video frames best, where the best performing algorithms were random forests and SMOreg. The results were compared to a ZeroR baseline generated over the ground truth, where both algorithms beat the baseline over the tree predicted categories of the motility of sperm.

3.2.2.2 Embryo Analysis using Traditional Machine Learning

Building off the semen analysis experiments, we started analyzing time-lapse videos of human embryos of early embryonic development up to day 5 to predict the likelihood of a successful birth [53]. Using the same setup as for the semen, we again found that Tamura features together with random forests yielded the best predictive performance. In addition to the supervised algorithms, we applied unsupervised clustering methods to group the embryo videos using a set and dynamic number of classes. The results showed that both the supervised and unsupervised methods could correctly categorize the embryos with high accuracy. However, the study was performed on a very small dataset, so further experiments using more data collected from various sources are required before making a solid conclusion.

3.2.2.3 Semen analysis using Deep Learning

One of the advantages of using deep learning is that it automatically learns what features are associated with a given task. This allows us to directly insert the raw frames into the deep neural network without stripping out any potentially helpful information. However, as a single video can contain several million different values per sample, we need a strategy to compress this information before sending it through the model. We tested several different methods of preparing the video data for analysis [60].

First, we applied the simplest approach of predicting the semen quality in terms of sperm motility by using a single frame as input to the CNN and averaging the predictions across the video. This approach has some obvious limitations in that it does not take the temporal information into account when making its predictions, something that is important for motility prediction. This is reflected in the results, where we found that the single-frame approach provided limited predictive performance. Overall, the performance is similar to that of the traditional machine learning algorithms, which makes sense as they both base their predictions on a single frame.

Another approach we tried was to concatenate a sequence of video frames in the channel dimension or flatten the frames and concatenate them spatially (shown in Figure 3.8).

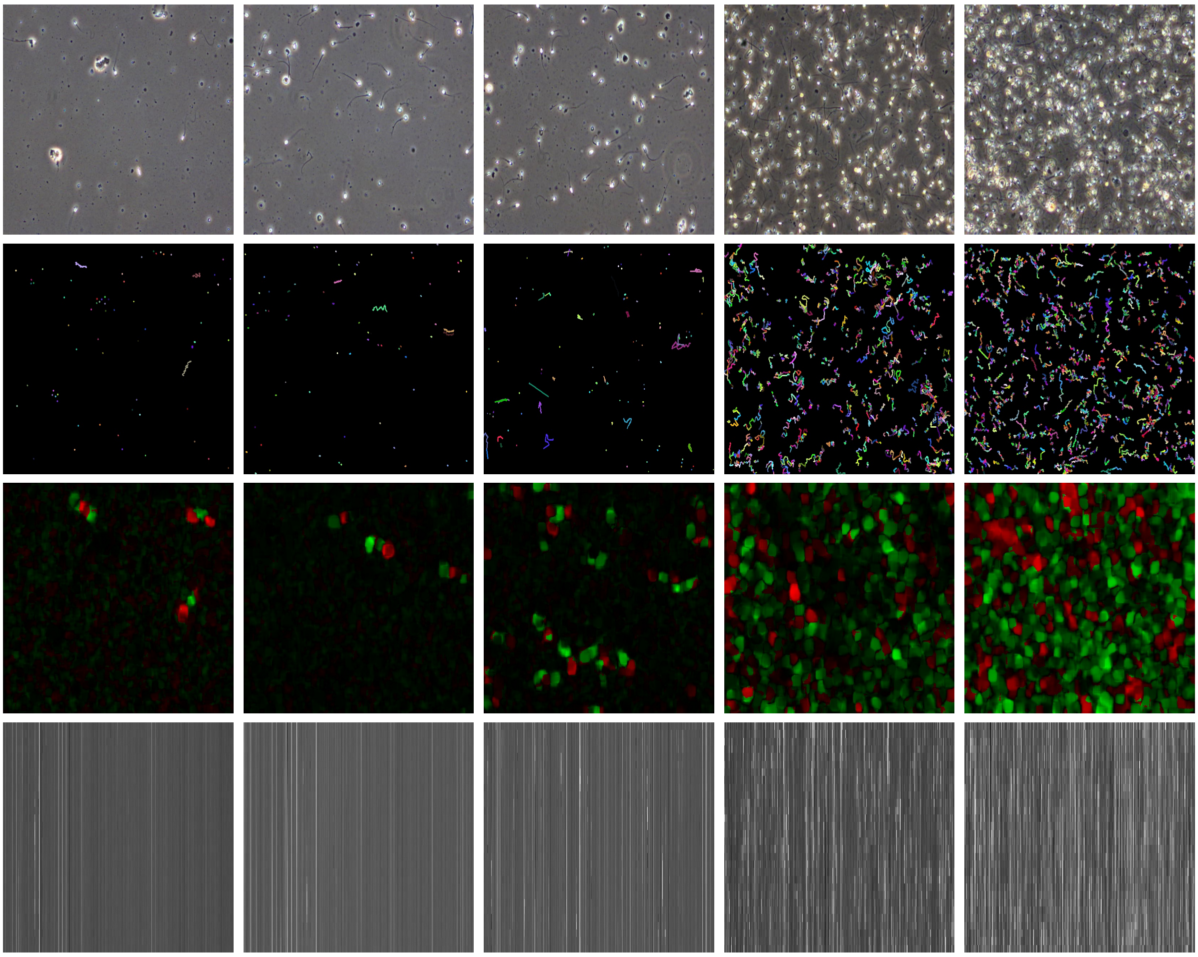


Figure 3.8: Examples of images from videos of semen samples with different concentrations (columns) and the four image representations used to train the sperm neural network-based algorithms (rows). Image representation by row; 1) original video, 2) sparse optical flow, 3) dense optical flow, and 4) vertical frame matrix.

This retains the temporal information present in the video sequence, which improved the results compared to the naive single-frame approach. Despite providing better performance, stacking multiple frames increases the size of the input drastically, making the overall model slower and more computationally hungry.

As an alternative to processing raw frame data, we explored using optical flow generated from extracted video sequences and using this as input to the model. Optical flow generates temporal representations of a sequence of frames into a single image. We use two different methods of generating optical flow, one method based on sparse optical flow and one method based on dense optical flow. For sparse optical flow, we used the Lucas–Kanade method [88] for optical flow estimation, which assumes that the flow is always in a local neighborhood of the tracked feature. We use Gunner Farneback's algo-

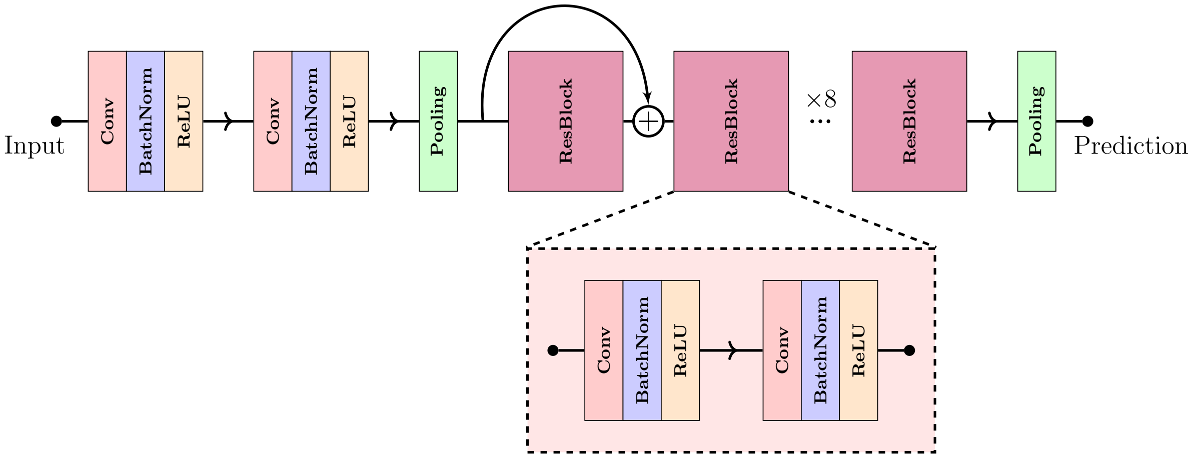


Figure 3.9: The convolutional neural network architecture used to analyze ECGs.

rithm [35] for dense optical flow, which compares two images and measures the overall change between one frame and another. The results showed comparable results to that of concatenating multiple frames but at a fraction of the input size.

3.2.2.4 Multimodal semen analysis

All aforementioned semen analysis methods were also tested with the insertion of participant-related data into the analysis. Overall, the results of the experiments showed that deep learning is considerably better at analyzing video. Moreover, for both the traditional machine learning and deep learning experiments, the addition of participant-related data seemed to confuse the models, making the models perform worse than without the additional information.

3.2.3 Electrocardiogram Analysis

For the ECG case study, we developed a system for automatically predicting specific attributes of a given ECG like the QT interval, QRS duration, PR interval, R-peak amplitude, T-peak amplitude, and J-point elevation. These waves and intervals should be part of any healthy patient and can be determined from a single median ECG complex. A median ECG complex is a representative heart beat calculated as the median complex over a set interval visualized in Figure 2.4. We also tried to predict other parameters such as the heart rate and the sex of the patient. To measure the hear rate, we used 10-second rhythm strips that include several heart beats. For sex classification, the median complex was used. We developed a novel CNN-based architecture that takes either a 12-lead median ECG complex or a 10-second rhythm strip to make a prediction (architecture shown in Figure 3.9). We evaluated the model by using quantitative metrics and a qualitative evaluation by comparing the prediction of a subset of the ECG against expert cardiologists. When predicting the waves and intervals, the results showed that the neural network was overall more precise and consistent in the predictions when compared to real world experts. Using the model to predict sex from a given ECG showed an even larger discrepancy, where the neural network was much more accurate than the experts.

3.2.4 Lessons Learned

The process of designing and implementing the presented systems for analysis taught us the importance of experimenting across data modalities and classes of algorithms. Although deep learning is currently a trending topic in AI, classical algorithms can still be preferable due to them often being less computationally expensive and more interpretable. With regards to the experiments performed for automatic semen analysis, the classical machine learning algorithms were able to outperform the simple baselines but performed worse than the deep learning-based methods. Choosing an approach for analysis should be done based on the requirements of the task at hand, where the choice should take data properties, interpretability, efficiency, and predictive performance into account.

3.3 Transparency in Evaluation and Results

Measuring the performance of a model is critical when deciding whether it is safe to use or not. Models that perform well in an experimental setting often do not shown the same level of success when deployed in real medical practice. Therefore, thoroughly understanding the vulnerabilities and weaknesses of a model before it reaches this stage is essential for preventing inaccurate, biased, and unintelligible predictions. Through the process of collecting data and implementing machine learning algorithms for medical use cases, we observed a large amount of medical AI research that lack thorough quantitative evaluation, miss related work to compare against, and make little effort to explain the results and potential biases of the developed model. This section focuses on transparency in the evaluation and presentation of results for medical AI systems. Our work can be broken down into three main areas of model evaluation, namely evaluating models using quantitative metrics, evaluation through challenges and benchmarks, and explaining model predictions using explainable AI. These three areas are further motivated in the following.

Quantitative Metrics Assessing a model through the metrics achieved on a given task is among the simplest forms of evaluation. The metrics used differ depending on the task; for example, precision and recall are two metrics commonly used in medicine to evaluate the model's predictive performance. Using multiple metrics to measure the performance of a model is an important step in understanding how a model will perform once deployed in a real-world setting.

Reproducible and Comparable Results One way of assessing the current state of machine learning is through organizing challenges or benchmarks that gather data scientists and engineers in order to solve a specific task. Although not directly tied to evaluating one specific model, these events may help future researchers have a set benchmark to compare against and can encourage researchers to contribute to a new field.

Explainable Artificial Intelligence As mentioned before, the most common approach in machine learning today is using deep neural networks, as they have shown a profound ability to perform well on almost any given task. However, despite these promising results, deep neural networks are not easy to interpret and are generally labeled a ”black box”. Explainable AI aims to open this black box by providing an explanation as to why the model produces a given prediction. Furthermore, these explanations can be used to discover new correlations between an input and outcome.

3.3.1 Quantitative Metrics

The most common way of evaluating the correctness of a machine learning model is using quantitative metrics that measure the direct performance in terms of a number. Different metrics tell a different story about how we can expect a model to perform when deployed into the real world. For example, a high precision indicates that when a model predicts a specific class, it is quite certain about its prediction. On the other hand, a model that achieves a high recall with low precision will make a lot of positive prediction but most incorrect. Evaluating a model with a wide range of different metrics is essential to get a full understanding of the model's potential. It is simple to make a model look like it performs better than it actually does by only presenting certain metrics and excluding others. While going through the related research for our selected medical case studies, we realized that a lot of studies that apply machine learning to medical problems do not evaluate their methods as thoroughly as research published in machine learning journals. Medical machine learning studies often show a few metrics to measure the performance of their methods, and with most data being private, any chance of reproducing the results to further check the performance is quite slim. To bring more awareness to this issue and provide more information about why different metrics are useful, we performed a study where we looked at research within GI machine learning applications and recalculated the results to see if the performance was as reported [62]. This study was focused on binary classification, but the same principles apply to all sets of metrics used to evaluate, for example, multi-classification and segmentation tasks. Together with this study, we developed a web-based tool that can be used to reverse-engineer the missing metrics of a given study. The tool is called *Medimetrics*[6] and is made publicly available online, and the source code is published on GitHub[7].

3.3.2 Reproducible and Comparable Results

Machine learning benchmarks and challenges can help bring more awareness to specific problems in medicine and also establish a common standard for medical machine learning systems. These events are a good opportunity to introduce a new audience of expert data scientists to a field that is unknown and can potentially have a real impact on someones life. Throughout the three years of this Ph.D. project, we organized several challenges held at well-established multimedia conferences and workshops. The challenges fall into one of two categories; challenges for GI image analysis and challenges for semen video analysis.

[6] https://medimetrics.no

[7] https://github.com/simula/medimetrics

3.3.2.1 Gastrointestinal Image Analysis Challenges

The first group of challenges were on automatic GI findings detection, where we held a total of six different competitions across four different venues. The first set of GI challenges were held at the MediaEval Multimedia Benchmark, which is a long-running workshop meant to challenge participants with solving different tasks using multimedia data. We organized the Medico Multimedia Task, which proposes several sub-tasks involving efficient detection of different findings in the GI tract. Overall, three editions of Medico were held at MediaEval, each with a different focus. First, in 2018, the main focus of Medico was to efficiently classify images collected from colonoscopies [111]. Second, in 2020, the main focus was to efficiently segment colon polyps [71]. Third, in 2021, the main focus was transparency in the development of colon polyp segmentation systems [55]. In addition to the GI-related challenges at MediaEval, we also organized a competition at ACM Multimedia in 2019 called BioMedia [59]. This was an extension of the task held in 2018 at MediaEval and provided a larger training and testing dataset. After organizing these challenges, we performed a meta-analysis study across the GI-related challenges held from 2017 to 2019. The results of this study showed that the best submissions improved year over year, showing that the field is progressing [70].

3.3.2.2 Semen Video Analysis Challenges

The second group of challenges were on automatic semen analysis using the VISEM dataset [50]. Here, we organized two a total of two challenges, one at the MediaEval Benchmark in 2019 [52] and one at ACM Multimedia in 2020 [59]. Both years, we proposed three different subtasks, one for both motility and morphology prediction, and one for automatic sperm tracking. The sperm motility task asked participants to predict the percentage of progressive, non-progressive, and immotile sperm in a given semen sample. This would account for every sperm contained in the video, making the sum of all prediction 100%. The sperm morphology task asked participants to predict the number of sperm that contain head defects, tail defects, and midpiece defects. Here, the model should only predict the sperms that contain defects, making each category independent. The sperm tracking task asked participants to track the individual sperm and calculate the highest and average speed of a sperm in a given sample. The participants were provided with a pre-partitioned version of the VISEM dataset, for which they were asked to train their model over three-fold cross-validation.

3.3.3 Explainable Artificial Intelligence

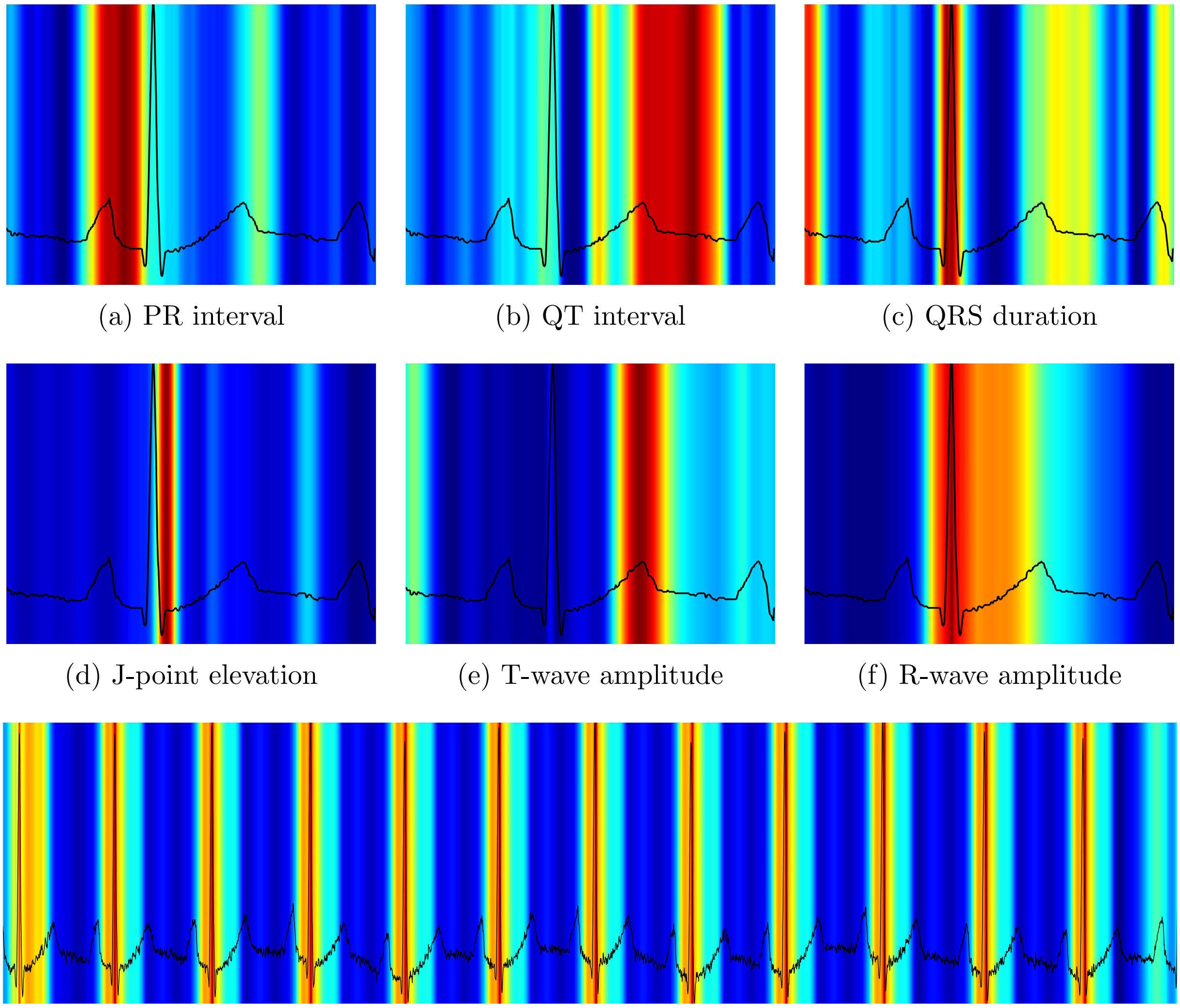
One part of making an AI system more transparent is providing explanations together with a prediction. Explainable AI is a field that has been gaining a lot of attention recently, especially in high-risk domains such as medicine and law. Researchers often cite a lack of trust in black-box models and that a better explanation of how a model makes a prediction would help discover biases and increase adoption [140]. Some models are inherently explainable, which means that one can directly infer how the input is associated with a produced output. These models are often categorized as white-box models and include algorithms such as decision trees and linear regression. Complex and difficult to explain models are often referred to as black-box models, which encompass the now popular deep neural networks. Explaining these models requires specialized methods that aim to interpret the model's inner workings to make an explanation that its end users can understand. There are several approaches to do this [129, 89, 3, 30], and the methods vary depending on the type of data used as input. A question may arise in that if complex models like deep neural networks work so well, why is it important to understand why they work? There are many reasons why explainable AI is essential for any domain, not just high-risk fields such as medicine. Blindly trusting models to perform the same in the real world as in a controlled experimental environment is irresponsible and could lead to many unforeseen consequences. In the following, we identify a few of the primary reasons why explainable AI has become such a booming area of research.

Model development Developing a high-performing model is difficult. This is especially true for complex models where we are unsure of which architectural and parameter choices lead to better results. Somewhat jokingly, deep learning development has been likened to randomly guessing different configurations until something works. As we do not know what features are learned from the data, it is difficult to assess what changes should be made to further improve the model. By understanding a model's internals, we can make more educated decisions based on the desired outcome. Furthermore, even if a model seems to perform well, it is essential to verify that its decisions are based on rational grounds. An *urban legend* in AI research describes a model trained to detect tanks in rough terrain. After the model was fully developed and showed an almost perfect performance score, the researchers realized that all images of tanks were taken on a cloudy day. In contrast, images not containing tanks were taken on a sunny day, meaning that the model had learned to rely on the lighting to make its decision. By visualizing the features detected by the model, one would be able to forgo such simple mistakes as we could verify that model does indeed focus on the object in question and not other artifacts.

Knowledge discovery As humans, we do not have a perfect view of the world. There are certain patterns and structures that are hard for us to discover and exploit for efficiency gain. Models such as deep neural networks automatically find and extract features that most efficiently lead them to their end goal. Sometimes, these may be unwanted artifacts, as previously described. Other times, the model may find unknown features that we humans have not yet discovered. By interpreting the features extracted by a model, we may gain new insights into what causes certain diseases or relationships in seemingly unstructured data. Through model interpretations, complex models may teach us some of the undiscovered laws of biology, physics, or chemistry.

Legal requirements As AI is applied to fields that may cause serious physical, mental, or financial harm, we need ways to assign responsibility to all affected parties. For example, if an autonomous car crashes into a pedestrian, who is to blame? The *right to an explanation* has also become a legal requirement in many countries. For example, in the United States, if a person is denied credit, they are legally entitled to an explanation of why they were denied. The right to an explanation has also been incorporated into the General Data Protection Regulation (GDPR), where businesses are legally required to give an explanation on why certain decisions were made. The legal aspects of AI are still in their infancy and will most likely be a long and rigorous process. Nevertheless, identifying why certain decisions are made will be an essential part of lawful judgment.

Ethics and transparency Explanations play a significant role in avoiding unintended biases towards certain demographics such as age, race, or sex. Based on the data used to train a model, biases that discriminate against certain demographics could potentially



(a) PR interval

(b) QT interval

(c) QRS duration

(d) J-point elevation

(e) T-wave amplitude

(f) R-wave amplitude

(g) Heart rate

Figure 3.10: Visualizations generated for the interval and amplitude prediction models. As we can see from the plots, the model learns to inspect the waves and and intervals that are related to the predicted variable.

occur. These issues also motivate for more transparency regarding how a model has trained apart from solely relying on explanations produced after a model is developed. However, it is important to mention that even though model explanation methods may give us some insight into why a model makes a certain prediction, it could also provide a false sense of security. Trust and confidence in a model should not only depend on whether the explanations fit the mental model of the user interpreting them. Explanations should always be interpreted in the context of their application, where the users understand what the explanation is based on and how it is produced.

3.3.3.1 Using Model Explanations for Knowledge Discovery

Using the ECG case study, we aimed to get a better understanding of how the model analyzes the ECG by explaining the predictions [61]. The motivation behind this was two-fold. First, we wanted to visualize how the deep neural network analyzes the ECG to predict the various waves and intervals in order to verify that it had learned the correct features. Visualizations were generated using a method based on the GradCAM technique [129], which was modified to work with the dimensionality constraints of the ECG. The results showed that the network does highlight the expected areas when making a prediction. For example, the QRS complex is highlighted when we predict QRS duration, and the end of the T-wave is delineated along with the beginning of the QRS complex for QT interval measurement. Figure 3.10 shows some example visualizations of the wave and interval predictions. The second motivation behind explaining the ECG model predictions was to understand what features that correlate to sex. As mentioned in Section 3.2.3, predicting the sex of an ECG is very difficult for even expert cardiologists, but our neural network was able to do this with ease. When visualizing the predictions for sex prediction, we discovered that the model used the downslope of the R-wave to determine sex. This finding was further verified through a set of logistic regression experiments and wave-blocking experiments, where the results showed that the R-wave is an essential feature in predicting the sex of an ECG.

3.3.3.2 Evaluating Explanations

As with any research contribution, we need a way of assessing the quality of said contribution against previous work. This is no different when it comes to evaluating machine learning model explanations. Attempts at coming to a standard explanation metric have been made, but there is still no quantitative metric used to measure whether one explanation is better than another. Measuring the quality of an explanation is quite different than evaluating predictive performance as the quality explanations may mean different things depending on the context. For example, in the context of polyp detection, a layman may find certain explanations more useful than a trained gastroenterologist that has more pre-existing knowledge. This adds a human component to the evaluation that may differ between fields and backgrounds. Furthermore, drawing a line where one measures the quality of the explanation or the quality of the model can be difficult. Using visual explanations as an example, one may postulate that if the explanation highlights areas of an image that correspond to a given class, like the tail of a cat when classifying cat images, that it is a good explanation. However, the explanation method is merely a proxy to explaining what the model is *looking* at, without any knowledge regarding what is present in the image. By this standard, if the explanation highlights nonsensical regions of an image, it may not be a fault of the explanation method but a fault of the model itself.

3.3.4 Lessons Learned

Diving into the current literature in medical applications of AI, we learned that several areas are lacking in the evaluation part of the research. This observation can perhaps be explained by a lack of experience in developing AI systems by the medical research community, but can also be attributed to the general level of confidentiality applied to the medical research. This secrecy extends across the entire pipeline of an AI system, from the data used to build the model, the specific methods used for analysis, and the metrics and methods used to evaluate and interpret the results.

The numerous benchmarks and challenges we organized taught us that there is immense interest in building machine learning models for medical applications among the computer science community. Given an open dataset and task, researchers with little experience within medical image and video analysis are able to produce high-quality models.

Our work on producing explanations for the ECG predictions showed that even domain experts could learn something new from the deep neural networks. What makes this work especially interesting is that we found that the doctors were continuously surprised about how the neural network performed its analysis, likening it to how experts in the filed perform their analysis. As computer scientists, this was very motivating as we could directly see the value that the explanation brought to the analysis.

3.4 Summary

This chapter presented our work in developing transparent AI systems within four areas of medicine. The development of medical AI systems can generally be split into three distinct steps; dataset development, model development, and model evaluation.

The data used to train and evaluate a model will substantially affect how it will perform in a real-world setting. Medical data is usually protected by laws and regulations, making it difficult to publish. In the context of AI systems, this is a problem as any research published on private data will not be reproducible. Furthermore, without having access to the underlying data used to train and evaluate a system, one can not assess the potential biases and shortcomings that may have affected the presented results. This motivated us to develop several public medical datasets to encourage transparent research and public benchmarks. Overall, we developed and published seven medical datasets across GI endoscopy [17, 69, 137], assisted reproductive technology [50], and mental health [67, 58].

To understand the intricacies of medical AI systems, we used our developed datasets to implement several AI systems using different medical case studies. First, we looked at analyzing data collected from the GI track to detect different findings such as lesions, anatomical landmarks, and instruments. Several different approaches were implemented, including methods for both standard classification [141] and segmentation [150]. Overall, our findings show that the systems are able to perform the given tasks with high accuracy. Second, we developed a system for automatically determine the quality of a human semen sample. The system used a deep neural network to analyze frames from a microscopic recording of semen to predict the motility and morphology of the sperm contained within [60]. Lastly, we implemented a system for detecting the various waves and intervals of a given ECG. The system used median and 10-second rhythm strips from standard 12-lead ECGs as input and showed comparable results to that of expert cardiologists [61].

Proper evaluation and testing of any AI system is essential before being deployed in a real-world environment. In medicine, this is especially important as biases and imprecision in a model could have fatal consequences. We examined the evaluation of medical AI systems from three perspectives; evaluation through quantitative metrics, evaluation through challenges and benchmarks, and explaining the predictions of complex deep neural networks. We looked at the current landscape of evaluating AI systems in medicine, where we discovered that a lot of studies present incomplete evaluation metrics of their system. In response to this, we selected five research papers and reverse-engineered the metrics to show how one may gain a different perspective of a model if presented with a more robust set of evaluation metrics [62]. The study was supported by a tool we developed that can automatically calculate missing binary classification metrics using the ones included in a given study. As a means to promote open and collaborative research within medical AI systems, we organized eight medical machine learning challenges across four different venues. Finally, we explored using explainable AI to understand the predictions of our ECG prediction model, which helped us discover new features in the prediction of sex for a give ECG [61].

Chapter 4

Conclusion

The conception of this thesis began with a recognition of the many transparency issues plaguing medical AI research. Although the initial plan for this work was to target explainability in medical AI applications, the transition to focus on transparency was merely an expansion on the subject matter. Throughout this research project, we published several medical datasets, performed extensive experiments across several medical application scenarios, looked at the current state of machine learning in different medial fields through AI-based challenges and reviews, and contributed to explainable methods in medical AI. Much of this work has been published at top-tier conferences and journals, and we are currently in the process of submitting several more.

This chapter concludes this thesis by recollecting the main research question and objectives, where we look back on the work done during this project and tie it back to the original ambitions. We present our plans for future work and discuss potential research directions that build on values of transparency in medical AI.

4.1 Main Contributions

This thesis presents our work on developing transparent AI systems in different medical domains. We developed and published several datasets, organized multiple medical AI challenges, benchmarked existing and novel AI-based methods, and looked at different methods of evaluating AI systems. The main contributions of this thesis are supported by publications in top-tier conferences or journals. In the following, we detail the contributions in relation to the research question and research objectives defined in Section 1.3.

Contributions to Objective 1 Objective one is supported by the collection and publication of several medical datasets in the field of gastroenterology [17, 137], assisted reproductive technology [50], mental health [67, 58], and cardiology [152]. Each dataset was developed in collaboration with health experts within each respective field and is made publicly available and free to use for research and educational purposes. The datasets were used as a basis for most of the research presented in this thesis and have been used to organize several challenges and benchmarks [56, 52, 59, 54, 71, 55]. Each dataset is accompanied by a research article published in a top-tier conference or journal, where we also include a set of benchmark experiments and possible research directions. The contribution to transparency in medical AI systems comes from the open and publicly available datasets published under non-restrictive licenses that can be used to develop and benchmark machine learning models. We also include information on how the data was collected and verified by the medical experts, including details on the equipment used to collect the data.

Contributions to Objective 2 Objective two is supported by our work on developing and training machine learning models to support medical doctors and clinicians in performing different tasks within medicine. This includes methods for automatically determining the quality of a given semen sample [60], predicting the sex, waves, and intervals of a standard 12-lead ECG [61], detecting disease and other findings in the GI tract [98, 75], and recognizing disorders such as ADHD or schizophrenia from activity data [58, 67, 37]. The contribution to transparency comes from the implementation descriptions, where we detail how the models are trained, which hyperparameters are used, and usually include an open implementation that is freely available online.

Contributions to Objective 3 Objective three is supported by our work on through evaluation and post validation of machine learning models after training. This work includes using explainable AI to discover new features related to the relationship between sex and ECG signals [61], exploring how evaluation metrics may be used to give an incomplete view of a models predictive performance [62, 121], and the organization of several machine learning challenges for benchmarking and comparability purposes [56, 55, 52, 59, 54, 71, 111]. Openness and accurate reporting of evaluation methods is a key attribute of transparency within the evaluation of a AI system. Therefore, we aimed to show both the negative and positive sides of our developed systems, reporting all relevant metrics and measuring performance against several benchmarks.

Additional Contributions We also contributed to areas that fall outside the aforementioned research question and objectives. Still, these contributions follow the principles of transparency and helped us better understand the requirements for the medical-based case studies. We researched and developed systems for automatically detecting and clipping events from soccer games [101]. These systems were built on open datasets and gave us experience with analyzing video data using deep learning. We also used model interpretation methods to generate explanations over the time-dimension to explain the predictions of the deep neural network. We looked at how machine learning can be used to predict latency in mobile networks, where we used both traditional machine learning methods and deep learning. Through this work we gained experience in analyzing large time-series data. We collected and published a dataset for performing sentiment analysis on disaster-related images and are currently organizing a challenge using this dataset [49, 168]. We collected and published a dataset on developing emotional intelligence machines using the video game Super Mario Bros. as an initial use case for building emotional intelligent machines [143]. This work spawned from the idea of incorporating human ethics into machine learning algorithms. We collected and published a dataset for analyzing sports activity in relation to other personal attributes like sleep and weight [151]. We collected and published a dataset containing activity data from everyday tasks like brushing teeth or watching television together with audio recordings [39]. We collected and published a dataset for predicting cloud fractional cover using satellite observations.

Through our work developing medical AI systems, we gained a better understanding of the problems medical doctors face in their everyday work. Collecting and publishing datasets for objective 1 gave us a better understanding of the intricacies of making medical data public and the potential obstacles that make this difficult. Developing AI systems for different tasks within medicine for objective 2 helped us recognize the issues and requirements that medical doctors and clinicians face every day. Researching the evaluation of AI systems in medicine and organizing several machine learning challenges for objective 3 gave us the experience of realizing that there is still a lot of work to be done in medical AI research, and a good way to accelerate this process is through public events that encourage collaboration, reproducibility, and comparability.

4.2 Future Work

There are several future directions we want to explore that we did not have time to during this Ph.D. project, some of which have been mentioned throughout this thesis. First, for the assisted reproductive technology case study, we would like to further explore the analysis of embryo data in conjunction with the sperm data to find what features between the two produce a healthy child. We would also like to extend the sperm dataset, VISEM, with tracking data to add an additional challenge of counting and measuring attributes of specific sperm contained within a video. Furthermore, we would like to publish a dataset containing time-lapse videos of human embryos. This is currently a popular research direction, and not many public datasets exist. Analyzing both the sperm and embryo data together would be an excellent scenario to apply model explanation methods to learn what features contribute to a healthy child.

For the ECG case study, we would like to apply our models on abnormal ECGs to detect to diagnose different diseases and disorders. Furthermore, we would like to research how to incorporate different parameters into the analysis, like adding information about the person's age and sex. Another aspect of ECG analysis we would like to explore is ECG reconstruction from a given set of parameters. For example, given a set of human traits like age, sex, BMI, and information about the person's genes, can one reconstruct that person's heart rhythm? This could potentially help us emulate the heart rate of a person under different scenarios.

For the mental health case study, we would like to apply more advanced time-series analysis to the activity data contained in the published schizophrenia and ADHD datasets. Furthermore, we want to combine the data from the two datasets as mentioned earlier and one containing data from patients with depression to perform a more comprehensive study on mental health analysis. This is also an area where we were not able to host a challenge or benchmark, something we wish to do in the future.

Another area we would like to explore more thoroughly is the evaluation of model interpretation and explanation methods. Unlike standard predictive performance metrics, an explanation is subjective and dependant on the context for which they are presented. Having quantitative metrics to measure the quality of an explanation against others would be very useful in developing new explanation methods. We are also working on a survey on evaluating different explanation methods on domain experts, where the current case study is explanations for a model that detects polyps in GI-related images. We want to expand this study to evaluate how different medical fields can use model explanation methods.

4.3 Final Remarks

This thesis presented our work on making machine learning-based research in medicine more open and transparent. I think machine learning has and will continue to profoundly impact our everyday lives, medicine being no exception. Applying these methods without having any insight into the development and evaluation of the proposed models is irresponsible and could potentially have fatal consequences. I hope that this thesis inspires more transparent research and wish for an overall more open scientific process in medical AI going forward.

Chapter 5

Papers and Author's Contributions

In this chapter, we list each paper published during the span of this Ph.D. project, and discuss my contributions to each paper and how it relates to the research objectives defined in Section 1.3. The articles themselves have been included in the appendix of this thesis.

5.1 Paper I - HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy

Auhtors Hanna Borgli, Vajira L 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, and 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.

Candidate contributions Steven contributed to the dataset development by cleaning, organizing, and preparing the online repositories (GitHub and OSF). This includes compiling data from different sources, ensuring that the dataset contains a minimal number of duplicates, creating image/video annotation files, and writing various splits to aid in the data preparation. He contributed to the initial experiments using a ResNet 50 implemented in TensorFlow and contributed to analyzing the unlabeled clustering experiments.

For the writing of the paper, he contributed by drafting and reviewing all parts of the article, but with a particular focus on the technical validation section. He also contributed by making Figure 1, Figure 8, and a figure that was used in the initial submission but excluded from the final manuscript.

Thesis objectives Objective 1.

5.2 Paper II - Medico multimedia task at mediaeval 2018

Auhtors Konstantin Pogorelov, Michael A Riegler, Pål Halvorsen, Steven Hicks, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange.

Abstract The Medico: Multimedia for Medicine Task, running for the second time as part of MediaEval 2018, focuses on detecting abnormalities, diseases, anatomical landmarks and other findings in images captured by medical devices in the gastrointestinal tract. The task is described, including the use case and its challenges, the dataset with ground truth, the required participant runs and the evaluation metrics.

Published Working Notes Proceedings of the MediaEval 2018 Workshop.

Candidate contributions Steven presented the task at the workshop and contributed to the paper by drafting and revising of the manuscript.

Thesis objectives Objective 3.

5.3 Paper III - Machine learning-based analysis of sperm videos and participant data for male fertility prediction

Auhtors Steven A Hicks, Jorunn M Andersen, Oliwia Witczak, Vajira L Thambawita, Pål Halvorsen, Hugo L Hammer, Trine B Haugen, and 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.

Candidate contributions Steven contributed to the conceptualization, design, and development of the study. He performed several deep learning-based experiments using optical flow and raw frame analysis. The experiments used a variety of different data, including patient information, sensor data, and image data. For the manuscript, he contributed to all parts of the paper.

Thesis objectives Objective 2.

5.4 Paper IV - Medico multimedia task at mediaeval 2020: Automatic polyp segmentation

Auhtors Debesh Jha, Steven A Hicks, Krister Emanuelsen, Håvard D Johansen, Dag Johansen, Thomas de Lange, Michael A Riegler, and Pål Halvorsen.

Abstract Colorectal cancer is the third most common cause of cancer worldwide. According to Global cancer statistics 2018, the incidence of colorectal cancer is increasing in both developing and developed countries. Early detection of colon anomalies such as polyps is important for cancer prevention, and automatic polyp segmentation can play a crucial role for this. Regardless of the recent advancement in early detection and treatment options, the estimated polyp miss rate is still around 20%. Support via an automated computer-aided diagnosis system could be one of the potential solutions for the overlooked polyps. Such detection systems can help low-cost design solutions and save doctors time, which they could for example use to perform more patient examinations. In this paper, we introduce the 2020 Medico challenge, provide some information on related work and the dataset, describe the task and evaluation metrics, and discuss the necessity of organizing the Medico challenge.

Published Working Notes Proceedings of the MediaEval 2020 Workshop.

Candidate contributions Steven contributed to the organization of the challenge and presented it at the workshop. He also contributed to drafting and revising the manuscript.

Thesis objectives Objective 3.

5.5 Paper V - Visem: A multimodal video dataset of human spermatozoa

Auhtors Trine B Haugen, Steven A Hicks, Jorunn M Andersen, Oliwia Witczak, Hugo L Hammer, Hanna Borgli (Rune Borgli), Pål Halvorsen, and Michael A Riegler.

Abstract Real multimedia datasets that contain more than just images or text are rare. Even more so are open multimedia datasets in medicine. Often, clinically related datasets only consist of image or videos. In this paper, we present a dataset that is novel in two ways. Firstly, it is a multi-modal dataset containing different data sources such as videos, biological analysis data, and participant data. Secondly, it is the first dataset of that kind in the field of human reproduction. It consists of anonymized data from 85 different participants. We hope this dataset paper will inspire people to apply their knowledge in this important field, generate shareable results in the domain, and ultimately improve human infertility investigation and treatment.

Published The ACM Multimedia Systems Conference, 2019.

Candidate contributions Steven performed all the experiments and contributed to drafting and revising the manuscript.

Thesis objectives Objective 1.

5.6 Paper VI - The EndoTect 2020 Challenge: Evaluation and Comparison of Classification, Segmentation and Inference Time for Endoscopy

Auhtors Steven Alexander Hicks, Debesh Jha, Vajira L Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler.

Abstract The EndoTect challenge at the International Conference on Pattern Recognition 2020 aims to motivate the development of algorithms that aid medical experts in finding anomalies that commonly occur in the gastrointestinal tract. Using HyperKvasir, a large dataset containing images taken from several endoscopies, the participants competed in three tasks. Each task focuses on a specific requirement for making it useful in a real-world medical scenario. The tasks are (i) high classification performance in terms of prediction accuracy, (ii) efficient classification measured by the number of images classified per second, and (iii) pixel-level segmentation of specific anomalies. Hopefully, this can motivate different computer science researchers to help benchmark a crucial component of a future computer-aided diagnosis system, which in turn, could potentially save human lives.

Published ICPR International Workshops and Challenges, 2020.

Candidate contributions Steven was responsible for being the main lead of the challenge. He prepared the development and testing datasets, communicated with the participants, evaluated the submissions, presented the task at the workshop, and handeled the logistics related to organizing the challenge. He also contributed to drafting and revising the overview paper.

Thesis objectives Objective 3.

5.7 Paper VII - Kvasir-Capsule, a video capsule endoscopy dataset

Auhtors Pia H Smedsrud, Vajira L Thambawita, Steven A Hicks, Henrik L Gjestang, Oda O 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.

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.

Candidate contributions Steven contributed to the dataset development by organizing, structuring, cleaning, and preparing/uploading the data to online repositories (OSF and GitHub). He contributed to the analysis, conception, and interpretation of the baseline experiments, where he made the scripts that generated the confusion matrices and evaluation results. Concerning the paper writing, he drafted and reviewed all parts of the paper, but with a particular focus on the technical aspects.

Thesis objectives Objective 1.

5.8 Paper VIII - Deep Learning Based Disease Detection Using Domain Specific Transfer Learning

Auhtors Steven Alexander Hicks, Pia H Smedsrud, Pål Halvorsen, and Michael A Riegler.

Abstract In this paper, we present our approach for the Medico Multimedia Task as part of the MediaEval 2018 Benchmark. Our method is based on convolutional neural networks, where we compare how fine-tuning, in the context of transfer learning, from different source domains (general versus medical domain) affect classification performance. The preliminary results show that fine-tuning models trained on large and diverse datasets is favorable, even when the model's source domain has little to no resemblance to the new target.

Published Working Notes Proceedings of the MediaEval 2018 Workshop.

Candidate contributions Steven contributed to the conceptualization and design of the study. He aided i the analysis and interpretation of ther data and results, performed the machine learning experiments, and contributed to the drafting and revision of the manuscript.

Thesis objectives Objective 2.

5.9 Paper IX - ACM Multimedia BioMedia 2020 Grand Challenge Overview

Auhtors Steven A Hicks, Vajira L 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.

Candidate contributions Steven was responsible for organizing the challenge and communicating with the participants. He prepared the development and testing datasets and evaluated the results for the submitted runs. He created the website (https://biomediachallenge.com created the public repositories (GitHub), made the video presentations, and chaired the session at ACM MultiMedia 2020. He drafted, revised, and submitted the manuscript.

Thesis objectives Objective 3.

5.10 Paper X - Unsupervised preprocessing to improve generalisation for medical image classification

Auhtors Mathias Kirkerød, Rune Borgli, Vajira L Thambawita, Steven Hicks, Michael A Riegler, and 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 approximately7.

Published Proceedings of the 13th International Symposium on Medical Information and Communication Technology (ISMICT), 2019.

Candidate contributions Steven contributed to the conception, design, and drafting, and revising of the manuscript.

Thesis objectives Objective 2.

5.11 Paper XI - Medico Multimedia Task at MediaEval 2019

Auhtors Steven Alexander Hicks, Pål Halvorsen, Trine B Haugen, Jorunn M Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L Hammer, Duc-Tien Dang-Nguyen, Mathias Lux, and Michael A Riegler.

Abstract The Medico: Multimedia for Medicine Task is running for the third time as part of MediaEval 2019. This year, we have changed the task from anomaly detection in images of the gastrointestinal tract to focus on the automatic prediction of human semen quality based on videos. The purpose of this task is to aid in the assessment of male reproductive health by providing a quick and consisted method of analyzing human semen. In this paper, we describe the task in detail, give a brief description of the provided dataset, and discuss the evaluation process and the metrics used to rank the submissions of the participants.

Published Working Notes Proceedings of the MediaEval 2019 Workshop.

Candidate contributions Steven was responsible for being the main lead of the challenge. He prepared the development and testing datasets, communicated with the participants, evaluated the submissions, presented the task at the workshop, and handeled the logistics related to organizing the challenge. He also contributed to drafting and revising the paper.

Thesis objectives Objective 3.

5.12 Paper XII - Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy

Auhtors Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira L Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, Dag Johansen, and 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.

Candidate contributions Steven made the Simula dataset page and revised the manuscript.

Thesis objectives Objective 1.

5.13 Paper XIII - Explaining deep neural networks for knowledge discovery in electrocardiogram analysis

Auhtors Steven A Hicks, Jonas L Isaksen, Vajira L 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.

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.

Candidate contributions Steven contributed to the conception, implementation, and design of the study. He wrote the pipeline for which the different models were trained and evaluated. He was part of the development of the proposed neural network and implemented the visualization methods described in the paper. Furthermore, he also did many preliminary experiments that did not end up in the final paper but were used to guide the study. He contributed to the drafting and revision of the manuscript.

Thesis objectives Objective 2 and Objective 3.

5.14 Paper XIV - Deep learning for automatic generation of endoscopy reports

Auhtors Steven Hicks, Pia H Smedsrud, Michael A Riegler, Thomas de Lange, Andreas Petlund, Sigrun L Eskeland, Konstantin Pogorelov, Peter T Schmidt, and Pål Halvorsen.

Abstract In an effort to achieve consistent, high-quality endoscopy reports, the World Endoscopy Organization (WEO) recommends using both a minimal standard for reporting (MSR) and a minimal standard terminology (MST) for describing anatomical landmarks and mucosal lesions found in the gastrointestinal (GI) tract. But creating reports which adhere to these standards is often time-consuming, and with opinions varying vastly between endoscopists, there is still a large issue of inconsistencies found in endoscopy reports worldwide. Methods within Artificial Intelligence (AI), like neural networks, have a proven capability of automatically detecting GI mucosal lesions and has shown much potential in eliminating the inherent human variation of GI disease diagnosis. However, deep neural networks are commonly known as black boxes, where the underlying decision process which led to a conclusion is relatively unknown, especially among end-users. Thus, we aim to open this black box and use the gained knowledge to develop a technology for automatically generating standardized endoscopy reports.

Published Gastrointestinal Endoscopy, 2019.

Candidate contributions Steven built the system that the study is based on and performed the experiments. He presented the work at the conference and contributed to drafting and revising the paper.

Thesis objectives Objective 3.

5.15 Paper XV - One-dimensional convolutional neural networks on motor activity measurements in detection of depression

Auhtors Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen, and Michael A Riegler.

Abstract Nowadays, it has become possible to measure different human activities using wearable devices. Besides measuring the number of daily steps or calories burned, these datasets have much more potential since different activity levels are also collected. Such data would be helpful in the field of psychology because it can relate to various mental health issues such as changes in mood and stress. In this paper, we present a machine learning approach to detect depression using a dataset with motor activity recordings of one group of people with depression and one group without, i.e., the condition group includes 23 unipolar and bipolar persons, and the control group includes 32 persons without depression. We use convolutional neural networks to classify the depressed and non-depressed patients. Moreover, different levels of depression were classified. Finally, we trained a model that predicts Montgomery-Åsberg Depression Rating Scale scores. We achieved an average F1-score of 0.70 for detecting the control and condition groups. The mean squared error for score prediction was approximately 4.0.

Published Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care.

Candidate contributions Steven contributed to drafting and revising the paper.

Thesis objectives Objective 2.

5.16 Paper XVI - ACM MM BioMedia 2019 Grand Challenge Overview

Auhtors Steven Hicks, Michael A Riegler, Pia H Smedsrud, Trine B Haugen, Kristin Ranheim Randel, Konstantin Pogorelov, Håkon K Stensland, Duc-Tien Dang-Nguyen, Mathias Lux, Andreas Petlund, Thomas de Lange, Peter T Schmidt, and Pål Halvorsen.

Abstract The BioMedia 2019 ACM Multimedia Grand Challenge is the first in a series of competitions focusing on the use of multimedia for different medical use-cases. In this year's challenge, the participants are asked to develop efficient algorithms which automatically detect a variety of findings commonly identified in the gastrointestinal tract (a part of the human digestive system). The purpose of this task is to develop methods to aid medical doctors performing routine endoscopy inspections of the GI tract. In this paper, we give a detailed description of the four different tasks of this year's challenge, present the datasets used for training and testing, and discuss how each submission is evaluated both qualitatively and quantitatively.

Published Proceedings of the 27th ACM International Conference on Multimedia.

Candidate contributions Steven was responsible for being the main lead of the challenge. He prepared the development and testing datasets, communicated with the participants, and evaluated the submissions. He also contributed to drafting and revising the paper.

Thesis objectives Objective 3.

5.17 Paper XVII - PSYKOSE: A Motor Activity Database of Patients with Schizophrenia

Auhtors Petter Jakobsen, Enrique Garcia-Ceja, Lena Antonsen Stabell, Ketil Joachim Oedegaard, Jan Øystein Berle, Vajira L Thambawita, Steven Alexander Hicks, Pål Halvorsen, Ole Bernt Fasmer, and Michael A 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 Proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS).

Candidate contributions Steven made the Simula dataset website for the dataset and contributed to drafting and revising the manuscript

Thesis objectives Objective 1.

5.18 Paper XVIII - Divergentnets: Medical image segmentation by network ensemble

Auhtors Vajira L Thambawita, Steven A Hicks, Pål Halvorsen, and Michael A Riegler.

Abstract Detection of colon polyps has become a trending topic in the intersecting fields of machine learning and 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 advantage of being more accurate than per-frame classification or object detection as it can show the affected area in greater detail. For our contribution to the EndoCV 2021 segmentation challenge, we propose two separate approaches. First, a segmentation model named TriUNet composed of three separate UNet models. Second, we combine TriUNet with an ensemble of well-known segmentation models, namely UNet++, FPN, DeepLabv3, and DeepLabv3+, into a model called DivergentNets to produce more generalizable medical image segmentation masks. In addition, we propose a modified Dice loss that calculates loss only for a single class when performing multi-class segmentation, forcing the model to focus on what is most important. Overall, the proposed methods achieved the best average scores for each respective round in the challenge, with TriUNet being the winning model in Round I and DivergentNets being the winning model in Round II of the segmentation generalization challenge at EndoCV 2021. The implementation of our approach is made publicly available on GitHub.

Published Proceedings of EndoCV 2021.

Candidate contributions Steven contributed to the conception, development, and design of the work presented in this manuscript. He was part of the development models used in the study. He also contributed by drafting and revising the manuscript and presented the work at the 3rd International Endoscopy Computer Vision Challenge and Workshop.

Thesis objectives Objective 2.

5.19 Paper XIX - Pyramid-Focus-Augmentation: Medical Image Segmentation with Step-Wise Focus

Auhtors Vajira L Thambawita, Steven Hicks, Pål Halvorsen, and 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 Working Notes Proceedings of the MediaEval 2020 Workshop.

Candidate contributions Steven contributed to the conception of pyramid-focus-augmentation, drafting, and revising the manuscript.

Thesis objectives Objective 2.

5.20 Paper XX - Data Augmentation Using Generative Adversarial Networks for Creating Realistic Artificial Colon Polyp Images: Validation Study by Endoscopists

Auhtors Vajira L Thambawita, Inga Strümke, Steven Hicks, Michael A Riegler, Pål Halvorsen, and 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. Method: Using training data from the Kvasir dataset, a large colonoscopy dataset, an image inpainting GAN is used to generate artificial colon polyp images. The GAN is pre-trained with colon images and fine-tuned to generate synthetic polyps using colon images as input. The discriminator of the GAN is used to assess the global and local quality of generated images, in addition to discriminating real from generated. The quality of the generated images is evaluated by 2 expert endoscopists, 3 non-expert endoscopists, and 3 internal medicine residents. The experience of the physicians ranges from 0 to 20 years. Five synthesized and five real images are selected for the evaluation. For each image, the physicians assessed whether the polyp appeared real or generated on a scale from 1-10. Results: To measure the agreement among the raters, we calculate Fleiss' kappa for all questions regarding visual appearance across all participants. For all questions, over all, only generated and only real instances, respectively, the Fleiss kappa values are (0.0352, 0.0206, 0.0347) with p-values of (0.00034, 0.176, 0.00909). Similarly, the Fleiss kappa values for the question "Does the polyp appear generated?" are (0.0115, -0.0159, -0.0222). Limiting the included responses to only our two gastroenterologists, the Fleiss' kappa reduces to Cohen's kappa, and the respective values are (-0.235, -0.316, -0.282) with p-values (0.108, 0.193, 0.208). Landis and Koch (1977) provide guidelines for interpreting Fleiss' kappa, and according to these, values in the range 0.01-0.2 indicate only slight agreement between the raters. Moreover, we observe higher reported confidences on generated polyps than real ones. We clearly see that the participants do not find a strong agreement for real or generated, even not the most experienced gastroenterologists. Conclusion: We develop and validate a GAN generating high-quality synthetic polyp images. Our evaluation by medical experts indicates only little assessors agreement, even among the most experienced gastroenterologists. We also observe higher reported confidences on generated polyps than real ones. This does not mean that generated polyps are indistinguishable from real ones, but that they share visual and anatomical properties. These promising results show GANs could contribute synthetic data for training and unrestricted sharing.

Published Gastrointestinal Endoscopy, 2021.

Candidate contributions Steven contributed to the testing of the study, and drafting and revision of the abstract.

Thesis objectives Objective 2.

5.21 Paper XXI - Impact of Image Resolution on Convolutional Neural Networks Performance in Gastrointestinal Endoscopy

Auhtors Vajira L Thambawita, Steven Hicks, Inga Strümke, Michael A Riegler, Pål Halvorsen, and 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. This study examines the changes in the performance of CNNs when trained with different resolutions. For all experiments, we rely on the Kvasir data set, consisting of 10,662 GI images from 23 different findings. We evaluate two state-of-the-art CNN models, ResNet-152 and DenseNet-161, for classification under quality distortions with image resolutions for training and testing ranging from 32×32 to 512×512 pixels as shown in Figure 1. For training the models transfer learning is performed with ImageNet weights. The model performance is evaluated using two-fold cross-validation and F1-score, MCC, precision, and sensitivity as metrics. Increased performance was observed with higher image resolution for all findings in the data set. Lower resolution has a significantly lower performance with an MCC of 0.34 for the lowest and 0.9 for the highest. Table 1 shows the evaluation results in terms of precision, sensitivity, F1-score and MCC for the evaluated ResNet-152 and DenseNet-161 models. The presented numbers are the average over both folds in the cross-validation. Increasing the resolution leads to increased performance measured in almost all metrics. There is a slight decrease in sensitivity for the highest resolution, but taking MCC into account, there is still an overall improvement. For both CNNs, we observe the same behavior. Different

image resolutions and their effect on CNNs are explored. We show that image resolution has a clear influence on the performance which calls for standards in the field in the future. Currently, CNNs usually operate on low to mid-level resolutions. Higher resolution data sets might require new methods, architectures and hardware. As hardware improvements and algorithmic advances continue to occur, developing deep learning applications for endoscopy at higher image resolutions becomes increasingly feasible. Nevertheless, although the full potential of high-resolution data sets might not be exploitable yet, it is evidently important to collect data with the highest resolution possible.

Published Gastrointestinal Endoscopy, 2021.

Candidate contributions Steven contributed to drafting and revising the abstract.

Thesis objectives Objective 2.

5.22 Paper XXII - On evaluation metrics for medical applications of artificial intelligence

Auhtors Steven Hicks, Inga Strümke, Vajira L Thambawita, Malek Hammou, Pål Halvorsen, Michael A Riegler, and 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.

Candidate contributions Steven contributed to the conception, design, analysis, and development of the work presented in this paper. He developed the initial prototype of the app (medimetrics) and led the development of the final version. He was part of the selections of papers that were to be used for the main analysis in this paper. He also contributed by drafting and revising the manuscript.

Thesis objectives Objective 3.

5.23 Paper XXIII - Using Deep Learning to Predict Motility and Morphology of Human Sperm

Auhtors Steven Alexander Hicks, Trine B Haugen, Pål Halvorsen, and Michael A Riegler.

Abstract In the Medico Task 2019, the main focus is to predict sperm quality based on videos and other related data. In this paper, we present the approach of team LesCats which is based on deep convolution neural networks, where we experiment with different data preprocessing methods to predict the morphology and motility of human sperm. The achieved results show that deep learning is a promising method for human sperm analysis. Out best method achieves a mean absolute error of 8.962 for the motility task and a mean absolute error of 5.303 for the morphology task.

Published Working Notes Proceedings of the MediaEval 2019 Workshop.

Candidate contributions Steven contributed to designing and performing the experiments, and by drafting and rivising the manuscript.

Thesis objectives Objective 2.

5.24 Paper XXIV - Predicting Sperm Motility and Morphology Using Deep Learning and Handcrafted Features

Auhtors Steven Alexander Hicks, Pål Halvorsen, Trine B Haugen, Jorunn M Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L Hammer, Duc-Tien Dang-Nguyen, Mathias Lux, and Michael A Riegler.

Abstract This paper presents the approach proposed by the organizer team (SimulaMet) for MediaEval 2019 Multimedia for Medicine: The Medico Task. The approach uses a data preparation method which is based on global features extracted from multiple frames within each video and then combines this with information about the patient in order to create a compressed representation of each video. The goal is to create a less hardware expensive data representation that still retains the temporal information of the video and related patient data. Overall, the results need some improvement before being a viable option for clinical use.

Published Working Notes Proceedings of the MediaEval 2019 Workshop.

Candidate contributions Steven contributed to designing and performing the experiments, and by drafting and rivising the manuscript.

Thesis objectives Objective 2.

5.25 Paper XXV - Artificial intelligence in the fertility clinic: status, pitfalls and possibilities

Auhtors Michael A Riegler, Mette H Stensen, Oliwia Witczak, Jorunn M Andersen, Steven Hicks, Hugo L Hammer, Erwan Delbarre, Pål Halvorsen, Anis Yazidi, Nicolai Holst, and Trine B Haugen.

Abstract In recent years, the amount of data produced in the field of ART has increased exponentially. The diversity of data is large, ranging from videos to tabular data. At the same time, artificial intelligence (AI) is progressively used in medical practice and may become a promising tool to improve success rates with ART. AI models may compensate for the lack of objectivity in several critical procedures in fertility clinics, especially embryo and sperm assessments. Various models have been developed, and even though several of them show promising performance, there are still many challenges to overcome. In this review, we present recent research on AI in the context of ART. We discuss the strengths and weaknesses of the presented methods, especially regarding clinical relevance. We also address the pitfalls hampering successful use of AI in the clinic and discuss future possibilities and important aspects to make AI truly useful for ART.

Published Human Reproduction, 2021.

Candidate contributions Steven contributed to the tables, figures, literature review, writing and revision of the text and tables.

Thesis objectives Objective 3.

5.26 Paper XXVI - Assessment of sperm motility according to WHO classification using convolutional neural networks

Auhtors Trine B Haugen, Steven Hicks, Oliwia Witczak, Jorunn M Andersen, Lars Björndahl, and Michael A Riegler.

Abstract Manual sperm motility assessment according to WHO guidelines is regarded as the gold standard. To obtain reliable and reproducible results, comprehensive training is essential as well as running internal and external quality control. Prediction based on artificial intelligence can potentially transfer human-level performance into models that perform the task faster and can avoid human assessor variations. CNNs have been groundbreaking in image processing. To develop AI models with high predictive power, the data set used should be of high quality and sperm motility assessment based on WHO guidelines.

Published Human Reproduction, 2021.

Candidate contributions Steven designed and performed the deep learning experiments, and contributed to drafting and revising the manuscript.

Thesis objectives Objective 2.

5.27 Paper XXVII - SinGAN-Seg: Synthetic Training Data Generation for Medical Image Segmentation

Auhtors Vajira L Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and 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.

Candidate contributions Steven contributed to drafting and revising the manuscript.

Thesis objectives Objective 1 and Objective 2.

5.28 Paper XXVIII - HYPERAKTIV: An Activity Dataset from Patients with Attention-Deficit/Hyperacti Disorder (ADHD)

Auhtors Steven A Hicks, Andrea Stautland, Ole Bernt Fasmer, Wenche Førland, Hugo L Hammer, Pål Halvorsen, Kristin Mjeldheim, Ketil Joachim Oedegaard, Berge Osnes, Vigdis Elin Giæver Syrstad, Michael A Riegler, and Petter Jakobsen.

Abstract Machine learning research within healthcare frequently lacks the public data needed to be fully reproducible and comparable. Datasets are often restricted due to privacy concerns and legal requirements that come with patient-related data. Consequentially, many algorithms and models get published on the same topic without a standard benchmark to measure against. Therefore, this paper presents HYPERAKTIV, a public dataset containing health, activity, and heart rate data from adult patients diagnosed with attention deficit hyperactivity disorder, better known as ADHD. The dataset consists of data collected from 51 patients with ADHD and 52 clinical controls. In addition to the activity and heart rate data, we also include a series of patient attributes such as their age, sex, and information about their mental state, as well as output data from a computerized neuropsychological test. Together with the presented dataset, we also provide baseline experiments using traditional machine learning algorithms to predict ADHD based on the included activity data. We hope that this dataset can be used as a starting point for computer scientists who want to contribute to the field of mental health, and as a common benchmark for future work in ADHD analysis.

Published The ACM Multimedia Systems Conference, 2021.

Candidate contributions Steven contributed by performing the machine learning experiments, presenting the study at the conference, and drafting and revising the manuscript.

Thesis objectives Objective 1.

5.29 Paper XXIX - A self-learning teacher-student framework for gastrointestinal image classification

Auhtors Henrik L Gjestang, Steven A Hicks, Vajira L Thambawita, Pål Halvorsen, and 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 F1-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 Proceedings of the IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS).

Candidate contributions Steven contributed to the conception and design of the experiments, and contributed to the drafting and revision of the manuscript. He also presented the paper at CBMS.

Thesis objectives Objective 2.

5.30 Paper XXX - DeepSynthBody: the beginning of the end for data deficiency in medicine

Auhtors Vajira L Thambawita, Steven A Hicks, Jonas L Isaksen, Mette H Stensen, Trine B Haugen, Jørgen K Kanters, Sravanthi Parasa, Thomas de Lange, Håvard D Johansen, Dag Johansen, Hugo L Hammer, Pål Halvorsen, and 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 Proceedings of the International Conference on Applied Artificial Intelligence (ICAPAI), 2021.

Candidate contributions Steven contributed to the conception of DeepSynthBody. He contributed to drafting and revising the manuscript. Also, he came up with the name.

Thesis objectives Objective 1.

5.31 Paper XXXI - DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine

Auhtors Vajira L 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.

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.

Accepted Nature Scientific Reports, 2021.

Candidate contributions Steven contributed in the inception and discussion of the experiments and work. He contributed to drafting and revising the manuscript.

Thesis objectives Objective 1.

5.32 Paper XXXII - Vid2Pix-A Framework for Generating High-Quality Synthetic Videos

Auhtors Oda O Nedrejord, Vajira L Thambawita, Steven A Hicks, Pål Halvorsen, and 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 Proceedings of the IEEE International Symposium on Multimedia (ISM), 2020.

Candidate contributions Steven contributed to the conception, design, discussion, and analysis of the models and results presented in the paper. He also contributed to drafting and revising the manuscript.

Thesis objectives Objective 2.

5.33 Paper XXXIII - Big data is not always better-prediction of live birth using machine learning on time-lapse videos of human embryos

Auhtors Steven Hicks, Trine B Haugen, Mario Iliceto, Hugo L Hammer, Jorunn M Andersen, Oliwia Witczak, Michael A Riegler, and Mette H Stensen.

Abstract Time-lapse technology is considered an exceptional tool when observing the dynamic processes of early embryonic development. However, there is not enough evidence to conclude that the introduction of this technology has improved live birth rates after ART. Machine learning has proven its capability in analyzing images at a level above many humans and may uncover unseen patterns of predictive value when assessing embryos from time-lapse videos. Few studies using ML have been done. However, it is not known whether these methods may be used to predict live births.

Published Human Reproduction, 2020.

Candidate contributions Steven contributed to the experiments and to drafting and revising the manuscript.

Thesis objectives Objective 2.

5.34 Paper XXXIV - Artificial intelligence as a tool in predicting sperm motility and morphology

Auhtors Michael A Riegler, Jorunn M Andersen, Hugo L Hammer, Steven Hicks, Oliwia Witczak, and Trine B Haugen.

Abstract Although computer-aided sperm analysis (CASA) has been available for several decades, manual semen analysis according to WHO guidelines is still regarded as the gold standard. The assessment of sperm motility by CASA systems is rapidly performed, however, the tracking for spermatozoa in fresh semen is prone to error, and results may differ from manual analysis. Assessment of sperm morphology is performed on stained cells for both manual and CASA and is time-consuming. AI methods may have a large potential in classification and interpretation of sperm imaging and thereby replace the subjective and time-consuming methods.

Published Human Reproduction, 2019.

Candidate contributions Steven contributed by performing the experiments, and drafting and revising the manuscript.

Thesis objectives Objective 2.

5.35 Paper XXXV - Artificial intelligence predicts sperm motility from sperm fatty acids

Auhtors Oliwia Witczak, Jorunn M Andersen, Steven Hicks, Hugo L Hammer, Michael A Riegler, and Trine B Haugen.

Abstract Omega-3 FAs are abundant in the sperm and are positively associated with sperm motility, especially progressive motility. Other sperm FAs present in lower levels may also be associated with sperm characteristics. AI may have the potential to predict sperm motility based on FA composition and thereby provide more insight into FAs impact on sperm function.

Published Human Reproduction, 2019.

Candidate contributions Steven contributed by performing the experiments, and drafting and revising the manuscript.

Thesis objectives Objective 2.

5.36 Paper XXXVI - Using 2D and 3D Convolutional Neural Networks to Predict Semen Quality

Auhtors Jon-Magnus Rosenblad, Steven Alexander Hicks, Håkon K Stensland, Trine B Haugen, Pål Halvorsen, and Michael A Riegler.

Abstract In this paper, we present the approach of team Jmag to solve this year's Medico Multimedia Task as part of the MediaEval 2019 Benchmark. This year, the task focuses on automatically determining quality characteristics of human sperm through the analysis of microscopic videos of human semen and associated patient data. Our approach is based on deep convolutional neural networks (CNNs) of varying sizes and dimensions. Here, we aim to analyze both the spatial and temporal information present in the videos. The results show that the method holds promise for predicting the motility of sperm, but predicting morphology appears to be more difficult.

Published Working Notes Proceedings of the MediaEval 2019 Workshop.

Candidate contributions Steven supervies Jon-Magnus in his work and contributed by aiding in the conception and design of the experiments, and by drafting and revising the manuscript.

Thesis objectives Objective 2.

5.37 Paper XXXVII - Artificial Intelligence in Medicine - Gastroenterology

Auhtors Inga Strümke, Steven Alexander Hicks, Vajira L Thambawita, Debesh Jha, Sravanthi Parasa, Michael A Riegler, and Pål Halvorsen.

Abstract The holy grail in endoscopy examinations has for a long time been assisted diagnosis using Artificial Intelligence (AI). Recent developments in computer hardware are now enabling technology to equip clinicians with promising tools for computer-assisted diagnosis (CAD) systems. However, creating viable models or architectures, training them, and assessing their ability to diagnose at a human level, are complicated tasks. This is currently an active area of research, and many promising methods have been proposed. In this chapter, we give an overview of the topic. This includes a description of current medical challenges followed by a description of the most commonly used methods in the field. We also present example results from research targeting some of these challenges, and a discussion on open issues and ongoing work is provided. Hopefully, this will inspire and enable readers to future develop CAD systems for gastroenterology.

Published Artificial Intelligence in Medicine, 2021

Candidate contributions Steven contributed to drafting and revising the manuscript.

Thesis objectives Objective 3.

Bibliography

[1] Steven A Hicks, Pia H Smedsrud, Pål Halvorsen, and Michael Riegler. Deep learning based disease detection using domain specific transfer learning. In *MediaEval*. MediaEval, 11 2018.

[2] U Rajendra Acharya, Hamido Fujita, Oh Shu Lih, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. *Information Sciences*, 405:81–90, 2017.

[3] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[4] Sang Bong Ahn, Dong Soo Han, Joong Ho Bae, Tae Jun Byun, Jong Pyo Kim, and Chang Soo Eun. The Miss Rate for Colorectal Adenoma Determined by Quality-Adjusted, Back-to-Back Colonoscopies. *Gut and liver*, 6(1):64–70, jan 2012.

[5] Sharib Ali, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer, Amar Hekalo, Yun Bo Guo, Bogdan Matuszewski, Mourad Gridach, Irina Voiculescu, Vishnusai Yoganand, Arnav Chavan, Aryan Raj, Nhan T. Nguyen, Dat Q. Tran, Le Duy Huynh, Nicolas Boutry, Shahadate Rezvy, Haijian Chen, Yoon Ho Choi, Anand Subramanian, Velmurugan Balasubramanian, Xiaohong W. Gao, Hongyu Hu, Yusheng Liao, Danail Stoyanov, Christian Daul, Stefano Realdon, Renato Cannizzaro, Dominique Lamarque, Terry Tran-Nguyen, Adam Bailey, Barbara Braden, James E. East, and Jens Rittscher. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis*, 70:102002, 2021.

[6] Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Kim V Anonsen, Michael A Riegler, et al. Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment. *arXiv preprint arXiv:2106.04463*, 2021.

[7] Sharib Ali *et al.* Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376*, 2020.

[8] Soroush Javadi and] Seyed Abolghasem Mirroshandel. A novel deep learning method for automatic assessment of human sperm images. *Computers in Biology and Medicine*, 109:182–194, 2019.

[9] Jorunn M. Andersen, Hilde Herning, Elin L. Aschim, Jøran Hjelmesæth, Tom Mala, Hans Ivar Hanevik, Mona Bungum, Trine B augen, and Oliwia Witczak. Body mass index is associated with impaired semen characteristics and reduced levels of anti-müllerian hormone across a wide weight range. *PLOS ONE*, 10(6):1–12, 06 2015.

[10] Neelima Bagri and Punit Kumar Johari. A comparative study on feature extraction using texture and shape for content based image retrieval. *International Journal of Advanced Science and Technology*, 80(4):41–52, 2015.

[11] Debarag Narayan Banerjee and Sasanka Sekhar Chanda. AI failures: A review of underlying issues. *CoRR*, abs/2008.04073, 2020.

[12] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12. Association for Computing Machinery, 2020.

[13] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(1):118, 2020.

[14] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

[15] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[16] Jens Peter E Bonde, Erik Ernst, Tina Kold Jensen, Niels Henrik I Hjollund, Henrik Kolstad, Thomas Scheike, Aleksander Giwercman, Niels Erik Skakkebæk, Tine Brink Henriksen, and Jørn Olsen. Relation between semen quality and fertility: a population-based study of 430 first-pregnancy planners. *The Lancet*, 352(9135):1172–1177, 1998.

[17] 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, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[18] Patrick Brandao, Evangelos B. Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Medical imaging: Computer-aided diagnosis - fully convolutional neural networks for polyp segmentation in colonoscopy. *Medical Imaging 2017: Computer-Aided Diagnosis*, 10134, March 2017.

[19] Michael Bretthauer, Lars Aabakken, Evelien Dekker, Michal F Kaminski, Thomas Rösch, Rolf Hultcrantz, Stepan Suchanek, Rodrigo Jover, Ernst J Kuipers, Raf Bisschops, and Others. Requirements and standards facilitating quality improvement for reporting systems in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy*, 48(3):291–294, 2016.

[20] Varun H Buch, Irfan Ahmed, and Mahiben Maruthappu. Artificial intelligence in medicine: current trends and future possibilities. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 68(668):143–144, mar 2018.

[21] Elisabeth Carlsen, Aleksander Giwercman, Niels Keiding, and Niels E Skakkebæk. Evidence for decreasing quality of semen during past 50 years. *British Medical Journal*, 305(6854):609–613, 1992.

[22] Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21 2, 2019.

[23] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[24] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[25] D. E. Comer, David Gries, Michael C. Mulder, Allen Tucker, A. Joe Turner, and Paul R. Young. Computing as a discipline. *Communications of the ACM*, 32(1):9–23, January 1989.

[26] K Cullen, N S Stenhouse, K L Wearne, and G N Cumpston. Electrocardiograms and 13 year cardiovascular mortality in busselton study. *Heart*, 47(3):209–212, 1982.

[27] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):54, 2019.

[28] Christopher De Jonge. Semen analysis: looking for an upgrade in class. *Fertility and Sterility*, 97(2):260–266, feb 2012.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[30] Filip Karlo Došilović, Mario Brcic, and Nikica Hlupic. Explainable Artificial Intelligence: A Survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.

[31] Ahmad El Hajjar and Jean-François Rey. Artificial intelligence in gastrointestinal endoscopy: general overview. *Chinese medical journal*, 133(3):326–334, 2020.

[32] Edvarda Eriksen, Steven Hicks, Michael A. Riegler, Pål Halvorsen, and Valentina Carapella. A web-based software for training and quality assessment in the image analysis workflow for cardiac t1 mapping mri. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 216–2164, 2019.

[33] Gianni L Faedda, Kyoko Ohashi, Mariely Hernandez, Cynthia E McGreenery, Marie C Grant, Argelinda Baroni, Ann Polcari, and Martin H Teicher. Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. *Journal of Child Psychology and Psychiatry*, 57(6):706–716, 2016.

[34] Tingting Fang and Risto Lahdelma. Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. *Applied Energy*, 179:544–552, 2016.

[35] Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Image Analysis*, pages 363–370, 2003.

[36] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[37] Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen, and Michael Alexander Riegler. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In *Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care*, HealthMedia '19, page 9–15, New York, NY, USA, 2019. Association for Computing Machinery.

[38] Enrique Garcia-Ceja, Michael Riegler, Petter Jakobsen, Jim Tørresen, Tine Nordgreen, Ketil J. Oedegaard, and Ole Bernt Fasmer. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, page 472–477, 2018.

[39] 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*, pages 196–205, 2021.

[40] 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, Allan Linneberg, Oluf Pedersen, Stig Haunsø, Jesper Hastrup Svendsen, Torben Hansen, Jørgen Kim Kanters, and Morten Salling Olesen. Rare genetic variants previously associated with congenital forms of long QT syndrome have little or no effect on the QT interval. *European heart journal*, 36(37):2523–2529, Oct 2015.

[41] 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 *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 539–544, 2021.

[42] Jeremy R Glissen Brown and Tyler M Berzin. Adoption of New Technologies: Artificial Intelligence. *Gastrointestinal Endoscopy Clinics*, 31(4):743–758, 2021.

[43] Robert J Goldberg, James Bengtson, Zuoyao Chen, Keaven M Anderson, Emanuela Locati, and Daniel Levy. Duration of the qt interval and total and cardiovascular mortality in healthy persons (the framingham heart study experience). *The American Journal of Cardiology*, 67(1):55–58, 1991.

[44] David S Guzick, James W Overstreet, Pam Factor-Litvak, Charlene K Brazil, Steven T Nakajima, Christos Coutifaris, Sandra Ann Carson, Pauline Cisneros, Michael P Steinkampf, Joseph A Hill, et al. Sperm morphology, motility, and concentration in fertile and infertile men. *New England Journal of Medicine*, 345(19):1388–1393, 2001.

[45] H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia

Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. Immune-related pathologic response criteria.

[46] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[47] Peter Riis Hansen, Christian Rimer Juhl, Jonas Lynggaard Isaksen, Gregor Borut Jemec, Christina Ellervik, and Jørgen Kim Kanters. Frequency of Electrocardiographic Abnormalities in Patients With Psoriasis. *The American journal of cardiology*, 121(8):1004–1007, Apr 2018.

[48] Daniel A Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R Meireles. Artificial Intelligence in Surgery: Promises and Perils. *Annals of surgery*, 268(1):70–76, jul 2018.

[49] Syed Zohaib Hassan, Kashif Ahmad, Steven Hicks, Pål Halvorsen, Ala Al-Fuqaha, Nicola Conci, and Michael Riegler. Visual sentiment analysis from disaster images in social media, 2020.

[50] Trine B. Haugen, Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Hugo L. Hammer, Rune Borgli, Pål Halvorsen, and Michael Riegler. Visem: A multimodal

video dataset of human spermatozoa. In *Proceedings of the 10th ACM Multimedia Systems Conference (MMSys)*, MMSys '19, page 261–266, New York, NY, USA, 2019. Association for Computing Machinery.

[51] Claire Henderson, Sara Evans-Lacko, and Graham Thornicroft. Mental illness stigma, help seeking, and public health programs. *American journal of public health*, 103(5):777–780, may 2013.

[52] 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 *Proc. of MediaEval 2019 CEUR Workshop*, 2019.

[53] Steven Hicks, Trine B. Haugen, Mario Iliceto, Hugo L. Hammer, Jorunn M. Andersen, Oliwia Witczak, Michael A. Riegler, and Mette H. Stensen. Big data is not always better-prediction of live birth using machine learning on time-lapse videos of human embryos. In *Human Reproduction*, volume 35, pages I235–I235, 2020.

[54] Steven Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo Hammer, and Michael Riegler. An Overview of the EndoTect Challenge at ICPR 2020. In *ICPR2020*, 2020.

[55] Steven Hicks, Debesh Jha, Vajira Thambawita, Michael Riegler, Pål Halvorsen, Bjørn-Jostein Singstad, Sachin Gaur, Klas Pettersen, Morten Goodwin, Sravanthi Parasa, and Thomas de Lange. MedAI: Transparency in Medical Image Segmentation. *Nordic Machine Intelligence*, 2021.

[56] Steven Hicks, Michael Riegler, Pia Smedsrud, Trine B. Haugen, Kristin Ranheim Randel, Konstantin Pogorelov, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Mathias Lux, Andreas Petlund, Thomas de Lange, Peter Thelin Schmidt, and Pål Halvorsen. Acm multimedia biomedia 2019 grand challenge overview. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2563–2567, New York, NY, USA, 2019. Association for Computing Machinery.

[57] Steven Hicks, Pia H Smedsrud, Michael A Riegler, Thomas de Lange, Andreas Petlund, Sigrun L Eskeland, Konstantin Pogorelov, Peter T Schmidt, and Pål

Halvorsen. Deep learning for automatic generation of endoscopy reports. *Gastrointestinal Endoscopy*, 89(6), jun 2019.

[58] Steven Hicks, Andrea Stautland, Ole Bernt Fasmer, Wenche Førland, Hugo Lewi Hammer, Pål Halvorsen, Kristin Mjeldheim, Ketil Joachim Oedegaard, Berge Osnes, Vigdis Elin Giæver Syrstad, Michael Riegler, and Petter Jakobse. HYPERAKTIV: An Activity Dataset from Adult Patients with Attention-Deficit/Hyperactivity Disorder (ADHD). In *Proceedings of the 12th ACM Multimedia Systems Conference*, 2021.

[59] Steven Hicks, Vajira Thabawita, Hugo L. Hammer, Trine B. Haugen, Pål Halvorsen, and Michael Riegler. ACM MM BioMedia 2020 Grand Challenge Overview. In *ACMMM2020*, ACM MM '20, New York, NY, USA, 2020. Association for Computing Machinery.

[60] 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. *Scientific Reports*, 9(1):16770, 2019.

[61] 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. *Scientific Reports*, 11(1):10949, 2021.

[62] 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. *medRxiv*, 2021.

[63] Steven Alexander Hicks, Sigrun Eskeland, Mathias Lux, Thomas de Lange, Kristin Ranheim Randel, Mattis Jeppsson, Konstantin Pogorelov, Pål Halvorsen, and Michael Riegler. Mimir: An automatic reporting and reasoning system for deep learning based analysis in the medical domain. In *Proceedings of the 9th ACM Mul-*

timedia Systems Conference, page 369–374. Association for Computing Machinery, 2018.

[64] Steven Alexander Hicks, Pia H Smedsrud, Pål Halvorsen, and Michael Riegler. Deep learning based disease detection using domain specific transfer learning. *MediaEval*, 18:29–31, 2018.

[65] Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):13, 2020.

[66] Hamza O Ilhan, I Onur Sigirci, Gorkem Serbes, and Nizamettin Aydin. A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. *Medical & Biological Engineering & Computing*, 58(5):1047–1068, 2020.

[67] Petter Jakobsen, Enrique Garcia-Ceja, Lena Antonsen Stabell, Ketil Joachim Oedegaard, Jan Oystein Berle, Vajira Thambawita, Steven Alexander Hicks, Pål Halvorsen, Ole Bernt Fasmer, and Michael Alexander Riegler. Psykose: A motor activity database of patients with schizophrenia. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 303–308, 2020.

[68] Thangasamy Jeyapoovan and M Murugan. Surface roughness classification using image processing. *Measurement*, 46(7):2065–2072, 2013.

[69] 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, and Pål Halvorsen. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling*, pages 218–229, 2021.

[70] Debesh Jha, Sharib Ali, Steven Hicks, Vajira Thambawita, Hanna Borgli, Pia H. Smedsrud, Thomas de Lange, Konstantin Pogorelov, Xiaowei Wang, Philipp Harzig, Minh-Triet Tran, Wenhua Meng, Trung-Hieu Hoang, Danielle Dias, Tobey H. Ko, Taruna Agrawal, Olga Ostroukhova, Zeshan Khan, Muhammad Atif Tahir, Yang

Liu, Yuan Chang, Mathias Kirkerød, Dag Johansen, Mathias Lux, Håvard D. Johansen, Michael A. Riegler, and Pål Halvorsen. A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging. *Medical Image Analysis*, 70:102007, 2021.

[71] Debesh Jha, Steven Hicks, Krister Emanuelsen, Håvard Johansen, Dag Johansen, Thomas de Lange, Michael Riegler, and Pål Halvorsen. Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation. In *Proc. of MediaEval 2020 CEUR Workshop*, 2020.

[72] Niels Jørgensen, Elisabeth Carlsen, Ingrid Nermoen, Margus Punab, Jyrki Suominen, Anne-Grethe Andersen, Anna-Maria Andersson, Trine B Haugen, Antero Horte, Tina Kold Jensen, et al. East–west gradient in semen quality in the nordic–baltic area: a study of men from the general population in denmark, norway, estonia and finland. *Human Reproduction*, 17(8):2199–2208, 2002.

[73] Geetha K and Rajan C. Automatic Colorectal Polyp Detection in Colonoscopy Video Frames. *Asian Pacific journal of cancer prevention : APJCP*, 17(11):4869–4873, nov 2016.

[74] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning, 2020.

[75] Rabindra Khadga, Debesh Jha, Sharib Ali, Steven A. Hicks, Vajira Thambawita, Michael A. Riegler, and Pål Halvorsen. Few-shot segmentation of medical images based on meta-learning with implicit gradients. *CoRR*, abs/2106.03223, 2021.

[76] 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)*, pages 1–6, 2019.

[77] Anastasios Koulaouzidis, Dimitris K. Iakovidis, Diana E. Yung, Emanuele Rondonotti, Uri Kopylov, John N. Plevris, Ervin Toth, Abraham Eliakim, Gabrielle Wurm Johansson, Wojciech Marlicz, Georgios Mavrogenis, Artur Nemeth,

Henrik Thorlacius, and Gian Eugenio Tontini. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open*, 5(6), 2017.

[78] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, NIPS'12, page 1097–1105, 2012.

[79] Naina Kumar and Amit Kant Singh. Trends of male factor infertility, an important cause of infertility: A review of literature. *Journal of human reproductive sciences*, 8:191–196, 01 2016.

[80] Piyush Kumar, Rishi Chauhan, Thompson Stephan, Achyut Shankar, and Sanjeev Thakur. A machine learning implementation for mental health care. application: Smart watch for depression detection. In *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 568–574, 2021.

[81] Vinayshekhar Bannihatti Kumar, Sujay S Kumar, and Varun Saboo. Dermatological disease detection using image processing and machine learning. In *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, pages 1–6, 2016.

[82] Artúr István Károly, Péter Galambos, József Kuti, and Imre J. Rudas. Deep learning in robotics: Survey on model structures and training strategies. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):266–279, 2021.

[83] T. K. Larsen, S. Friis, U. Haahr, I. Joa, J. O. Johannessen, I. Melle, S. Opjordsmoen, E. Simonsen, and P. Vaglum. Early detection and intervention in first-episode schizophrenia: a critical review. *Acta Psychiatrica Scandinavica*, 103(5):323–334, 2001.

[84] Andreas Leibetseder, Stefan Petscharnig, Manfred Jürgen Primus, Sabrina Kletz, Bernd Münzer, Klaus Schoeffmann, and Jörg Keckstein. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, pages 357–362. ACM, 2018.

[85] David Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, June 2019.

[86] Hagai Levine, Niels Jørgensen, Anderson Martino-Andrade, Jaime Mendiola, Dan Weksler-Derri, Irina Mindlis, Rachel Pinotti, and Shanna H Swan. Temporal trends in sperm count: a systematic review and meta-regression analysis. *Human Reproduction Update*, 23(6):646–659, 2017.

[87] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125, 2017.

[88] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) - Volume 2*, pages 674–679. Morgan Kaufmann Publishers Inc., 1981.

[89] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, 2017.

[90] Gang Luo. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):1–16, 2016.

[91] Mathias Lux. Content based image retrieval with lire. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, page 735–738, New York, NY, USA, 2011. Association for Computing Machinery.

[92] Justin Mateo and José J Rieta. Application of artificial neural networks for versatile preprocessing of electrocardiogram recordings. *Journal of Medical Engineering & Technology*, 36(2):90–101, 2012.

[93] Andreas Mayr, Dominik Kißkalt, Moritz Meiners, Benjamin Lutz, Franziska Schäfer, Reinhardt Seidel, Andreas Selmaier, Jonathan Fuchs, Maximilian Metzner, Andreas

Blank, and Jörg Franke. Machine learning in production – potentials, challenges and exemplary applications. *Procedia CIRP*, 86:49–54, 2019.

[94] Christopher McCallum, Jason Riordon, Yihe Wang, Tian Kong, Jae Bem You, Scott Sanner, Alexander Lagunov, Thomas G Hannam, Keith Jarvi, and David Sinton. Deep learning-based selection of human sperm with high DNA integrity. *Communications biology*, 2:250, jul 2019.

[95] Sharon T Mortimer, Gerhard van der Horst, and David Mortimer. The future of computer-aided sperm analysis. *Asian journal of andrology*, 17(4):545–553, 2015.

[96] Lourdes García Murillo, Samuele Cortese, David Anderson, Adriana Di Martino, and Francisco Xavier Castellanos. Locomotor activity measures in the diagnosis of attention deficit hyperactivity disorder: Meta-analyses and new findings. *Journal of neuroscience methods*, 252, 2015.

[97] Julio Murra-Saca. El salvador atlas of gastrointestinal video endoscopy. `http://www.gastrointestinalatlas.com/index.html`. Accessed: 2021-09-29.

[98] 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)*, pages 25–26, 2020.

[99] Jonas B Nielsen, Claus Graff, Peter V Rasmussen, Adrian Pietersen, Bent Lind, Morten S Olesen, Johannes J Struijk, Stig Haunsø, Jesper H Svendsen, Lars Køber, Thomas A Gerds, and Anders G Holst. Risk prediction of cardiovascular death based on the QTc interval: evaluating age and gender differences in a large primary care population. *European Heart Journal*, 35(20):1335–1344, Mar 2014.

[100] Jonas Bille Nielsen, Claus Graff, Adrian Pietersen, Bent Lind, Johannes Jan Struijk, Morten Salling Olesen, Stig Haunsø, Thomas Alexander Gerds, Jesper Hastrup Svendsen, Lars Køber, and Anders Gaarsdal Holst. J-shaped association between qtc interval duration and the risk of atrial fibrillation. *Journal of the American College of Cardiology*, 61(25):2557–2564, 2013.

[101] 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)*, pages 135–144, 2020.

[102] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[103] Alberto Ochoa, Luis J Mena, and Vanessa G Felix. Noise-Tolerant Neural Network Approach for Electrocardiogram Signal Classification. In *Proceedings of the International Conference on Compute and Data Analysis*, ICCDA '17, pages 277—282, New York, NY, USA, 2017. Association for Computing Machinery.

[104] Chul-Hyun Park, Joon-Jae Lee, Sang-Keun Oh, Young-Chul Song, Doo-Hyun Choi, and Kil-Houm Park. Iris feature extraction and matching based on multiscale and directional image representation. In *International Conference on Scale-Space Theories in Computer Vision*, pages 576–583. Springer, 2003.

[105] Krushi Patel, Kaidong Li, Ke Tao, Quan Wang, Ajay Bansal, Amit Rastogi, and Guanghui Wang. A comparative study on polyp classification using convolutional neural networks. *PLOS ONE*, 15(7):1–16, 07 2020.

[106] Sandip Patil and Atul Dusane. Use of color feature extraction technique based on color distribution and relevance feedback for content based image retrieval. *International Journal of Computer Applications*, 52:9–12, 08 2012.

[107] Matthias Pierce, Holly Hope, Tamsin Ford, Stephani Hatch, Matthew Hotopf, Ann John, Evangelos Kontopantelis, Roger Webb, Simon Wessely, Sally McManus, and Kathryn M Abel. Mental health before and during the covid-19 pandemic: a longitudinal probability sample survey of the uk population. *The Lancet Psychiatry*, 7(10):883–892, 2020.

[108] Dong ping Tian et al. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4):385–396, 2013.

[109] 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 ACM Multimedia Systems Conference (MMSYS)*, pages 170–174, 2017.

[110] 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 ACM Multimedia Systems Conference (MMSYS)*, pages 164–169, 2017.

[111] 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 *Proc. of MediaEval 2018 CEUR Workshop*, 2018.

[112] Suranai Poungponsri and Xiao-Hua Yu. An adaptive filtering approach for electrocardiogram (ecg) signal noise reduction using neural networks. *Neurocomputing*, 117:206–213, 2013.

[113] Bahareh Pourbabaee, Mehrsan J Roshtkhari, and Khashayar Khorasani. Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(12):2095–2104, Dec 2018.

[114] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.

[115] Bhupendra Ramani, Kamini Solanki, and Warish Patel. Anxiety Detection Using Physiological Data and Wearable IoT Devices: A Survey. In *Handbook of Research on Applied Intelligence for Health and Clinical Informatics*, pages 31–43. IGI Global, 2021.

[116] Qing Rao and Jelena Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, pages 35–38, 2018.

[117] Colin J Rees, Roisin Bevan, Katharina Zimmermann-Fraedrich, Matthew D Rutter, Douglas Rex, Evelien Dekker, Thierry Ponchon, Michael Bretthauer, Jaroslaw Regula, Brian Saunders, Cesare Hassan, Michael J Bourke, and Thomas Rösch. Expert opinions and scientific evidence for colonoscopy key performance indicators. *Gut*, 65(12):2045–2060, dec 2016.

[118] Alessandro Repici, Matteo Badalamenti, Roberta Maselli, Loredana Correale, Franco Radaelli, Emanuele Rondonotti, Elisa Ferrara, Marco Spadaccini, Asma Alkandari, Alessandro Fugazza, Andrea Anderloni, Piera Alessia Galtieri, Gaia Pellegatta, Silvia Carrara, Milena Di Leo, Vincenzo Craviotto, Laura Lamonaca, Roberto Lorenzetti, Alida Andrealli, Giulio Antonelli, Michael Wallace, Prateek Sharma, Thomas Rosch, and Cesare Hassan. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology*, 159(2):512–520.e7, 2020.

[119] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

[120] Michael Riegler, Jorunn Andersen, Hugo Hammer, Steven Hicks, Oliwia Witczak, and Trine Haugen. Artificial intelligence as a tool in predicting sperm motility and morphology. In *Human Reproduction*, volume 34, pages 199–199, 2019.

[121] Michael A. Riegler, Mette H. Stensen, Oliwia Witczak, Jorunn M. Andersen, Steven A. Hicks, Hugo L. Hammer, Erwan Delbarre, Pål Halvorsen, Anis Yazidi, Nikolai Holst, and Trine B. Haugen. Artificial intelligence in the fertility clinic: status, pitfalls and possibilities. *Human Reproduction*, 36(9):2429–2442, 07 2021.

[122] Michael J. Rigby. Ethical Dimensions of Using Artificial Intelligence in Health Care, 2019.

[123] Olav A. Nergård Rongved, Steven A. Hicks, Vajira Thambawita, Håkon K. Stensland, Evi Zouganeli, Dag Johansen, Cise Midoglu, Michael A. Riegler, and Pål Halvorsen. Using 3d convolutional neural networks for real-time detection of soccer events. *International Journal of Semantic Computing*, 15(02):161–187, 2021.

[124] Frank Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, pages 65–386, 1958.

[125] Alessio Rossi, Eleonora Da Pozzo, Dario Menicagli, Chiara Tremolanti, Corrado Priami, Alina Sîrbu, David A. Clifton, Claudia Martini, and Davide Morelli. A public dataset of 24-h multi-levels psycho-physiological responses in young healthy adults. *Data*, 5(4), 2020.

[126] Parvaneh Saeedi, Dianna Yee, Jason Au, and Jon Havelock. Automatic identification of human blastocyst components via texture. *IEEE Transactions on Biomedical Engineering*, 64(12):2968–2978, 2017.

[127] Jan Scott, Greg Murray, Chantal Henry, Gunnar Morken, Elizabeth Scott, Jules Angst, Kathleen R Merikangas, and Ian B Hickie. Activation in bipolar disorders: a systematic review. *JAMA psychiatry*, 74(2):189–196, 2017.

[128] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[129] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, 2016.

[130] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical Physics*, 47(5):e148–e167, 2020.

[131] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[132] Fariba Shaker, S. Amirhassan Monadjemi, Javad Alirezaie, and Ahmad Reza Naghsh-Nilchi. A dictionary learning approach for human sperm heads classification. *Computers in Biology and Medicine*, 91:181–190, 2017.

[133] Wei Shi, Zhuozhuo Shen, Siyuan Wang, and Brian J. Hall. Barriers to professional mental health help-seeking among chinese adults: A systematic review. *Frontiers in Psychiatry*, 11:442, 2020.

[134] 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. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[135] Niels E Skakkebaek, Niels Jørgensen, Anna-Maria Andersson, Anders Juul, Katharina M Main, Tina Kold Jensen, and Jorma Toppari. Populations, decreasing fertility, and reproductive health. *The Lancet*, 393(10180):1500–1501, 2019.

[136] Remy Slama, Florence Eustache, B Ducot, Tina Kold Jensen, Niels Jørgensen, Antero Horte, S Irvine, J Suominen, AG Andersen, J Auger, et al. Time to pregnancy and semen parameters: a cross-sectional study among fertile couples from four european cities. *Human Reproduction*, 17(2):503–515, 2002.

[137] Pia H. Smedsrud, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Vajira Thambawita, Steven Hicks, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc-Tien D. Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T Schmidt, Michael A de Lange, Thomas Riegler, and Pål and Halvorsen. Kvasir-capsule, a video capsule endoscopy dataset, Aug 2020.

[138] Friedemann W Stallmann and Hubert V Pipberger. Automatic recognition of electrocardiographic waves by digital computer. *Circulation research*, 9(6):1138–1143, 1961.

[139] D. Steinkraus, I. Buck, and P.Y. Simard. Using gpus for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005.

[140] Hannah Stower. Transparency in Medical AI. *Nature Medicine*, 26(12):1804–1805, 2020.

[141] Inga Strümke, Steven A. Hicks, Vajira Thambawita, Debesh Jha, Sravanthi Parasa, Michael A. Riegler, and Pål Halvorsen. *Artificial Intelligence in Medicine*, pages 1–20. Springer International Publishing, 2020.

[142] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[143] 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*, page 309–314, New York, NY, USA, 2020. Association for Computing Machinery.

[144] Bilal Taha, Naoufel Werghi, and Jorge Dias. Automatic polyp detection in endoscopy videos: A survey. In *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*, pages 233–240, 2017.

[145] Masoud Tahmasian, Habibolah Khazaie, Sanobar Golshani, and Kristin T Avis. Clinical Application of Actigraphy in Psychotic Disorders: A Systematic Review. *Current Psychiatry Reports*, 15(6):359, 2013.

[146] Nima Tajbakhsh, Changching Chi, Suryakanth R. Gurudu, and Jianming Liang. Automatic polyp detection from learned boundaries. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 97–100, 2014.

[147] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks.

In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 79–83, 2015.

[148] Joanna Talarczyk-Desole, Anna Berger, Grażyna Taszarek-Hauke, Jan Hauke, Leszek Pawelczyk, and Piotr Jedrzejczak. Manual vs. computer-assisted sperm analysis: can casa replace manual assessment of human semen in clinical practice? *Ginekologia Polska*, 88(2):56 – 60, 2017.

[149] Vajira Thambawita, Steven Hicks, Pål Halvorsen, and Michael A Riegler. Pyramid-focus-augmentation: Medical image segmentation with step-wise focus. *arXiv preprint arXiv:2012.07430*, 2020.

[150] Vajira Thambawita, Steven A Hicks, Pål Halvorsen, and Michael A Riegler. Divergentnets: Medical image segmentation by network ensemble. *arXiv preprint arXiv:2107.00283*, 2021.

[151] 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*, page 231–236, 2020.

[152] 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, and Jørgen K. Kanters. Deepfake electrocardiograms: the beginning of the end for privacy issues in medicine. *medRxiv*, 2021.

[153] 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. 1(3), June 2020.

[154] Vajira L. Thambawita, Inga Strümke, Steven A. Hicks, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. Data augmentation using generative adversarial networks for creating realistic artificial colon polyp images: validation study by endoscopists. *Gastrointestinal Endoscopy*, 93(6):AB190, jun 2021.

[155] Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. Natural language processing advancements by deep learning: A survey. *CoRR*, abs/2003.01200, 2020.

[156] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, 10 1950.

[157] Jean M Twenge, A Bell Cooper, Thomas E Joiner, Mary E Duffy, and Sarah G Binau. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of Abnormal Psychology*, 128(3):185–199, 2019.

[158] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018, 2018.

[159] Jonas Wacker, Marcelo Ladeira, and Jose Eduardo Vaz Nascimento. Transfer learning for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham, 2021. Springer International Publishing.

[160] Zi Ying Wee, Samantha Wei Lee Yong, Qian Hui Chew, Cuntai Guan, Tih Shih Lee, and Kang Sim. Actigraphy studies and clinical and biobehavioural correlates in schizophrenia: a systematic review. *Journal of Neural Transmission*, 126(5):531–558, 2019.

[161] Oliwia Witczak, Jorunn M. Andersen, Steven Hicks, Hugo L. Hammer, Michael A. Riegler, and Trine B. Haugen. Artificial intelligence predicts sperm motility from sperm fatty. In *Human Reproduction*, volume 34, pages 200–201, 2019.

[162] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[163] Heather Cleland Woods and Holly Scott. #sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of Adolescence*, 51:41–49, 2016.

[164] World Health Organization, Department of Reproductive Health and Research. *WHO laboratory manual for the examination and processing of human semen.* Geneva: World Health Organization, 2010.

[165] Tong Yu and Hong Zhu. Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*, 2020.

[166] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[167] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional Recurrent Neural Networks for Electrocardiogram Classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4, Sep 2017.

[168] Syed Zohaib Hassan, Kashif Ahmad, Steven Riegler, Michael Hicks, Nicola Conci, Pål Halvorsen, and Ala Al-Fuqaha. Visual Sentiment Analysis: A Natural Disaster Use-case Task at MediaEval 2021. In *Proc. of MediaEval 2021 CEUR Workshop*, 2021.