LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

12th Conference on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Druskininkai, Lithuania, Hotel "Europa Royale"
http://www.mii.lt/DAMSS

December 2–4, 2021

VILNIUS UNIVERSITY PRESS
Vilnius, 2021

Co-Chairmen:

Dr. Saulius Maskeliūnas (Lithuanian Computer Society)
Prof. Gintautas Dzemyda (Vilnius University, Lithuanian Academy of Sciences)

Programme Committee:

Prof. Juris Borzov (Latvia)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Yuriy Kharin (Belarus)
Prof. Tomas Krilavičius (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Dr. Olga Kurasova
Dr. Viktor Medvedev
Dr. Martynas Sabaliauskas
Laima Paliulionienė

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Dr. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. +370 5 2109 315

https://doi.org/10.15388/DAMSS.12.2021
ISBN 978-609-07-0673-2 (print)
ISBN 978-609-07-0674-9 (digital PDF)

On the Computed Tomography Image Data to Diagnose Pancreatic Cancer Using Machine Learning

Aušra Šubonienė[1], Olga Kurasova[1], Viktor Medvedev[1], Aistė Kielaitė-Gulla[2], Artūras Samuilis[3], Džiugas Jagminas[4], Kęstutis Strupas[2], Gintautas Dzemyda[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Institute of Clinical Medicine, Faculty of Medicine
Vilnius University
[3] Institute of Biomedical Sciences, Department of Radiology, Nuclear Medicine and Medical Physics, Faculty of Medicine
Vilnius University
[4] Faculty of Medicine
Vilnius University
ausra.suboniene@mif.vu.lt

Medical imaging data, which is suitable for solving segmentation problems, are difficult to obtain due to data sensitivity issues and the effort that is required to make ground truth segmentations. In order to increase the robustness of results by including more medical images, multiple datasets are often combined. However, challenges arise when trying to combine such datasets from different sources. Populations of patients that differ by age and other conditions could affect the results. Also, there might be different approaches to segmentation and its accuracy. Experts can segment medical images as true to anatomical structures as possible, or they might include some surrounding tissues in order to speed up manual segmentation. Also, rough region boundaries can be used instead of segmentations. Lastly, there can be different diagnostic devices used, which might result in different pre-processing of images.

Data sources. Due to data sensitivity and effort that is needed to anonymise images, there is a lack of publicly available pancreatic cancer data. Most publicly available medical data is without segmentation by experts, and images that do have some anatomical structures segmented are scarce. Currently, the largest public collections of computer tomography (CT) images of pancreatic cancer are available are the Can-

cer Imaging Archive (TCIA) dataset and Medical Segmentation Decathlon dataset. The Medical Segmentation Decathlon dataset consists of 421 portal-venous phase 3D CT scans. Segmentations of both the pancreatic parenchyma and pancreatic mass (cyst or tumour) are provided, although done as ROI only, which makes the segmentation process quicker but less accurate. TCIA dataset consists of 82 abdominal contrast enhanced 3D CT scans with slice thickness between 1.5−2.5 mm. Manual segmentations were done only to segment the pancreas. Here we also analyse the dataset which was acquired in the Vilnius University Hospital Santaros Klinikos. Segmentations that were provided by the experts consist of healthy pancreas, pancreatic cancer and pancreatic duct, which can be confused with pancreatic cancer by machine learning algorithms due to similar intensity of pixels. When combining publically available datasets segmentation differences need to be taken into account and some additional manual segmentation might be needed.

Differences in populations of patients. Due to the limited availability of medical images and pancreatic cancer being more common later in life, it is difficult to achieve equal coverage of all age categories of pancreatic cancer patients. This can be partially solved by increasing the dataset size of categories by retrospective analysis of computer tomography images provided by the medical institution, that have been filtered by the desired conditions.

Issues in the quality of the segmentation. Manual segmentations of computer tomography images can be prepared using different approaches to segmentation and its accuracy. Experts can segment medical images as true to anatomical structures as possible, or they might include some surrounding tissues in order to speed up manual segmentation. Also, rough region boundaries can be used instead of segmentations. Data pre-processing might be done by removing fat tissue based on the values of Hounsfield units. This results in holes and irregular edges of segmented regions. This creates additional problems when trying to unify segmentations across multiple datasets and might reduce the accuracy of segmentations when using machine learning.

Different diagnostic devices. Lastly, there can be different diagnostic devices used, which might result in different pre-processing of images. Even after pre-processing, these different diagnostic devices can have unique artefacts in computer tomography images that can reduce segmentation accuracy when combining multiple datasets.