Summary of the contribution of the Ph.D. Thesis - DeepSynthBody: the beginning of the end for data deficiency in medicine

In this thesis, we have contributed to the development of Artificial Intelligence (AI) in medicine by proposing Machine Learning (ML) methods, datasets (real and synthetic), and tools by addressing the data deficiency problem, which has been identified as an obstacle for producing well-performing and generalizable ML models within the field of medicine.

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

In this regard, we have introduced our main objective and four sub-objectives to address our research question, "**What are the problems that emerge from data in computer-aided diagnosis systems, and how can these problems be tackled?**". Our main objective was "Research and develop ML models which are the main component of CAD systems for different medical applications, focusing on the problems of limited availability of biomedical data".

To achieve the main objective, four sub-objectives were introduced. They are, 1) research and develop ML models for CAD systems to assist doctors, 2) collect, research, and develop datasets to develop ML models for CAD systems for biomedical applications, 3) research and develop benchmark analysis with the medical datasets to identify the problems for producing well-performing ML solutions in the medical domain, and 4) research and develop deep GAN that can produce synthetic data to address the data deficiency problem, the major obstacle for developing medical AI-based solutions.

To achieve the objectives of the thesis, we investigated case studies from three different medical branches, namely cardiology, gastroenterology, and andrology. Using data from these case studies, we developed ML models to achieve Sub-objective I. Addressing the scarcity of medical data, we collected, analyzed, and developed medical datasets to achieve Sub-objective II and performed benchmark analyses to achieve Sub-objective III. A framework for generating synthetic medical data has been developed using Generative Adversarial Networks (GANs) as a solution to address the data deficiency problem to achieve Sub-objective IV. Our results indicate that our generated synthetic data may be a solution to the data challenge. As an overarching concept, we introduced the DeepSynthBody as a basis for structured and centralized synthetic medical data generation. The studies presented in the thesis, such as generating synthetic Electrocardiogram (ECG), Gastrointestinal (GI) tract images and videos with and without polyps, and sperm samples, showed that DeepSynthBody can help to overcome data privacy concerns, the time-consuming and costly data annotation process, and the data imbalance problem in the medical domain. Our experiments showed that our generative models generate realistic synthetic data providing comparable results to experiments using real data to tackle the identified problems. In the end, we have achieved the main objective by achieving Sub-objectives I, II, III, and IV and published 28 publications in total.

The DeepSynthBody framework is available as an open-source project (www.deepsynthbody.org) that allows researchers, industry, and practitioners to use the system and contribute to future developments.

In conclusion, our proposed CAD systems show good performance for all three case studies. For example, one of our segmentation models introduced in this thesis won first place in EndoCV 2021. In addition to collecting and publishing medical domain data sets, we identified that generating synthetic medical data to train ML models is an alternative solution to overcome this data deficiency problem. Well-performing GAN architectures can generate realistic synthetic data. These synthetic data can represent real medical data when the real datasets are not permitted to share. Moreover, we showed that conditional GAN architectures can generate synthetic datasets with the corresponding ground truth data, which domain experts normally do. In the future, this DeepSynthBody concept can be improved to use as a model to represent the human body. Furthermore, the data compression ability of GANs is a solution for storing medical data in a limited space avoiding privacy concerns.