Combining deep learning and radiomics-based machine learning to optimize predictions on medical images

COMBINING DEEP LEARNING AND RADIOMICS-BASED MACHINE LEARNING TO OPTIMIZE PREDICTIONS ON MEDICAL IMAGES

Manon Beuque

Lay-out and printing: ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 978-94-6469-423-9

The research presented in this thesis was conducted within GROW - School for Oncology and Reproduction, Maastricht University.

This thesis was accomplished with financial support from the Marie Skłodowska-Curie grant, project “PREDICT”.

COMBINING DEEP LEARNING AND RADIOMICS-BASED MACHINE LEARNING TO OPTIMIZE PREDICTIONS ON MEDICAL IMAGES

Dissertation

To obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public
on Friday 8th of September 2023 at 10:00 hours

by

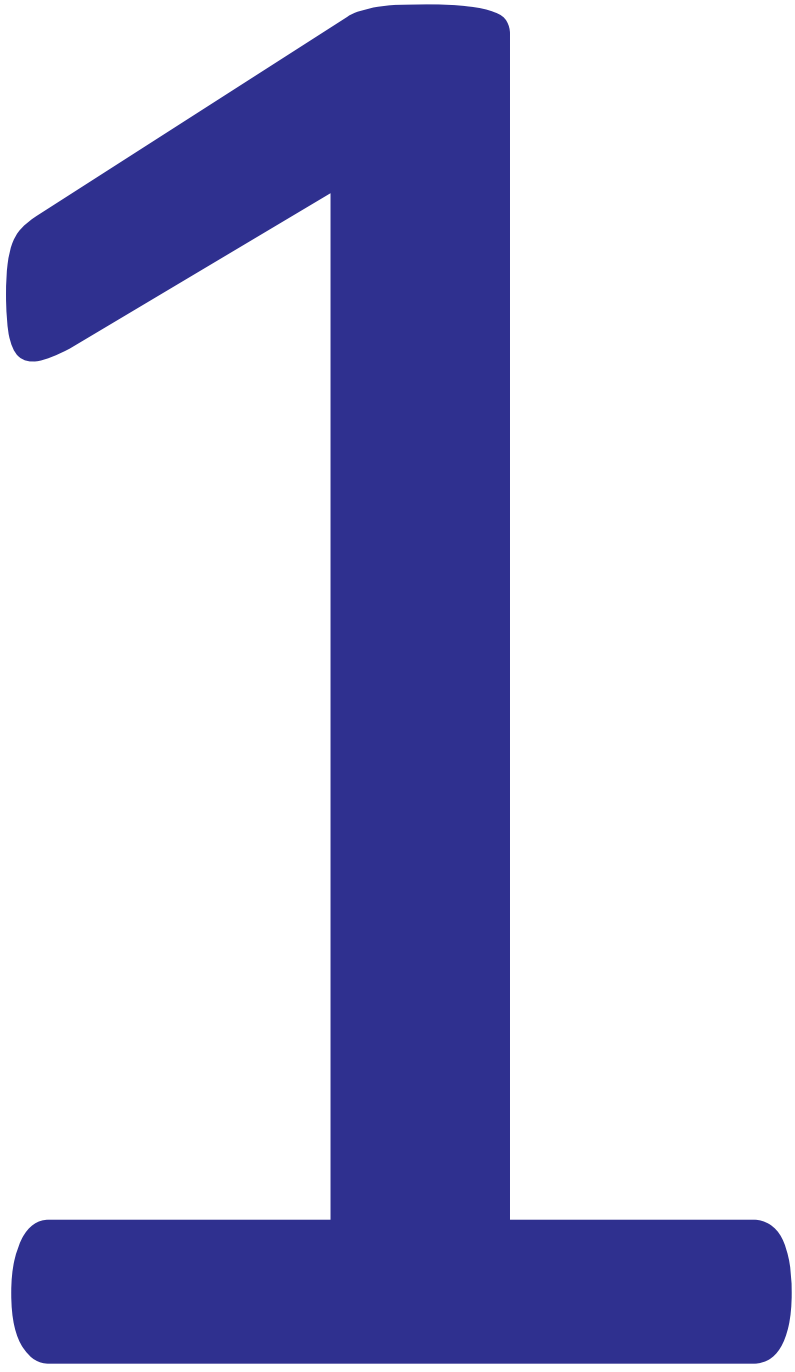
Manon Paola Luce Beuque

Supervisors
Prof. Dr. P. Lambin
Prof. Dr. H.I. Grabsch

Co-supervisor
Dr. H. Woodruff

Assessment Committee
Prof. Dr. D. Keszthelyi (chair)
Prof. Dr. J.N. Kather, Dresden University of Technology
Prof. Dr. W.J. Niessen, Erasmus University Medical Centre Rotterdam
Prof. Dr. M.L. Smidt

Table of Contents



Chapter 1

Introduction

The role of medical imaging in cancer patients

Cancer is one of the leading causes of mortality worldwide (1). Cancer can take many forms and invade any organ, making this disease highly diverse, requiring multiple diagnostic tools and treatment strategies. For almost 200 years clinicians have worked with medical imaging to diagnose and treat patients with cancers. First with histopathology, applying cell theory introduced in 1838 (2), which consists of monitoring changes in cell architecture using chemically stained biopsies (for example by using Haematoxylin & Eosin (H&E)). In the 1970s, other medical imaging modalities started to be implemented and are still used today for the anatomical (e.g. computer tomography (CT) and magnetic resonance imaging (MRI)), metabolic (e.g. positron emission tomography (PET)), and analytical (e.g. mass spectrometry imaging (MSI)) evaluation of cancer. Imaging facilitates understanding of the disease at different levels and monitoring disease progression, which help establishing the most efficient treatment strategy for patients. Thanks to this multitude of available medical imaging modalities and their diverse properties, monitoring cancer patients can be performed at every stage of patient treatment, despite the dynamic and complex nature of this disease (3).

Diagnosis and staging

To diagnose cancer at an early stage, screening programs have been implemented to test symptom-free populations at high risk for a particular cancer: Currently, in the western world, there are screening programs for early detection of cancer in different organs such as breast, cervix, and bowel (4) which utilise various imaging modalities. Digital mammography is used to detect suspicious lesions in the breast and is sometimes accompanied with other imaging modalities if appropriate (5). The examination for cervical cancer includes digital colposcopy after a positive cytology test (6).

If there is suspicion of cancer either through screening programs or due to the presentation of symptoms, a number of tests might be performed such as physical examination of the patient and examination of fluid or tissue samples in a pathology laboratory. Imaging using CT, PET-CT, MRI, PET-MRI, ultrasound and/or X-ray might be necessary to locate the primary tumour and stage the disease including assessment of metastatic spread (7) (8).

Disease staging takes into consideration different information such as the size and location of the primary tumour, the number of lymph nodes potentially invaded by tumour in the immediate surroundings of the primary tumour and presence and localisation of distant metastases. The main classification system used to describe the disease stage in cancer patients is referred to as TNM: extent of primary tumour (T), presence or absence of regional lymph node metastasis (N) and presence or absence of distant metastasis (M). Disease staging includes the analysis of medical images and is used to decide the most appropriate treatment option for the patient (9).

Active surveillance, treatment planning, and disease progression monitoring

Active surveillance might be an option for patients with certain cancer types which are not progressing or progressing only slowly such as some types of prostate cancer (10). Active surveillance for prostate cancer usually includes some form of medical imaging as well as a digital rectal exam at regular intervals and taking of a biopsy if appropriate (11).
Monitoring disease progression in patients undergoing treatment is performed by imaging at regular intervals using MRI, CT, ultrasound or other imaging modalities depending on the location and type of cancer.

If the treatment plan includes radiotherapy, pre-treatment three dimensional dose calculation within the mapped out radiation field is performed on a planning CT. The dose delivery is calculated with advanced simulations to most efficiently deliver the necessary treatment dose to the tumour avoiding irradiating too much of the healthy tissue surrounding the tumour, especially protecting organs at risk such as the heart if for example the patient receives radiation for a mass located in the left breast (12).

Follow-up

Follow-up after initial treatment can be performed through medical imaging depending on the cancer type to identify recurrent disease at an early stage, usually combined with laboratory tests and symptoms surveillance.

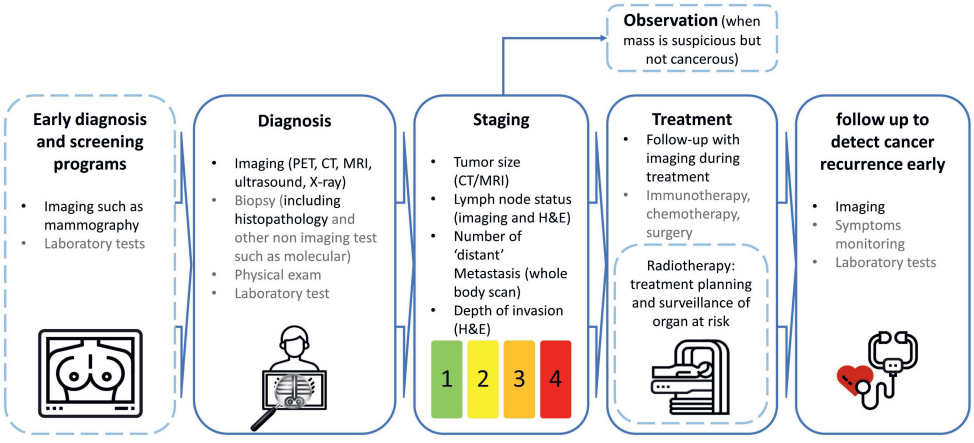


Figure 1: Steps of the cancer patient pathway – in grey non-imaging modalities, light blue dashed boxes are not systematically part of the treatment.

All of the above mentioned imaging modalities are in theory suitable for analysis by means of artificial intelligence (AI) in order to support the different tasks which need to be routinely performed by the specialist in charge of the image analysis. AI tools have the potential to go a step further, improving prognosis prediction and supporting the selection of the optimal treatment based on the analysis of medical images. In the context of this work, AI

refers to extracting complex statistical correlations from large amounts of data and is synonymous with machine learning. As such, AI is a quantitative approach aiming to remove subjectivity, which contrasts with the mostly qualitative approach of the clinicians. Thus, assessment of images with AI will remain consistent over time and has the potential to reduce inter- and intra-variabilities of the expert observers. Another advantage of using AI could be speed of assessment allowing clinicians to make decisions faster, helping to prioritize individual patient's care.

AI assistance in the clinical setting

AI can be defined broadly as a system which learns from past data to automatically make predictions and decisions based on new, real world data. The use of AI was made possible in recent years thanks to advances made in computer science. This includes access to large amounts of data for model training, improvements in processor performance and the development of new AI model architectures. AI has already had a large impact in a variety of fields, from social media to cyber security, automatically performing tasks which otherwise are labour-intensive tasks requiring human input (13). Specifically in the field of medical imaging, AI has been utilized for the creation of clinical decision support systems (CDSS), aiming to support clinicians in making diagnoses or treatment choices for their patients.

Machine learning and deep learning solutions already implemented in the clinical routine

Some CDSS are already used in clinical routine and show potential to help clinicians at every step of the cancer patient pathway (see Figure 1). The field of computer aided detection (CADe) and diagnosis (CADx) systems emerged in the 1980's and took shape in 1998 with the approval of the first commercially used CAD by the US Food and Drug Administration (FDA), which was used to assist radiologists in the detection of breast cancer on mammograms (14). The resolution and availability of digital medical images improved over time and so did the software to assist clinicians. CDSS include a variety of tools which received CE mark and/or FDA approval: The FDA reported 343 devices using AI in medical imaging with FDA approval as of September 2021 (15) and the website www.aiforradiology.com (visited August 2022) contained the description of 202 CE-marked AI-based products used for medical imaging (16) mainly assisting clinicians with automatic measurements/segmentations (33%), detections (27%) or being used for diagnosis predictions (22%) (17).

Only seven software packages using machine learning specifically for histopathological image analysis are referenced in the FDA report (15), although computational pathology research exists since the 1960s (18). Compared to radiologists, the tasks of pathologists differ: radiologists mostly have to detect a lesion, give a preliminary assessment of what

was found, whereas pathologists have to provide a definitive diagnosis, which will directly impact on clinical decision, which possibly makes FDA or CE approval more difficult to obtain for dedicated software. Furthermore, histopathology data is more complex, at higher definition than radiology images but suffer from the same pitfalls: shortage of high quality data, lack of good quality annotations, high disease heterogeneity, etc.

Radiomics models for clinical image analysis

Radiomics models can be categorised into two types: (1) machine learning models based on features crafted by an expert (i.e. predefined mathematical formulas) and (2) deep learning models where the model learns, or crafts, a set of features on its own.

Handcrafted radiomics features are a set of quantitative features extracted from a region of interest (usually a suspicious mass within a radiology image: CT, MRI, X-ray, ultrasound…) which are used as input for a machine learning (ML) model to predict a particular outcome (19). Those quantitative features aim to optimally characterise the lesion of interest and can be classified into different categories such as texture, intensity, and shape features. These features can be used to train a ML model to classify masses (benign versus malignant), grade suspicious lesions, or predict survival. A handcrafted radiomics-based signature or biomarker can be found after model training and externally validated on a new dataset.

Although handcrafted radiomics studies have the potential to extract relevant information from medical images, robust features are difficult to obtain due to the variability of segmentations, differences in image acquisition and quality (20). Moreover, the lack of external validation of the identified signature remains a major challenge in the field. Only 41% of handcrafted radiomics studies published in 2018 included results from external validation dataset (21).

Similarly to handcrafted radiomics, other quantitative imaging biomarkers are investigated in the medical research setting by mining datasets such as proteomics (proteins), dosiomics (radiation therapy dose distributions), histomics (histopathology) which differ by their dataset type. For histopathology images, the predictive value of different histomic features can be tested such as texture-features, pattern, and histogram features (22). Combining features extracted from different types of data sets can potentially improve predictions of ML models, an hypothesis which was tested in a multimodal data integration study exploring risk stratification for ovarian cancer patients (23).

Deep learning (DL) is a relatively new machine learning approach, which is increasingly used in the medical imaging field in recent years with numbers of publications rising from 384 in 2015 to 14,669 in 2021 according to the data retrieved from PubMed, searching “deep learning” (24). Compared to handcrafted feature based analysis, DL models can learn the most efficient set of features automatically from the images in the training dataset (i.e. without handpicking the relevant features). To do so, multiple convolutions are applied to

the images which are then reduced to a set of features used to predict a certain outcome. The models learn a set of weights for the different layers of the model by backpropagation, a method which compares the predicted outcomes with the ground truth and makes changes to the network weights in order to minimise the difference. In medical imaging analysis, DL has been used to predict different outcomes: the model can be trained to detect a suspicious lesion, segment it, and label it. It can also be trained to predict the grade of a tumour, treatment, or prognosis.

Machine learning challenges for predictions using medical imaging data

Handcrafted feature-based ML models and DL models face the same challenges when trained on medical images, one of which is heterogeneity of the medical datasets. This is very different to the data used to train non-medical deep learning models such as COCO or ImageNet (25) (26), which are regular photographs, less complex to analyse and more widely available. The heterogeneity of the medical imaging data can be partially explained by preferred acquisition and reconstruction settings chosen by individual clinicians and healthcare centres, making images look different. Other factors can also occurs such as differences in spatial resolution of the images according to the hardware used, different reconstruction parameters used by different vendors, etc. All those differences make the use of DL in different fields of image analysis such as classification or segmentation more challenging than tasks on regular images (27). Moreover, dataset shifts also depends on the origin of the data: the characteristics of a disease in a medical image from one part of the world might look different in another part of the world due to genetic and environmental factors, making it difficult to obtain a model usable for any vendor and patient cohort.

To increase reproducibility of the results obtained with radiomics features, the Image Biomarker Standardization Initiative (IBSI) (28) initiated standardization of radiomics features, consisting in identification of radiomics features which were stable when extracted with different tools. Robust pre-processing methods can also help homogenising the datasets. Some initiatives to standardize pre-processing in an attempt to make results better reproducible have been described by (29) and (30).

For histopathology datasets, homogenisation of the data is also necessary. Because the staining protocols used for the tissue samples can differ and the scanner can be different, the colour of the tissue samples originating from different sites but also from the same site can appear very different. Moreover, due to the high resolution of histopathology data, fully annotating those tissue samples is very time consuming. Therefore, high resolution datasets require a well defined analysis strategy and a lot of computational power to adopt a ML or DL solution (31).

Another challenge for the medical imaging analysis field is the shortage of publicly accessible sufficiently large datasets (at least hundreds of samples) and/or well annotated datasets to train ML models in order to achieve a good performance. This lack of data is

particularly significant for studies which therefore cannot externally validate their models. Lack of external validation makes it difficult for the potential user to understand whether the developed models are sufficiently robust and independent of the training and testing datasets (32). Although large quantities of medical imaging are gathered every day in every hospitals, certain diseases or subtypes of diseases are very rare leading to hugely imbalanced datasets which is challenging to overcome.

Types of images analysed in this thesis

Mass spectrometry imaging (MSI) with matrix-assisted laser desorption/ionization (MALDI)

Mass spectrometry imaging is an imaging technique which measure per raster of a few micrometres the mass-to-charge ratio of ions within a sample. To analyse the composition of a tissue section, a mass spectrometer ionizes molecules of the sample and collects the mass spectrum for each location. As the spatial information is also saved, it is possible to analyse the distribution of the mass-to-charge (m/z) values of a tissue per location (33). The m/z values can then be processed and analysed with specific software usually supplied by the vendor of the instrument, superimposed onto a consecutive tissue section stained with H&E, fluorescence markers, etc. and digitalized to complete the study if necessary. MSI is used in cancer research to discover biomarkers to improve tumour classification and potentially identify new treatment options (34).

Haematoxylin and eosin (H&E) stained digitised tissue sections

H&E are two dyes used in histology to allow visualisation of cellular structures in a tissue sample: haematoxylin is used to stain the cell nuclei in blue/dark blue, while eosin stains in pink/red the rest of the tissue (cytoplasm, connective tissue, and matrices) (35). The results of the staining can be observed through a microscope or digitally after scanning of the slides with the stained tissue section. This staining method is the routine staining for all tissue samples in histopathology laboratories worldwide including biopsies suspicious to contain cancer or resection specimens after the cancer was surgically removed.

Magnetic resonance imaging (MRI)

MR imaging is a medical imaging technique which probes the atomic and molecular structure of tissues by aligning then disturbing the spin of protons and measuring the radio-frequencies resulting from realignment in a magnetic field generated by the scanner, converting them into an image of the body part under scrutiny. This non-invasive imaging technique is used to preferably analyse soft tissue but has also been used in cancer detection and staging (36).

Contrast enhanced mammography (CEM)

CEM is a medical imaging technique which is performed after intravenous injection of an iodinated contrast agent. Image acquisition is made at two time points and two energies, leading to one scan called low-energy image, similar to a conventional mammogram, and one high-energy image which is more sensitive to the contrast agent. These two images are combined to form an image where the contrast enhancement becomes visible (37). This image modality has a better specificity for classifying lesions than a regular mammogram, which means that with CEM, less patients would be recalled for an additional examination if they have a benign lesion (38). This technique is mainly used during recall of patients who underwent breast screening with regular mammography and had a possible malignant lesion detected or the exam was inconclusive.

Hypothesis: model predictions can be improved combining handcrafted features from various sources and deep learning

DL can be used for detecting and segmenting images, whereas handcrafted features were not designed to do this as they are computed on a predefined region of interest. However, a handcrafted feature-based model needs less data to train, can more easily be implemented for regression tasks (for example to predict best treatment or survival), and although radiomics feature can be abstract, feature importance is easy to retrieve and helps interpreting the findings. DL can learn optimal feature representation from a dataset without *a priori* knowledge but lacks explainability and interpretability making it sometimes impossible to understand false negative or false positive results. Moreover, DL outperforms handcrafted feature-based ML only on very large datasets (as a rule of thumb, more than one thousand samples) (39).

In comparison, handcrafted features extracted from a region of interest are predetermined from a list of features. This list might not be comprehensive, resulting in a final model which might not necessarily be trained on the most relevant features which could be extracted from the images. However, the results are easier to explain once a set of features are pre-selected before training a model. Whilst this method needs less input data for a classification task, it nevertheless requires delineations of the dataset by experts, a time and resource-consuming task.

Based on the above mentioned information, we formulated the following hypothesis for this thesis: **Feature-based ML models and DL models capture information from medical imaging datasets which is complementary and their combined use can result in more accurate classification predictions.**

Scope and table of contents

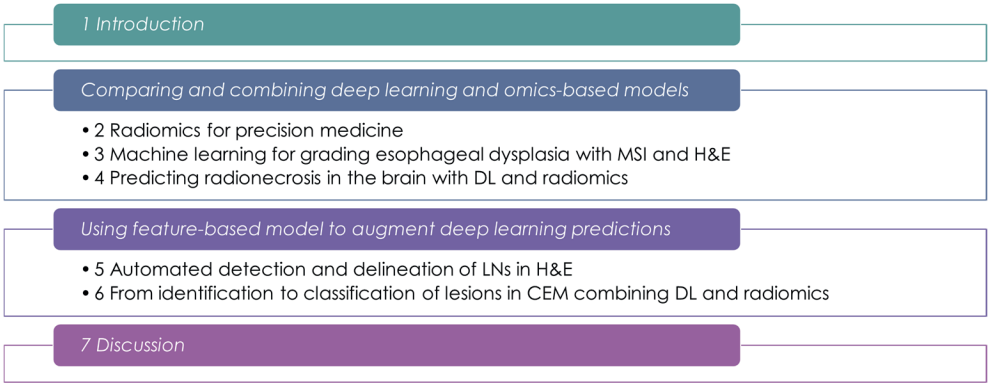


Figure 2: Roadmap of this thesis.

We divided this thesis in two parts.

Part 1: Comparing and combining deep learning and feature-based machine learning

Feature-based ML and DL might learn complementary classifiers, thus, combining the results obtained from those models could potentially lead to a more accurate and robust overall model.

In **Chapter 2**, we reviewed the literature on the current use of ML with handcrafted radiomics and DL for medical image analysis. This review also explored the challenges faced by using ML for medical imaging analyses, including issues with reproducibility of models and lack of explainability, and suggests a potential framework to overcome those issues.

The study presented in **Chapter 3** focussed on datasets from patients with Barrett's oesophagus. This disease is a known precursor of oesophageal cancer and is characterized by a change in the composition of the lining epithelium where squamous epithelial cells are replaced by intestinal-type columnar cells. Progression of Barrett's oesophagus towards high grade dysplasia and cancer is not yet predictable based on histology alone (40). Our work explored the prediction capacity of two set of data using ML: mass spectrometry imaging (MSI) and images of Haematoxylin and Eosin (H&E)-stained tissue sections to first grade dysplasia in Barrett's oesophagus and then predict disease progression for patients with low grade dysplasia. In this study, MSI and H&E-stained images were acquired in parallel and co-registered to allow comparison.

Our study presented in **Chapter 4** was performed using ML with handcrafted radiomics and DL on pre- stereotactic radiotherapy brain MRIs of patients with brain metastases. Our goal

was to test the prediction power of radiomics, patient characteristics, and deep learning first individually and then in combination for predicting adverse radiation effect risk. Different pre-processing methods were first tested independently for the radiomics and deep learning predictions and the pre-processing methods giving best results on the test dataset were kept.

Part 2: Using feature-based models to augment deep learning predictions

Another use for feature-based models is to possibly improve and explain predictions computed with deep learning models:

In **Chapter 5**, we explored the potential added value of using a feature-based machine learning model based on the predictions of a regular deep learning model which was trained to find and segment lymph nodes (LNs) within histopathology images of oesophageal cancer resections. We wanted to see whether the DL model performance would improve when adding a feature-based machine learning model. Thus, to recognize whether LNs are present in the image, we gave a prediction score per potential LN. We also studied whether the score could be used as an uncertainty measurement. Our model was tested on an external validation dataset to test whether the results were reproducible.

In **Chapter 6**, we tested the combination of radiomics-based models and deep learning models to automatically classify suspicious lesion within contrast enhanced mammography images. Our goal was to first train a deep learning model which would identify, delineate and classify suspicious lesions automatically. We then added a radiomics-based model based on the ground truth contours classifying the lesions (benign versus malignant) and compared and combined the prediction results with the predictions of the deep learning model. Finally, we trained a new radiomics-based model on the contours automatically generated by the deep learning model, classifying the findings as malignant or other (benign and false positive).

In **Chapter 7**, we discuss our findings and the future prospects of medical imaging analysis.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 2021;71(3):209-249. doi: https://doi.org/10.3322/caac.21660
2. Hussein I, Raad M, Safa R, Jurjus RA, Jurjus A. Once upon a microscopic slide: the story of histology. Journal of Cytology & Histology 2015;6.
3. Cancer Imaging Program (CIP). https://imaging.cancer.gov/imaging_basics/cancer_imaging/uses_of_imaging.htm. Published 2016.
4. Cancer - Screening and early detection. https://www.who.int/europe/news-room/fact-sheets/item/cancer-screening-and-early-detection-of-cancer. Published 2010.
5. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. Annals of Internal Medicine 2020;172(1):46-56. doi: 10.7326/m19-2125 %m 31766052
6. Bedell SL, Goldstein LS, Goldstein AR, Goldstein AT. Cervical Cancer Screening: Past, Present, and Future. Sexual Medicine Reviews 2020;8(1):28-37. doi: https://doi.org/10.1016/j.sxmr.2019.09.005
7. Cancer - Diagnosis and treatment - Mayo Clinic. Mayoclinic.org. https://www.mayoclinic.org/diseases-conditions/cancer/diagnosis-treatment/drc-20370594. Published 20-05-2022.
8. How Cancer Is Diagnosed - NCI. National Cancer Institute. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis. Published 2015. Updated 2015/03/09/08:00.
9. Cancer Staging - NCI. National Cancer Institute. https://www.cancer.gov/about-cancer/diagnosis-staging/staging. Published 2015. Updated 2015/03/09/08:00.
10. active surveillance - NIH. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/active-surveillance. Published 2011. Updated 2011/02/02/07:00.
11. Active Surveillance for Prostate Cancer. https://www.hopkinsmedicine.org/health/conditions-and-diseases/prostate-cancer/active-surveillance-for-prostate-cancer. Published 2021. Updated 2021/08/08/.
12. Pereira GC, Traughber M, Muzic RF. The Role of Imaging in Radiation Therapy Planning: Past, Present, and Future. BioMed Research International 2014;2014:231090. doi: 10.1155/2014/231090
13. Shabbir J, Anwer T. Artificial intelligence and its role in near future. arXiv preprint arXiv:180401396 2018.
14. Giger ML, Chan H-P, Boone J. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. Medical Physics 2008;35(12):5799-5820. doi: https://doi.org/10.1118/1.3013555

15. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Published 2021. Updated 2021/09/22/Wed, - 12:25.
16. van Leeuwen KG. AI for Radiology. //www.AIforRadiology.com/. Published 2022.
17. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology 2021;31(6):3797-3804. doi: 10.1007/s00330-021-07892-z
18. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nature Medicine 2021;27(5):775-784. doi: 10.1038/s41591-021-01343-4
19. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer (Oxford, England : 1990) 2012;48(4):441-446. doi: 10.1016/j.ejca.2011.11.036
20. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging 2020;11(1):91. doi: 10.1186/s13244-020-00887-2
21. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. European Journal of Radiology 2020;127:108991. doi: https://doi.org/10.1016/j.ejrad.2020.108991
22. Kather JN, Weis C-A, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, Zöllner FG. Multi-class texture analysis in colorectal cancer histology. Scientific Reports 2016;6(1):27988. doi: 10.1038/srep27988
23. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, Zamarin D, Roche KL, Liu Y, Patel D, Aukerman A, Pasha A, Rose D, Selenica P, Causa Andrieu PI, Fong C, Capanu M, Reis-Filho JS, Vanguri R, Veeraraghavan H, Gangai N, Sosa R, Leung S, McPherson A, Gao J, Lakhman Y, Shah SP, Consortium MM. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature Cancer 2022;3(6):723-733. doi: 10.1038/s43018-022-00388-9
24. PubMed.gov search. PubMed.gov: National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/?term=deep+learning&filter=years.2014-2021&timeline=expanded. Published 2022. Accessed 2022.
25. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. European conference on computer vision: Springer, 2014; p. 740-755.
26. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 2015;115(3):211-252. doi: 10.1007/s11263-015-0816-y
27. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML. Deep learning in medical imaging and radiation therapy. Medical Physics 2019;46(1):e1-e36. doi: https://doi.org/10.1002/mp.13264

28. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit M-C, Dinapoli N, Dinh CV, Echegaray S, Naqa IE, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Fernandez JS, Spezi E, Steenbakkers RJHM, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, Dijk LVv, Griethuysen Jv, Velden FHPv, Whybra P, Richter C, Löck S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295(2):328-338. doi: 10.1148/radiol.2020191145

29. Pérez-García F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine 2021;208:106236. doi: https://doi.org/10.1016/j.cmpb.2021.106236

30. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, Choyke PL, Turkbey B. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. J Med Imaging (Bellingham) 2021;8(1):010901. doi: 10.1117/1.Jmi.8.1.010901

31. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal 2018;16:34-42. doi: https://doi.org/10.1016/j.csbj.2018.01.001

32. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP. Preparing Medical Imaging Data for Machine Learning. Radiology 2020;295(1):4-15. doi: 10.1148/radiol.2020192224

33. Buchberger AR, DeLaney K, Johnson J, Li L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. Anal Chem 2018;90(1):240-265. doi: 10.1021/acs.analchem.7b04733

34. Berghmans E, Boonen K, Maes E, Mertens I, Pauwels P, Baggerman G. Implementation of MALDI Mass Spectrometry Imaging in Cancer Proteomics Research: Applications and Challenges. J Pers Med 2020;10(2):54. doi: 10.3390/jpm10020054

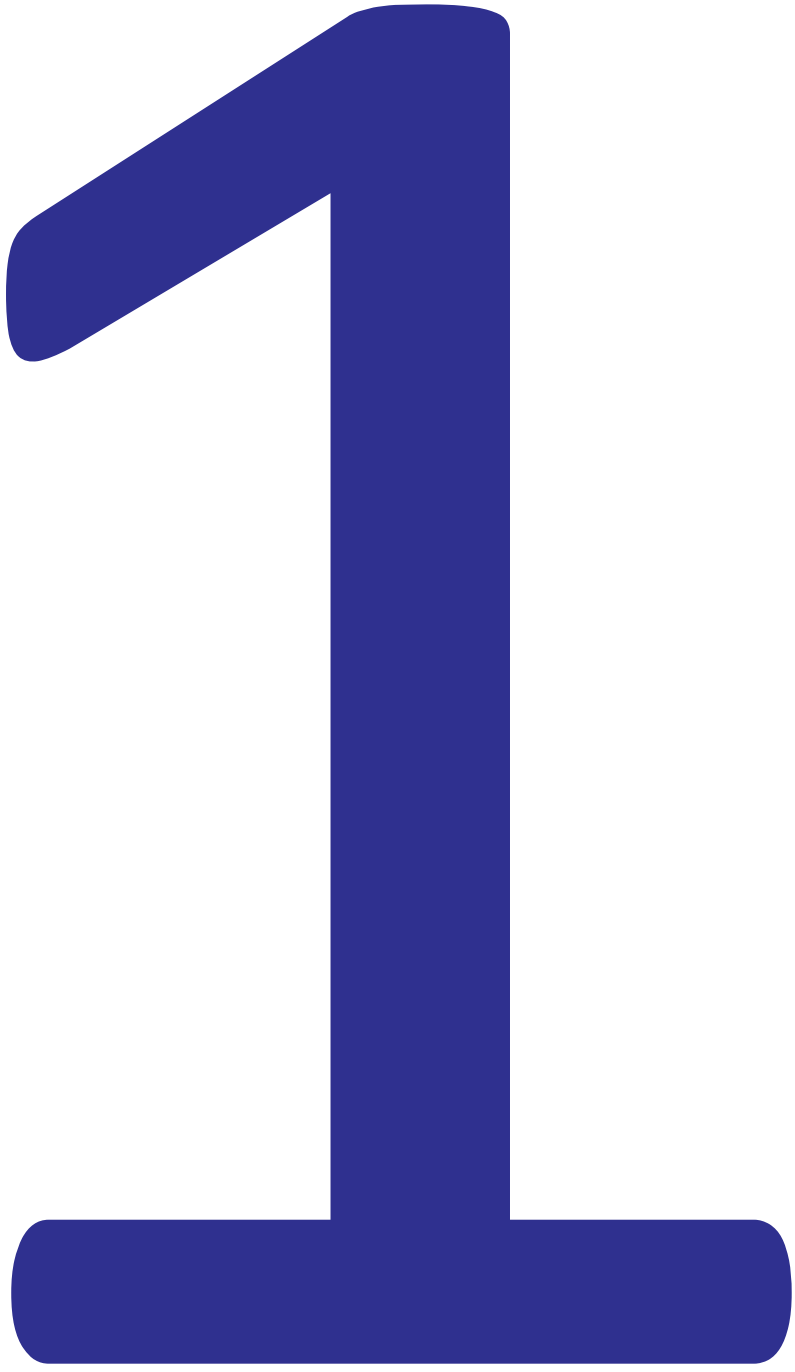
35. Bancroft JD, Layton C. The hematoxylins and eosin. Bancroft's theory and practice of histological techniques 2012:173-186.

36. Magnetic Resonance Imaging (MRI). Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/magnetic-resonance-imaging-mri. Published 2021. Updated 2021/12/06/.

37. Lobbes MBI, Lalji U, Houwers J, Nijssen EC, Nelemans PJ, van Roozendaal L, Smidt ML, Heuts E, Wildberger JE. Contrast-enhanced spectral mammography in patients referred from the breast cancer screening programme. European Radiology 2014;24(7):1668-1676. doi: 10.1007/s00330-014-3154-5

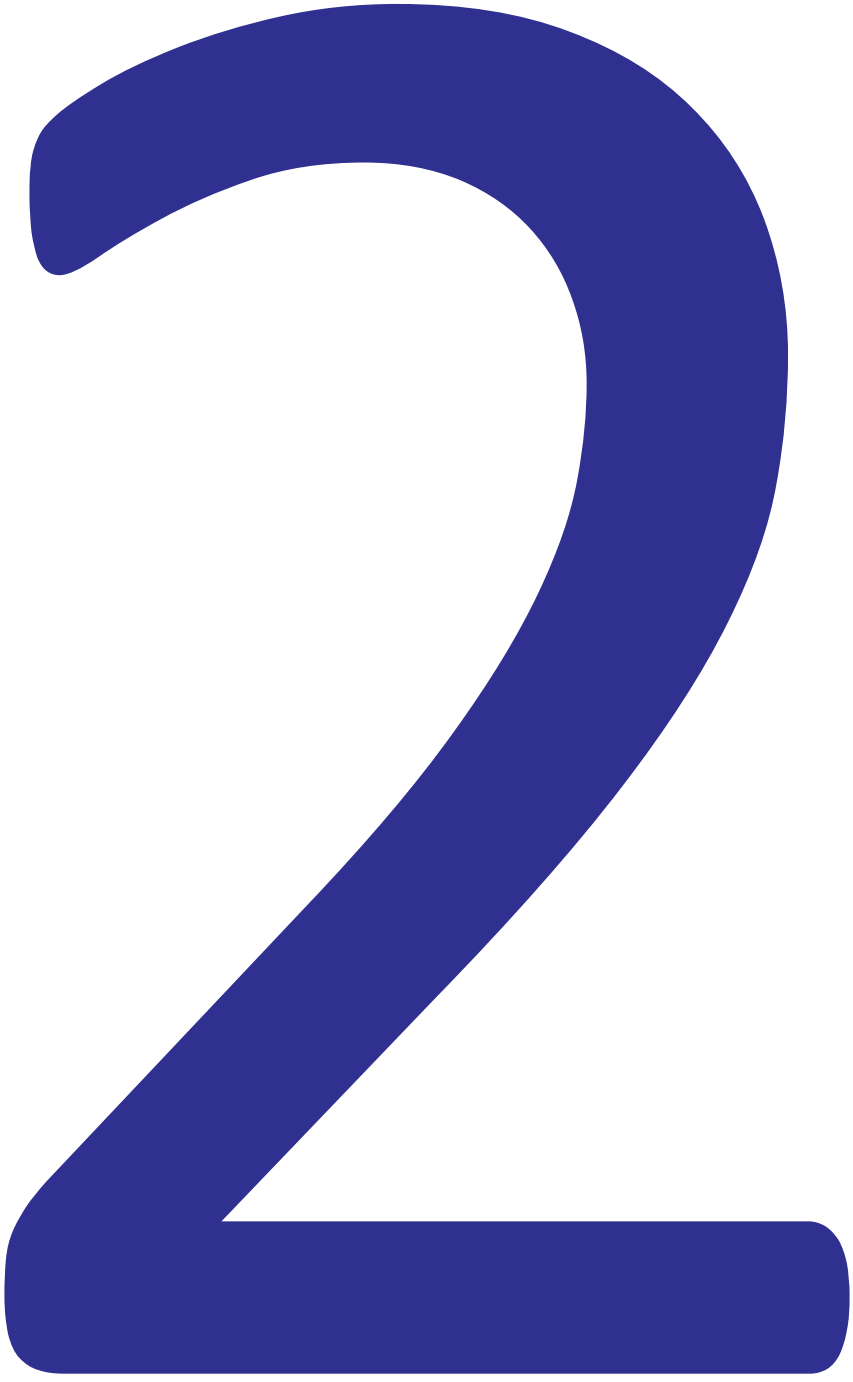
38. Cozzi A, Magni V, Zanardo M, Schiaffino S, Sardanelli F. Contrast-enhanced Mammography: A Systematic Review and Meta-Analysis of Diagnostic Performance. Radiology 2022;302(3):568-581. doi: 10.1148/radiol.211412

39. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F. AI applications to medical images: From machine learning to deep learning. Physica Medica 2021;83:9-24. doi: https://doi.org/10.1016/j.ejmp.2021.02.006
40. Gross SA, Kingsbery J, Jang J, Lee M, Khan A. Evaluation of dysplasia in Barrett esophagus. Gastroenterol Hepatol (N Y) 2018;14(4):233.



PART 1

Comparing and combining deep learning and feature-based machine learning



Chapter 2

Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework

Abdalla Ibrahim, Sergey Primakov[1] , Manon Beuque[1] , Henry C. Woodruff, Iva Halilaj, Guangyao Wu, Turkey Refaee, Renee Granzier, Yousif Widaatalla, Roland Hustinx , Felix M. Mottaghy, Philippe Lambin

[1] These authors contributed equally

Adapted from:
Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, Refaee T, Granzier R, Widaatalla Y, Hustinx R, Mottaghy FM, Lambin P. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 2021; 188:20-29. doi: https://doi.org/10.1016/j.ymeth.2020.05.022

Abstract

The advancement of artificial intelligence concurrent with the development of medical imaging techniques provided a unique opportunity to turn medical imaging from mostly qualitative, to further quantitative and mineable data that can be explored for the development of clinical decision support systems (cDSS). Radiomics, a method for the high throughput extraction of handcrafted features from medical images, and deep learning the data driven modeling techniques based on the principles of simplified brain neuron interactions, are the most researched quantitative imaging techniques. Many studies reported on the potential of such techniques in the context of cDSS. Such techniques could be highly appealing due to the reuse of existing data, automation of clinical workflows, minimal invasiveness, three-dimensional volumetric characterization, and the promise of high accuracy and reproducibility of results and cost-effectiveness. Nevertheless, there are several challenges that quantitative imaging techniques face, and need to be addressed before the translation to clinical use. These challenges include, but are not limited to, the explainability of the models, the reproducibility of the quantitative imaging features, and their sensitivity to variations in image acquisition and reconstruction parameters. In this narrative review, we report on the status of quantitative medical image analysis using radiomics and deep learning, the challenges the field is facing, propose a framework for robust radiomics analysis, and discuss future prospects.

1. Introduction

Advances in artificial intelligence applications, combined with those in medical imaging, have led to the gradual conversion of digital medical images into high-dimensional data appropriate for data mining and data science techniques (1). Meanwhile, computing power and quantitative image analysis (QIA) techniques have made enormous progress, and the application of quantitative imaging techniques on medical imaging gained exponential momentum (2). Currently, radiomics and deep learning are the most researched techniques on medical imaging.

Broadly, radiomics refers to the use of computational or statistical approaches to extract large numbers of quantitative features from a number of medical imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), to develop predictive models ultimately aiming to enable personalized clinical management (3–5). Radiomics features are quantitative descriptions of the intensity, shape, volume, and texture of the region of interest (ROI), with the recent addition of more abstract features such as radial gradient and radial deviation (6). Radiomics features are broadly divided into histogram-based and texture features. Different statistical methods are used to calculate the radiomics features. The methods include first-order statistics, which depends on the values of single voxels (histogram-based features for e.g. maximum and minimum intensity); second-order statistics, which depends on the relation between two voxels (for e.g. grey-level co-occurrence matrix (GLCM) features), and higher-order statistics (relations among three or more voxels, for e.g. neighborhood grey-tone difference matrices (NGTDM) features) (7,8). The main hypothesis behind radiomics analysis is that radiomics features decode or correlate with the molecular characteristics, phenotype, and genotype of the region of interest (ROI) under study. This information can be used in combination with other patient information to improve patient management. Moreover, as the tumours are of heterogeneous nature (9,10), clinical approaches, such as tissue biopsies, might fail to characterize the entirety of the tumour (11). In contrast, Radiomics takes the whole tumour region (or even the surrounding or healthy tissue) into account, which enables a better characterization (3). Furthermore, frequent clinical imaging can transform radiomics into a non-invasive, easily repeatable, and cost-effective longitudinal approach for cDSS (12).

Deep learning (DL) is a field of data driven modelling techniques that utilizes the principles of simplified neuron interactions (13). Using artificial neurons started to draw attention decades ago (14), but it only became a major research focus recently (15–17). The artificial neuron model is used as a foundation unit to create complex chains of interactions – DL layers. These layers are used to generate even more complex structures DL architectures (see Figure 1). The neural network (NN) training procedure is typically a cost-function minimization process. The cost function measures the error of predictions based on the ground truth labels (18). Due to the high complexity of the network architectures, computational limitations are reached when trying to solve the optimization task analytically. Henceforth, iterative algorithms are used to overcome this issue. Commonly,

these algorithms are variations of the gradient descent (GD). GD iteratively moves in the direction of steepest descent of the cost function, in order to find a local minimum. During the model training process, every image from the training dataset contributes to the cost minimization process. Thereby, a DL network learns how to solve a problem directly from existing data, and apply it to data it has never seen. These complex models contain the parameters (weights) for millions of neurons, which can be trained for the recognition of problem-related patterns in the data being analyzed. DL has been shown to be efficient in other fields, such as face recognition (19) and autonomous cars (20).

Since the introduction of the field, many studies have reported on the potential of such techniques for predicting patient outcomes (5,21,22). The successful translation of QIA techniques into cDSS will have a significant impact on the clinical workflow and current patient management protocols. Clinicians will be able to non-invasively obtain a more detailed and accurate tumour characterization, in a shorter amount of time. Patients will have to go through less invasive procedures, while having treatment optimized based on their individual characteristics. Furthermore, patient-specific informed decisions can be made with more confidence. However, QIA is still developing in the field of medical imaging and several challenges, including the stability and reproducibility of imaging biomarkers, as well as the interpretability of the developed algorithms, need to be addressed before QIA can be translated to clinical applications.

In this narrative review, we focus on the current status of the potential of radiomics and deep learning to be incorporated in clinical decision support systems (cDSS), their challenges, as well as future prospects for these methods. We further propose a workflow to guide robust radiomics analysis.

2. Quantitative image analysis for precision medicine

The need for personalizing the management of patients has been widely reported (23,24). QIA represents a suitable candidate to be incorporated into the body of personalized medicine due to the non-invasive three-dimensional characterization of the ROIs, the availability of vast amounts of medical images, the longitudinal capabilities, and the cost-effectiveness of the method.

The currently implemented imaging biomarker development workflow is generalizable across different imaging modalities. The workflow can be described as consecutive steps divided into the main categories of data collection, image segmentation, features extraction, development of the signature, and evaluation of the performance (Figure 2), with the segmentation step being optional in the case of deep learning. The workflow has been previously extensively described (22,25).

Many studies have investigated and reported on the added clinical value of radiomics features for predicting various clinical outcomes, such as overall survival, tumour histology, response to therapy, and genetic profiling, among other endpoints. Furthermore, these studies were performed on various imaging modalities, including CT, MR, and PET.

While the handcrafted radiomics pipeline necessitates the use of machine learning or statistical algorithms after feature extraction for modeling, DL techniques perform feature

extraction and modelling internally without the need for further user interaction. DL has its own advantages and drawbacks compared to traditional radiomics. One of the key benefits of using DL is avoiding the contouring problem, the bottleneck of a traditional radiomics pipeline. However, due to the complexity of DL models, it is easier to overfit the model to the training data. As a result, a larger data set is needed for DL compared to handcrafted radiomics. Furthermore, DL is considered a 'black box', i.e the models and features generated are not (or barely) interpretable. This is currently one of the major challenges of the application of artificial intelligence (AI) in medical image analysis. Efforts are being made towards providing explainable AI algorithms, by investigating the correlation of the chosen features with biologic or semantic characteristics. Such correlations would provide an understanding about how the algorithm makes the decision, and ease its incorporation into cDSS.

QIA techniques have a great potential for involvement in developing classification, prognostic and predictive clinical tools. In comparison, classification tasks (for e.g classifying tissue histology) seem to yield a better performance than predictive tasks (for e.g survival prediction). This is in part due to the unaccounted for variables when trying to predict future events. In 2.1 and 2.2, we report on some examples that highlighted the potential of radiomics and deep learning to predict various clinical endpoints, acknowledged or addressed the challenges of QIA techniques used, and/or applied the techniques on a relatively large sample size compared to other studies addressing the same clinical endpoint.

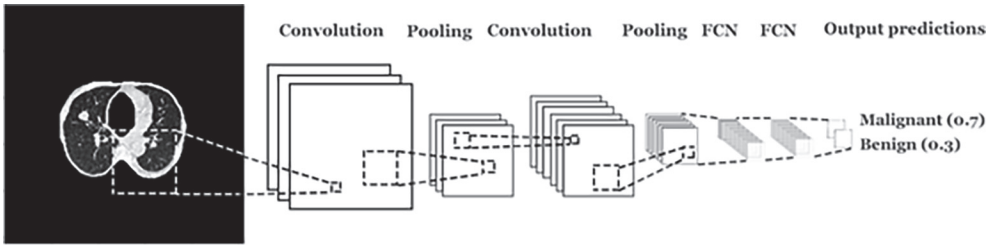


Figure 1. Graphical depiction of DL architectures. * FCN: fully connected network.

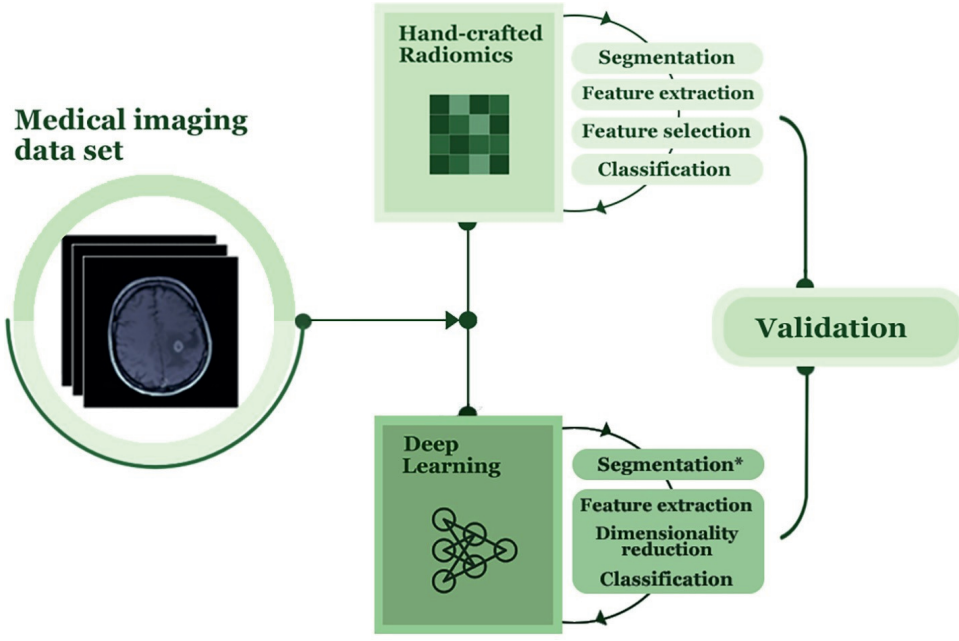


Figure 2. Development of imaging biomarkers using quantitative image analysis. * Segmentation is not a necessity in the automated radiomics pipeline.

2.1. Handcrafted radiomics

Overall survival

Wang et al. (26) investigated the potential of radiomics signatures to predict overall survival in patients with locally advanced rectal cancer. The authors tried to address the current clinical need for a risk stratification tool for such patients to safely forgo surgical resection, due to the high comorbidities associated. The study included 411 treatment planning CT-scans of patients treated with neoadjuvant chemotherapy followed by surgery. The authors developed a radiomics signature that could stratify patients into low- and high-risk survival groups. The radiomics features included in the signature were found to be independent of the clinical features. Adding radiomics features to the clinical model resulted in an improvement of the predictive power (c- index) of the clinical only model from 0.67 (0.62–0.73) to 0.73 (0.66–0.80) (26). The authors used two investigations to ensure the selection of stable radiomics features, namely test–retest and contour- recontour robustness analysis. The results signifies the added value of properly using radiomics analysis on CT scans in improving patients' risk stratification. Yet, the authors did not externally validate their signature, casting doubt on the generalizability of their signature. It is expected to be of value in cases where the scanning parameters are identical to those used in the study.

Another study by Bae et al. (27) investigated the potential of MR- based radiomics to improve the survival prediction of patients diagnosed with glioblastoma multiforme. The study is an effort to address the unmet clinical need for assessing the survival of the target group following therapy. The authors extracted radiomics features from 217

multiparametric MR scans of patients with glioblastoma. The authors identified 18 radiomics features to build a radiomic signature, and reported that the addition of radiomics features to clinical and genetic profiles of the patients significantly improves the stratification of patients (27). The authors in this study applied a unique approach for the analysis by simultaneously analyzing radiomics features extracted from different co-registered MR sequences. The identified features were independent of the clinical and genetic factors, and the improvement in the survival prediction following their addition, supports the hypothesis of radiomics. Pitfalls in the study include the lack of assessment of radiomic feature stability before modeling, and as often seen in these studies, a lack of an external validation of the signature. However, their results support the hypothesis that radiomics are of great use when applied on scans acquired using identical settings.

Oikonomou et al. (28) reported on the potential of PET/CT-based radiomics to improve the survival stratification of patients with lung cancer treated with stereotactic body radiotherapy. The aim was to identify radiomics features that can improve the prognostication of patients following treatment. The authors extracted radiomics features from 150 PET/CT scans, and built radiomics signatures using 10 radiomics features. The authors reported that the radiomics signature was the sole predictor in the case of overall survival, and provided complementary information for the prediction of regional control (28). The uniqueness in this study is the joint use of radiomics features extracted from the CT-component and PET-component of the PET/CT scans. The authors show how other currently used clinical parameters fail to predict overall survival, while only radiomics could. While the study highlights the potential of radiomics to improve risk stratification, no external validation of the signature was performed.

Progression free survival

Kirienko et al. (29) investigated the role of PET/CT-based radiomics to predict disease free survival in patients with non-small cell lung cancer undergoing surgery. The authors extracted radiomics features from PET, CT, and combined PET/CT images. The authors developed Cox regression models using only CT, only PET, and combined PET/CT radiomics features. They reported that the radiomic signatures they developed improve the current clinical stratification of the targeted patients (29). The authors in this study investigated the reproducibility of radiomics features across the different imaging parameters in their dataset. This ensured selecting the comparable features before proceeding with signature building. The authors also provide evidence of the added value of combining radiomics features extracted from different imaging modalities. Furthermore, the ability to predict disease free survival from the time of diagnosis -which radiomics offer improves physicians and patients decision making. However, the authors in this study did also not perform an external validation of their signature. Further validation of the signature can prompt a prospective validation trial, before incorporation into cDSS.

Another study by Kickingereder et al. (30) investigated the role of MR-based radiomics in predicting survival in patients with glioblastoma multiforme. The authors extracted radiomics features from 119 MR scans, and developed a radiomic signature using 11

features. The developed signature performed significantly better than the radiologic and clinical risk models, and its addition to those resulted in an overall improvement of progression-free survival stratification (30). The finding that the radiomics signature performed better than the clinical and radiologic models supports the findings reported by Bae at al. (27), and adds more evidence that radiomics features decode complementary biologic information. However, the study did not address the issues of the reproducibility and generalizability sufficiently, leaving a room for improving the performance of radiomics.

Tumour histology

Wu et al. (31) explored the role of radiomics in differentiating between the histologic subtypes of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. The study was an effort to address the clinical need for less invasive and easily repeatable methods to determine tumour histology. The authors extracted radiomics features from 350 CT scans of NSCLC patients for whom the tumour histology has been determined from surgical specimens. The developed signature included 5 radiomics features, and they reported an area under the receiver characteristics curve (AUC) of 0.72 (31). This study reflected on the potential of non-invasive radiomic signatures to differentiate between adenocarcinoma and squamous cell carcinoma. They also investigated different machine learning methodologies for building the radiomics signature. While this study generates evidence for the potential of radiomics, the performance of the developed signature is significantly lower than the current gold standard -tissue biopsy. However, there is a great room for improving the development and performance of the signature. The authors did not address the acknowledged challenges in radiomics, nor did they validate their signature on an external dataset. Preselection of reproducible features, external and prospective validation of the signature are necessary steps in the development of radiomics biomarkers.
In another study, Wu et al. (32) investigated the added value of MR- based radiomics features for the prediction of hepatocellular carcinoma (HCC) grade. The authors extracted radiomics features from 170 MRI scans of HCC patients, whose tumour grade was identified through pathological samples. The radiomics-only signature (AUC of 0.74) outperformed the clinical model (AUC of 0.60), and the combination of both significantly improved the prediction (AUC of 0.80) (32). The authors in this study also combined radiomics features extracted from two different MR sequences and analyzed them simultaneously. The significant improvement of the predictions following the combination of clinical and radiomics features supports the independence of radiomics features from other clinical information. However, external validation of the developed signature is still a necessity before confidently performing prospective validation.
Valleries et al. (33) explored the potential of the combination of FDG-PET- and MR- based radiomics features to classify lung nodules. The authors extracted radiomics features from 51 PET and MR scans of histologically confirmed lung lesions in patients with soft-tissue sarcoma. The authors achieved a sensitivity of 0.96 and specificity of 0.93 in diagnosing metastatic nodules using a model with combined radiomics features from both PET and MR modalities. The authors used a novel interesting approach by simultaneously analyzing the

features extracted from FDG-PET and MR scans, and were the first to show the potential of this method. The performance of the developed signature makes it a suitable alternative for patients for whom tissue biopsy is contraindicated. Its possible translation to cDSS might significantly improve patient outcomes, as treatment is based on the histologic diagnosis. Yet, further external and prospective validation of the signature is needed.

Response to therapy

Trebeschi et al. (34) explored the role of radiomics in predicting response to anti-PD1 immunotherapy in patients diagnosed with advanced melanoma and NSCLC patients. Immunotherapy has shown promising results. Yet, there is still a need for a tool to determine which patients will benefit from receiving anti-PD-1 antibodies. The authors extracted radiomics features from 1055 ROIs segmented on 203 CT scans. The authors developed a radiomic signature that could predict the response to therapy with an AUC of 0.76; showing the potential of radiomics to predict response to therapy in such patients (34). Interestingly, the authors found correlations between the radiomic biomarker and the genes associated with cell cycle progression and mitosis. Radiomics can become a tool for assisting decision making in immunotherapy, a great unmet clinical need. The study however did not externally validate the signature, and did not sufficiently address the issues of feature stability and reproducibility. Therefore, the application of the developed signature is also limited to the patients who are scanned with the same scanning parameters as used in the training.

In a study by Horvat et al. (35), the authors investigated the role of radiomics in assessing complete clinical response (cCR) after neoadjuvant chemoradiotherapy (CRT) in patients with locally advanced rectal cancer. The guidelines of treating these patients include surgery, but evidence showed recently that a select group of patients can be safely treated with only CRT. The authors extracted radiomics features from 114 MR scans, and developed a radiomics signature with a sensitivity of 1.00, and a specificity of 0.91, which outperformed qualitative assessment of the response performed by two radiologists. The current clinical standard evaluation of cCR includes digital rectal examination and endoscopy, with an accuracy ranging between 0.71 and 0.88 (35). The developed radiomic signature showed the highest accuracy among the available compared-with tools. Nonetheless, several steps to improve the methodology and performance of the radiomics signature could be made. The sound cCR evaluation following RCT can improve the patient management by eliminating surgical risks, time and money.

2.2. Deep learning

The application of deep learning on medical imaging could potentially fulfil more complicated tasks than handcrafted radiomics, especially when large amounts of data are available. Furthermore, as definition of the ROIs is not a necessity in the automated deep learning workflows, the algorithm will learn patterns from the whole image and possibly

make connections with the habitat of the ROIs. The applications of neural networks on medical imaging are also not limited to classification and prediction of clinical end points, but can extend to include other tasks, such as the detection and segmentation of abnormalities or target volumes, which have been investigated for decades (36). Especially the detection and segmentation of lesions can be easily incorporated into the radiomics workflow, further automating the process. In the following paragraphs, we give examples of different applications of DL on medical imaging to perform various tasks on datasets acquired with one of the three main medical images modalities: CT, MRI, and PET.

Automatic segmentation of target structures

Jiang et al. (37) tried to develop a DL model that is able to accurately perform volumetric lung tumour segmentation on CT images. The authors used two versions of multiple resolution residual network models for the delineation of the ROIs. The authors used 377 tumours from the open source dataset available on The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net) to train the model, and two independent datasets of 304 and 529 lung tumours to validate it. The dice similarity coefficient (DSC), which measures the spatial overlap of the segmentations, was computed to evaluate the performance of the model. The DSCs of the model on the two validation datasets were 0.75 and 0.68, respectively. The authors reported that there was no significant difference between the DL-generated mask and experts' segmentations (37). The new approach for segmenting medical images used in this study shows to be superior to the traditional use of UNet. The approach generalizes well on external data and overcomes the multiple sizes problem. The major pitfalls is that the authors did not use the 3D geometry of the CTs to compute the results, which would probably increase the performance significantly. The translation of such a tool to clinical practice will significantly reduce the time spent by the clinicians to plan the treatment, or evaluate the response to therapy. Moreover, from a research perspective, it can significantly reduce the time needed for radiomics research, and it will address the issue of inter-observer sensitivity of radiomics features.

In the study by Yi et al. (38), the authors developed a DL model for the segmentation of brain tumours based on 274 brain MRIs extracted from the Brain Tumour Image Segmentation Benchmark (BRATS) dataset (39). Segmentation of brain Glioblastoma on MRI is a time-exhaustive process, and an automated, accurate and reproducible tool for this purpose is considered a clinical need. The model was trained using four different MRIs sequences. The particularity of their convolutional neural network (CNN) model is a fixed difference of Gaussian filters as a first convolution layer, as it was proven to be the most efficient for 3D segmentation. The DSC for the model was 0.89 on the BRATS dataset when compared to ground truth segmentations (38). This article shows the superiority of 3D CNN compared to 2D CNN. The algorithm generated segmentations with a volumetric overlap of 0.89 with the experts' segmentations, which shows the potential of these tools for clinical use. However, the lack of external validation in the study limits the applicability of the algorithm to scanning parameters in the training set. The clinical practice can benefit from

such tools, as it significantly reduces the time the clinicians spend, and can provide more accurate evaluation of tumour response than the current clinical routine.

Chen et al. (40) explored the possibility of developing a DL model that is able to detect and segment cervical tumours on PET imaging. The authors proposed prior information constraint CNN (PIC-CNN), which integrates a CNN with prior information of cervical tumour. The authors reported a DSC of 0.84, which was superior to the other methods tin the comparison, including transfer learning based on fully convolutional neural networks (FCN) (DSC of 0.77), automatic thresholding (DSC of 0.59), and region growing method (DSC of 0.52) (40). The study highlights the potential of deep learning to perform well-defined and robust segmentations on PET imaging. The novelty of the approach is the use of prior information as input of the model, with delineation of the bladder. This extra information seems to give the traditional model an advantage compared to models that solely segment the tumours. However, the results were not validated on an external dataset. The application of the developed algorithm -after validating it would decrease the need for tissue biopsy, as well as the time spent on segmenting the tumours manually or semi-automatically.

Oncologic classification tasks

Ardila et al. (41) tried to predict the risk of lung cancer using screening low-dose CTs. The algorithm is trained on screening low-dose CT scans of patients who were known to be at risk. The authors trained their DL model on approximately 7000 scans, and validated its performance on 1139 cases. The authors reported that the model achieved the “state-of-the-art” performance (AUC of 0.944). Furthermore, the model outperformed all the radiologists (n = 6) who were asked to give predictions. The model resulted in a significant reduction in the false positive (11%), and false negative rates (5%) (41). While the current low-dose CT screening protocol has substantially improved in terms of consistency, it still faces major limitations represented in the inter-observer variability and incomplete characterization of image findings. The authors in (41) developed an algorithm that achieved significantly better performance than the current protocol, highlighting the potential of DL algorithms to revolutionize the field of lung cancer screening. Other advantages of the algorithm are that it eliminates the current clinical practice limitations.

Ismael et al. (42) investigated the ability of DL algorithms to classify different brain tumours. The algorithm predicts if the lesion is either a meningioma, glioma, or pituitary tumour. The authors developed the algorithm on 3064 T1 MRI images from 233 cancer patients. As input to the algorithm, the 2D images were considered independent from each other, and were split into 80% training and 20% testing, with strictly different patient data. The classifier used is ResNet50, a classic deep learning network, and the resultant balanced accuracy was 0.99 on a slice level and 0.97 at a patient level. This study shows that deep learning can very accurately classify brain tumours based solely on MRI data. However, the data to be used should be acquired using the same scanning parameters, as no external validation was performed in this study. There is a great clinical significance from the development of such a cDSS, as it eliminates the need for risky brain biopsies, while maintaining high accuracy.

In another study by Sibille et al. (43), the authors used the combination of CT, fluorine 18-fluorodeoxyglucose PET, atlas and PET maximum intensity projection (MIP) imaging to classify lung nodules. The study included a set of 629 patients who were diagnosed with either lung cancer or lymphoma. The authors developed models using each of imaging modalities separately, as well as in combination. The recommended algorithm achieved an AUC of 0.98 when both CT and PET were combined (43). This study shows that the combination of CT and PET can achieve an outstanding performance in terms of predictions. The current clinical practice requires the clinician to review and classify all of the increased-uptake foci in a PET/CT scan. The algorithm could help the clinicians to quickly read those images, after highlighting the suspicious areas and their most likely classification using DL.

Non-oncologic classification tasks

Walsh et al. (44) explored the potential of DL to classify fibrotic lung diseases using high resolution CT scans. The current clinical guidelines for classifying fibrotic lung diseases are based on high resolution scans, and diagnoses are made based on the semantic features identified by the radiologists. While these guidelines are the current gold-standard, it suffers greatly from inter-observer variability. The authors tried to address this unmet clinical need using DL approaches. The authors trained their DL model on 929 CT scans, and tested it on 139 scans. The authors reported a performance with human-level accuracy (0.76) (44). Of interest, the algorithm developed had a better agreement with expert radiologists than among them. The ease of application of such methods in clinical settings could benefit clinical practice, especially in centers where such clinical expertise is scarce.

In the study by Ding et al. (45), the authors tried to develop a DL model that is able to diagnose Alzheimer's disease (AD), using 18F-FDG PET scans of the brain. The current clinical guidelines to diagnose AD necessitate the interpretation of scans by an expert, and there is no definitive biomarker. To investigate the potential of DL, the authors collected two datasets: one used for training and testing the model (n = 2109 scans), which was split into 90% training and 10% testing; and an independent dataset (n = 40) for the validation of the model. The authors reported an AUC of 0.98, sensitivity of 1.00 and specificity of 0.82, using scans acquired 75.8 months on average before establishing the diagnosis. The model further outperformed the readers' performance (sensitivity of 0.57 and specificity of 0.91) (45). The significance in this study lies within the novelty of developing a biomarker for AD that is currently an unmet clinical need. In addition to the significantly better performance compared to human experts, the model can predict that the patient has AD in progression significantly earlier (~6 years). Such an application will revolutionize the clinical management of AD. However, prospective validation of this signature is needed before its translation to clinical practice.

Oh et al. (46) applied a DL based approach in order to classify the neuroimaging data related to AD. Authors used 694 MRI scans (T1- weighted MP-RAGE sequence) for solving several binary classification problems: AD vs. normal control (NC), progressive mild cognitive impairment (pMCI) vs. NC, stable mild cognitive impairment (sMCI) vs. NC and pMCI vs. sMCI. The authors utilized convolutional autoencoder- based unsupervised learning

algorithms in order to classify the AD vs. NC. Following that, the authors applied a supervised transfer learning approach to classify the pMCI vs. sMCI. The developed algorithms achieved accuracies of 0.87, 0.77, 0.63, and 0.73 for the AD, pMCI, sMCI and pMCI vs. sMCI classifications, respectively. In comparison to Ding et al. (45), the authors in this study used different DL approaches, and less numbers of patients were available for training and testing the algorithm. Furthermore, the difference in the imaging modality analysed in each study could justify the variation in performance, as AD begins with functional impairment rather than structural changes. Although the model developed by oh et al. (46) was outperformed by human experts, the authors demonstrated the possibility of end-to-end DL algorithms, which could be translated to clinical use after further optimization and prospective validation.

Response to therapy

Lou et al. (47) reported on the potential of DL models to predict response to radiotherapy in patients with lung cancer (primary or metastatic) using CT scans. Currently, all patients are treated similarly, while personalizing radiotherapy remains a desired, but unmet clinical need. The authors in this study collected a total of 849 scans for training the DL algorithm, and 95 scans to validate it. The authors developed a deep learning model (deep profiler) that computes and includes radiomics features in the deep-profiling process. A model combining the deep profiler and clinical variables is then used to calculate a risk score that is used to predict the response to treatment. The algorithm classifies patients into high and low risk groups, with a high performance (c-index of 0.72), which is significantly better compared to the results obtained with solely handcrafted radiomic models (c-index between 0.65 and 0.68) (47). The algorithm developed in this study opens new potentials for individualizing radiotherapy based on patients' sensitivity. Thereby, avoiding over- or under-treatment, and the side-effects of unnecessary treatment. Nevertheless, proper prospective validation of the developed algorithm remains a necessity.

Ypsilantis et al. (48) used convolutional neural networks to develop a model that is capable of predicting response to neo-adjuvant chemotherapy (NAC) in patients with esophageal cancer using PET scans. NAC is considered a standard of care in some cancers. While NAC has favourable outcomes in patients who respond, patients who do not end up with worse outcomes. To investigate the potential of QIA techniques, the authors collected 107 PET scans of patients diagnosed with esophageal cancer, treated with NAC, and followed-up to determine response. The authors compared the performance of handcrafted radiomics with deep learning approaches. The authors reported that the developed deep learning algorithm outperformed the handcrafted radiomics model, and achieved a sensitivity of 0.81 and specificity of 0.82 (48). The algorithm developed in this study highlights the potential of using DL to predict patients' response to therapy at baseline, which is considered a substantial clinical added value.

3. Challenges and future directions

Biomarkers are defined as "objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly" (49). The core of choosing a biomarker is the ability to measure it objectively. The reproducibility of imaging quantitative features across different imaging parameters is currently the steepest hurdle in QIA. As more research is being performed, other challenges, such as the sensitivity of QIA features to variations in the segmentation of the ROIs; and the lack of feature reproducibility across different implementations of radiomics toolboxes, are becoming increasingly clear.

3.1. The stability and reproducibility of quantitative features

Since the first landmark study in radiomics by Aerts et al. (50), the sensitivity of radiomics features to repeated acquisitions has been acknowledged. The authors performed a test-retest stability investigation and used 100 out of 440 calculated radiomics features based on the stability rank of the features. The authors also acknowledged the sensitivity of features to differences in segmentations, and performed a primary feature selection based on the features' robustness with regards to differences in both test-retest and segmentations. More recently, several studies reported on the sensitivity of radiomics features to temporal changes in test-retest studies across different modalities, including CT, MRI, and PET.

Anatomical imaging

Anatomical imaging (CT and MRI) is used to explore the underlying anatomical structures. CT imaging is standardized using the hounsfield units (HU) (51). On the other hand, MR imaging has no such standardized intensity measurements (52). Even though CT imaging uses standardized measurements, CT-based radiomics are not necessarily reproducible. Several studies reported that a significant number of CT- based radiomics features are not reproducible in test-retest settings, where the scans are acquired using the same scanning parameters (53–55). Other studies that investigated the reproducibility of CT-based radiomics features across different imaging acquisition and reconstruction parameters reported that the majority of radiomics features are significantly affected by such differences (53,56,57). Unreproducible radiomics features should be removed before starting the modeling of radiomics signatures. Therefore, it is always necessary to perform preselection of stable radiomics features based on the data under study, before starting the modeling.

MR-based radiomics is even more complex and challenging to standardize compared to CT based radiomics, as more factors -in addition to lack of standardized intensity measurements affect MR imaging (58). Some studies reported on the stability of various MR-based features. Fiset et al. (59) investigated the reproducibility of T2- weighted MRI of cervical cancer in three different settings: (i) test–retest; (ii) diagnostic MRI versus simulation MRI; (iii) interobserver variability. The authors reported that 22.6%, 6.2% and 74.4% of 1761 extracted radiomics features were reproducible across test-retest, diagnostic

versus simulation MRI, and different observers, respectively. Semi-parametric maps derived from specialized MRI sequences suffer less from the lack of stability: Peerlings et al. (60) reported on the stability of radiomics features extracted from apparent diffusion coefficient (ADC) map in test-retest and across different cancer types, centers, and vendors. The authors reported that out of 1322 extracted radiomics features, 122 features were stable across all cancers, centers, and vendors.

On top of these challenges, using contrast agents for imaging adds another level of complexity to the reproducibility of features, due to the differences in the cardiac function of patients being scanned. Changes in cardiac function can affect the time the distribution of the contrast in the body takes (61). Another factor in contrast-enhanced images is the difference in time between the injection of the contrast and scan acquisition, which might be slightly different across centers and protocols.

Functional imaging

Functional imaging is used to assess the metabolic activity of a region of interest, and includes the injection of radiopharmaceuticals. Some standardized measurements in PET are already being extracted and used in clinical practice, such as the standardized uptake value (SUV), and the metabolically active tumour volume (MTV) (7).

The challenges of radiomics for functional imaging are similar to the challenges of contrast-enhanced anatomical imaging radiomics, where the variability in the injected radiopharmaceutical activity, the time between injection and image acquisition, and acquisition time per bed position have profound implications on the reproducibility of radiomics features (62). In addition, functional imaging lacks anatomical specificity and suffers from low resolution, which could be addressed by the use of hybrid imaging (22). Tixier et al. (63) investigated the reproducibility of SUV measurements, intensity histogram features, intensity-size zone features, and co-occurrence matrices features. The authors acquired two 18F-FDG PET scans of 16 patients, with a 4-days' time interval. In contrast to further studies, the authors reported that these features were insensitive to the discretization range. Hatt et al. (64) investigated the robustness of PET based heterogeneity textural features with respect to the delineation of functional volumes and partial volume effects correction. The authors reported that these features were significantly affected by the differences in the delineation. The authors further reported that local features, e.g entropy and heterogeneity, were more robust when compared to regional features, e.g intensity variability and size-zone variability. Leijenaar et al. (65) investigated the role of SUV discretization on radiomics features. The authors used two different methods for SUV discretization, and reported that differences in SUV discretization significantly affect the reproducibility of 18F-FDG PET based radiomics features. The authors recommended the standardization of methodology for radiomics analysis. Altazi et al. (66) investigated the reproducibility of PET based radiomics features in cervical cancer patients. The authors investigated the reproducibility in three different scenarios: (i) manual versus computer-aided segmentations, (ii) gray-level discretization, and (iii) reconstruction algorithms. The authors extracted 79 PET radiomics features, and reported that the percentage of stable

features in the three scenarios were 13%, 5%, and 1% respectively. Shiri et al. (67) explored the effects of different reconstruction on 18F-FDG PET radiomics. The authors studied the effects of several factors including number of sub- iterations, number of subsets, full width at half maximum (FWHM) of Gaussian filter, and scan time per bed position and matrix size. The authors reported that 47% of the features were found to be robust, and these include shape, 44% of the intensity based features, and 41% of the texture based features. However, with changes in matrix size, the authors reported that only 6% of the features were robust.

The discrepancies in the reported percentages of stable/reproducible features across the reported studies are most likely linked to the variations between the datasets used in each of the studies in the scanners, and scans acquisition and reconstruction parameters combinations. However, these discrepancies are expected because of the different complexity of radiomics features, as well as the interaction between the different scanning parameters. All of the above mentioned studies reported that a variable percentage of radiomics features are affected, which highlights the necessity of performing feature stability/ reproducibility studies based on the data under analysis before performing radiomics analysis.

3.2. Sensitivity of quantitative imaging features to variations in the segmentation of the ROIs

In QIA, the medical images are converted to numerical arrays before feature calculation. Consequently, it is intuitive that differences in segmentations affect the quantitative imaging feature values variably, depending on the feature definition. Many studies have identified lists of radiomics features that are robust to variability in segmentations (50,68,69). Furthermore, with the inclusion of deep learning methods in image analysis, efforts are being made to develop reliable and reproducible automatic segmentations of various regions of interest as described in 3.2.1. Deep learning suffers less in this aspect, as the provision of ROIs is not obligatory.

3.3. The different implementations of radiomics feature extraction toolboxes

It is common knowledge in the radiomics community that different radiomics toolboxes use different pre-processing techniques and/or feature definitions, which lead(s) to variations in estimation of radiomics feature values when different software solutions are used. To address this issue, the radiomics community started an initiative – Imaging Biomarkers Standardization Initiative (IBSI) – that aims at standardizing radiomics feature extraction using different toolboxes (70). To date, the IBSI standardized the extraction of 169 radiomics features (71). Limiting the radiomics analysis to the IBSI standardized features can facilitate radiomics features interchangeability across platforms.

3.4. Future directions

To address the issue of radiomics features reproducibility, some harmonization methods have been investigated in the literature. Of the trending methods is Combine Batches (ComBat). ComBat is a statistical method that was developed to remove the batch effects in microarray expressions (72). While several studies have reported on the application of ComBat harmonization in radiomics analysis as a means to remove batch effects (73,74), its direct application on radiomics data is not in concordance with the mathematical definition of ComBat (72), or with the hypothesis that radiomics correlate with biology. This is because ComBat assumes that the differences between batches are attributed to two groups of factors, the first group refers to the biological covariates, which radiomics features are investigated for correlations with. Moreover, adding biologic covariates for ComBat in the training of radiomics signatures will hinder its prospective use, because it will be the outcome the radiomic signature tries to predict. The second group refers to the "non-biologic" factors, such as image acquisition and reconstruction parameters. ComBat was defined to handle one batch effect at a time. In contrast to gene expression arrays for which ComBat was designed, radiomics features have different complexity levels, which are expected to be non-uniformly affected by the variations in imaging parameters. In addition, the differences in image acquisition and reconstruction settings in a given retrospective imaging dataset are usually in more than one imaging parameter. The proper use of ComBat would require the assessment of the reproducibility of radiomics features after applying ComBat on representative objects with no biologic variations, such as phantoms. Then, radiomics features extracted from patients' scans acquired with the same imaging parameters can be transformed based on the location/scale parameters estimated by the application of ComBat on the phantom data. We here propose a framework for performing robust radiomics analysis (Figure 3). Nonetheless, a radiomics-specific harmonization method is still needed to eliminate the need for phantom studies, as the performance of ComBat is expected to be dependent on the variations in scanning parameters in the data. The workflow consists of consecutive steps, and can be used to preselect reproducible and harmonizable radiomics features. The first step in the workflow is the collection of retrospective patient imaging data to be analyzed. In the second step, scan acquisition and reconstruction parameters must be extracted from the collected patient data. The next step includes scanning a phantom with the parameters extracted from the patient imaging data. This allows the assessment of the reproducibility of radiomics features across the different scan acquisition and reconstruction parameters, and the selection of those features for performing robust radiomics analysis.

Based on our review of existing literature and our own experience, in order to use ComBat in the context of radiomics analysis (steps 5–7), two extra steps are needed. After selecting the features that are insensitive to the variations in the scanning parameters extracted from the patient data, features that are reproducible in test-retest in each of the combinations of those scanning parameters must be identified. ComBat is then applied on the features that are reproducible in test- retest but not across different scanning parameters. The concordance of radiomics features is assessed following the application of ComBat. The

location/scale shift parameters estimated by performing ComBat on the phantom data are then applied to the radiomics features extracted from patient data to harmonize them. The combination of the identified stable and harmonizable features can be further used to build the radiomics signature.

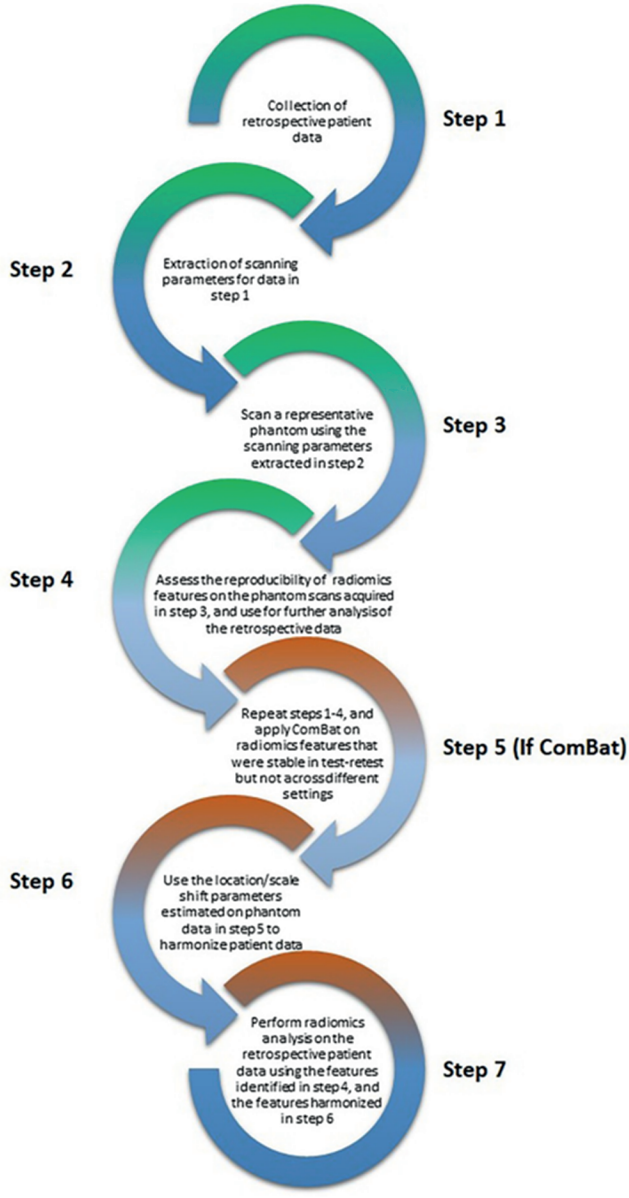


Figure 3. Proposed workflow for robust radiomics analysis.

The challenges discussed above raise questions about the future applications of radiomics, and the development of radiomic signatures as clinical biomarkers. To begin with, how to

approach the concept of external validation in radiomics studies. Do radiomic signatures need to be externally validated as is the case with other biomarkers, given all the challenges of reproducibility across different imaging settings? Or would the observatory prospective validation of a given signature in a specific image setting suffice? Does the development of radiomic signatures need to be specific for a scanner model and imaging settings? The ultimate solution will be the development of specific quantitative imaging parameters, as there is currently a clinical direction to personalize imaging settings per patient, which will have its toll on radiomics analysis. The direct application of radiomics analysis on data acquired heterogeneously could lead to spurious results, and inability of translating the results in a meaningful manner.

4. Conclusion

Quantitative imaging techniques (radiomics and deep learning) present a perfect candidate for personalizing patients' management. Applying these techniques in a sound manner can provide highly accurate and reproducible tools that minimize costs and time loss. However, to incorporate QIA in cDSS, the quantitative features should fulfil the definition of a biomarker, namely the stability and reproducibility. The future of quantitative image analysis in general lies within harmonizing the imaging protocols across centers and scanners, or within the development of a unique global protocol for quantitative analysis scans. Hence, the development of radiomics-specific tools to harmonize medical images and facilitate meaningful quantitative image analysis of the currently available retrospective data remains a necessity. Our proposed framework is expected to improve the robustness of radiomics analysis. Nevertheless, the benefits of the proper application and translation of QIA on medical imaging are undoubted. QIA techniques will be a valuable asset for both: the clinicians and patients. QIA can become an efficient means for aiding clinicians in risk stratification, early diagnosis, and improved management of patients.

Competing interests

Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr. Lambin has shares in the company Oncoradiomics SA and Convert pharmaceuticals SA and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/ NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patentable invention (softwares) licensed to ptTheragnostic/ DNAmito, Oncoradiomics and Health Innovation Ventures. Dr. Woodruff has (minority) shares in the company Oncoradiomics.

CRediT authorship contribution statement

A. Ibrahim: Conceptualization, Methodology, Formal analysis, Data curation, Writing - original draft, Project administration. S. Primakov: Formal analysis, Data curation, Writing - original draft, Visualization.
M. Beuque: Formal analysis, Data curation, Writing - original draft.
H.C. Woodruff: Supervision, Writing - review & editing. I. Halilaj:
Visualization. G. Wu: Resources, Data curation. T. Refaee: Resources.
R. Granzier: Resources. Y. Widaatalla: Resources. R. Hustinx:
Supervision. F.M. Mottaghy: Supervision, Writing - review & editing.
P. Lambin: Conceptualization, Methodology, Writing - review & editing, Project administration, Supervision.

Acknowledgements

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), ERC-2018-PoC (n° 81320– CL-IO). We further acknowledge the financial support from Maastricht-Liege imaging valley grant. This research is also supported by the Dutch technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL – n°673780), EUROSTARS (DART, DECIDE), the European Program H2020-2015-17 (BD2Decide - PHC30-689715, ImmunoSABR – n° 733008, PREDICT – ITN – n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 'CLEARLY'- n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine ('Euradiomics'). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2.

References

1. S. Walsh, E.E.C. de Jong, J.E. van Timmeren, A. Ibrahim, I. Compter, J. Peerlings, S. Sanduleanu, T. Refaee, S. Keek, R.T.H.M. Larue, Y. van Wijk, A.J.G. Even, A. Jochems, M.S. Barakat, R.T.H. Leijenaar, P. Lambin, Decision support systems in oncology, JCO Clin. Cancer Inform. 3 (2019) 1–9, https://doi.org/10.1200/CCI.18. 00001.

2. P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J.W.L. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (2012) 441–446, https://doi.org/10.1016/j.ejca. 2011.11.036.

3. R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2016) 563–577, https://doi.org/10.1148/radiol. 2015151169.

4. P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (2017) 749–762, https://doi.org/10.1038/nrclinonc.2017.141.

5. T. Refaee, G. Wu, A. Ibrahim, I. Halilaj, R.T.H. Leijenaar, W. Rogers, H.A. Gietema, L.E.L. Hendriks, P. Lambin, H.C. Woodruff, The emerging role of radiomics in COPD and lung cancer, Respiration 99 (2020) 99–107, https://doi.org/10.1159/ 000505429.

6. R.C. Hardie, S.K. Rogers, T. Wilson, A. Rogers, Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs, Med. Image Anal. 12 (2008) 240–258, https://doi.org/10.1016/j.media.2007.10. 004.

7. G.J.R. Cook, M. Siddique, B.P. Taylor, C. Yip, S. Chicklore, V. Goh, Radiomics in PET: principles and applications, Clin. Transl. Imaging 2 (2014) 269–276, https:// doi.org/10.1007/s40336-014-0064-0.

8. S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, G.J.R. Cook, Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis, Eur. J. Nucl. Med. Mol. Imaging 40 (2013) 133–140, https://doi.org/10.1007/s00259-012- 2247-0.

9. C. Swanton, Intratumor heterogeneity: evolution through space and time, Cancer Res. 72 (2012) 4875–4882, https://doi.org/10.1158/0008-5472.CAN-12-2217.

10. M. Gerlinger, A.J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N.Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C.R. Santos, M. Nohadani, A.C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P.A. Futreal, C. Swanton, Intratumour heterogeneity and branched evolution revealed by multiregion sequencing, N. Engl. J. Med. 366 (2012) 883–892, https://doi.org/10.1056/NEJMoa1113205.

11. T.M. Soo, M. Bernstein, J. Provias, R. Tasker, A. Lozano, A. Guha, Failed stereotactic biopsy in a series of 518 cases, Stereotact. Funct. Neurosurg. 64 (1995) 183–196, https://doi.org/10.1159/000098747.

12. S.S.F. Yip, H.J.W.L. Aerts, Applications and limitations of radiomics, Phys. Med. Biol. 61 (2016) R150–66, https://doi.org/10.1088/0031-9155/61/13/R150.
13. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.
14. W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133, https://doi.org/10.1007/ BF02478259.
15. L. Hongtao, Z. Qinchuan, Applications of Deep Convolutional Neural Network in Computer Vision, J. Data Acquisition Process. (2016). http://en.cnki.com.cn/ Article_en/CJFDTotal-SJCJ201601001.htm.
16. H. Shirani-Mehr, Applications of deep learning to sentiment analysis of movie reviews, Tech. Rep. NAVTRADEVCEN (2014) 1–8.
17. L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Information Processing 3 (2014), https://doi.org/ 10.1017/atsip.2013.9.
18. K. Janocha, W.M. Czarnecki, On loss functions for deep neural networks in classification, Schedae Informaticae. 1/2016 (2017). doi: 10.4467/20838476si.16.004. 6185.
19. A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018) 7068349, https://doi.org/10.1155/2018/7068349.
20. R. Simhambhatla, K. Okiah, S. Kuchkula, R. Slater, Self-Driving Cars: Evaluation of Deep Learning Techniques for Object Detection in Different Driving Conditions, SMU Data Science Review. 2 (2019) 23. https://scholar.smu.edu/datasciencereview/vol2/iss1/23/ (accessed May 14, 2020).
21. D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, https://doi.org/10.1146/annurev-bioeng- 071516-044442.
22. A. Ibrahim, M. Vallières, H. Woodruff, S. Primakov, M. Beheshti, S. Keek, T. Refaee, S. Sanduleanu, S. Walsh, O. Morin, P. Lambin, R. Hustinx, F.M. Mottaghy, Radiomics analysis for clinical decision support in nuclear medicine, Semin. Nucl. Med. 49 (2019) 438–449, https://doi.org/10.1053/j.semnuclmed.2019.06.005.
23. L.R. Cardon, H. Watkins, Waiting for the working draft from the human genome project. A huge achievement, but not of immediate medical use, BMJ 320 (2000) 1223–1224, https://doi.org/10.1136/bmj.320.7244.1223.
24. N.J. Schork, Personalized medicine: time for one-person trials, Nature 520 (2015) 609–611, https://doi.org/10.1038/520609a.
25. C. Parmar, J.D. Barry, A. Hosny, J. Quackenbush, H.J.W.L. Aerts, Data analysis strategies in medical imaging, Clin. Cancer Res. 24 (2018) 3492–3499, https://doi. org/10.1158/1078-0432.CCR-18-0385.
26. J. Wang, L. Shen, H. Zhong, Z. Zhou, P. Hu, J. Gan, R. Luo, W. Hu, Z. Zhang, Radiomics features on radiotherapy treatment planning CT can predict patient survival in locally advanced rectal cancer patients, Sci. Rep. 9 (2019) 15346, https://doi.org/10.1038/s41598-019-51629-4.

27. S. Bae, Y.S. Choi, S.S. Ahn, J.H. Chang, S.-G. Kang, E.H. Kim, S.H. Kim, S.-K. Lee, Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction, Radiology 289 (2018) 797–806, https://doi.org/10.1148/radiol.2018180200.
28. A. Oikonomou, F. Khalvati, P.N. Tyrrell, M.A. Haider, U. Tarique, L. Jimenez-Juan, M.C. Tjong, I. Poon, A. Eilaghi, L. Ehrlich, P. Cheung, Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy, Sci. Rep. 8 (2018) 4003, https://doi.org/10.1038/ s41598-018-22357-y.
29. M. Kirienko, L. Cozzi, L. Antunovic, L. Lozza, A. Fogliata, E. Voulaz, A. Rossi, A. Chiti, M. Sollini, Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery, Eur. J. Nucl. Med. Mol. Imaging 45 (2018) 207–217, https://doi.org/10.1007/s00259-017- 3837-7.
30. P. Kickingereder, S. Burth, A. Wick, M. Götz, O. Eidel, H.-P. Schlemmer, K.H. Maier-Hein, W. Wick, M. Bendszus, A. Radbruch, D. Bonekamp, Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models, Radiology 280 (2016) 880–889, https://doi.org/10.1148/radiol.2016160845.
31. W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, H.J.W.L. Aerts, Exploratory study to identify radiomics classifiers for lung cancer histology, Front. Oncol. 6 (2016) 71, https://doi.org/10.3389/fonc.2016.00071.
32. M. Wu, H. Tan, F. Gao, J. Hai, P. Ning, J. Chen, S. Zhu, M. Wang, S. Dou, D. Shi, Predicting the grade of hepatocellular carcinoma based on non-contrast-enhanced MRI radiomics signature, Eur. Radiol. 29 (2019) 2802–2811, https://doi.org/10. 1007/s00330-018-5787-2.
33. M. Vallières, C.R. Freeman, S.R. Skamene, I. El Naqa, A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft- tissue sarcomas of the extremities, Phys. Med. Biol. 60 (2015) 5471–5496, https:// doi.org/10.1088/0031-9155/60/14/5471.
34. S. Trebeschi, S.G. Drago, N.J. Birkbak, I. Kurilova, A.M. Calin, A. Delli Pizzi, F. Lalezari, D.M.J. Lambregts, M. W. Rohaan, C. Parmar, E.A. Rozeman, K.J. Hartemink, C. Swanton, J.B.A.G. Haanen, C.U. Blank, E.F. Smit, R.G.H. Beets-Tan, H.J.W.L. Aerts, Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers, Annals of. (2019). https://academic.oup.com/annonc/article-abstract/30/6/998/5416144.
35. N. Horvat, H. Veeraraghavan, M. Khan, I. Blazic, J. Zheng, M. Capanu, E. Sala, J. Garcia-Aguilar, M.J. Gollub, I. Petkovska, MR imaging of rectal cancer: radiomics analysis to assess treatment response after neoadjuvant therapy, Radiology 287 (2018) 833–843, https://doi.org/10.1148/radiol.2018172300.
36. J. Alirezaie, M.E. Jernigan, C. Nahmias, Automatic segmentation of cerebral MR images using artificial neural networks, IEEE Trans. Nucl. Sci. 45 (1998) 2174–2182, https://doi.org/10.1109/23.708336.
37. J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M.D. Hellmann, J.O. Deasy, G. Mageras, H. Veeraraghavan, Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images, IEEE Trans. Med. Imaging 38 (2019) 134–144, https://doi.org/10.1109/TMI.2018.2857800.

38. D. Yi, M. Zhou, Z. Chen, O. Gevaert, 3-D Convolutional Neural Networks for Glioblastoma Segmentation, arXiv [cs.CV]. (2016). http://arxiv.org/abs/1611. 04534.
39. B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendia, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C.R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K.M. Iftekharuddin, R. Jena, N.M. John, E. Konukoglu, D. Lashkari, J.A. Mariz, R. Meier, S. Pereira, D. Precup, S.J. Price, T.R. Raviv, S.M.S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C.A. Silva, N. Sousa, N.K. Subbanna, G. Szekely, T.J. Taylor, O.M. Thomas, N.J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D.H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. Van Leemput, The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), IEEE Trans. Med. Imaging 34 (2015) 1993–2024, https://doi. org/10.1109/TMI.2014.2377694.
40. L. Chen, C. Shen, S. Li, G. Maquilan, K. Albuquerque, M.R. Folkert, J. Wang, Automatic PET cervical tumor segmentation by deep learning with prior information, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics, 2018: p. 1057436. doi: 10.1117/12.2293926.
41. D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse,
M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, Author Correction: End-to- end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (2019) 1319, https://doi.org/10.1038/ s41591-019-0536-x.
42. S.A. Abdelaziz Ismael, A. Mohammed, H. Hefny, An enhanced deep learning approach for brain cancer MRI images classification using residual networks, Artif. Intell. Med. 102 (2020) 101779, https://doi.org/10.1016/j.artmed.2019.101779.
43. L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, M. Schäfers, 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks, Radiology 294 (2020) 445–452, https://doi.org/10.1148/radiol.2019191114.
44. S.L.F. Walsh, L. Calandriello, M. Silva, N. Sverzellati, Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study, Lancet Respir. Med. 6 (2018) 837–845, 30286-8.
45. Y. Ding, J.H. Sohn, M.G. Kawczynski, H. Trivedi, R. Harnish, N.W. Jenkins, D. Lituiev, T.P. Copeland, M.S. Aboian, C. Mari Aparici, S.C. Behr, R.R. Flavell, S.-Y. Huang, K.A. Zalocusky, L. Nardo, Y. Seo, R.A. Hawkins, M. Hernandez Pampaloni, D. Hadley, B.L. Franc, A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain, Radiology 290 (2019) 456–464, https://doi.org/10.1148/radiol.2018180958.
46. K. Oh, Y.-C. Chung, K.W. Kim, W.-S. Kim, I.-S. Oh, Author Correction: Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning, Sci. Rep. 10 (2020) 5663, https://doi.org/10.1038/ s41598-020-62490-1.
47. B. Lou, S. Doken, T. Zhuang, D. Wingerter, M. Gidwani, N. Mistry, L. Ladic, A. Kamen, M.E. Abazeed, An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction, The Lancet Digital Health 1 (2019) e136–e147, https://doi.org/10.1016/s2589- 7500(19)30058-5.

48. P.-P. Ypsilantis, M. Siddique, H.-M. Sohn, A. Davies, G. Cook, V. Goh, G. Montana, Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks, PLoS ONE 10 (2015) e0137036, , https://doi.org/10.1371/journal.pone.0137036.
49. K. Strimbu, J.A. Tavel, What are biomarkers? Curr. Opin. HIV AIDS 5 (2010) 463 https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3078627/.
50. H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006, https://doi.org/10.1038/ncomms5006.
51. U. Schneider, E. Pedroni, A. Lomax, The calibration of CT Hounsfield units for radiotherapy treatment planning, Phys. Med. Biol. 41 (1996) 111–124, https://doi.org/10.1088/0031-9155/41/1/009.
52. L.G. Nyúl, J.K. Udupa, On standardizing the MR image intensity scale, Magn. Reson. Med. 42 (1999) 1072–1081. https://doi.org/3.0.co;2-m.“ > 10.1002/(sici)1522-2594(199912)42:6 < 1072::aid-mrm11 > 3.0.co;2-m.
53. R. Berenguer, M.D.R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M.V. Villas, F. Mansilla Legorburo, S. Sabater, Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters, Radiology. 288 (2018) 407–415. doi: 10.1148/radiol.2018172361.
54. J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test-retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography. 2 (2016) 361–365, https://doi.org/10.18383/j.tom. 2016.00208.
55. L. Lu, R.C. Ehmke, L.H. Schwartz, B. Zhao, Assessing agreement between radiomic features computed for multiple CT imaging settings, PLoS ONE 11 (2016) e0166550, , https://doi.org/10.1371/journal.pone.0166550.
56. D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring computed tomography scanner variability of radiomics features, Invest. Radiol. 50 (2015) 757–765, https://doi.org/10.1097/RLI.0000000000000180.
57. I. Zhovannik, J. Bussink, A. Traverso, Z. Shi, P. Kalendralis, L. Wee, A. Dekker, R. Fijten, R. Monshouwer, Learning from scanners: bias reduction and feature correction in radiomics, Clin. Transl. Radiat. Oncol. 19 (2019) 33–38, https://doi.org/10.1016/j.ctro.2019.07.003.
58. A. Webb, G.C. Kagadis, Introduction to Biomedical Imaging, Med. Phys. 30 (2003) 2267–2267. doi: 10.1118/1.1589017.
59. S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, J. Xie, K. Han, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiother. Oncol. 135 (2019) 107–114, https://doi.org/10.1016/j.radonc.2019.03.001.
60. J. Peerlings, H.C. Woodruff, J.M. Winfield, A. Ibrahim, B.E. Van Beers, A. Heerschap, A. Jackson, J.E. Wildberger, F.M. Mottaghy, N.M. DeSouza, P. Lambin, Stability of radiomics

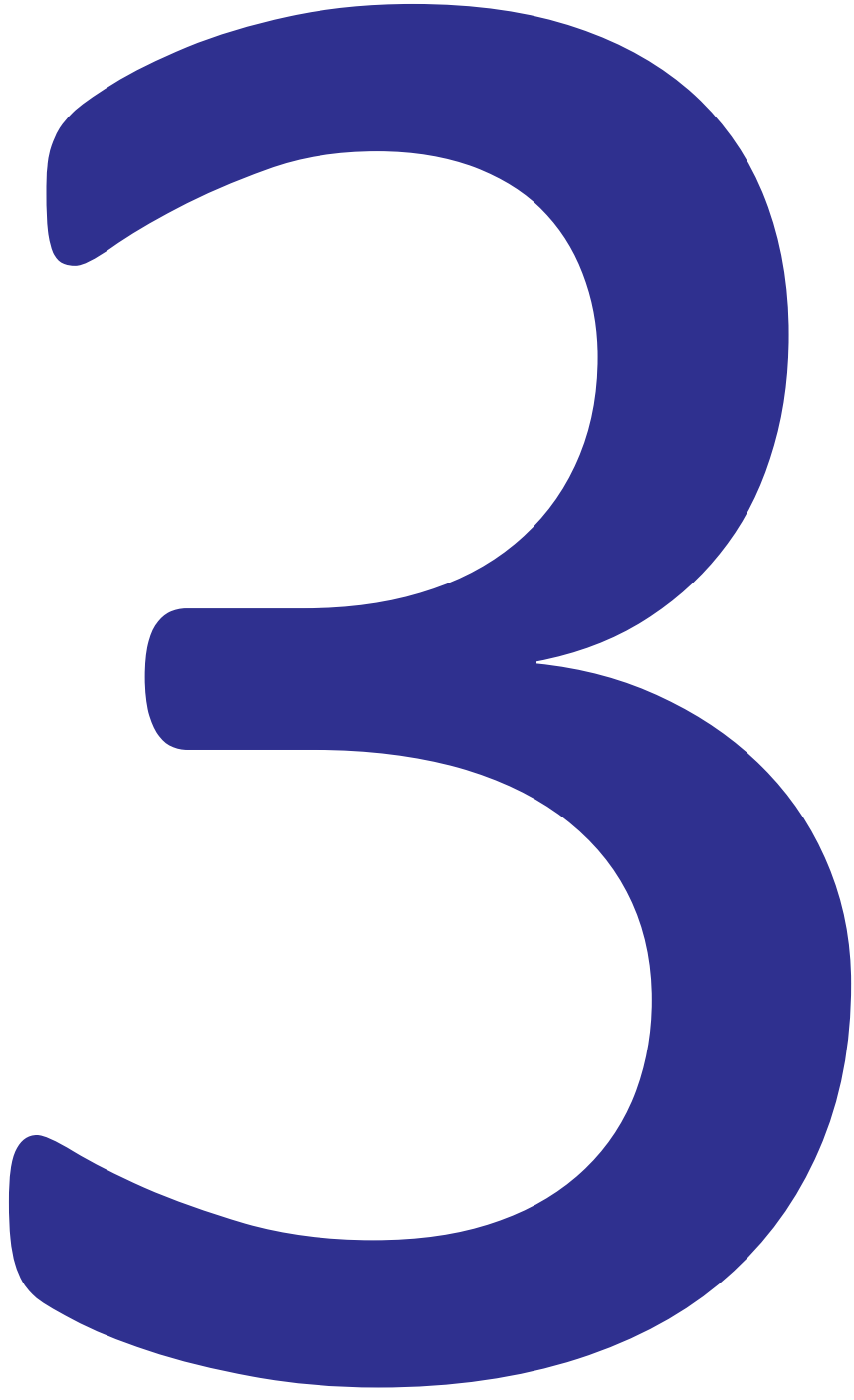
features in apparent diffusion coefficient maps from a multi-centre test-retest trial, Sci. Rep. 9 (2019) 4800, https://doi.org/10. 1038/s41598-019-41344-5.
61. K.T. Bae, Intravenous contrast medium administration and scan timing at CT: considerations and approaches, Radiology 256 (2010) 32–61, https://doi.org/10. 1148/radiol.10090908.
62. G.J.R. Cook, G. Azad, K. Owczarczyk, M. Siddique, V. Goh, Challenges and promises of PET radiomics, Int. J. Radiat. Oncol. Biol. Phys. 102 (2018) 1083–1089, https:// doi.org/10.1016/j.ijrobp.2017.12.268.
63. F. Tixier, M. Hatt, C.C. Le Rest, A. Le Pogam, L. Corcos, D. Visvikis, Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET, J. Nucl. Med. 53 (2012) 693–700, https://doi.org/10.2967/jnumed. 111.099127.
64. M. Hatt, F. Tixier, C.C. Le Rest, O. Pradier, Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma, European Journal of. (2013). https://link.springer.com/article/10. 1007/s00259-013-2486-8.
65. R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, P. Lambin, The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis, Sci. Rep. 5 (2015) 11075, https://doi.org/10.1038/ srep11075.
66. B.A. Altazi, G.G. Zhang, D.C. Fernandez, M.E. Montejo, D. Hunt, J. Werner, M.C. Biagioli, E.G. Moros, Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms, J. Appl. Clin. Med. Phys. 18 (2017) 32–48 https://aapm. onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12170.
67. I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, A. Bitarafan-Rajabi, The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies, Eur. Radiol. 27 (2017) 4498–4509, https://doi.org/10.1007/s00330-017-4859-z.
68. R.T.H. Leijenaar, S. Carvalho, E.R. Velazquez, W.J.C. van Elmpt, C. Parmar, O.S. Hoekstra, C.J. Hoekstra, R. Boellaard, A.L.A.J. Dekker, R.J. Gillies, H.J.W.L. Aerts, P. Lambin, Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability, Acta Oncol. 52 (2013) 1391–1397, https://doi.org/10.3109/0284186X.2013.812798.
69. M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky, M. Friess, P. Veit-Haibach, M. Huellner, I. Opitz, W. Weder, T. Frauenfelder, M. Guckenberger, S. Tanadini-Lang, Influence of inter-observer delineation variability on radiomics stability in different tumor sites, Acta Oncol. 57 (2018) 1070–1074, https://doi.org/10.1080/0284186X.2018.1445283.
70. M. Hatt, M. Vallieres, D. Visvikis, A. Zwanenburg, IBSI: an international community radiomics standardization initiative, J. Nucl. Med. 59 (2018) 287–287. http://jnm. snmjournals.org/content/59/supplement_1/287.abstract.
71. A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R.

Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M. Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J. H.M. Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L.V. van Dijk, J. van Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, Radiology. (2020) 191145. doi: 10.1148/radiol.2020191145.

72. W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127, https://doi.org/10.1093/biostatistics/kxj037.

73. F. Orlhac, F. Frouin, C. Nioche, N. Ayache, I. Buvat, Validation of A method to compensate multicenter effects affecting CT radiomics, Radiology 291 (2019) 53–59, https://doi.org/10.1148/radiol.2019182023.

74. F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, I. Buvat, A postreconstruction harmonization method for multicenter radiomic studies in PET, J. Nucl. Med. 59 (2018) 1321–1328, https://doi.org/10.2967/jnumed.117.199935.



Chapter 3

Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging

Manon Beuque[1], Marta Martin-Lorenzo[1], Benjamin Balluff, Henry C. Woodruff, Marit Lucas, Daniel M. de Bruin, Janita E. van Timmeren, Onno J.de Boer, Ron MA. Heeren, Sybren L. Meijer[2], Philippe Lambin[2]

[1] These authors contributed equally
[2] Share senior authorship

Adapted from:
Beuque M, Martin-Lorenzo M, Balluff B, Woodruff HC, Lucas M, de Bruin DM, van Timmeren JE, Boer OJd, Heeren RMA, Meijer SL, Lambin P.
Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging. Computers in Biology and Medicine 2021; 138:104918. doi: https://doi.org/10.1016/j.compbiomed.2021.104918

Abstract

Background: Barrett's esophagus (BE) is a precursor lesion of esophageal adenocarcinoma and may progress from non-dysplastic through low-grade dysplasia (LGD) to high-grade dysplasia (HGD) and cancer. Grading BE is of crucial prognostic value and is currently based on the subjective evaluation of biopsies. This study aims to investigate the potential of machine learning (ML) using spatially resolved molecular data from mass spectrometry imaging (MSI) and histological data from microscopic hematoxylin and eosin (H&E)-stained imaging for computer-aided diagnosis and prognosis of BE.

Methods: Biopsies from 57 patients were considered, divided into non-dysplastic (n = 15), LGD non-progressive (n = 14), LGD progressive (n = 14), and HGD (n = 14). MSI experiments were conducted at 50 × 50 µm spatial resolution per pixel corresponding to a tile size of 96x96 pixels in the co-registered H&E images, making a total of 144,823 tiles for the whole dataset.

Results: ML models were trained to distinguish epithelial tissue from stroma with area-under-the-curve (AUC) values of 0.89 (MSI) and 0.95 (H&E)) and dysplastic grade (AUC of 0.97 (MSI) and 0.85 (H&E)) on a tile level, and low-grade progressors from non-progressors on a patient level (accuracies of 0.72 (MSI) and 0.48 (H&E)).

Conclusions: In summary, while the H&E-based classifier was best at distinguishing tissue types, the MSI-based model was more accurate at distinguishing dysplastic grades and patients at progression risk, which demonstrates the complementarity of both approaches. Data are available via ProteomeXchange with identifier PXD028949.

1. Introduction

Esophageal adenocarcinoma (EAC) remains one of the deadliest cancers with a 5-year survival rate of less than 20% (1) and Barrett's esophagus (BE) is the only known precursor lesion. BE is a condition of the distal esophagus where the stratified squamous epithelium is replaced by columnar epithelium with goblet cells due to gastroesophageal reflux disease (2). BE may progress from non-dysplastic metaplasia (NDBE) through low-grade dysplasia (LGD), to high-grade dysplasia (HGD) and esophageal adenocarcinoma (EAC). A histopathology diagnosis of LGD is an important independent risk factor to develop EAC (3). This diagnosis however is hampered by inter- and intra-observer variability and international guidelines therefore mandate a second opinion. The individual rate of progression from BE patients with LGD to HGD/EAC is difficult to evaluate on hematoxylin and eosin (H&E) slides using light microscopy and ranges between 0.6 and 13.4% per patient per year (4). Due to the lack of reliable indicators of progression, current clinical treatment guidelines for LGD patients are not well defined and range from immediate local treatments to further endoscopic surveillance (5). Currently no objective biomarkers exist to identify BE patients with LGD that quickly progress to HGD and EAC from LGD lesions that remain stable for years.

Computer-aided diagnostics of histological images and new molecular imaging modalities are therefore needed to assist the pathologist in grading BE lesions and give a reliable prediction of the disease progression.

For the molecular analysis of histological tissue section, mass spectrometry imaging (MSI) is a young technique in expansion. MSI enables the acquisition of spatially resolved molecular profiles from tissue sections without any labelling. MSI has demonstrated during the past decade to be a powerful tool to extract clinically relevant information beyond histology from the molecular setup of different cancer types (6). In the context of EAC, the group of Walch and coworkers has already used MSI to find several proteins to be indicative of poor survival, metastasis, and chemosensitivity (7).

MSI and histology can be used in combination and we hypothesize that the complementarity of both can potentially reinforce the accurate grading of BE and prognosis. From a technical point of view, both imaging modalities (optical microscopy and MSI) provide copious amounts of data: histological images are usually high resolution whereas MSI data is high-dimensional in its feature space, making them both suited for machine learning (ML) approaches (8).

Our general objective is to investigate ML solutions applied to MSI and H&E data and analyse its ability to discriminate epithelial from stromal tissue and to classify BE samples according to the grade of dysplasia. We propose here a workflow based on ML, which can classify the tissue between epithelial tissue and stroma and display where the classifier identifies dysplastic areas of interest in the epithelial region. This way the experts can focus on the specific region of the H&E stained slides. This would provide a cheap and fast auxiliary observation and would help the experts to give a faster and more accurate

diagnosis. The second aim of the study is to use ML solutions to distinguish LGD-lesions at risk to progress from those that display stable disease.

2. Methods

2.1. Patient material

Formalin-fixed paraffin-embedded (FFPE) esophagus tissue biopsies were retrieved from the archives of the Department of Pathology of the Amsterdam UMC, location Meibergdreef. A total of 57 biopsy samples from 57 patients were collected and covered the complete spectrum of BE, ranging from NDBE (n = 15) to LGD (n = 28) and HGD (n = 14). Based on the patients' follow-up LGD samples were sub-classified into LGD non-progressors (n = 14, no progression to HGD or EAC within a period of two years) and LGD progressors (n = 14, developed HGD or EAC within 2 years). All samples were anonymized for further use and did not require approval from the relevant Institutional Ethics Committee under applicable local regulatory law ('Code of conduct', FEDERA).

2.2. Mass spectrometry imaging experiments

For this unique dataset, FFPE oesophagus tissue samples were cut at 5 µm thickness and randomly distributed on a total of 19 indium tin oxide- coated conductive glass slides (Delta Technologies). For MSI peptide measurements, samples were prepared as previously described by Vos et al. in (9). Briefly, samples were deparaffinised with xylene, exposed to antigen-retrieval and on-tissue tryptic digested using the Antigen Retriever 2100 (Aptum Biologics, UK) and a SunCollect pneumatic sprayer (SunChrom GmbH, Germany), respectively. After a 17h long incubation, alpha-cyano-4-hydroxycinnamic acid was applied using the same SunCollect sprayer. Before MSI, optical images of the glass slides were taken with a high-quality film scanner (Nikon LS-5000) with a true optical resolution of 4000 dpi (i.e. one pixel is 6.35 µm) in order to define the measurement region. This image therefore acts as anchor image and is later also used to co-register high-resolution H&E images to the MSI data. MSI experiments were performed at 50 µm lateral pixel size (40 40 µm laser beam scan range) on a rapifleX MALDI-ToF mass spectrometer (Bruker Daltonics) in reflectron and positive-ion mode within an m/z range of 800–3000. The instrument was calibrated beforehand using Red Phosphorus. Line scan sequence was non-random (i.e. spectra are acquired sequentially from upper left to lower right). Random walk within one pixel was deactivated and 700 spectra were averaged per pixel with a MALDI laser repetition rate of 10 kHz. All individual spectra underwent on-the-fly smoothing (Savitzky-Golay 5%) and baseline subtraction (TopHat). Digitization rate was 1.25 GS/s resulting in 55,000 data points per spectrum (i.e. per MSI pixel), which was reduced to 80% of its original size in FlexImaging (Bruker Daltonics). MSI datasets contained on average 4500 MSI pixels (min 1370; max 11,647) and were exported separately as imzML files from FlexImaging.

2.3. Haematoxylin and eosin staining

The same tissue sections analysed by MSI were concurrently stained with H&E to minimize possible staining differences. For this, the matrix was first washed-off from the slides using 70% ethanol for 2–3 min, followed by a 3 min wash with Milli Q water. Slides were stained with haematoxylin (3 min), washed for 3 min with tap water to remove excess haematoxylin, then stained with eosin (30 s), washed again with tap water for 3 min to remove excess eosin, followed by a 1 min ethanol wash and a 30 s xylene wash before attaching coverslips to the slides using Entellan as a mounting medium. The stained slides were scanned with a digital slide scanner at 20x magnification (Mirax Desk, Carl Zeiss MicroImaging, Göttingen, Germany). Tissue scans were exported using Pannoramic Viewer (3DHistech, Hungary) in JPG file format (at 90% quality compression and at original resolution), resulting in pixel sizes of 0.52x0.52 µm2. The images were superposed to the MALDI-MSI data in FlexImaging using a 3-control-point co-registration of previously applied fiducial markers. Stained tissue sections were annotated by an expert pathologist according to the tissue type (epithelial/stroma) and BE grade. In order to evaluate the co-registration quality, all datasets were individually inspected visually. In all cases, the eye-estimated average error (<10 µm) was significantly smaller than the laser spot size (50 µm) (Supplementary Figure 1). An example of the aligned information comprising histological images, MSI, and annotations is shown in Figure 1.

Table 1: Number of tiles distributed over the different grades and tissue types Abbreviations used: NDBE, non-dysplastic Barrett Esophagus; LGD, low-grade dysplasia; HGD: high-grade dysplasia.

Data	Portion of the total tiles [%]	Portion of epithelial tissue [%]	Total number of tiles
Stroma	50%	–	72516
NDBE	36%	73%	52704
LGD	8%	16%	11615
HGD	6%	11%	7988
Total	100%	100%	144823

2.4. MSI data pre-processing

A recalibration was performed in Flex Analysis v3.4 using the lock masses m/z 842.510 and 1045.564 with a peak assignment tolerance of 500 ppm and m/z 1303.615, 1508.750, 1833.954, 1835.957, 2104.190, 2105.190, 2106.190 with a tolerance of 250 ppm. All recalibrated MSI data, coregistered H&E images, and annotations were imported to SCiLS Lab (Bruker Daltonik) where each spectrum was normalized to its total- ion-count (TIC). From SCiLS Lab, the overall spectra from on- and off- tissue regions were exported to mMass (http://www.mmass.org/) for peak picking using the following parameters: (1) Baseline correction precision = 40; (2) Peak-picking: S/N 5.0; Picking height = 90; (3) Deisotoping: maximum charge = 1, isotope mass tolerance m/z = 0.15, isotope intensity tolerance = 70%,

isotope mass shift = 0.0. The picking lists from on- and off-tissue were subsequently compared with a tolerance of 0.2 Da and common peaks were removed from the on-tissue peak list after visual inspection and confirmation (Supplementary Table 1).
This final peak list (Supplementary Table 2) was then imported, together with the imzML files and the histological images of every patient, into Python 3.7. All of these data are available via ProteomeXchange (https://www.ebi.ac.uk/pride/) using the identifier PXD028949. In Python, the mass spectrometry pixels were normalized to their total-ion-count before extracting the maximum intensity for every peak in its 0.5 m/z interval across all MSI spectra.

Table 2: Tile-based classifier performance for predicting tissue type and grade on the test sets. Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

Prediction of …	Data type	Labels	Precision	Recall	f1-score	Support (number of tiles)
Tissue type	MSI	Epithelial tissue	0.80	0.81	0.81	10602
		Stroma	0.81	0.82	0.80	11121
	H&E	Epithelial tissue	0.89	0.87	0.88	10602
		Stroma	0.88	0.90	0.89	11121
Grade	MSI	Non dysplastic Barrett's Esophagus	0.97	0.84	0.90	7836
		Low-grade dysplasia	0.68	0.90	0.77	1787
		High-grade dysplasia	0.70	0.98	0.82	1224
	H&E	Non dysplastic Barrett's Esophagus	0.93	0.70	0.80	7836
		Low-grade dysplasia	0.37	0.75	0.50	1787
		High-grade dysplasia	0.44	0.47	0.45	1224

2.5. MSI and histology data extraction

Data extraction was performed using Python 3.7 with ImzMLParser and OpenCV libraries. Affine geometric transformations were performed in order to spatially link MSI and H&E based on the co-registrations previously done in FlexImaging, which were accessible via the respective .mis XML files. These files also contained the annotations as sets of polygonal coordinates (Supplementary Figure 2). For each of the MSI pixels (see Figure 1 c), the corresponding histological patch was extracted with a size of 96x96 pixels (see Figure 1 a). Henceforth, the word "tile" will be used to describe both the MSI pixel and the matching histological patch. The histology tiles were labelled according to the annotation of the centre pixel.

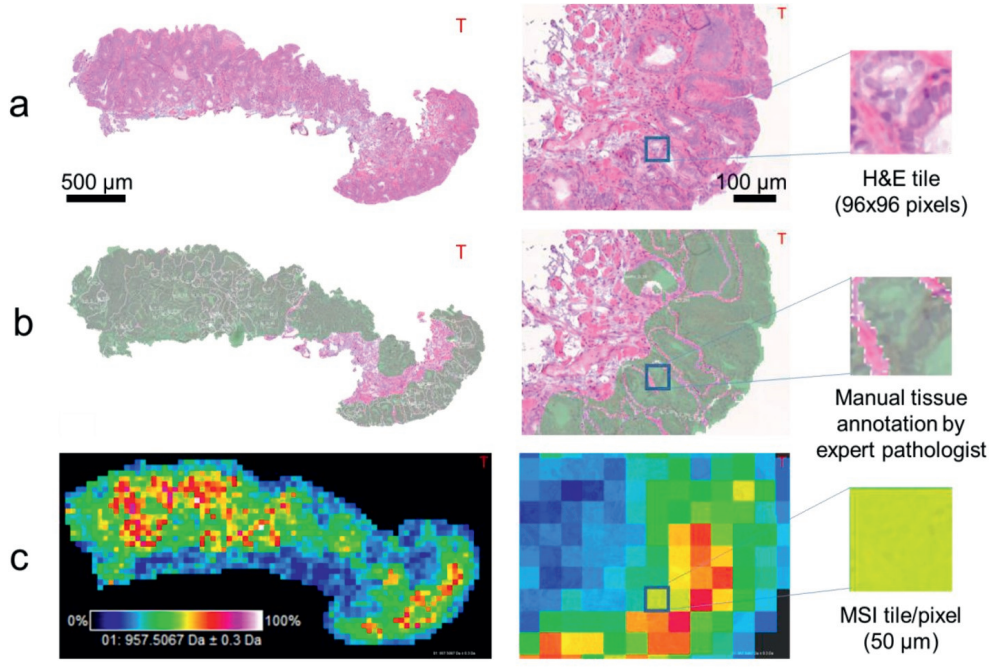


Figure 1. Illustration of the multimodal imaging data used in this study. Three increasing magnification levels (left to right) of the three spatially co-registered layers of information and their size-matched representation on a tile level are shown in one of the Barrett's esophagus tissue biopsies: (a) The H&E- stained microscopic image (b) the manual pathological annotations made by an expert pathologist on the same H&E image in this case indicating the glandular areas (grey colour) and (c) the MSI data, here represented by the visualization of a particular mass channel (m/z 957.5) which co-localizes with the glandular areas shown in (b). Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging; TIC, total-ion- count.

2.6. Machine learning models

For each modality (MSI and H&E), three tile-based ML classifiers were trained: the first classified the tiles between epithelial tissue and stroma (tissue type prediction), the second predicted the dysplastic grade of a tile, and the third the progression of dysplasia on a patient level. The data was randomly split into training, validation, and test datasets in a ratio of 0.70/0.15/0.15, respectively. For the patient-level classifiers, we performed a leave-one-patient-out cross validation (LOPOCV) by excluding the data of one patient from the training/validation sets and splitting the training dataset and validation dataset in a ratio of 0.90/0.10 and repeat the process for all the patients. The model was computed using Pytorch 1.2.0 on Python 3.7, on a GPU-clusters of 10 GPUs (NVIDIA GeForce RTX 2080 Ti). The implementation of the machine learning models can be found at: https://github.com/precision-medicine-um/ML-and-MS-in-esophageal-cancer.git.

2.7. Tissue type prediction

The signals of each MS-tile were rescaled by multiplying all values by a factor 105 for a better compliance in the software. Then, the features were mapped with a Gaussian distribution per patient with Box-Cox transformation in order to obtain a similar range of signal intensities in all datasets and to remove possible patient/acquisition biases. After pre-processing, the parameters of three models were optimized on the training dataset with three independent grid searches: The impurity measure, the number of estimators, and the maximum depth for random forest (RF), the weight decay, the batch size, the hidden layer sizes, the maximum iteration, and the optimizer for multi-layer perceptron (MLP), and the learning rate, the number of estimators, the maximum depth, the minimum child weight, and subsample for XGBoost. The rest of the parameters were the default parameters from the python library scikit- learn 0.24.2. The three models were then merged with an ensemble modelling method, a voting classifier, which used argmax function to obtain a final probability class prediction. The pipeline can be found as flowchart in Supplementary Figure 3.

As the H&E and MSI datasets were co-registered, we could use the equivalent split of H&E data to form the training dataset and validation dataset and all the tiles were resized to 224x224 pixels to match the required input size of the DL model. A data augmentation step using the library albumentations within Python was applied to the training dataset where transformations were applied on the images with a probability of 0.5 for each augmentation: rotation by 90◦, transposition, flipping around the horizontal/vertical or both axes, random intensity filtering, and random affine transformations. All the tiles were normalized on the three channels individually (red, green, and blue). We used a Convolutional Block Attention Module (CBAM) (10) with Resnet50 as the backbone (11) as proposed by the work of Tomita et al. (12) with minor modifications. The CBAM was added between two convolutional blocks, aggregating max pooling and average pooling into a channel attention module and a spatial attention module to focus on representation. The parameters chosen were binary cross-entropy and Adam optimizer with a learning rate of 4x10—4. The model was trained on

batches of 600 tiles. The training stopped once the loss on the validation dataset stopped decreasing after one epoch. Then, the model was evaluated on the test dataset using test time augmentation. Ten different augmentation functions were used on the test dataset, such as a 90◦ rotation, transposition, horizontal and vertical flip. The model gave a probability per class for each transformation following which the tissue type predictions were averaged, and the class predicted with the highest probability was chosen. The workflows for the MSI and H&E data are presented in Figure 2.

2.8. Grade of dysplasia prediction

The workflow for grade prediction on a tile-level follows the same workflow as described for tissue type prediction (see in Figure 2) for MSI and H&E with the difference that the analysis focuses only on tiles from the epithelial regions since BE grading is based on morphological changes in the epithelial structures (13). As the dataset was highly unbalanced, undersampling was performed in the proportion of the high-grade tiles (lowest number of tiles), choosing randomly the tiles among the 3 classes to not overfit during the training phase.

2.9. Multi-modal classifier

For the two different tasks, the same approach was pursued to establish a multi-modal classifier: we extracted the last layer of features from the trained DL model (2048 features) and combined them to the mass spectrometry features. Then we used grid-search in combination with a MLP to obtain an optimized classifier.

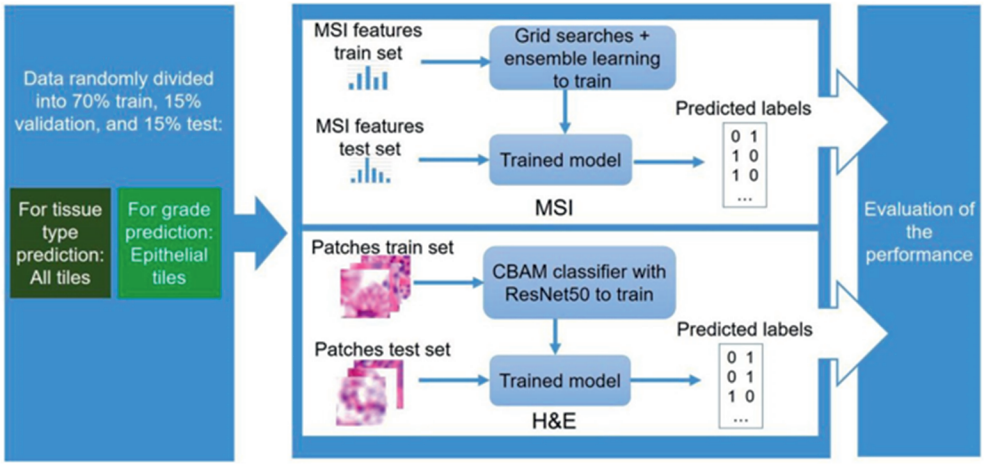


Figure 2. Workflow for the prediction of tissue type (using all tiles) and grading (using tiles belonging to epithelial tissue only) on a tile-level using the MSI data (top row) and H&E data (bottom row). Abbreviations used: CBAM, Convolutional Block Attention Module; H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

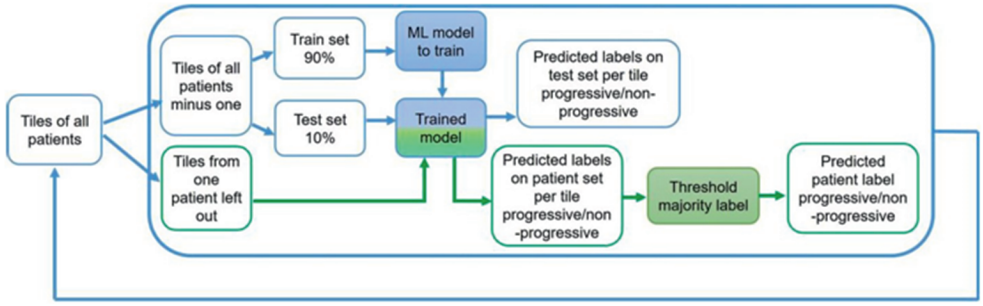


Figure 3. Workflow for the prediction of low-grade dysplasia progression on a patient-level which is used for both the MSI as well as the H&E data. Abbreviations used: MSI, mass spectrometry imaging; H&E, hematoxylin and eosin-stained tissue scans.

2.10. Identify low grade dysplasia progression

To predict the progression of BE dysplasia, only the tiles belonging to epithelial tissue from 25 patients annotated with LGD regions were considered. We used a LOPOCV where, at each iteration, a new model was trained on 90% of the remaining data. We used 10% for the validation dataset to make sure that the classifiers didn't overfit on the training dataset. The signals of each tile in the MSI dataset were mapped with a Gaussian distribution per patient with the Box-Cox method and all the m/z features were used. The classifier used was the same MLP classifier with the same hyperparameters computed during the training of grade classification.

To build a classifier using the H&E data, a CBAM architecture was used with ResNet50 as the backbone, trained on 2 epochs, re-trained, and validated using LOPOCV as described above. The workflow is presented in Figure 3. In both MSI and H&E, a patient was classified as progressive or stable according to the majority vote of the predicted tiles.

2.11. Evaluation of the generated models

We reported the confusion matrices, the precision, the recall, the f1- score, and the number of samples per category for both validation and test datasets as calculated with the libraries sklearn and matplotlib within Python 3.7. For the classification of the tiles, the receiving operating characteristic (ROC) and the area under the ROC curves (AUC) were calculated. The confidence intervals of the AUC at 95% were computed with the DeLong algorithm, using the pROC library in R 3.6.3. The significance of every feature was calculated with the feature_importances function of sklearn based on the mean Gini decrease. The dice coefficient per tile was calculated for the test dataset to evaluate the grading performance: when the grade was correctly predicted, the dice was calculated with the delineations made by the pathologist and the full shape of the tile. In any other case, the dice coefficient was considered zero.

3. Results

Processing of the MSI data led to the detection of 321 on-tissue peptide signals, which were all used for training the models. Matching the MSI pixel size of 50 50 µm to the tile size in the histological images (96x96 pixels) made up a total of 144,823 tiles for the whole dataset. Table 1 inventories all the extracted tiles from all the images with the corresponding annotated tissue type and grade.

The epithelial tissue and stroma tiles dataset were equally balanced. However, the grades were highly unbalanced with the NDBE at 73%, LGD class 16%, and HGD class 11% of the epithelial dataset. These results were expected since NDBE and stroma classes can be found in all the samples, but dysplastic regions can only be found in specific areas.

3.1. Tissue type classification

The first task of our study was to use all 144,823 annotated pixels and classify the tiles as either epithelial or stroma regions.

For this, three models were trained on the MSI dataset and optimized with a grid search: the best parameters for MLP were a weight decay of 10—5, with a batch size of 32, hidden layer sizes of (20, 10), maximum iteration of 1100, using Adam as optimizer. The best parameters for RF were the entropy as criterion with a maximum depth of 16, number of estimators at 200. For XGBoost we found that a learning rate of 0.1, number of estimators at 140, maximum depth of 9, a minimum child weight of 1 with a subsample at 0.8 worked best. Then the results were combined using a voting classifier with argmax function. This model gave an AUC of 0.89 (95% confidence interval (CI): 0.89–0.90) on the test dataset. The list of features importance computed on random forest, xgboost and their average is provided in supplementary material (Supplementary Table 3). The model using H&E data as an input was trained for 2 epochs. The model achieved an AUC of 0.95 (95% CI: 0.94–0.95) on the test dataset.

The performances of both models were assessed with normalized confusion matrices on both validation (Supplementary Figure 4) and test dataset (see Figure 4 a), and with the display of ROC curves (both Supplementary Figure 5). The classification reports (Table 2) show the balance of precision and recall in both models.

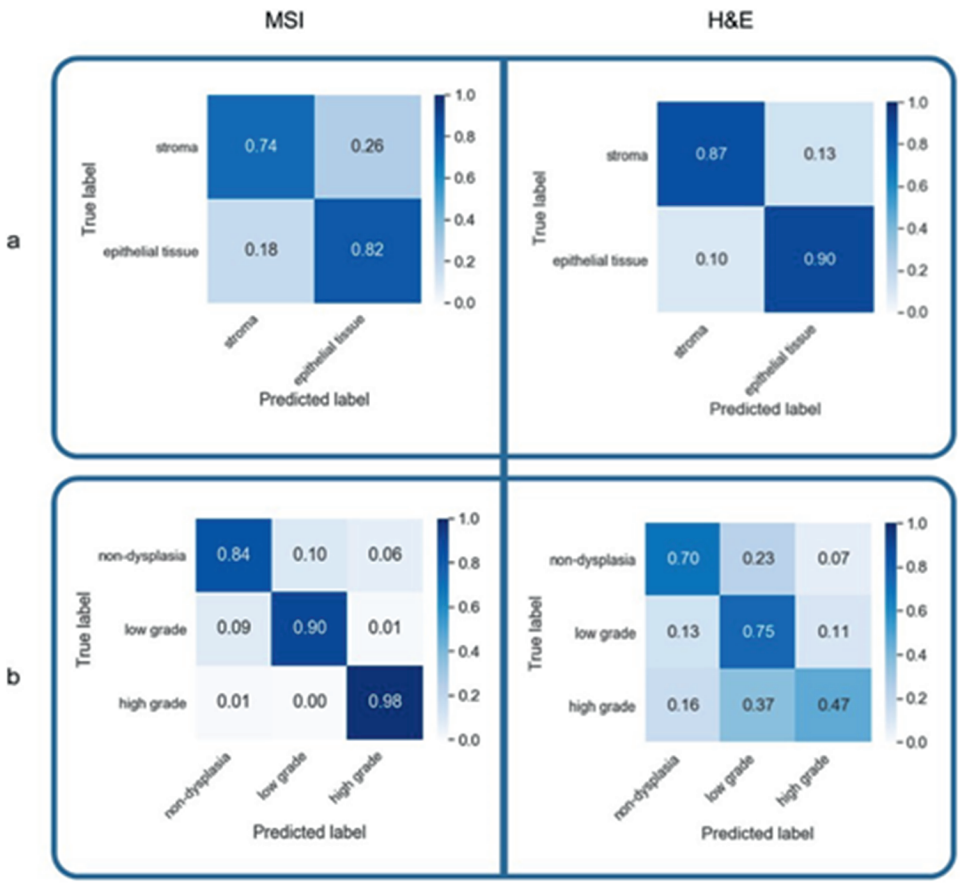


Figure 4. Normalized confusion matrices of the test datasets for a prediction of the tissue type (a) and grade (b) using the mass spectrometry imaging (MSI) data (left) and the H&E data (right). Abbreviations used: MSI, mass spectrometry imaging; H&E, hematoxylin and eosin-stained tissue scans.

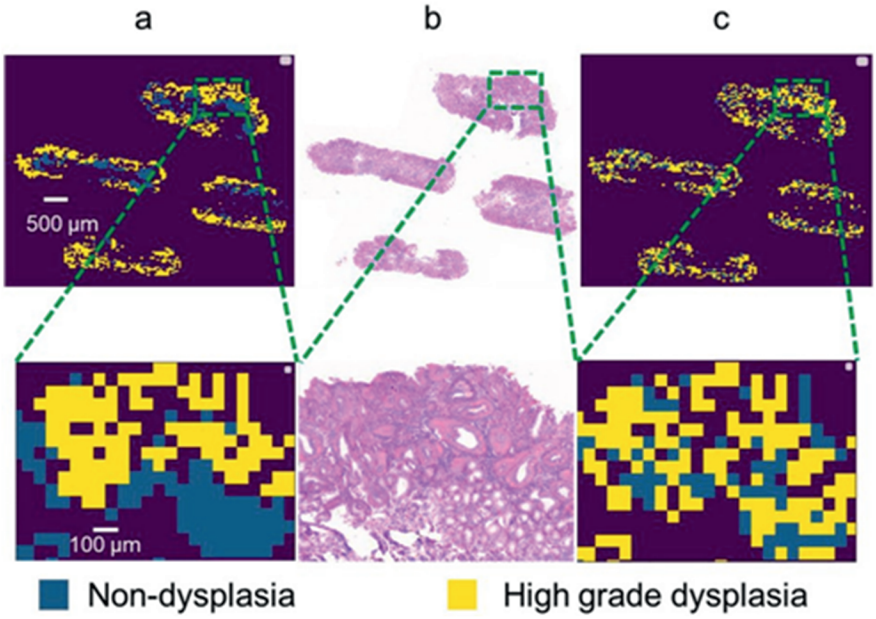


Figure 5. Example of tile-based classification of the grade of a BE tissue: (a) ground truth label per tile of the full slide, (b) H&E of the full slide, and (c) prediction of tile labels on the full slide based on the classification made with MSI data Magnifications are shown in the lower row. Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

3.2. Tile-based dysplastic grade prediction

The second aim consisted in determining the grade of the tiles belonging to the epithelial regions (n=16,920). Also, here 321 MSI features were used to train the model for grade prediction. The optimal weight decay was 10—6 with a batch size of 32, hidden layer sizes of (20,20), a maximum iteration of 1100, and Adam as optimizer for the MLP model. The best parameters found for RF were entropy as criterion, a maximum depth of 16, number of estimators at 200 and for XGBoost we found that a learning rate at 0.1, a number of estimators at 140, a maximum depth at 9, a minimum child weight at 5, and using a subsample at 0.8 worked best. The results were combined using a voting classifier with argmax function.

The model performance as per micro-average AUC was 0.97 (95% CI: 0.96–0.97) on the test dataset. The comparison of performance on both the test and validation datasets allows us to observe that the model does not overfit. Because the test dataset was unbalanced but the training dataset was balanced, we can observe that the f1-scores are not consistent, achieving a worse performance on the low-grade tiles and the high- grade tiles (Table 2). The ROC curves of the micro-average and the macro-average ROC curves were also computed to give a better understanding of the overall prediction performance (Supplementary Figure 6). The list of features importance computed on random forest, xgboost and their average is provided in Supplementary Table 3.

For the H&E-based classifier, the weights from the 17th epoch gave the best average accuracy on the validation dataset and were selected to evaluate the model. The micro-average AUC was 0.85 (95% CI: 0.85–0.86) on the test dataset. The results are visualized by confusion matrices of both the test (Figure 4 b) and validation dataset (Supplementary Figure 4b). The ROC curves of the different grade predictions are provided in Supplementary Figure 6. The average dice coefficients calculated on the test dataset are given in Supplementary Table 4. An example of a tile-wise full slide prediction for a patient diagnosed with LGD using the H&E classifier is given in Figure 5.

3.3. Multi-modal prediction

We trained an MLP model by combining the features extracted from the DL model trained to distinguish the epithelial tissue from stroma with the MSI features. The model was trained with an L2-regularization coefficient of 0.1, a batch size of 32, hidden layer sizes of (10,10), a maximum of 1100 iterations and Adam as optimizer. Using this configuration, we obtained an AUC of 0.95 (95% CI: 0.95–0.95) on the test dataset. We repeated the same logic for the grade prediction and we obtained an optimal MLP with the following parameters: The optimal weight decay was 10—6, with a batch size of 64, hidden layer sizes of (10, 10), maximum 1100 iterations, and Adam optimizer. The micro-average gave an AUC of 0.96 (95% CI: 0.96–0.96) on the test dataset. The corresponding confusion matrices are provided in Supplementary Figure 7.

3.4. Prediction of disease progression

Finally, models were trained for both modalities with the aim of forecasting the progression of low-grade dysplasia to a higher grade. The predictions of the models were thereby assumed to make statements on progression on a patient-level. The accuracies of the MSI-model and H&E-model were 0.72 (95% CI: 0.54–0.90) and 0.48 (95% CI:0.28–0.68), respectively (Supplementary Figure 8).

4. Discussion

As a use-case, we chose the task of classifying the grade of dysplasia in BE and identify LGD lesions at high risk of progression. While deep learning has recently been used to detect neoplasia in BE using endoscopy (14), the application to histopathology is novel and ML-based classification of dysplasia grade or risk of progression has not been done with the combination of the two modalities as far as we know. When automatically classifying the tissue into epithelial and stromal structures, we could observe that the results are comparable between the validation (Supplementary Table 5) and the test datasets (Table 2), indicating that the model was not overfitting the validation dataset very strongly for both models. Analysing the precision and recall scores, we observe similar results, which indicate well-balanced models. The model based on H&E data (AUC: 0.95 (95% CI: 0.94–0.95)) obtained a better performance at classifying epithelial tissue versus stroma than the model

based on MSI data (AUC: 0.89 (CI: 0.89–0.90)). Given the clear visual differences between these tissue types, the superior performance of the H&E data is not surprising.

When predicting the grade of the dysplasia on a pixel-level, we observed similar prediction scores between the validation and the test datasets. The different grading implementation exhibited similar predictions with the model based on MSI features but the model based on H&E images obtained poor results for the classification of high-grade tiles. This allowed us to conclude that the models did not overfit on the validation dataset and the model based on MSI did not over-predict one class rather than another, but the model based on the H&E images did over-predict low grade tiles. With the classification reports (Supplementary Table 5) we observed that the precision/recall was unbalanced on the validation dataset and on the test dataset (Table 2). This was caused by unbalanced data. In contrast with the previous task, we found that the model based on MSI data (AUC: 0.97 (CI: 0.96–0.97)) outperformed the model based on H&E data (AUC: 0.85 (CI: 0.85–0.86)). Moreover, the average dice coefficients obtained on the grades (Supplementary Table 4) were lower but close to the true positive values obtained on the test dataset; thus confirming the capability of our model to reliably identify dysplastic regions. Indications for the potential of MSI data in similar scenarios can be found in previous MSI literature, where Elsner et al. (7) were able to distinguish metaplasia from carcinoma with an accuracy of 91% using a pattern of 31 proteins, albeit on a sample-level.

The use of multi-modal classifiers didn't improve the results obtained by using the modalities separately. H&E is better than MSI to distinguish the tissue type and MSI is better at predicting the grade of the tiles.

Despite endoscopic surveillance of patients with non-dysplastic BE or BE with low-grade dysplasia, up to 25% of EACs and HGDs are diagnosed within one year after last screening (15). Our study shows the potential of MSI coupled to DL to identify patients that are higher at risk to progress to HGD with 72% accuracy. The performance of the model using MSI was similar to other studies such as the study of Kate and co-workers who used clinical features in combination with p53 immunohistochemistry and histology criteria to obtain an AUC of 0.77 (16). Our method was independent of clinical features and still obtained similar results. Our classifier could be therefore a useful addition to the existing surveillance strategies.

At the moment, the size of the cohort (57 in total) limits any strong clinical conclusions. A larger external validation sample dataset is therefore required to evaluate and confirm the predictions made by both approaches. In such a follow-up study, the predictive values of already known biomarkers in BE or EAC could be evaluated and compared to the MSI/H&E based approach. As mentioned, p53 is a biomarker for progression observed in 75% of the patients with multifocal aggregates of positive cells (13). Another indicator for progression could be alpha-methyl-CoA racemase, an enzyme with high specificity and low sensitivity for the progression of indefinite for dysplasia towards dysplasia (17).

When comparing the classificatory power of H&E and MSI, an

explanation of the improved classification capability of the model based on MSI data might be the fact that the dataset was annotated based on the H&E staining making it compile information from both approaches. Furthermore, the H&E dataset was not exploited to its full potential. We restricted the optimal parameter space of the H&E classification models

by fixing the size of the patches (96x96 pixels) to the size of the MSI pixel (50 µm lateral pixel size), although similar tile sizes are being used in this field at 20x magnification (18). In this context, it would be interesting to use multiple magnification levels for the classification of the data as done by Han et al. (19). In contrast, we believe that using histomics to extract features at a cellular level combined with a ML model which classifies tissues instead of DL would help the pathologists to understand better what characteristics of the H&E are important for the classification (20).

Nevertheless, this study reveals the strength of each modality and their complementarity to address diagnostic and prognostic challenges in pathology using advanced ML. In our study, H&E provided higher accuracies for diagnostic purposes where the information is visually located in the histological phenotype. MSI, conversely, seems better suited for purposes where clinically relevant molecular alterations are present but still not morphologically visible at the microscopic level (21). One can, therefore, envision a cascade-like application of the presented ML classifiers, where the optimum data and model are used for different tasks. In our example, the sequence would start with the H&E-based classifier for the detection of epithelial tissue regions. The MSI-based grade classifier would be applied to determine the grade of these epithelial structures, followed by the second MSI-based classifiers to predict if a patient's lesion is at risk of progressing.

In summary, the intention of our work was to investigate the complementarity and suitability of histological and molecular images using ML approaches for different clinical tasks in BE (tissue annotation, pathological grading, and patient prognosis). We have found that MSI can add valuable prognostic information beyond the histological level, whereas histology remains strong at the tissue annotation level. Based on these results we conclude that both, histological imaging and MSI, can complement each other for different clinical questions, which could ultimately help pathologists in diagnosing BE patient biopsies.

Declaration of competing interest

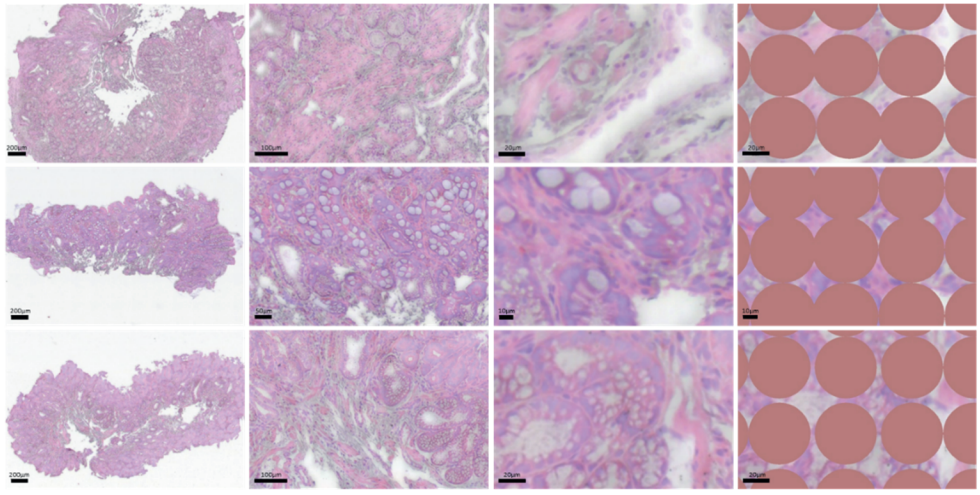
Dr. Lambin is co-inventor of two non-issues, non-licensed patents on Deep Learning-Radiomics-Histomics (N2024482, N2024889). Dr. Woodruff and Dr. Lambin have (minority) shares in the company Oncoradiomics. The rest of the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

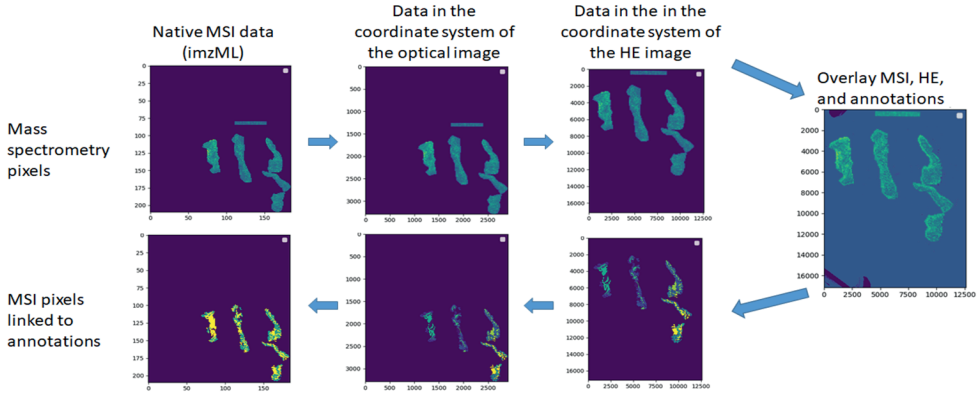
This work was made possible through the support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276). BB and MML acknowledge the financial support of the European Union and the Dutch Cancer Society (ERA-NET TRANSCAN 2; Grant No. 643638). MML acknowledges the financial support of CAM (2018-T2/BMD-11561). Part of this work

was conducted with financial support of the Province of Limburg through the LINK program. Authors furthermore acknowledge financial support from ERC advanced grant (ERC-ADG-2015 no. 694812 - Hypoximmuno, ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO.DISTINCT). Authors also acknowledge financial support from SME Phase 2 (RAIL no.673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR no. 733008, FETOPEN- SCANnTREAT no. 899549, CHAIMELEON no.952172, EuCanImage no. 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY no. UM 2017–8295) and Interreg V-A Euregio Meuse-Rhine (EURADIOMICS no. EMR4). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018–2.

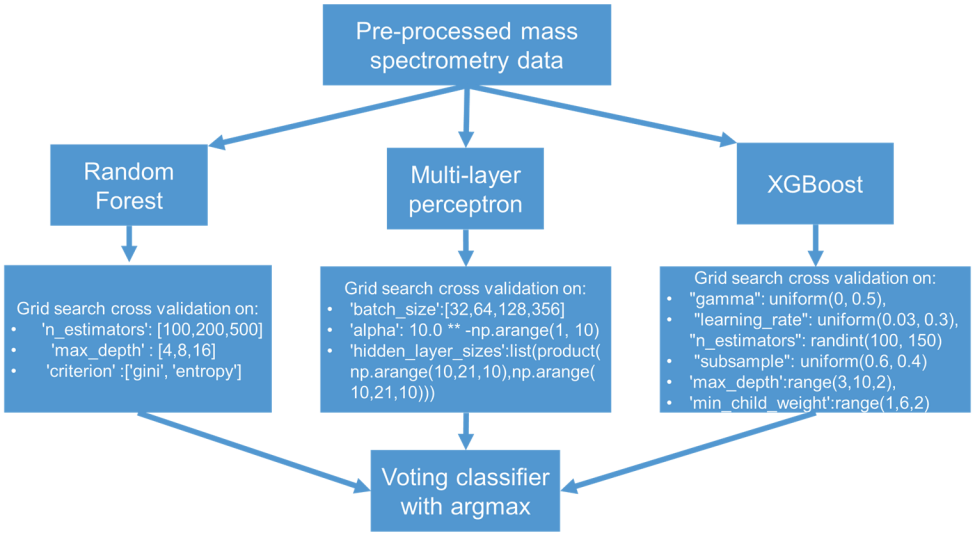
Supplementary Figures



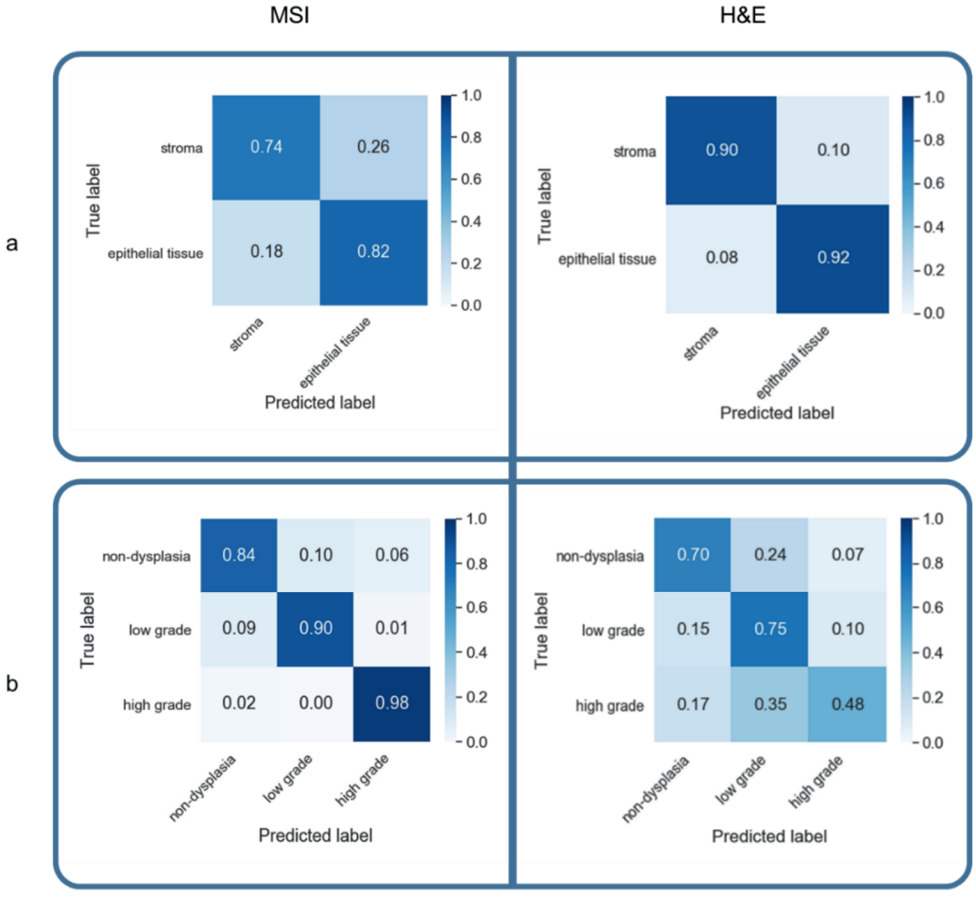
Supplementary Figure 1: three randomly selected samples at different zoom-in levels for a visual evaluation of the accuracy and for comparison the size of the MSI pixels in the last column.



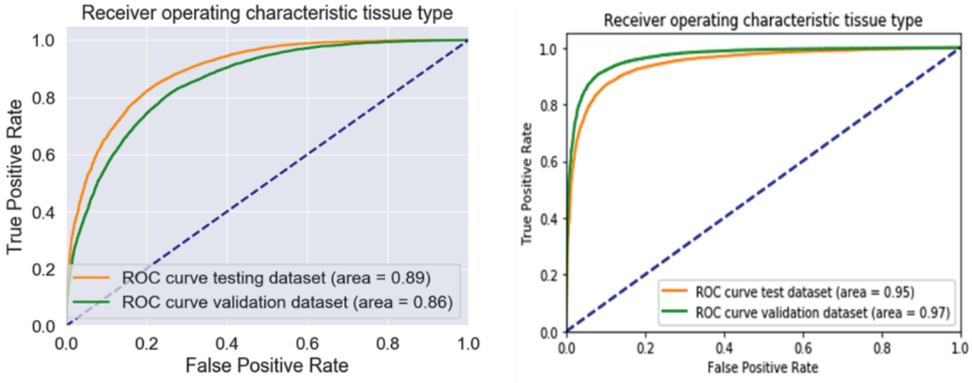
Supplementary Figure 2: example of superposition of the mass spectrometry data with the H&E data and annotation masks using affine geometric transformations for image coregistration. Abbreviations used: H&E, hematoxylin and eosin stained tissue scans; MSI, mass spectrometry imaging



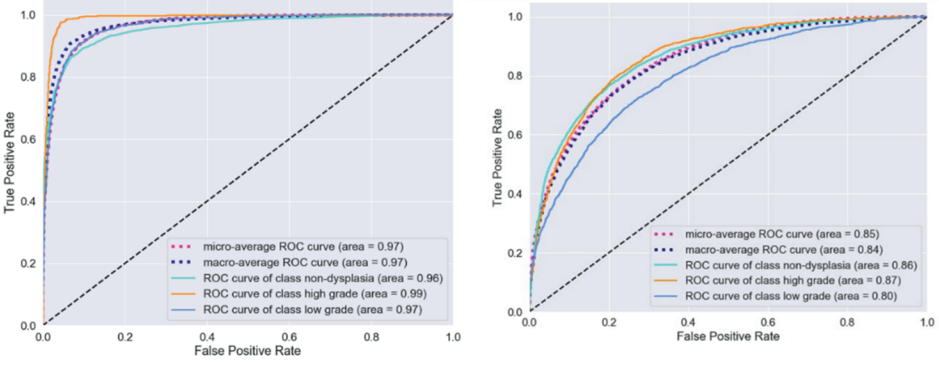
Supplementary Figure 3: Machine learning diagram for tissue type classification based on mass spectrometry features.



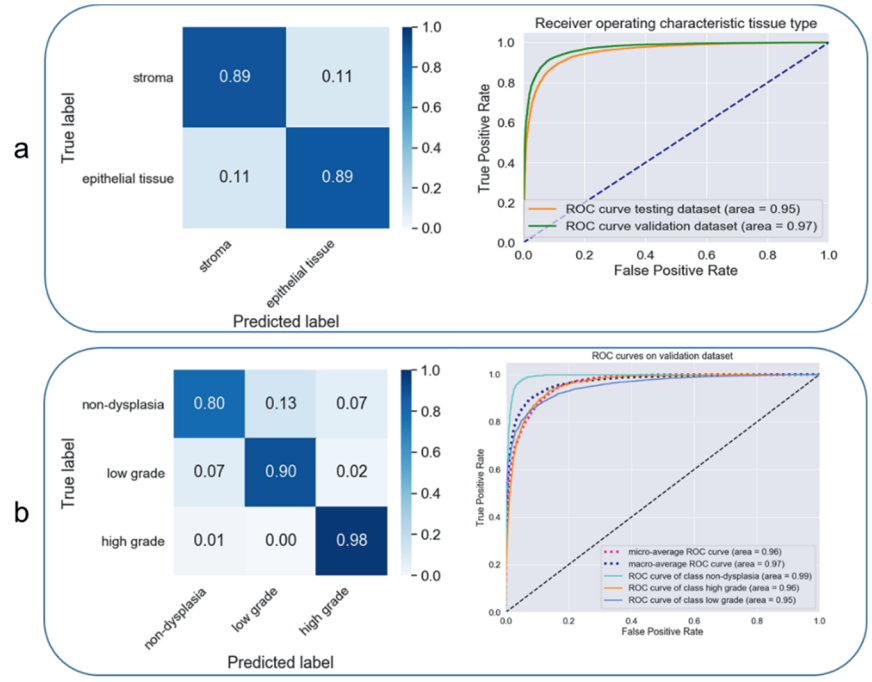
Supplementary Figure 4: Normalized confusion matrices of the validation sets for: a- prediction of the tissue type using the mass spectrometry imaging (MSI) data (left) and the H&E data (right); b- prediction of the grade using the MSI dataset balanced (left) and H&E dataset balanced (right). Abbreviations used: H&E, hematoxylin and eosin stained tissue scans; MSI, mass spectrometry imaging.



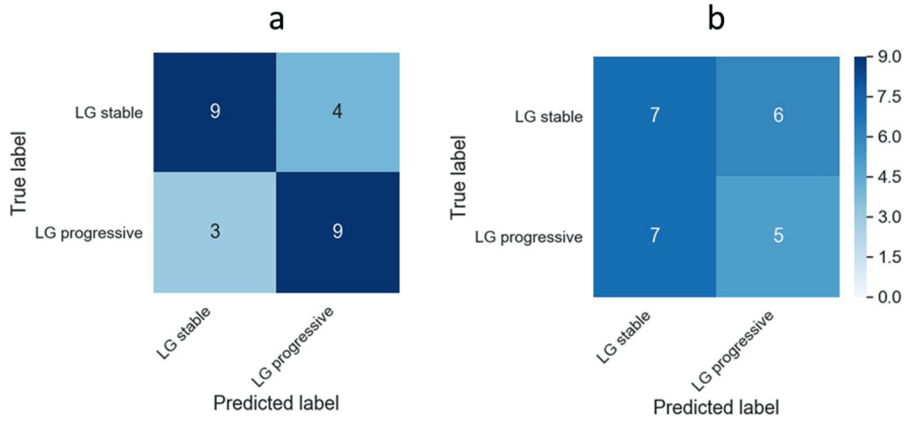
Supplementary Figure 5: ROCs curves for tissue type predictions: test and validation datasets predictions based on the MSI data (on the left); test and validation datasets predictions based on the H&E data (on the right)



Supplementary Figure 6: ROCs curves on the test sets for grading predictions: based on the MSI data (on the left); based on the H&E data (on the right)



Supplementary Figure 7: Results obtained with the combined features: a- Results for the classification between stroma and epithelial tissue. On the left, normalized confusion matrix obtained on the test dataset, on the right the comparison between the ROC curves based on the validation and the test dataset; b- Results for the prediction of the grade. On the left, normalized confusion matrix obtained on the test dataset, on the right the comparison between the ROC curves



Supplementary Figure 8: Confusion matrix for the prediction of progression in LGD patients based on: a- the MSI data; b- the H&E data. Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; LG, low grade; MSI, mass spectrometry imaging

Supplementary Tables

Supplementary Table 1: Common peaks between on- and off-tissue regions

Tissue [m/z]	Control [m/z]	Tentative identity
825,11	825,1	CHCA matrix cluster
842,52	842,51	Trypsin autolysis product (VATVSLPR) + H
845,11	845,09	CHCA matrix cluster
861,08	861,07	CHCA matrix cluster
864,51	864,49	Trypsin autolysis product (VATVSLPR) + Na
870,49	870,51	
880,46	880,47	Trypsin autolysis product (VATVSLPR) + K
1011,54	1011,65	
1045,58	1045,56	Trypsin autolysis product (LSSPATLNSR) + H
1049,54	1049,57	
1067,56	1067,54	Trypsin autolysis product (LSSPATLNSR) + Na
2713,61	2713,48	

Supplementary Table 2: MSI Peak list

Tissue m/z								
800,42 - 862,46	**866,47 2- 921,50**	**922,50 - 965,53**	**966,54- 1012,51**	**1014,54 - 1088,55**	**1090,56 - 1155,58**	**1158,58 3- 1337,70**	**1339,68 2- 1867,93**	**1869,92 3- 2982,91**
800,42	866,47	922,50	966,54	1014,54	1090,56	1158,58	1339,68	1869,92
801,43	867,46	923,49	967,54	1015,54	1092,55	1160,57	1340,65	2084,03
805,42	868,46	924,45	969,51	1018,50	1093,59	1171,59	1342,67	2104,16
806,42	871,48	926,49	971,57	1020,51	1094,60	1173,56	1347,64	2115,21
807,41	872,45	927,50	972,53	1021,53	1095,59	1176,58	1356,66	2126,17
808,44	874,44	928,48	974,51	1022,53	1098,57	1177,59	1359,69	2137,19
809,43	876,45	929,52	975,51	1024,52	1099,56	1184,57	1364,66	2159,20
810,44	878,48	930,48	976,48	1027,57	1101,56	1196,61	1366,66	2181,19
814,46	879,48	931,49	977,51	1028,60	1105,58	1198,72	1381,70	2208,13
815,44	882,48	933,47	978,52	1030,54	1107,58	1214,62	1459,72	2216,16
816,45	883,48	934,48	979,55	1032,56	1109,57	1217,63	1465,71	2219,11
817,42	884,47	936,50	980,55	1034,53	1110,56	1220,70	1481,72	2303,23
819,43	886,46	937,51	981,53	1035,54	1111,60	1229,58	1487,71	2305,25
822,44	888,46	938,49	982,51	1036,55	1113,59	1234,68	1508,73	2318,23
823,45	889,47	939,49	983,52	1039,55	1115,58	1235,64	1510,74	2321,25
825,41	890,45	940,51	984,51	1040,53	1116,58	1237,63	1530,73	2338,23
828,44	892,47	942,50	985,57	1042,55	1117,57	1239,62	1542,78	2461,34
829,43	894,47	943,54	986,55	1044,54	1119,58	1242,69	1546,80	2568,49
830,45	896,42	944,55	987,53	1048,53	1120,57	1249,67	1552,73	2690,59
831,45	898,50	945,53	988,53	1050,57	1125,59	1251,60	1562,81	2695,61
834,44	900,51	947,49	990,50	1052,54	1126,59	1257,63	1564,80	2705,59
836,45	901,51	948,49	993,56	1054,56	1127,57	1267,68	1568,79	2727,61
837,45	902,48	949,49	994,51	1059,56	1129,59	1269,69	1580,79	2750,66
839,40	903,48	950,48	995,51	1062,55	1131,57	1271,67	1584,81	2958,98
840,44	904,48	952,49	996,52	1065,55	1133,59	1275,64	1585,80	2982,91
844,49	905,47	953,50	997,52	1066,53	1135,59	1279,62	1607,80	
845,45	906,46	954,48	998,50	1069,58	1136,59	1280,62	1612,82	
846,45	908,46	955,51	999,53	1070,57	1137,58	1289,67	1629,79	
850,46	909,47	956,51	1000,52	1072,57	1138,57	1297,63	1653,81	
851,45	911,48	957,57	1001,53	1074,55	1139,56	1302,65	1655,82	
852,44	912,45	958,57	1002,50	1076,58	1141,58	1303,63	1677,82	

854,45	914,48	960,50	1004,51	1077,57	1143,58	1307,62	1706,82
856,47	915,49	961,49	1006,51	1079,56	1147,59	1319,65	1751,86
857,47	917,47	962,50	1007,55	1080,55	1149,56	1320,68	1833,95
858,45	918,47	963,50	1009,54	1081,57	1151,57	1324,64	1850,92
859,45	920,49	964,50	1010,54	1087,57	1154,58	1325,65	1855,94
862,46	921,50	965,53	1012,51	1088,55	1155,58	1337,70	1867,93

Supplementary Table 3: Features importance of random forest, xgboost and the average score based on MSI data for tissue type classification and grading.

available in addendum

Supplementary Table 4: Average dice coefficients per tile computed on the test dataset

grades	**MSI**	**H&E**
non-dysplasia	0,8	0,66
low grade	0,87	0,73
high grade	0,94	0,45
Overall average	0,82	0,65

Supplementary Table 5: Tile-based classifier performance for predicting tissue type and grade on the validation sets. Abbreviations used: H&E, hematoxylin and eosin stained tissue scans; MSI, mass spectrometry imaging

Prediction of ...	Data type	labels	precision	recall	f1-score	support (number of tiles)
tissue type	MSI data	Epithelial tissue	0,76	0,82	0,79	11053
		stroma	0,8	0,74	0,76	10671
	H&E data	Epithelial tissue	0,92	0,9	0,91	11053
		stroma	0,9	0,92	0,91	10671
grade	MSI data	non dysplastic Barrett's Esophagus	0,97	0,84	0,9	7978

		low grade dysplasia	0,67	0,9	0,77	1744
		high grade dysplasia	0,68	0,98	0,8	1124
	H&E data	non dysplastic Barrett's Esophagus	0,92	0,7	0,8	7978
		low grade dysplasia	0,37	0,75	0,49	1744
		high grade dysplasia	0,43	0,48	0,45	1124

Author contribution statement

Manon Beuque performed all the ML analysis, analysed the results of the classifiers and wrote the manuscript. Marta Martin-Lorenzo performed the MSI experiments and H&E staining. Benjamin Balluff designed the MSI experiments, supervised the progression of the project, helped to pre-process the mass spectrometry data and made the high- resolution data handling and machine ML compatible. Henry Woodruff supervised the progression of the project and the writing of this article and guarantees the integrity of the analysis and results presented. Marit Lucas transferred the tissue an-notations from the H&E staining to the MSI data. Daniel M. de Bruin supervised the work of Marit Lucas. Janita van Timmeren helped with the analysis. Onno de Boer managed and organized the BE histopathology dataset (together with Sybren Meijer). Ron MA Heeren supervised the progression of the project. Sybren Meijer devised the project's aim, collected and provided the samples, and annotated the tissues. Philippe Lambin supervised the progression of the project. All authors have participated in writing the manuscript.

References

1. R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics 70, CA Cancer J. Clin., 2020, 7–30, 1 2020.
2. N.J. Shaheen, G.W. Falk, P.G. Iyer, L.B. Gerson, ACG clinical guideline: diagnosis and management of barrett's esophagus, Am. J. Gastroenterol. 111 (2016) 30–50.
3. Y. Zhang, Epidemiology of esophageal cancer, World J. Gastroenterol. 19 (2013) 5598.
4. L.C. Duits, K.N. Phoa, W.L. Curvers, F.J.W. Ten Kate, G.A. Meijer, C.A. Seldenrijk, G.J. Offerhaus, M. Visser, S.L. Meijer, K.K. Krishnadath, J.G.P. Tijssen, R. C. Mallant-Hent, J.J.G.H.M. Bergman, Barrett's oesophagus patients with low- grade dysplasia can be accurately risk-stratified after histological review by an expert pathology panel, Gut 64 (5 2015) 700–706.
5. S.A. Gross, J. Kingsbery, J. Jang, M. Lee, A. Khan, Evaluation of dysplasia in Barrett esophagus, Gastroenterol. Hepatol. 14 (4) (2018) 233–239.
6. P.-M. Vaysse, R.M.A. Heeren, T. Porta, B. Balluff, "Mass spectrometry imaging for clinical research–latest developments, applications, and current limitations, Analyst 142 (2017) 2690–2712.
7. M. Elsner, S. Rauser, S. Maier, C. Schone, B. Balluff, S. Meding, G. Jung, M. Nipp, H. Sarioglu, G. Maccarrone, M. Aichler, A. Feuchtinger, R. Langer, U. Jütting, M. Feith, B. Küster, M. Ueffing, H. Zitzelsberger, H. Ho¨fler, A. Walch, MALDI imaging mass spectrometry reveals COX7A2, TAGLN2 and S100-A10 as novel prognostic markers in Barrett's adenocarcinoma, J. Proteomics 75 (8 2012) 4693–4704.
8. R. Lazova, K. Smoot, H. Anderson, M.J. Powell, A.S. Rosenberg, F. Rongioletti, L. Pilloni, S. D'Hallewin, R. Gueorguieva, I. Tantcheva-Poo´r, O. Obadofin, C. Camacho, A. Hsi, H.H. Kluger, O. Fadare, E.H. Seeley, "Histopathology-guided mass spectrometry differentiates benign nevi from malignant melanoma, J. Cutan. Pathol. 47 (3 2020) 226–240.
9. D.R.N. Vos, I. Jansen, M. Lucas, M.R.L. Paine, O.J. de Boer, S.L. Meijer, C.D. SavciHeijink, H.A. Marquering, D.M. de Bruin, R.M.A. Heeren, S.R. Ellis, B. Balluff, Strategies for managing multi-patient 3D mass spectrometry imaging data, J. Proteomics 193 (2 2019) 184–191.
10. S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block Attention module, Computer Vision – ECCV (2018) 3–19, 2018.
11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 in Proceedings of the IEEE conference on computer vision and pattern recognition.
12. N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, S. Hassanpour, Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides, JAMA Netw Open 2 (2019) e1914645, 11.
13. F. Yin, D. Hernandez Gonzalo, J. Lai, X. Liu, Histopathology of Barrett's esophagus and early-stage esophageal adenocarcinoma: an updated review, Gastrointestinal Disorders 1 (2019) 147–163.
14. A.J. de Groof, M.R. Struyvenberg, J. van der Putten, F. van der Sommen, K.N. Fockens, W.L. Curvers, S. Zinger, R.E. Pouw, E. Coron, F. Baldaque-Silva, O. Pech, B. Weusten,

A. Meining, H. Neuhaus, R. Bisschops, J. Dent, E.J. Schoon, P.H. de With, J.J. Bergman, Deep-learning system detects neoplasia in patients with barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking, Gastroenterology 158 (e4) (3 2020) 915–929.
15. K. Visrodia, S. Singh, R. Krishnamoorthi, D.A. Ahlquist, K.K. Wang, P.G. Iyer, D.A. Katzka, "Magnitude of missed esophageal adenocarcinoma after Barrett's esophagus diagnosis: a systematic review and meta-analysis, Gastroenterology 150 (2016) 599–607, e7.
16. F.J.C.T. Kate, F.J.C. ten Kate, D. Nieboer, F.J.W. ten Kate, M. Doukas, M.J. Bruno, M.C.W. Spaander, L.H.J. Looijenga, K. Biermann, Improved progression prediction in barrett's esophagus with low-grade dysplasia using specific histologic criteria, Am. J. Surg. Pathol. 42 (2018) 918–926.
17. S.A. Sonwalkar, O. Rotimi, N. Scott, E. Verghese, M. Dixon, A.T.R.A. Axon, S.M. Everett, A study of indefinite for dysplasia in Barrett's oesophagus: reproducibility of diagnosis, clinical outcomes and predicting progression with AMACR (α-methylacyl-CoA-racemase): indefinite for dysplasia in Barrett's oesophagus, Histopathology 56 (5 2010) 900–907.
18. N. Dimitriou, O. Arandjelovi´c, P.D. Caie, Deep learning for whole slide image analysis: an overview, Front. Med. 6 (2019).
19. Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, S. Li, Breast cancer multi-classification from histopathological images with structured deep learning model, Sci. Rep. 7 (6 2017) 4172.
20. M. Nalisnik, M. Amgad, S. Lee, S.H. Halani, J.E. Velazquez Vega, D.J. Brat, D. A. Gutman, L.A.D. Cooper, Interactive phenotyping of large-scale histology imaging data with HistomicsML, Sci. Rep. 7 (11) (2017) 14588.
21. M. Aichler, M. Elsner, N. Ludyga, A. Feuchtinger, V. Zangen, S.K. Maier, B. Balluff, C. Scho¨ne, L. Hierber, H. Braselmann, S. Meding, S. Rauser, H. Zischka, M. Aubele, M. Schmitt, M. Feith, S.M. Hauck, M. Ueffing, R. Langer, B. Kuster, H. Zitzelsberger, H. Ho¨fler, A.K. Walch, Clinical response to chemotherapy in oesophageal adenocarcinoma patients is linked to defects in mitochondria, J. Pathol. 230 (8 2013) 410–419.

Chapter 4

Predicting adverse radiation effects in brain tumours after stereotactic radiotherapy with deep learning and handcrafted radiomics

Simon A. Keek[1], Manon Beuque1, Sergey Primakov, Henry C. Woodruff, Avishek Chatterjee, Janita E. van Timmeren, Martin Vallières, Lizza E. L. Hendriks, Johannes Kraft, Nicolaus Andratschke, Steve E. Braunstein, Olivier Morin[2] and Philippe Lambin[2]

[1] These authors have contributed equally
[2] Share senior authorship

Adapted from:
Keek SA, Beuque M, Primakov S, Woodruff HC, Chatterjee A, van Timmeren JE, Vallières M, Hendriks LEL, Kraft J, Andratschke N, Braunstein SE, Morin O, Lambin P.
Predicting Adverse Radiation Effects in Brain Tumors After Stereotactic Radiotherapy With Deep Learning and Handcrafted Radiomics. Frontiers in Oncology 2022; 12. doi: 10.3389/fonc.2022.920393

Abstract

Introduction: There is a cumulative risk of 20–40% of developing brain metastases (BM) in solid cancers. Stereotactic radiotherapy (SRT) enables the application of high focal doses of radiation to a volume and is often used for BM treatment. However, SRT can cause adverse radiation effects (ARE), such as radiation necrosis, which sometimes cause irreversible damage to the brain. It is therefore of clinical interest to identify patients at a high risk of developing ARE. We hypothesized that models trained with radiomics features, deep learning (DL) features, and patient characteristics or their combination can predict ARE risk in patients with BM before SRT.

Methods: Gadolinium-enhanced T1-weighted MRIs and characteristics from patients treated with SRT for BM were collected for a training and testing cohort (N = 1,404) and a validation cohort (N = 237) from a separate institute. From each lesion in the training set, radiomics features were extracted and used to train an extreme gradient boosting (XGBoost) model. A DL model was trained on the same cohort to make a separate prediction and to extract the last layer of features. Different models using XGBoost were built using only radiomics features, DL features, and patient characteristics or a combination of them. Evaluation was performed using the area under the curve (AUC) of the receiver operating characteristic curve on the external dataset. Predictions for individual lesions and per patient developing ARE were investigated.

Results: The best-performing XGBoost model on a lesion level was trained on a combination of radiomics features and DL features (AUC of 0.71 and recall of 0.80). On a patient level, a combination of radiomics features, DL features, and patient characteristics obtained the best performance (AUC of 0.72 and recall of 0.84). The DL model achieved an AUC of 0.64 and recall of 0.85 per lesion and an AUC of 0.70 and recall of 0.60 per patient.

Conclusion: Machine learning models built on radiomics features and DL features extracted from BM combined with patient characteristics show potential to predict ARE at the patient and lesion levels. These models could be used in clinical decision making, informing patients on their risk of ARE and allowing physicians to opt for different therapies.

1. Introduction

Brain metastases (BM) are the most common intracranial malignancies, accounting for more than 50% of all brain tumours and occurring in 10 to over 40% of patients with solid malignancies (1–3). BM occur most often in patients with lung cancer, breast cancer, and melanoma, which have a cumulative risk ranging from 20 to 40% of developing BM (4–7). BM can be treated locally by surgery or radiotherapy or with systemic anticancer therapy. Treatment depends on several factors, such as patient performance status, number and volume of metastases, presence of extracranial metastases, symptoms, and presumed efficacy of available systemic therapy ["Systemic therapy for brain metastases", (8, 9). The radiotherapy of BM can be either stereotactic radiotherapy (SRT) or whole brain radiotherapy (WBRT), with SRT being the guideline-recommended treatment for a limited number of BM. As WBRT is associated with neurocognitive deterioration, SRT is increasingly used in multiple BM as well (10–12). SRT is delivered either in a single fraction, with stereotactic radiosurgery (SRS), or as fractionated stereotactic radiotherapy (FSRT) and results in a high dose within the target volume with a steep dose gradient to the surrounding healthy tissue (13).

Even though most of the healthy brain is spared from high doses of radiation, a major shortcoming of SRT is a chance of high toxicity in the immediate surrounding tissues, which may lead to adverse radiation effects (ARE) such as radiation necrosis (RN), subacute edema, structural changes in the white matter, and vascular lesions (14). ARE are a relatively late reaction to irradiation of healthy tissues where either reversible or irreversible injury has occurred (15). The risk of ARE after SRT and SRS is found to be similar and ranges from 5 to 10% at patient level (16–19) or approximately 3% at lesion level (15). Known predictors of ARE are tumour volume, isodose volume, and previous SRT to the same lesion (15). ARE of the tumour area and tumour progression (TP) as two different post-therapeutic events require different treatment strategies: while steroids are often indicated for the initial treatment of ARE, true progression or relapse requires repeated radiotherapy, surgery, or effective intracranial systemic therapy for tumour control. Being able to differentiate between ARE and TP is therefore of utmost clinical interest.

Unfortunately, the (neurological) symptoms of ARE and TP are usually indistinguishable. Furthermore, the appearances of ARE and TP are very difficult to discern through qualitative radiological imaging, requiring multiple successive magnetic resonance images (MRI), specialized MRI sequences such as perfusion-weighted or MR spectroscopy, and trained experts to evaluate the findings (19, 20). The clinical workflow is time- and labor-intensive, and while it is unfeasible to perform for every lesion, a definitive confirmation of the presence of ARE requires tissue acquisition (19).

SRT requires routine pretreatment MRI for accurate target volume delineation. This imaging provides a source of non-invasively acquired information about BM and brain phenotypes that could be investigated for their potential to determine before treatment which patient has a high risk of developing ARE. The early identification of these patients is an unmet clinical need which may help in clinical decision making by informing the patients of the risk

of ARE, the early risk stratification of patients that may develop ARE, and the consideration of ARE risk mitigating strategies such as deferring radiotherapy for central nervous system-penetrant systemic therapy.

Advanced quantitative medical image analysis methods such as radiomics and deep learning (DL) extract large amounts of imaging features and associate these with biological and/or clinical outcomes using machine learning (ML) techniques (21–26). Thus, radiological images from routine imaging procedures could potentially be used to non-invasively quantify the lesion phenotype, providing clinically necessary information for patient management decisions. Several studies have indicated that MRI radiomics analysis is able to differentiate BM from glioblastoma (27, 28) to predict local recurrence (29, 30), to predict the origin of metastases (31, 32), and to predict overall survival (33, 34). DL has also shown potential in predicting treatment response on brain MRI (35). Moreover, DL and radiomics can have a complementary value, potentially establishing a more robust classifier (36).

We hypothesize that models trained with radiomics features, DL features, and patient characteristics or a combination thereof can predict the occurrence of ARE in patients with BM, both lesion specific and patient specific.

2. Materials and methods

2.1. Patient Characteristics

All data from patients with BM treated with SRT between 1997 and 2017 for which imaging, outcome data, and patient data were available were collected retrospectively from the University of California—San Francisco (UCSF) medical center's picture archiving and communication system. Available imaging data, outcome data, and patient data of all patients with BM treated with SRS/SRT between 2014 and 2019 at the University Hospital Zürich (USZ) were collected retrospectively. The data included clinical and biological information for both the patient and the lesion. The eligibility criteria included radical treatment for metastatic brain cancer using Gamma Knife SRS for the UCSF patients and SRS/FSRT for the USZ patients. The inclusion of patients was regardless of the number of BM, but pathohistological or imaging-based confirmation of ARE during the follow-up was required in addition to pathohistological confirmation of the primary tumour. For the USZ cohort, in case of imaging-based suspicion of RN, positron emission tomography imaging was additionally used to exclude TP. The effort obtained ethical approval for observational research using anonymized linked care data for supporting medical purposes that are in the interests of individuals and the wider public. UCSF Institutional Review Board (https://irb.ucsf.edu) and Cantonal Ethics Committee Zurich approval with waiver of informed consent was obtained.

The UCSF dataset was divided randomly into sub-cohorts for training (70%) and testing (30%) while maintaining the ratios of events to non-events equal in both groups. The USZ dataset was used as an independent external validation dataset, i.e., it was entirely unseen by the models during the training and testing phases. The binary outcome used in training and validation was ARE per lesion, defined as either pathologically or imaging-based confirmation of RN occurring at any time after treatment. For both the USCF and USZ

patients, ARE was confirmed by histopathology when treated with open surgery. In all other cases, ARE was confirmed either at routine re-staging 3 months after radiotherapy for asymptomatic patients or at the onset of new symptoms. When patients presented new symptoms, imaging was performed usually after awaiting the effects of cortisone administration. As the time of BM formation is unknown, the outcome was not defined as right-censored. As every lesion is able to independently develop ARE after treatment, every lesion was considered to be an independent sample. The probability of ARE occurring for any lesion within a patient as an outcome was also investigated, whereby each patient was treated as an independent sample instead.

2.2. MR Acquisition Parameters and Lesion Segmentation

All images were axial gadolinium-enhanced T1-weighted MRI acquired prior to the treatment of BM. All included lesions were three-dimensionally delineated for curative Gamma Knife SRS treatment purposes for the UCSF cohort and for curative SRS/ FSRT purposes for the USZ cohort according to local protocols by an experienced radiation oncologist. Figure 1 shows two T1-weighted gadolinium-enhanced MRI with lesions delineated for SRT purposes.

To perform segmentations of the brain and the ventricles on the entire dataset, an atlas-based segmentation strategy was chosen. To create the atlas in the MIM software package (MIM v. 6.9.4, MIM Software Inc., Cleveland, OH, USA), 50 randomly chosen MRI were manually segmented by an expert radiologist.

2.3. Pre-Processing of Brain MRI Data

Bias-field correction was performed in the MIM software package using the N4 algorithm, which required brain segmentations (37). A bias field is a low-frequency signal distributed over an MR image, which is caused by inhomogeneities in the magnetic field of the MRI scanner. This causes shifts of intensity value ranges across the image (38). The ventricle mask was subtracted from the brain mask to obtain a white- and gray-matter segmentation. This segmentation was used to determine and correct the bias field present in the image using the N4 algorithm (37) using the MIM software package.

Following the bias correction, all remaining pre-processing, feature extraction, model training, and evaluation were performed in Python (version 3.7). The different Python packages used during this study can be found in Supplementary Table S1. Pre-processing of MRI is essential for ML purposes, for reducing scanner dependence, and for ensuring reproducibility (39–41). As there is, to date, no consensus regarding the best way to pre-process MRI for our purposes, three different pre-processing workflows were applied and compared: “minimalist”, standardization, and “harmonization”. The descriptions of these pre-processing workflows can be found in the Supplementary Materials (Section 1 and in Figure 2).

Pre-processing for radiomics and feature extraction

Feature extraction was performed according to the Image Biomarker Standardization Initiative (IBSI) guidelines (42–44) on the three different sets of processed MRI scans using the BM segmentations. All images were resampled to uniform 1 × 1 × 1- mm3 voxels using the "sitkBSpline" interpolator to correct for differences in pixel size and slice spacing. The choice for voxel dimensions was made based on majority ruling, as it was found that most patients had a pixel spacing of ~1 mm. To achieve isotropic voxels, the choice for resampling in the z-direction was also chosen as 1 mm. Pixel intensity values were resampled to a fixed number of 64 bins, as the number of gray levels was found to affect the interchangeability of MRI radiomics features, and a fixed bin number of 64 has been found recommended in previous studies (42–44).

A total of 106 IBSI features were extracted from each segmentation. The features were extracted from the BM segmentations of the pre-processed images and can be divided into first-order intensity, histogram statistics, shape, and texture features. A full list and a description of the features can be found in the PyRadiomics documentation ([Radiomics features— PyRadiomics Documentation, (45)], and a description of the feature groups can be found in the Supplementary Materials (Section 2).

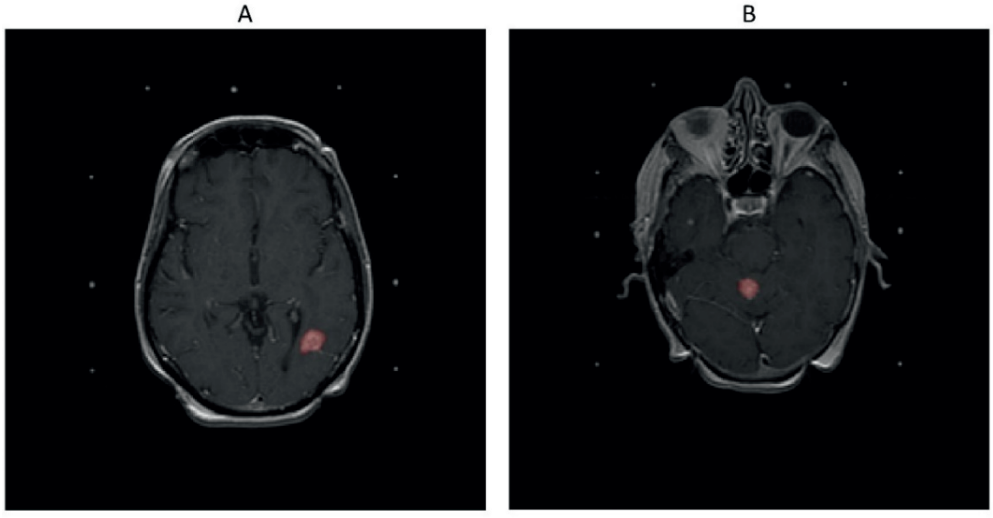


Figure 1: T1-weighted gadolinium-enhanced MRIs of the brain. Delineated in red (A) is a lesion that developed adverse radiation effects after stereotactic radiotherapy and (B) a lesion that did not develop adverse radiation effects after stereotactic radiotherapy.

Pre-processing for deep learning

To inform the DL model on the location and extension of the lesions, lesion masks were used to highlight the ROI. A Gaussian smoothing filter was applied to the image, gradually decreasing the intensity values around the lesion from a factor of 1.0 to 0.2 to still include information of the voxels immediately around the lesion masks.

Otsu thresholding was performed to create a mask containing the brain and the skull. This mask was used to determine the largest three-dimensional bounding box containing the brain and the skull to crop the images. Anything outside this mask was defined as the image background, for which all pixel values were set at 0. For the "minimalist" and the "standardization" datasets, the intensities were resampled in a range between 0 and 255. Finally, the scans were rescaled at 256 × 256 × 64 with spline interpolation order 3. As an example, the steps of the pre- processing workflow for the "minimalist" normalization are illustrated in Figure 3.

Figure 2: Pre-processing strategies for the "minimalist", "standardization", and "harmonization" approaches.

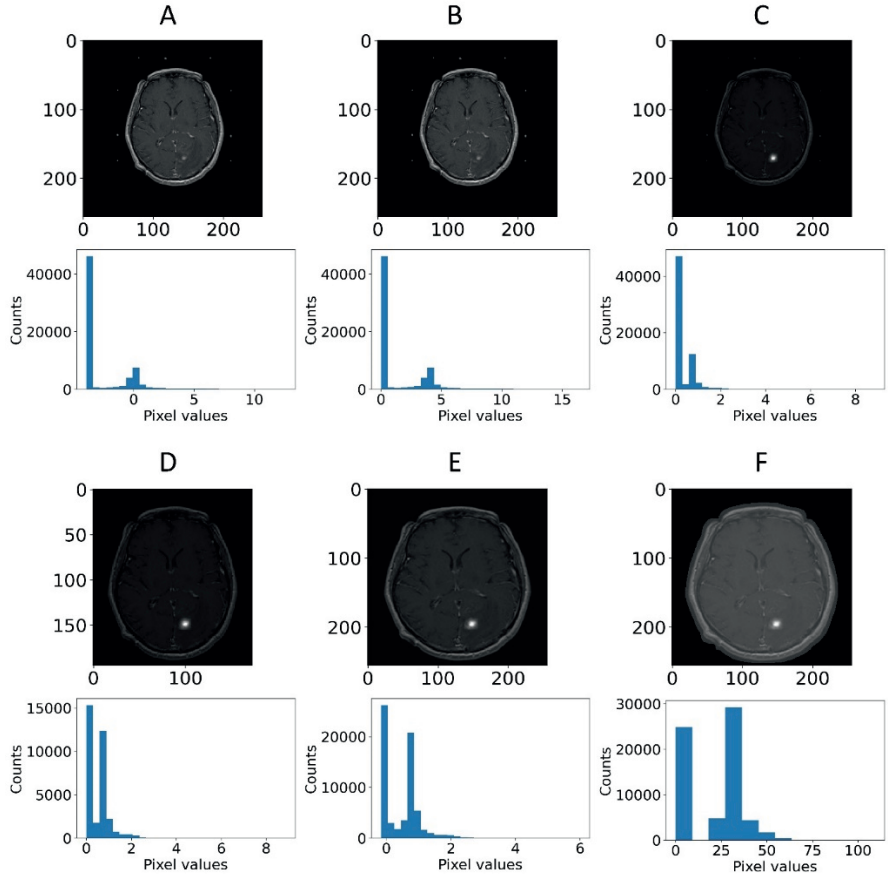


Figure 3: Example of pre-processing strategy: deep learning on the "minimalist" approach. The different steps of preprocessing were (A) z-score normalization, (B) shift to positive values only, (C) pixel attenuations with Gaussian smoothing filtering, (D) cropping around the largest bounding box and background set to 0, (E) resizing at 256 × 256, and (F) rescaling the pixel value range to 0–255.

2.4. Machine Learning Models

The mean and SD of each feature over the entire training population were determined. These values were used to apply z-score normalization to the features of the training, testing, and external validation datasets (46). Next, features with low variance (<0.01) were determined and excluded from the dataset. Lastly, the correlation between features was determined using absolute pairwise Spearman rank correlation. As highly correlated features (>0.85) were assumed to contain overlapping information about the outcome, the feature with the highest mean absolute correlation with the rest of the features was excluded. Lastly, supervised feature selection was performed through recursive feature elimination (RFE). RFE uses a ML algorithm to build a multivariate model and determine predictive performance using the currently selected features. It recursively drops and adds

features, determining the optimal number of features and the selection of most predictive features.
An extreme gradient boosting (XGBoost) model was used for RFE and ARE prediction. A description of the XGBoost architecture and the methodology to determine the optimal hyperparameters for the trained models can be found in the supplementary materials (Section 3).

2.5. Deep Learning Model

An Xception three-dimensional model was trained and tested on the same datasets as the handcrafted radiomics-based model. Xception is the extreme version of an Inception model (47), which uses depth-wise separable convolutions. The architecture can be found in Supplementary Figure S1. Adam optimization was used (48) with an initial learning rate of 10-5, which updated the learning rate during training, and used for loss function binary cross-entropy. This model produced a score ranging from 0 to 1, indicating the estimated probability that a lesion develops ARE. The area under the curve (AUC) of the receiver operating characteristic (ROC) was monitored on the test dataset. The ROC displays the discriminative performance of a model expressed through the sensitivity and specificity as the threshold for binary classification is shifted. The AUC of the ROC is a metric from 0 to 1, where 1 means that the model has perfect predictive performance and 0.5 is equivalent to guessing. To limit the imbalance of the outcomes to affect the model training, the model was only trained on lesions for those patients who had at least a single ARE and tested on the scans of the patients who had ARE in the test dataset. To combine DL and radiomics, the last fully connected layer consisting of 256 features obtained after training the model was extracted. These features were then used to train a ML model similarly to using radiomics features and used in models combining radiomics features and patient characteristics.

2.6. Clinical and Treatment-Related Feature Model

As the training and testing datasets contained patient characteristics not available in the external validation dataset, any feature not overlapping between these datasets was dropped. The list of the remaining features was as follows: primary tumour location, primary tumour histology, primary tumour controlled, extra-cranial metastases presence, patient age, patient sex, SRS to the same location, prior external beam radiotherapy (EBRT), prior radiosurgery (RS), neurological symptoms, headaches, seizures, hypertension, diabetes, connective tissue disorder, Karnofsky performance score (KPS) status, prescription dose, and isodose lines. For XGBoost to be able to handle categorical variables, one-hot encoding was performed on two categorical clinical features (primary tumour location and primary tumour histology).
Missing values were imputed using MissForest. MissForest is an imputation algorithm that uses RandomForest to train a model on the non-missing data for each feature with missing values to predict the missing values. In the first iteration, all values are set to the mean value present for each variable (i.e., each column). Then, over multiple iterations, each data

column with missing values will be predicted using all the data except for the rows containing the missing values in question. This process is repeated over several iterations.

2.7. Metrics Used for Data Analysis

The patient and tumour characteristics in the UCSF and USZ cohorts were assessed through a two-proportion z-test to test for significant differences in categorical variables between the cohorts or the unpaired two-sample t-test to test for significant differences in numerical variables. For the latter, the assumptions of the data having a normal distribution and possessing the same variance in both cohorts were tested through Shapiro–Wilk's test and f-test, respectively. The significance level was set at 5%.

To determine which method ensured best performance for the radiomics-based and DL models, models were trained on the three different pre-processed datasets, and the best AUC of the ROC on the testing set was used to determine the best pre- processing methods for ML and DL separately. The 95% confidence intervals (CI) displayed on the ROC curves were obtained using bootstrapping (n = 2,000). For the radiomics- based model, the results were reported on the full train dataset and the entire test dataset. For the DL model, the results were reported on the balanced train dataset (which served to train the different DL models) and the full test dataset.

Once the best models were selected, the models were validated on the external dataset. The predictive performance of each model was expressed through the ROC curve and its AUC on the training, testing, and external data. By determining an optimal threshold value using Youden's J statistic (49) based on the training dataset, a binary classification was performed on the external dataset. From this binary classification, the balanced accuracy, precision, recall, and F1-score were determined. The confusion matrices were also derived from the binary classification. To determine model performance and to compare between models, the recall was investigated specifically, which is the proportion of true positives of the total number of true cases. As the number of events was relatively low and not missing any patients at risk of ARE is crucial, a high recall of the models was desirable. The CI obtained for all metrics were obtained using bootstrapping, resampling the results 2,000 times. Moreover, an analysis of the agreement prediction between the DL model and the radiomics-based model was performed. To give a prediction per patient, the maximum prediction of ARE among the different lesion predictions of the patient was selected. The ground truth to which the prediction was compared with was the ARE status of the patient, meaning that the patient had at least one ARE lesion. An overview of the models tested can be found in Figure 4.

We evaluated on the external dataset for which cases the DL model and the best radiomics classifier obtained the same predictions and reported the number of cases for which those models agreed on the label. The metrics based on the data for which the models agreed was also reported.

3. Results

3.1 Patient Characteristics

A total of 1,404 patients with 7,974 lesions from UCSF and 237 patients with 646 lesions from USZ were included. Table 1 shows an overview of the patient characteristics of the UCSF and USZ data. Significant differences between the proportion of male and female patients between the datasets ($P < 0.01$), median age ($P = 0.03$), KPS status ($P < 0.01$), and the number of lesions per patient at treatment ($P < 0.01$) were found. Furthermore, the proportions of primary tumour (lung, melanoma, and breast) were different between the datasets, and the data from USZ did not have kidney, GI, sarcoma, or other types of primary locations that were present in the UCSF dataset. For the histology of the primary tumour, only the melanoma histology subtype was found to be present in a significantly different proportion.

3.2. Radiomics-Based Model and DL Model Results Based on the Three Different Preprocessing Methods of the Dataset

The best AUC on the test dataset for the radiomics-based models was found using the "harmonization" normalization, with an AUC of 0.76 (CI of 0.70–0.81), compared with 0.75 (CI of 0.70– 0.80) and 0.73 (CI of 0.67–0.79) for the "minimalist" and "standardization" methods, respectively.

The best AUC on the test dataset for the DL models was found using the "standardization" normalization, with an AUC of 0.72 (CI of 0.66–0.78), compared with 0.63 (CI of 0.57–0.70) and 0.65 (CI of 0.58–0.71) for the "minimalist" and "harmonization" methods, respectively. Figure 5 shows the ROC curves of the training and testing datasets for the three different pre-processing methods for radiomics-based ML and for DL.

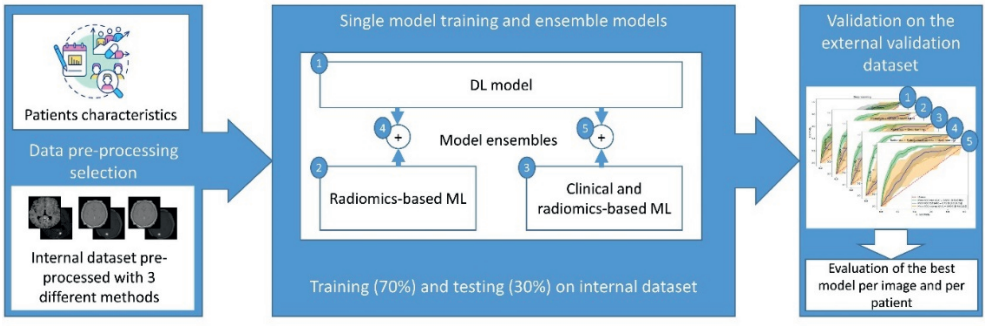


Figure 4: General workflow of the model training process: first, the MRI data was pre-processed using 3 pre-processing methods, the most suitable pre- processed set of images was selected according to the radiomics-based model or the DL model performance on the internal test dataset, then the models were ensembled or trained separately, and finally the performance of each model was computed on the external dataset.

Table 1: Patient characteristics of University of California—San Francisco (UCSF) and University Hospital Zurich (USZ) datasets.

Patient/Tumour Characteristic		Total UCSF data N = 1404	USZ data N = 237	*P*
Sex (%)	**Male**	571 (41)	128 (54)	<0.01
	Female	833 (59)	109 (46)	
Median Age ± SD		59 (13)	62 (12)	0.03
KPS (%)	**80-100**	1053 (75)	198 (83)	<0.01
	40-80	351 (25)	37 (16)	<0.01
	10-40	0 (0)	2 (1)	-
Primary tumour location (%)	**Lung**	530 (38)	136 (58)	<0.01
	Breast	357 (25)	27 (11)	<0.01
	Melanoma	272 (19)	74 (31)	<0.01
	Kidney	91 (7)	0 (0)	-
	Gastrointestinal	57 (4)	0 (0)	-
	Gynecologic	27 (2)	0 (0)	-
	Sarcoma	20 (1)	0 (0)	-
	Other	50 (4)	0 (0)	-
Histology primary tumour (%)	**Adenocarcinoma**	802 (57)	124 (52)	0.17
	Melanoma	272 (19)	74 (31)	<0.01
	Renal cell carcinoma	88 (6)	0 (0)	-
	Small cell carcinoma	44 (3)	0 (0)	-
	Squamous cell carcinoma	40 (3)	10 (4)	0.26
	Sarcoma	18 (1)	0 (0)	-
	Large cell carcinoma	9 (0.6)	2 (1)	0.72
	Bone carcinoma	8 (0.6)	0 (0)	-
	Adeno squamous carcinoma	6 (0.4)	0 (0)	-
	Broncho alveolar cell carcinoma	5 (0.4)	0 (0)	-
	Germ cell carcinoma	2 (0.1)	0 (0)	-
	Lymphoma	1 (0.1)	0 (0)	-
	Other/NOS	109 (8)	27 (11)	0.06
Primary controlled		974 (70)	149 (63)	0.05
ECM present		1097 (78)	190 (80)	0.48
#Lesions per patient at treatment	**Median ± SD**	3 (7)	2 (3)	<0.01
Symptoms	**Headaches**	437 (31)	31 (13)	<0.01
	Hypertension	407 (29)	0 (0)	< 0.01
	Seizures	134 (10)	16 (7)	0.17
	Diabetes	98 (7)	13 (6)	0.4
	CTD	21 (2)	2 (1)	0.43
#Lesions in total		7974	646	-
#ARE cases (% of total lesions)		217 (2.7)	20 (3.1)	0.61
#Patients with ARE (% of total patients)		155 (11)	19 (8)	0.16
Prescription dose ± SD (Gy)		18.5 (1.5)	20 (5.0)	-

Footnote: P value of two-proportion z-test or unpaired two-sample t-test for significant differences between datasets was reported for each characteristic if applicable. SD = standard deviation; KPS = Karnofsky performance score: 80-100 good performance, 50-70 medium performance, 10-40 bad performance; ECM = extracranial metastasis; BM = brain metastasis; CTD = connective tissue disorder; ARE = adverse radiation effect; Gy = gray.

3.3. Results of the Combined Best- Performing Models

We calculated the AUC and CI for each model combination on the external validation dataset. The DL model, built on images pre-processed with the “standardization” method,

achieved an AUC of 0.64 (CI of 0.50–0.76). The model built on radiomics features, extracted from the images pre-processed with the "harmonization" method, achieved an AUC of 0.73 (CI of 0.63–0.83). The model was built on 20 features selected through RFE. Supplementary Figure S2A provides an overview of the selected features and the corresponding importance in the XGBoost model. Supplementary Table S2 provides an overview of the hyperparameters determined through grid search cross-validation. The model based on the combination of the DL features extracted from the last layer and radiomics features achieved an AUC of 0.71 (CI of 0.60–0.82). The model was built on 10 features selected through RFE. Supplementary Figure S2B provides an overview of the selected features and the corresponding importance in the XGBoost model. The model built on radiomics features, extracted from images pre-processed with the "harmonization" method, combined with patient characteristic features achieved an AUC of 0.70 (CI of 0.57–0.80). The model was built on 19 features selected through RFE. Supplementary Figure S2C provides an overview of the selected features and the corresponding importance in the XGBoost model. Finally, the model built on radiomics features, extracted from images pre-processed with the "harmonization" method, combined with DL features, extracted from images pre-processed with the "standardization" method, and patient characteristics achieved an AUC of 0.69 (CI of 0.56–0.81). The model was built on 20 features selected through RFE. Supplementary Figure S2D provides an overview of the selected features and the corresponding importance in the XGBoost model. Figure 6 shows the ROC curves with CI of the training datasets, testing datasets, and validation datasets for these models.

The combination of radiomics and DL features achieved the highest combination of balanced accuracy and recall of 0.67 (CI of 0.56–0.76) and 0.80 (CI of 0.62–0.96), respectively, of the externally validated models for predictions per lesion. For a patient-level prediction, the DL model achieved an AUC of 0.70 (CI of 0.56–0.80) and that of the radiomics model an AUC of 0.72 (CI of 0.60–0.83). A combination of radiomics and DL achieved an AUC 0.71 (CI of 0.57–0.83), that of a combination of radiomics and patient characteristics an AUC of 0.71 (CI of 0.59–0.81), and that of a combination of radiomics features, DL features, and patient characteristics an AUC of 0.72 (CI of 0.58– 0.84). The model combining radiomics features, DL features, and patient characteristics achieved the highest combination of balanced accuracy and recall of 0.65 (CI of 0.55–0.74) and 0.84 (CI of 0.65–1.00), respectively, of the externally validated models for predictions per patient. The DL model predictions and the radiomics-based model predictions per lesion agreed for 32% of the external dataset. For the per-patient classification, the DL model predictions and the radiomics combined with clinical feature-based model predictions agreed for 19% of the external dataset. Because the number of patients for which the models agreed was low (47 patients, 6 with ARE), no CI could be derived. Table 2 provides an overview of the AUC, balanced accuracy, precision, recall, and F1 score metrics for all DL and ML models on both lesion and patient levels and for the agreed labels on the external validation. The corresponding confusion matrices are in Supplementary Figures S3, S4, respectively. Supplementary Tables S3, S4 contain the same metrics as that in Table 2 for the training and testing datasets, respectively.

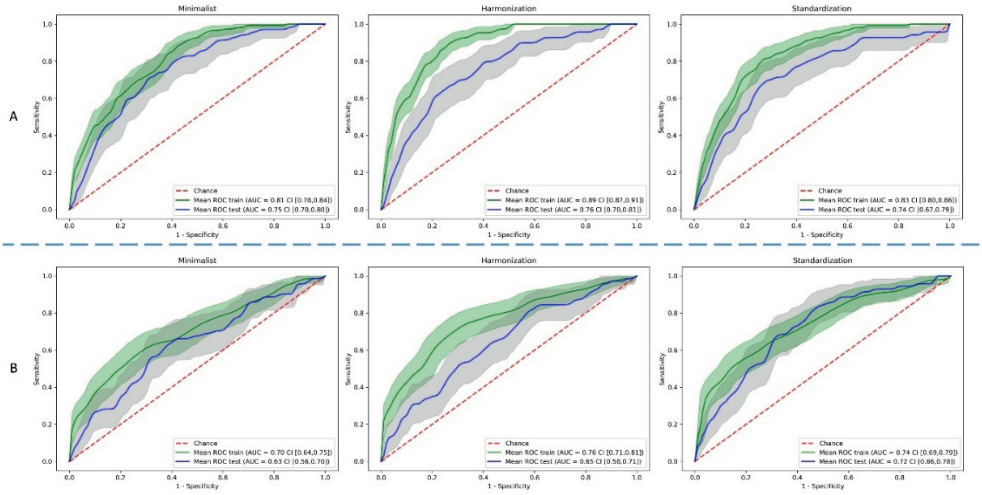


Figure 5: Comparison of predictive performance through receiver operating characteristic curves for (A) radiomics-based machine learning and (B) deep learning models using three different pre-processed image datasets. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

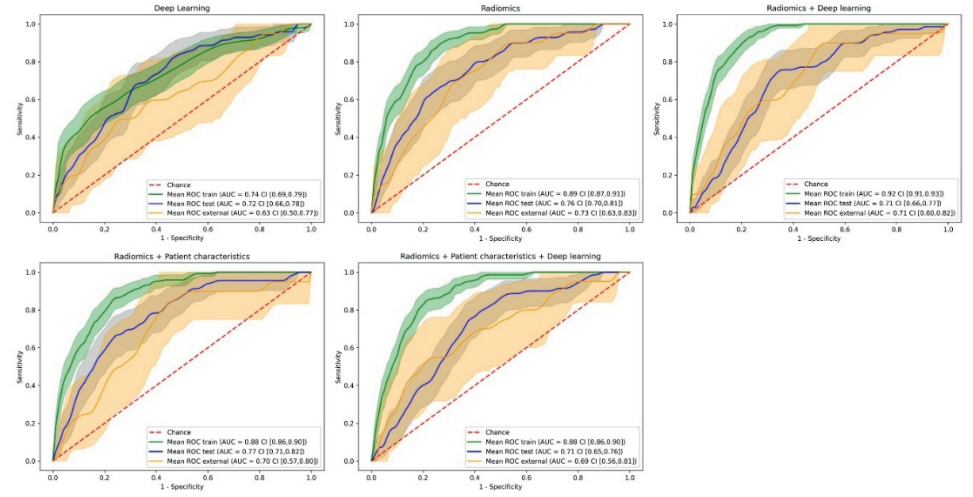


Figure 6: Receiver operating characteristic curves of the training, testing, and external validation datasets for the different model combinations. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

4. Discussion

Patients with BM treated with SRT are at risk of developing ARE, such as RN. Early identification of these patients can help in clinical decision making. The MRIs required for

SRT planning provide an opportunity to identify these patients through quantitative imaging methods. In this large-scale study, ML models that can successfully predict ARE were trained on T1- weighted MR imaging features from secondary brain tumours treated with SRT. As no consensus to harmonize MR images within and between centers exists, multiple methods were tested for the DL and ML pipeline, resulting in two optimal pre- processing methods ("harmonization" for the ML pipeline and "standardization" for the DL pipeline). A ML model trained with radiomics features combined with DL features yielded the highest predictive performance, with a combination of ROC AUC, balanced accuracy, and recall of 0.71, 0.67, and 0.80, respectively. At the patient level, the best-performing ML model was clearly a combination of radiomics, clinical (age at treatment, prior RS, and sex), and DL features achieving the highest predictive performance (AUC of 0.72), a balanced accuracy of 0.65, and recall of 0.84.

Performing an aggregate prediction (i.e., using only those predictions that agreed on the outcome) did not improve predictive performance for the lesion-level prediction (AUC of 0.67) nor the binary prediction (balanced accuracy of 0.65). However, using this method, the highest recall of 0.90 was achieved, making this method very robust in detecting true positives. The models pave the way for clinical decision making of patients at risk of ARE before treatment. The information on the risk of an individual patient may be used by clinicians to inform patients of the risk of ARE when SRT is used as treatment. Furthermore, this information may be used to perform an early stratification of those patients at high risk or may allow the patient and clinician to pursue alternative therapy, such as systemic therapy or alternate radiotherapy approaches (e.g., dose de-intensified SRT or WBRT), if the risk of ARE outweighs the possible benefits of SRT (50).

To our knowledge, this is the first study that performs a pre- treatment prediction of ARE using quantitative image analysis. Several studies have investigated the possibility of differentiating between tumour recurrence and RN after treatment, which is nominally similar in purpose to identify those patients who may have ARE. Zhang et al. (51) used radiomics features extracted from four different MR sequences [T1, T1 post-contrast, T2, and fluid-attenuated inversion recovery (FLAIR)] at two different time-points during follow-up to differentiate RN from TP as confirmed pathologically. A model was built on a dataset of 87 patients with 97 lesions using 5 delta-radiomics features from T1 and T2 sequences. The AUC and binary prediction accuracy of the model were both 0.73. However, this result was obtained using leave-one-out cross-validation, as no external validation was used. Similarly, Peng et al. created a model on radiomics features extracted from T1 and T2 FLAIR on 66 patients with 77 lesions in total (52). The model was compared with a neuroradiologist's performance. No external validation was used, and instead a leave- one-out cross-validation was performed, which gave an AUC of 0.81. The sensitivity and specificity of the neuroradiologist were 0.97 and 0.17, compared with 0.65 and 0.87 for the radiomics- based model. In Park etal. (53), the study compared the results obtained after training radiomics-based models using different MRI sequences [T1, T2, and apparent diffusion coefficient (ADC)]. The models were trained using the data from 86 patients and tested on an external dataset of 41 patients. The best AUC was found on the ADC-based data with 0.80, while the other sequences had AUCs of around 0.65. These results are

similar or higher than the results obtained with our model, though within the range of the confidence intervals for the model based on radiomics and DL, and the lack of an external dataset on two of the studies makes the validity of these models difficult to determine (52). Most other studies have a similar lack of external validation and total number of included patients, further making the results difficult to compare with the present study (54). These results show that the model presented in this study is able to perform similarly to or even outperform models that perform classification (post-treatment) instead of prediction (pre-treatment) of ARE.

Table 2: Area under the curve (AUC), balanced accuracy, precision, recall, and F1 metrics with CI on the external validation on patient and lesion levels.

Per-lesion classification						Per-patient classification					
Approaches	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**	**Approaches**	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**
Best deep learning model	0.64 CI [0.50, 0.76]	0.57 CI [0.48, 0.64]	0.04 CI [0.02, 0.05]	0.85 CI [0.67, 1.00]	0.07 CI [0.04, 0.10]	Best deep learning model	0.70 CI [0.56, 0.83]	0.63 CI [0.52, 0.73]	0.17 CI [0.09, 0.25]	0.60 CI [0.39, 0.78]	0.26 CI [0.16, 0.37]
Best radiomics model	0.73 CI [0.63, 0.83]	0.62 CI [0.51, 0.74]	0.07 CI [0.03, 0.11]	0.45 CI [0.23, 0.67]	0.12 CI [0.05, 0.19]	Best radiomics model	0.72 CI [0.60, 0.83]	0.59 CI [0.51, 0.69]	0.40 CI [0.09, 0.75]	0.21 CI [0.05, 0.43]	0.28 CI [0.07, 0.48]
Radiomics and DL	0.71 CI [0.60, 0.82]	0.67 CI [0.56, 0.76]	0.05 CI [0.03, 0.08]	0.80 CI [0.62, 0.96]	0.10 CI [0.06, 0.14]	Radiomics and DL	0.71 CI [0.57, 0.83]	0.66 CI [0.54, 0.77]	0.14 CI [0.07, 0.22]	0.63 CI [0.40, 0.84]	0.23 CI [0.13, 0.34]
Radiomics and patient characteristics	0.70 CI [0.57, 0.80]	0.62 CI [0.51, 0.74]	0.06 CI [0.03, 0.10]	0.50 CI [0.28, 0.73]	0.11 CI [0.05, 0.17]	Radiomics and patient characteristics	0.71 CI [0.59, 0.81]	0.57 CI [0.48, 0.68]	0.16 CI [0.04, 0.30]	0.26 CI [0.08, 0.47]	0.20 CI [0.05, 0.35]
Radiomics, DL, and patient characteristics	0.69 CI [0.56, 0.81]	0.64 CI [0.53, 0.74]	0.05 CI [0.03, 0.08]	0.70 CI [0.48, 0.89]	0.09 CI [0.05, 0.14]	Radiomics, DL, and patient characteristics	0.72 CI [0.58, 0.84]	0.65 CI [0.55, 0.74]	0.12 CI [0.07, 0.17]	0.84 CI [0.65, 1.00]	0.21 CI [0.13, 0.29]
Agreed labels	0.67 CI [0.53, 0.81]	0.65 CI [0.53, 0.73]	0.07 CI [0.03, 0.12]	0.90 CI [0.67, 1.00]	0.13 CI [0.06, 0.21]	Agreed labels	NA	NA	NA	NA	NA

One of the strengths of the present study is the large number of included patients and subsequent lesions, with 7,974 lesions (2.7% ARE) of 1,404 patients in training and testing and 646 lesions (3.1% ARE) of 237 patients in the external validation. This provides a large volume of data for our models to train on, ensuring that it covers the wide variability found between patients. In addition, the inclusion of an external validation is another strength, especially seeing the general lack of one in most other studies investigating ARE. This ensures that the reported result is not too optimistic and shows that our model can be generalizable to populations from a different hospital in a different country and even with different treatments from the training and testing sets. While the difference in treatment

between the training (exclusively SRS) and external validation (a mix of SRS and FSRT) may induce variability due to small differences in treatment planning for these methods, literature has shown that these methods carry the same risk of ARE and were therefore considered interchangeable (16, 17, 19).

The large confidence interval on the external validation is partially due to the low number of positive findings in this dataset (n = 20). This is because of the large imbalance in outcomes for both ARE and tumour failure. One of the major problems that may arise from this imbalance is a skewed view of predictive performance. However, this was addressed in the present study through multiple measures. The DL model was trained on a balanced subset of the data that only included patients that suffered at least 1 ARE. For ML, the XGBoost model was trained while scaling the weights of positive and negative classes and the respective proportion of the labels. Finally, through analysis of the confusion matrix, precision recall curves, and recall metric, we ensured that the performance of the model was not entirely driven by labeling the data as the majority class.

While the models have been successfully validated on a dataset from an external center, further validation on multiple centers is required to ensure that the models are generalizable. Future research could therefore focus on validating the present model on other datasets, potentially with recalibration of the model. At a later stage, a clinical trial to test the efficacy of the model is needed to be able to incorporate the model in a clinical setting. A model combining radiomics features, DL features, and patient characteristics with a high accuracy could help choose other treatment options such as surgery only, systemic therapy, or palliative care (55) if the predicted risk of developing ARE is high. The model could also predict if the patient would be at a low risk of developing ARE, in which case SRT could be preferred over other treatment options.

In the present study, only one sequence of the MRI scan was used. Previous studies showed that a combination of radiomics computed on T1 and T2 sequences performs best to differentiate ARE and TP (51, 52), and ADC sequence seems to also show a higher performance (53). Investigating more sequences in a future study may therefore improve the performance of the imaging-based models.

Lastly, for ARE (and, to a lesser degree, TP), treatment is one of the primary factors. In this study, multiple-dose-treatment- related variables have been included, such as prior treatments to the same patients as well as dose variables and the volumes encompassing certain dose levels. However, a more thorough “dosiomics” analysis would probably improve the prediction of ARE. Liang et al. (56) described a method to extract the spatial and texture radiomics features from dose maps (56). They found several radiomics features which have a significant predictive value of radiation pneumonitis. Using a similar method for ARE in BM may result in improved prediction results. Our predictions could also be combined with models automatically classifying tumours and RN on brain MRI, such as in Zhang et al. (51), potentially strengthening the results of those studies.

5. Conclusion

Radiomics is able to predict lesions at a high risk of ARE, especially when combined with DL features. When predicting ARE on a patient level, the highest performance was found using a combination of radiomics, DL, clinical, and treatment-related features. These models could potentially be used to aid clinical decision making for patients with BM treated with either gamma knife or EBRT.

Data availability statement

The corresponding author does not own the datasets used (acquired with DTAs). Requests to access the datasets should be directed to olivier.morin@ucsf.edu (for the data from UCSF); Nicolaus.Andratschke@usz.ch (for the data from USZ).

Ethics statement

The studies involving human participants were reviewed and approved by the cantonal ethics committee Zurich and University of California San Francisco (UCSF) Institutional Review Board (IRB). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

MB and SK performed all the ML/DL analysis and wrote the manuscript. SK, MV, SB, and OM collected and curated the imaging and patient data from UCSF. SP helped with the ML/DL analysis and study design. HW supervised the progression of the project and the writing of this article and guaranteed the integrity of the analysis and results presented. AC and MV helped with the ML analysis. JT, JK, and NA collected the imaging and patient data from USZ. LH and SB aided with the clinical aspects of the study. PL and OM devised the project's aim and supervised the progression of the project. All authors contributed to the article and approved the submitted version.

Funding

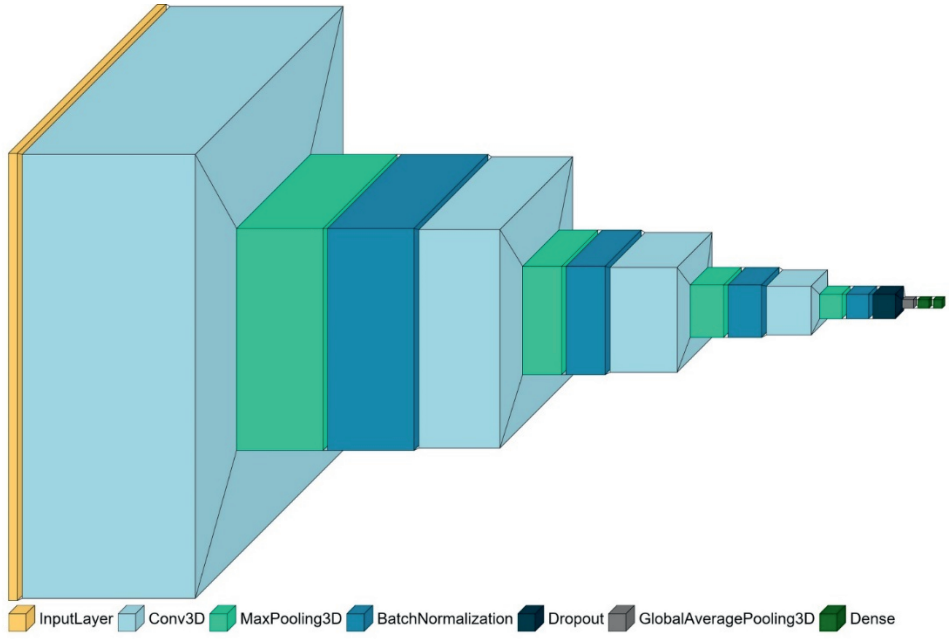
The research project has been partially funded by the Clinical Research Priority Program "Artificial Intelligence in Oncological Imaging" of the University of Zurich. PL, HW, MB, SK acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), the European Union's Horizon 2020 research and innovation programme under grant agreement: MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103 and IMI- OPTIMA n° 101034347.

Conflict of Interest

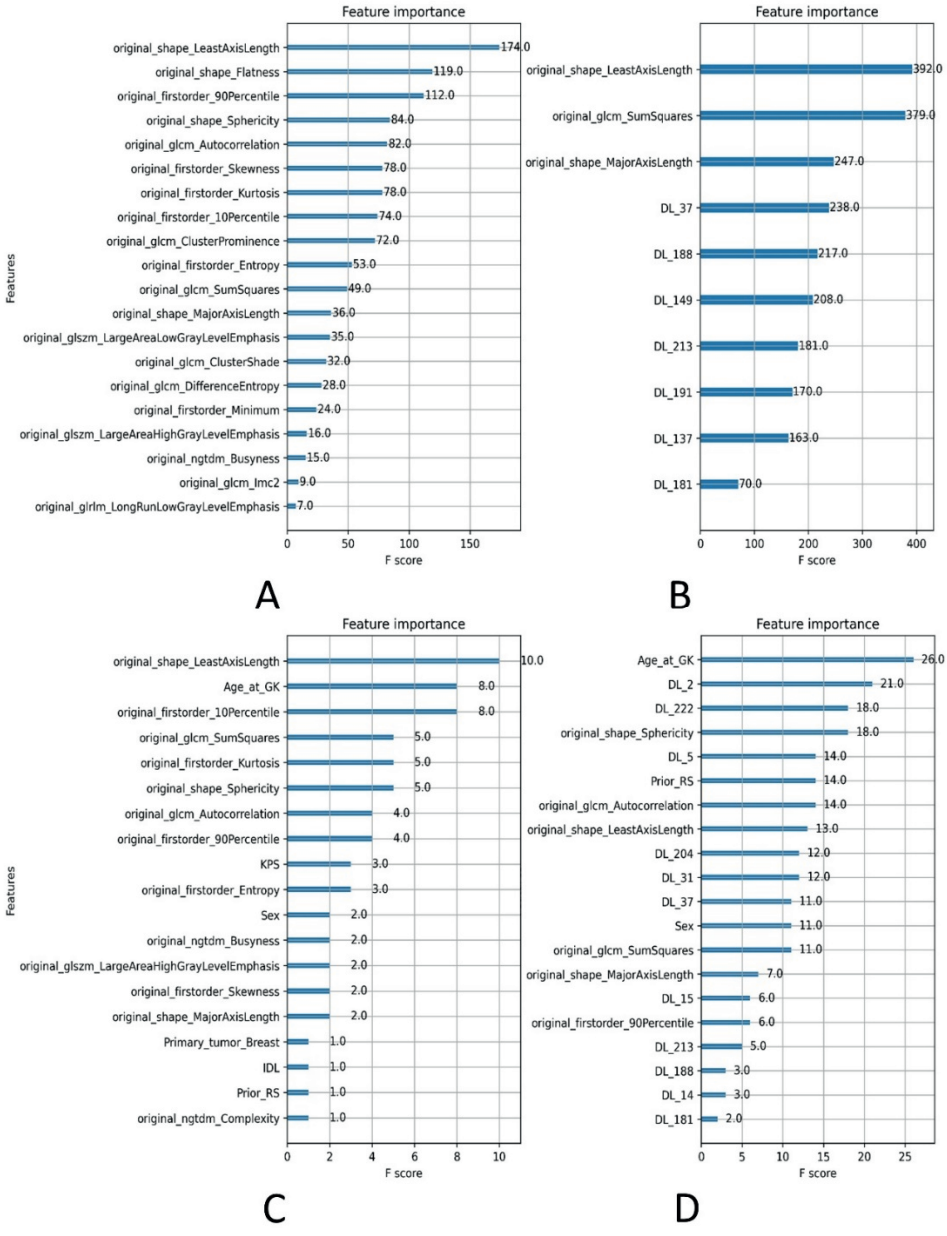
LH: none related to the current manuscript, outside of current manuscript: research funding Roche Genentech, Boehringer Ingelheim, AstraZeneca, Takeda (all institution, Beigene under negotiation); advisory board: BMS, Eli Lilly, Roche Genentech, Pfizer, Takeda, MSD, Merck, Novartis, Boehringer Ingelheim, Amgen, Janssen (all institution, Roche one time self); speaker: MSD, Lilly (institution); travel/conference reimbursement: Roche Genentech (self); mentorship program with key opinion leaders: funded by AstraZeneca; fees for educational webinars: Benecke, Medtalks, VJOncology (self), high5oncology (institution); interview sessions funded by Roche Genentech, Bayer, Lilly (institution); local PI of clinical trials: AstraZeneca, Novartis, BMS, MSD, Merck, GSK, Takeda, Blueprint Medicines, Roche Genentech, Janssen Pharmaceuticals, Mirati; PL: none related to the current manuscript; outside of current manuscript: grants/sponsored research agreements from Radiomics SA, Convert Pharmaceuticals and LivingMed Biotech. He received a presenter fee (in cash or in kind) and/or reimbursement of travel costs/consultancy fee (in cash or in kind) from Radiomics SA, BHV, Varian, Elekta, ptTheragnostic/DNAmito, BMS, and Convert pharma. PL has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/ 050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/ 059089), licensed to ptTheragnostic/DNAmito; one non-issued patent on LSRT (PCT/ P126537PC00), licensed to Varian; three non-patented inventions (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and two non-issued, non-licensed patents on Deep Learning-Radiomics (N2024482, N2024889). He confirms that none of the above entities or funding sources were involved in the preparation of this paper.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

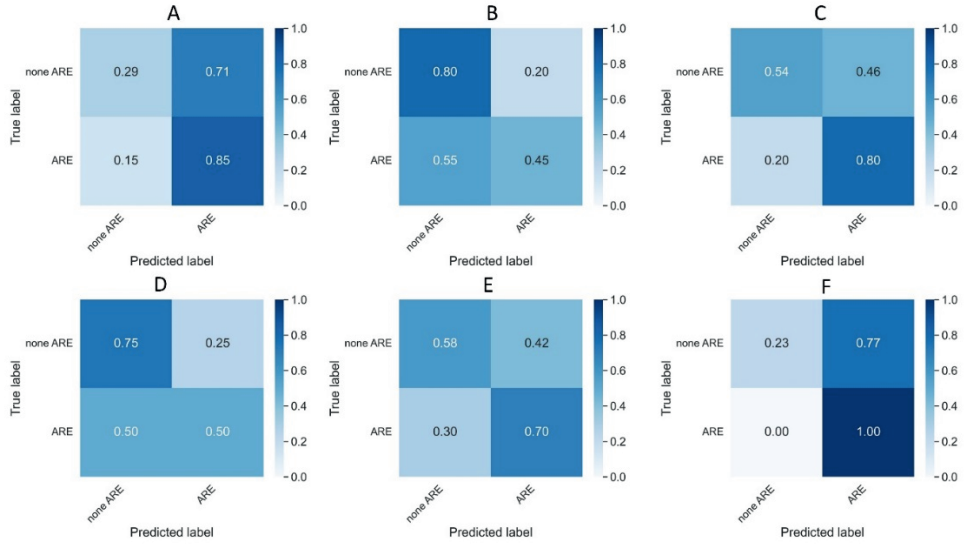
Supplementary Figures



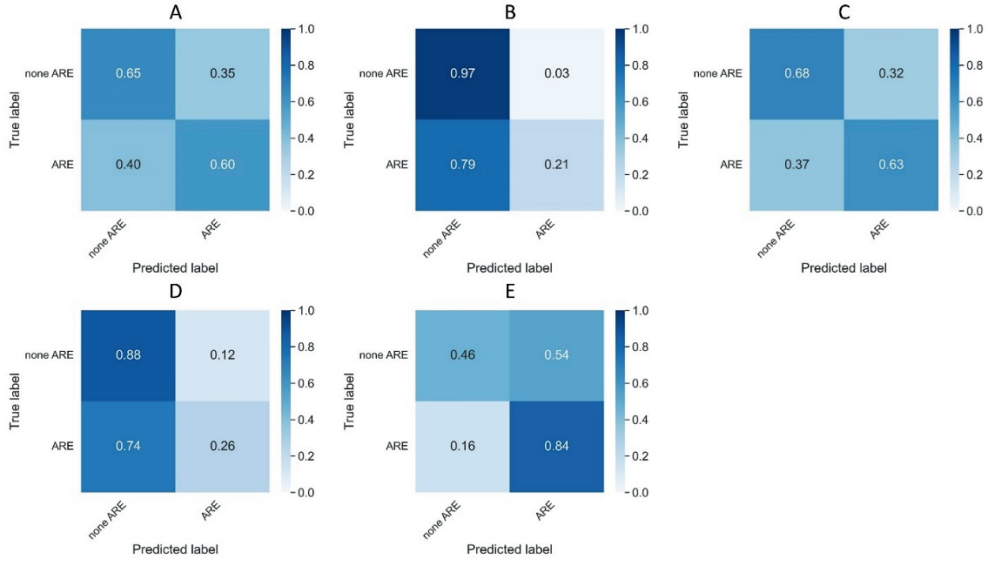
Supplementary Figure 1: architecture of Xception 3D



Supplementary Figure 2: Feature importance lists of the ML models, respectively: (A) radiomics, (B) radiomics and deep learning, (C) radiomics and patient characteristics, and (D) radiomics, patient characteristics, deep learning



Supplementary Figure 3: Normalized confusion matrices on the external validation dataset per target for the following approaches: (A) deep learning, (B) radiomics, (C) radiomics and deep learning, (D) patient characteristics and radiomics, (E) radiomics, deep learning and patient characteristics features, (F) agreed labels.



Supplementary Figure 4: Normalized confusion matrices on the external validation dataset per patient for the following approaches: (A) DL, (B) radiomics, (C) radiomics and DL, (D) patient characteristics and radiomics, (E) radiomics, DL and patient characteristics features.

Supplementary Tables

Supplementary Table 1. Python packages used and their versions.

purpose	packages	versions
pre-processing	imutils	0.5.4
	intensity-normalization	2.0.2
	numpy	1.19.2
	opencv	4.1.0.25
	os	n/a
	pandas	0.25.0
	pydicom	2.2.2
	scikit-image	0.17.2
	scikit-learn	0.24.2
	scipy	1.5.2
	simpleITK	2.1.1
deep learning	keras	2.3.1
	tensorflow-gpu	2.1.0
feature processing and calculation	precision-medicine-toolbox	0.0.0
	missingpy	0.2.0
	pyradiomics	3.0.1
machine learning	xgboost	1.5.1
statistics	statsmodels	0.13.0
visualisation	matplotlib	3.3.4

Supplementary Table 2. Overview of hyperparameters optimized through gridsearch cross-validation.

parameters/data	radiomics only	radiomics + patient characteristics	radiomics + deep learning	radiomics + patient characteristics + deep learning
gamma	0,3	0,3	0,3	0,3
learning rate	0,01	0,1	0,01	0,1
max depth	3	3	4	1
min child weight	1	1	1	5
n estimators	173	10	173	227
number of features selected	20	10	20	20

Supplementary Table 3. AUC, balanced accuracy, precision, recall, and F1 metrics with CI on the training on patient and lesion levels.

Per-lesion classification						Per-patient classification					
Approaches	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**	**Approaches**	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**
DL	0.70 [0.66, 0.75]	0.67 [0.63, 0.71]	0.06 [0.05, 0.08]	0.0.56 [0.48, 0.64]	0.11 [0.09, 0.14]	DL	0.58 [0.53, 0.64]	0.58 [0.54, 0.62]	0.11 [0.09, 0.13]	0.73 [0.65, 0.81]	0.19 [0.16, 0.23]
Rad	0.89 [0.87, 0.91]	0.81 [0.78, 0.84]	0.09 [0.08, 0.11]	0.86 [0.80, 0.01]	0.17 [0.14, 0.19]	Rad	0.76 [0.72, 0.80]	0.71 [0.67, 0.76]	0.22 [0.18, 0.26]	0.65 [0.55, 0.74]	0.33 [0.27, 0.38]
Rad + DL	0.92 [0.91, 0.93]	0.85 [0.83, 0.87]	0.10 [0.09, 0.12]	0.0.93 [0.88, 0.96]	0.18 [0.16, 0.21]	Rad + DL	0.81 [0.78, 0.84]	0.75 [0.71, 0.78]	0.19 [0.15, 0.22]	0.84 [0.77, 0.91]	0.31 [0.26, 0.35]
Rad + Clin	0.88 [0.86, 0.90]	0.81 [0.78, 0.84]	0.09 [0.08, 0.10]	0.86 [0.80, 0.91]	0.16 [0.14, 0.19]	Rad + Clin	0.78 [0.73, 0.82]	0.70 [0.66, 0.74]	0.18 [0.14, 0.21]	0.73 [0.64, 0.81]	0.0.29 [0.24, 0.33]
Rad + DL + Clin	0.88 [0.86, 0.90]	0.82 [0.79, 0.85]	0.10 [0.08, 0.11]	0.85 [0.79, 0.90]	0.17 [0.15, 0.20]	Rad + DL + Clin	0.77 [0.73, 0.81]	0.70 [0.66, 0.73]	0.15 [0.12, 0.18]	0.88 [0.82, 0.94]	0.25 [0.21, 0.29]
Agreed labels	0.88 [0.85, 0.90]	0.82 [0.77, 0.85]	0.09 [0.07, 0.11]	0.81 [0.73, 0.88]	0.16 [0.13, 0.19]	Agreed labels	0.74 [0.69, 0.78]	0.60 [0.58, 0.62]	0.13 [0.11, 0.16]	0.97 [0.93, 1.00]	0.23 [0.19, 0.27]

Supplementary Table 4. AUC, balanced accuracy, precision, recall, and F1 metrics with CI on the internal validation on patient and lesion levels.

Per-lesion classification						Per-patient classification					
Approaches	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**	**Approaches**	**AUC**	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**
DL	0.72 [0.66, 0.78]	0.61 [0.55, 0.67]	0.07 [0.04, 0.09]	0.37 [0.26, 0.49]	0.11 [0.07, 0.16]	DL	0.63 [0.55, 0.71]	0.59 [0.52, 0.66]	0.12 [0.09, 0.17]	0.63 [0.50, 0.77]	0.21 [0.15, 0.27]
Rad	0.76 [0.69, 0.81]	0.70 [0.64, 0.76]	0.07 [0.05, 0.09]	0.67 [0.55, 0.78]	0.13 [0.10, 0.16]	Rad	0.76 [0,70, 0.81]	0.70 [0.65, 0.76]	0.07 [0.05, 0.09]	0.67 [0.56, 0.78]	0.13 [0.10, 0.16]
Rad + DL	0.71 [0.66, 0.76]	0.64 [0.58, 0.70]	0.0.06 [0.04, 0.08]	0.0.53 [0.41, 0.64]	0.0.11 [0.08, 0.14]	Rad + DL	0.55 [0.47, 0.63]	0.51 [0.44, 0.58]	0.10 [0.06, 0.14]	0.84 [0.77, 0.91]	0.31 [0.26, 0.35]
Rad + Clin	0.77 [0.71, 0.82]	0.71 [0.65, 0.76]	0.07 [0.05, 0.09]	0.69 [0.58, 0.79]	0.13 [0.10, 0.16]	Rad + Clin	0.64 [0.55, 0.72]	0.60 [0.52, 0.67]	0.13 [0.09, 0.18]	0.55 [0.41, 0.69]	022. [0.14, 0.28]
Rad + DL + Clin	0.71 [0.65, 0.76]	0.63 [0.57, 0.69]	0.06 [0.04, 0.08]	0.53 [0.41, 0.64]	0.10 [0.07, 0.13]	Rad + DL + Clin	0.59 [0.51, 0.67]	0.65 [0.49, 0.63]	0.11 [0.07, 0.15]	0.65 [0.51, 0.78]	0.19 [0.13, 0.25]
Agreed labels	0.81 [0.73, 0.89]	0.73 [0.64, 0.81]	0.10 [0.06, 0.15]	0.55 [0.38, 0.71]	0.17 [0.11, 0.24]	Agreed labels	0.71 [0.62, 0.79]	0.61 [0.59, 0.63]	0.11 [0.08, 0.15]	1.00 [1.00, 1.00]	0.20 [0.14, 0.26]

Supplementary Material: Materials and methods

1. Pre-processing workflow

Conversion of the data from DICOM to NRRD was done using the "precision-medicine-toolbox" (1) to extract the three-dimensional images. For the first workflow, termed "minimalist", z-score normalization was applied per scan on the white matter only using the "intensity_normalization" package in python 3.7 (2). Z-score normalization refers to the process of normalizing an image by subtracting the mean intensity value from each pixel, and dividing each pixel by the standard deviation (SD) of the intensity histogram. The second pre-processing workflow, termed "standardization", had two steps: z-score normalization as described previously and three-dimensional contrast limited adaptive histogram equalization (CLAHE) applied on the brain using the python package "intensity_normalization", after having rescaled the image intensities to 256 bins. CLAHE applies histogram equalization in small patches of the images to increase image contrast, after which through bilinear interpolation any artificial borders between images are removed (3). The third approach, termed "harmonization", was designed to make the intensities comparable across scans for similar regions of the brain. For this the white stripe

normalization which is described in (4), was applied using the python package "intensity_normalization". It has the advantage of harmonizing the images based solely on data contained within the MRI, compared to piecewise linear histogram matching, which requires additional information about the dataset (5). An overview of the applied pre-processing methods can be seen in figure 2.

2. Description of the radiomics features

First order and histogram statistics describe the total distribution of voxel intensities over the MR image. Shape and size features describe the three-dimensional spatial dimensions of the tumor. Texture features describe the relative spatial distribution of intensity values within the tumor derived from 6 different matrices that are defined over the region of interest (ROI): gray-level co-occurrence matrix (GLCM) (6), gray-level run length matrix (GLRLM) (7), gray-level size-zone matrix (GLSZM) (8), gray-level distance-zone matrix (GLDZM) (9), neighborhood gray-level dependence matrix (NGLDM) (10), and neighborhood gray-tone difference matrix (NGTDM) (11).

3. Specifications of the XGBoost model classifying ARE versus no ARE

Gradient boosting creates a classification model built on ensemble decision trees. These decision trees make simple, weak predictions on an outcome. The XGBoost model sums up all the individual decision tree predictions to make a final overall prediction which is a measure ranging from 0 to 1 indicating the estimated probability that a lesion develops ARE. By additively adding new trees, and calculating what the gain of an added tree is by calculating a loss function of the overall model performance, trees are either selected or pruned.

The XGBoost model contains a number of hyperparameters that regulate training. To prevent the imbalance in outcome from affecting model training, the ratio of true events to controls was used as the weights for positive and negative classes of the model. For feature selection, the following default hyperparameter values were used to define an initial list of predictive features: the maximum depth of a single decision tree (6), the minimum sum of instance weight a node in a decision tree needs to be added (1), the number of decision trees to build the final model (100), the gamma or the minimum loss reduction needed to add a tree (0), and lastly the learning rate (0.3). To find the optimal values for feature selection a grid search with cross-validation (k=10) was performed using these initial features. Grid search is a tuning technique which trains different models, based on all the possible combinations of hyperparameters to test, to determine an optimal set of parameters. It finds this best estimator based on the combination of parameters that produces the highest score, optimizing the area under the precision recall-curve (AUCPR). A 10-fold cross validation grid search was performed for the following XGBoost parameters with corresponding value ranges: the maximum depth of a single decision tree (1-5), the minimum sum of instance weight a node in a decision tree needs to be added (1-6), the

number of decision trees to build the final model (10-500 in steps of 10), the gamma or the minimum loss reduction needed to add a tree (0.3-0.5), and lastly the learning rate (10-1, 10-2, and 10-3). Default parameters of the XGBoost model further included the learning task (logistic regression), the weights of positive and negative classes (set to proportion of ARE to non ARE = 0.03), the subsample ratio of columns when constructing each tree (0.7), and the evaluation metric (area under the PR curve). A total of 4800 (10 folds for 480 candidates) folds were fitted for each feature set. The optimized hyperparameters were then used to perform feature selection again, resulting in the optimized list of selected features. These features were subsequently used to perform another grid search with cross-validation, resulting in a final optimal XGBoost model.

Supplementary References

1. Primakov S, Lavrova E, Salahuddin Z, Woodruff HC, Lambin P. Precision-medicine-toolbox: An open-source python package for facilitation of quantitative medical imaging and radiomics analysis. arXiv [eessIV] (2022) http://arxiv.org/abs/2202.13965
2. Reinhold JC, Dewey BE, Carass A, Prince JL. Evaluating the Impact of Intensity Normalization on MR Image Synthesis. (2018) http://arxiv.org/abs/1812.04652
3. Zuiderveld K. Contrast Limited Adaptive Histogram Equalization. Graphics Gems (1994)474–485. doi: 10.1016/b978-0-12-336156-1.50061-6
4. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical (2014) 6:9–19. doi: 10.1016/j.nicl.2014.08.008
5. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging (2000) 19:143–150. doi: 10.1109/42.836373
6. Haralick RM, Shanmugam K, Dinstein I 'hak. Textural Features for Image Classification. IEEE Trans Syst Man Cybern (1973) SMC-3:610–621. doi: 10.1109/TSMC.1973.4309314
7. Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing (1975) 4:172–179. doi: 10.1016/S0146-664X(75)80008-6
8. Thibault, Fertil, Navarro, Pereira, Cau. Texture indexes and gray level size zone matrix. Nuclei Classification PRIP
9. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification. 2011 18th IEEE International Conference on Image Processing (2011) doi: 10.1109/icip.2011.6116401
10. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. Computer Graphics and Image Processing (1982) 20:297. doi: 10.1016/0146-664x(82)90093-4
11. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern (1989) 19:1264–1274. doi: 10.1109/21.44046

References

1. Walker AE, Robins M, Weinfeld FD. Epidemiology of Brain Tumors: The National Survey of Intracranial Neoplasms. Neurology (1985) 35:219–9. doi: 10.1212/wnl.35.2.219

2. Johnson JD, Young B. Demographics of Brain Metastasis. Neurosurg Clinics North America (1996) 7:337–44. doi: 10.1016/s1042-3680(18)30365-6

3. Wen PY, Loeffler JS. Management of Brain Metastases. Oncology (1999) 13 (7):941–54, 957–61.

4. Schouten LJ, Rutten J, Huveneers HAM, Twijnstra A. Incidence of Brain Metastases in a Cohort of Patients With Carcinoma of the Breast, Colon, Kidney, and Lung and Melanoma. Cancer (2002) 94(10):2698–705. doi: 10.1002/cncr.10541

5. Barnholtz-Sloan JS, Sloan AE, Davis FG, Vigneau FD, Lai P, Sawaya RE. Incidence Proportions of Brain Metastases in Patients Diagnosed, (1993 to 2001) in the Metropolitan Detroit Cancer Surveillance System. J Clin Oncol: Off J Am Soc Clin Oncol (2004) 22(14):2865–72. doi: 10.1200/JCO.2004.12.149

6. Rangachari D, Yamaguchi N, VanderLaan PA, Folch E, Mahadevan A, Floyd SR, et al. Brain Metastases in Patients With EGFR -Mutated or ALK -Rearranged non-Small-Cell Lung Cancers. Lung Cancer (2015) 88:108–11. doi: 10.1016/j.lungcan.2015.01.020

7. Huber RM, Hansen KH, Paz-Ares RL, West HL, Reckamp KL, Leighl NB, et al. Brigatinib in Crizotinib-Refractory ALK+ NSCLC: 2-Year Follow-Up on Systemic and Intracranial Outcomes in the Phase 2 ALTA Trial. J Thorac Oncol: Off Publ Int Assoc Study Lung Cancer (2020) 15(3). doi: 10.1016/ j.jtho.2019.11.004

8. Venur VA, Karivedu V, Ahluwalia MSSystemic Therapy for Brain Metastases. In: Handbook of Clinical Neurology. Elsevier. p. 137–53.

9. Vogelbaum MA, Brown PD, Messersmith H, Brastianos PK, Burri S, Cahill D, et al. Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline. J Clin Oncol: Off J Am Soc Clin Oncol (2022) 40(5):492–516. doi: 10.1200/JCO.21.02314

10. McTyre E, Scott J, Chinnaiyan P. Whole Brain Radiotherapy for Brain Metastasis. Surg Neurol Int (2013) 4(Suppl 4):S236–44. doi: 10.4103/2152- 7806.111301

11. Kraft J, Zindler J, Minniti G, Guckenberger M, Andratschke N. Stereotactic Radiosurgery for Multiple Brain Metastases. Curr Treat Options Neurol (2019) 21(2):6. doi: 10.1007/s11940-019-0548-3

12. Kraft J, Mayinger M, Willmann J, Brown M, Tanadini-Lang S, Wilke L, et al. Management of Multiple Brain Metastases: A Patterns of Care Survey Within the German Society for Radiation Oncology. J Neuro Oncol (2021) 152 (2):395–404. doi: 10.1007/s11060-021-03714-w

13. Badiyan SN, Regine WF, Mehta M. Stereotactic Radiosurgery for Treatment of Brain Metastases. J Oncol Pract / Am Soc Clin Oncol (2016) 12(8):703–12. doi: 10.1200/JOP.2016.012922

14. Walker AJ, Ruzevick J, Malayeri AA, Rigamonti D, Lim M, Redmond KJ, et al. Postradiation Imaging Changes in the CNS: How can We Differentiate Between Treatment Effect and Disease Progression? Future Oncol (2014) 10 (7):1277–97. doi: 10.2217/fon.13.271

15. Sneed PK, Mendez J, Vemer-van den Hoek JGM, Seymour ZA, Ma L, Molinaro AM, et al. Adverse Radiation Effect After Stereotactic Radiosurgery for Brain Metastases: Incidence, Time Course, and Risk Factors. J Neurosurg (2015) 123(2):373–86. doi: 10.3171/2014.10.JNS141610
16. Gerosa M, Nicolato A, Foroni R, Zanotti B, Tomazzoli L, Miscusi M, et al. Gamma Knife Radiosurgery for Brain Metastases: A Primary Therapeutic Option. J Neurosurg (2002) 97:515–24. doi: 10.3171/jns.2002.97.supplement_5.0515
17. Lawrence YR, Allen Li X, el Naqa I, Hahn CA, Marks LB, Merchant TE, et al. Radiation Dose–Volume Effects in the Brain. Int J Radiat Oncol Biol Phys (2010) 76:S20–7. doi: 10.1016/j.ijrobp.2009.02.091
18. Minniti G, D'Angelillo RM, Scaringi C, Trodella LE, Clarke E, Matteucci P, et al. Fractionated Stereotactic Radiosurgery for Patients With Brain Metastases. J Neuro Oncol (2014) 117(2):295–301. doi: 10.1007/s11060-014- 1388-3
19. Vellayappan B, Tan CL, Yong C, Khor LK, Koh WY, Yeo TT, et al. Diagnosis and Management of Radiation Necrosis in Patients With Brain Metastases. Front Oncol (2018) 8:395. doi: 10.3389/fonc.2018.00395
20. Petrovich Z, Yu C, Giannotta SL, Oâ€™Day S, Apuzzo MLJ. Survival and Pattern of Failure in Brain Metastasis Treated With Stereotactic Gamma Knife Radiosurgery. J Neurosurg (2002) 97(5 Suppl):499–506. doi: 10.3171/ jns.2002.97.supplement_5.0499
21. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting More Information From Medical Images Using Advanced Feature Analysis. Eur J Cancer (2012) 48(4):441–6. doi: 10.1016/j.ejca.2011.11.036
22. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. Nat Commun 5 (2014) p:4006. doi: 10.1038/ncomms5006
23. Zhou M, Scott J, Chaudhury B, Hall L, Goldgof D, Yeom KW, et al. Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches. Am J Neuroradiol (2018) 39:208–16. doi: 10.3174/ajnr.a5391
24. Morin O, Chen WC, Nassiri F, Susko M, Magill ST, Vasudevan HN, et al. Integrated Models Incorporating Radiologic and Radiomic Features Predict Meningioma Grade, Local Failure, and Overall Survival. Neuro Oncol Adv (2019) 1:vdz011. doi: 10.1093/noajnl/vdz011
25. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and Deep Learning Methods for Radiomics. Med Physics (2020) 47:185–202. doi: 10.1002/mp.13678
26. Rogers W, Thulasi Seetha S, Refaee TAG, Lieverse RIY, Granzier RWY, Ibrahim A, et al. Radiomics: From Qualitative to Quantitative Imaging. Br J Radiol (2020) 93(1108):20190948. doi: 10.1259/bjr.20190948
27. Abidin AZ, Dar I, D'Souza AM, Lin EP, Wismüller A. Investigating a Quantitative Radiomics Approach for Brain Tumor Classification. In. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging. Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE (2019). p. 36–45.

28. Dong F, Li Q, Jiang B, Zhu X, Zeng Q, Huang P, et al. Differentiation of Supratentorial Single Brain Metastasis and Glioblastoma by Using Peri- Enhancing Oedema Region-Derived Radiomic Features and Multiple Classifiers. Eur Radiol (2020) 30:3015–22. doi: 10.1007/s00330-019- 06460-w
29. Huang C-Y, Lee C-C, Yang H-C, Lin C-J, Wu H-M, Chung W-Y, et al. Radiomics as Prognostic Factor in Brain Metastases Treated With Gamma Knife Radiosurgery. J Neuro Oncol (2020) 146:439–49. doi: 10.1007/s11060- 019-03343-4
30. Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I, et al. Use of Radiomics for the Prediction of Local Control of Brain Metastases After Stereotactic Radiosurgery. Neuro-oncology (2020) 22:797–805. doi: 10.1093/ neuonc/noaa007
31. Ortiz-Ramón R, Larroza A, Ruiz-España S, Arana E, Moratal D. Classifying Brain Metastases by Their Primary Site of Origin Using a Radiomics Approach Based on Texture Analysis: A Feasibility Study. Eur Radiol (2018) 28(11):4514–23. doi: 10.1007/s00330-018-5463-6
32. Kniep HC, Madesta F, Schneider T, Hanning U, Schönfeld MH, Schön G, et al. Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type. Radiology (2019) 28:4514–23. doi: 10.1148/radiol.2018180946
33. Bhatia A, Birger M, Veeraraghavan H, Um H, Tixier F, McKenney AS, et al. MRI Radiomic Features are Associated With Survival in Melanoma Brain Metastases Treated With Immune Checkpoint Inhibitors. Neuro-oncology (2019) 21(12):1578–86. doi: 10.1093/neuonc/noz141
34. Della Seta M, Collettini F, Chapiro J, Angelidis A, Engeling F, Hamm B, et al. A 3D Quantitative Imaging Biomarker in Pre-Treatment MRI Predicts Overall Survival After Stereotactic Radiation Therapy of Patients With a Singular Brain Metastasis. Acta Radiol (2019) 60(11):1496–503. doi: 10.1177/ 0284185119831692
35. Cho J, Kim YJ, Sunwoo L, Lee GP, Nguyen TQ, Cho SJ , et al. Deep Learning- Based Computer-Aided Detection System for Automated Treatment Response Assessment of Brain Metastases on 3D MRI. Front Oncol (2021) 11:739639. doi: 10.3389/fonc.2021.739639
36. Parekh VS, Jacobs MA. Deep Learning and Radiomics in Precision Medicine. Expert Rev Precis Med Drug Dev (2019) 4(2):59–72. doi: 10.1080/ 23808993.2019.1585805
37. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging (2010) 29 (6):1310–20. doi: 10.1109/TMI.2010.2046908
38. Juntu J, Sijbers J, Dyck D, Gielen J. Bias Field Correction for MRI Images. Adv Soft Computing (2005), 543–51. doi: 10.1007/3-540-32390-2_64
39. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H. Impact of Image Preprocessing on the Scanner Dependence of Multi- Parametric MRI Radiomic Features and Covariate Shift in Multi- Institutional Glioblastoma Datasets. Phys Med Biol (2019) 64(16):165011. doi: 10.1088/1361-6560/ab2f44
40. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of Image Preprocessing Methods on Reproducibility of Radiomic Features in Multimodal Magnetic Resonance Imaging in

Glioblastoma. J Appl Clin Med Phys / Am Coll Med Phys (2020) 21(1):179–90. doi: 10.1002/acm2.12795

41. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, et al. Quick Guide on Radiology Image Pre-Processing for Deep Learning Applications in Prostate Cancer Research. J Med Imaging (Bellingham Wash) (2021) 8(1):010901. doi: 10.1117/1.JMI.8.1.010901

42. Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik J-C, et al. Gray-Level Discretization Impacts Reproducible MRI Radiomics Texture Features. PLoS One (2019) 14(3):e0213459. doi: 10.1371/journal. pone.0213459

43. CarréA, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of Brain MR Images Across Machines and Protocols: Bridging the Gap for MRI-Based Radiomics. Sci Rep (2020) 10(1):12340. doi: 10.1038/s41598-020-69298-z

44. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. Radiology (2020) 295:328–38. doi: 10.1148/radiol.2020191145

45. Radiomic Features — Pyradiomics V3.0.1.Post9+Gdfe2c14 Documentation (2019). Available at: https://pyradiomics.readthedocs.io/en/latest/features. html (Accessed 21 October 2021).

46. Chatterjee A, Vallieres M, Dohan A, Levesque IR, Ueno Y, Saif S, et al.) Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. IEEE Trans Radiat Plasma Med Sci (2019) 3:210–5. doi: 10.1109/trpms.2019.2893860

47. Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions. Proc IEEE Conf Comput Vision Pattern Recognition (2017), 1251–8. doi: 10.1109/CVPR.2017.195

48. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG] (2014). Available at: http://arxiv.org/abs/1412.6980.

49. Youden WJ. Index for Rating Diagnostic Tests. Cancer (1950) 3(1):32–5. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

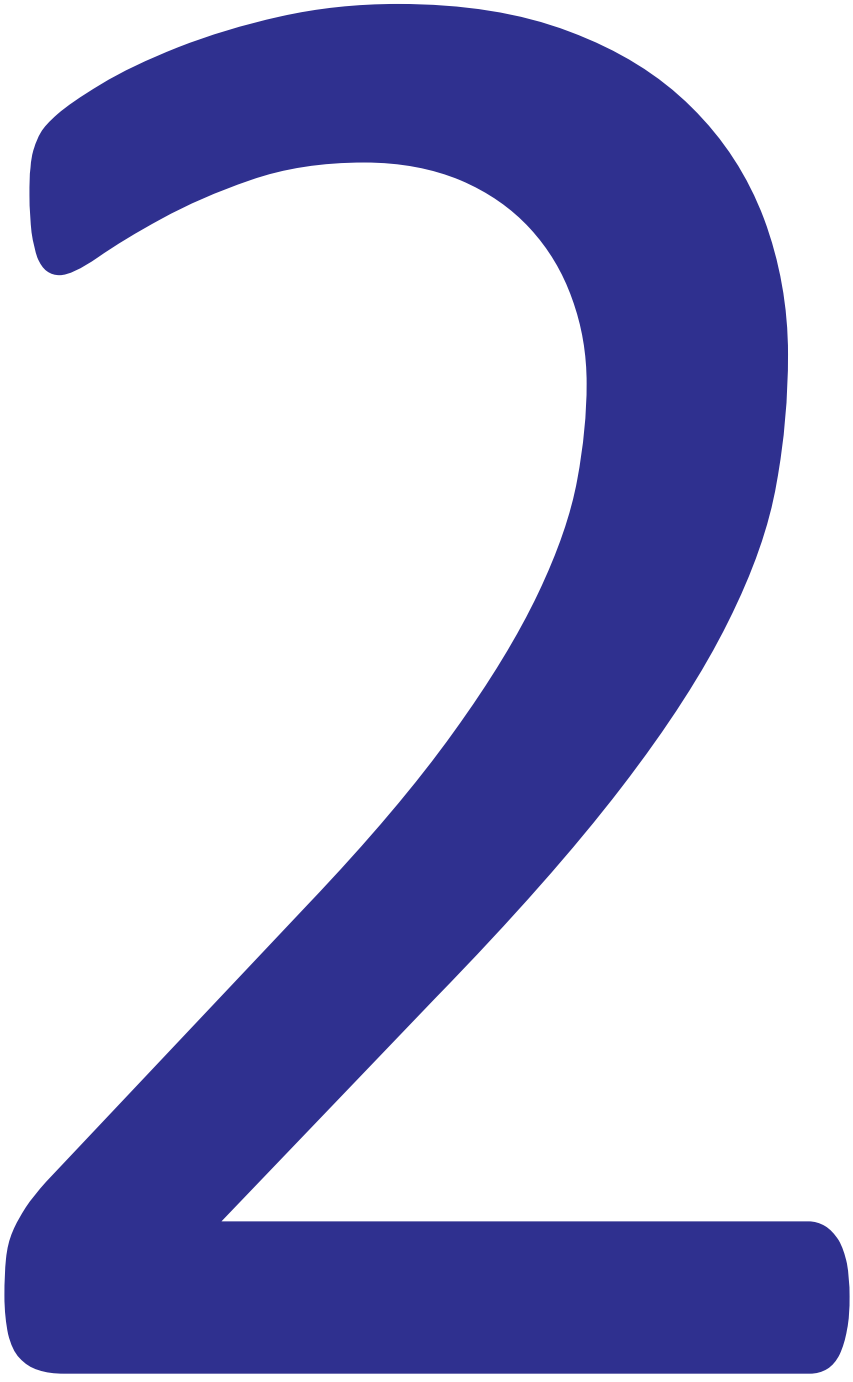
50. Alvarez-Breckenridge C, Remon J, Piña Y, Nieblas-Bedolla E, Forsyth P, Hendriks L, et al. Emerging Systemic Treatment Perspectives on Brain Metastases: Moving Toward a Better Outlook for Patients. Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meeting (2022) 42:1–19. doi: 10.1200/EDBK_352320

51. Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A Predictive Model for Distinguishing Radiation Necrosis From Tumour Progression After Gamma Knife Radiosurgery Based on Radiomic Features From MR Images. Eur Radiol (2018) 28:2255–63. doi: 10.1007/s00330-017-5154-8

52. Peng L, Parekh V, Huang P, Lin DD, Sheikh K, Baker B, et al. Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics. Int J Radiat Oncol Biol Phys (2018) 102:1236–43. doi: 10.1016/j.ijrobp.2018.05.041

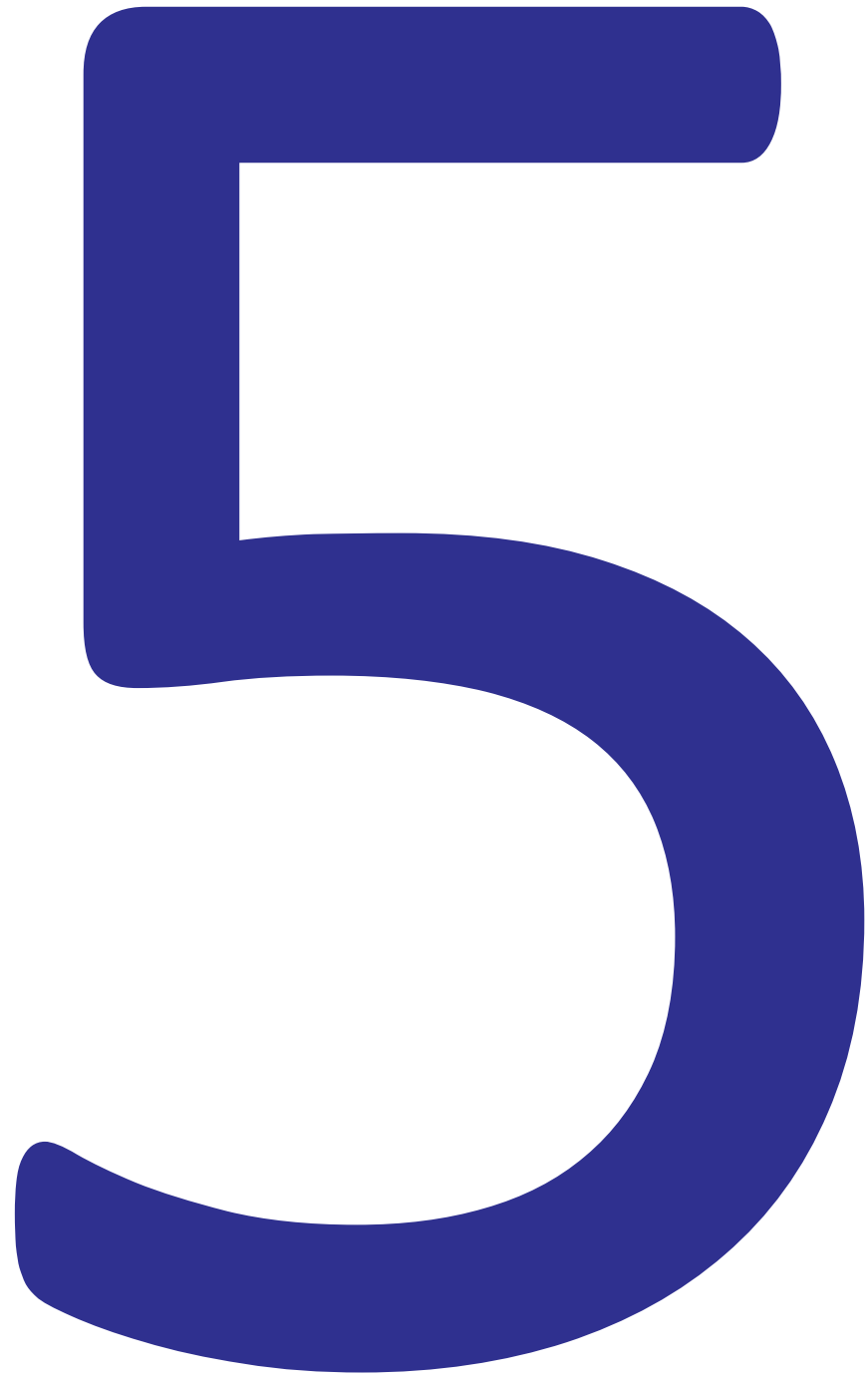
53. Park YW, Choi D, Park JE, Ahn SS, Kim H, Chang JH, et al. Differentiation of Recurrent Glioblastoma From Radiation Necrosis Using Diffusion Radiomics With Machine Learning

Model Development and External Validation. Sci Rep (2021) 11:2913. doi: 10.1038/s41598-021-82467-y
54. Salvestrini V, Greco C, Guerini AE, Longo S, Nardone V, Boldrini L, et al. The Role of Feature-Based Radiomics for Predicting Response and Radiation Injury After Stereotactic Radiation Therapy for Brain Metastases: A Critical Review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO). Trans Oncol (2022) 15:101275. doi: 10.1016/ j.tranon.2021.101275
55. Lin X, DeAngelis LM. Treatment of Brain Metastases. J Clin Oncol: Off J Am Soc Clin Oncol (2015) 33(30):3475–84. doi: 10.1200/JCO.2015.60.9503
56. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: Extracting 3d Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. Front Oncol (2019) 9:269. doi: 10.3389/fonc.2019.00269



PART 2

Using feature-based models to augment deep learning prediction



Chapter 5

Automated detection and delineation of lymph nodes in haematoxylin & eosin stained digitised slides

Manon Beuque, Derek R. Magee, Avishek Chatterjee, Henry C. Woodruff, Ruth E. Langley, William Allum, Matthew G. Nankivell, David Cunningham, Philippe Lambin, Heike I. Grabsch

Adapted from:
Manon Beuque, Derek R. Magee, Avishek Chatterjee, Henry C. Woodruff, Ruth E. Langley, William Allum, Matthew G. Nankivell, David Cunningham, Philippe Lambin, Heike I. Grabsch.
Automated detection and delineation of lymph nodes in haematoxylin & eosin stained digitised slides. Journal of Pathology Informatics 2023, doi: https://doi.org/10.1016/j.jpi.2023.100192

Abstract

Treatment of patients with oesophageal and gastric cancer (OeGC) is guided by disease stage, patient performance status and preferences. Lymph node (LN) status is one of the strongest prognostic factors for OeGC patients. However, survival varies between patients with the same disease stage and LN status. We recently showed that LN size from patients with OeGC might also have prognostic value, thus making delineations of LNs essential for size estimation and the extraction of other imaging biomarkers.

We hypothesized that a machine learning workflow is able to: (1) find digital H&E stained slides containing LNs, (2) create a scoring system providing degrees of certainty for the results, and (3) delineate LNs in those images. To train and validate the pipeline, we used 1695 H&E slides from the OE02 trial. The dataset was divided into training (80%) and validation (20%). The model was tested on an external dataset of 826 H&E slides from the OE05 trial. U-Net architecture was used to generate prediction maps from which predefined features were extracted. These features were subsequently used to train an XGBoost model to determine if a region truly contained a LN. With our innovative method, the balanced accuracies of the LN detection were 0.93 on the validation dataset (0.83 on the test dataset) compared to 0.81 (0.81) on the vali- dation (test) datasets when using the standard method of thresholding U-Net predictions to arrive at a binary mask. Our method allowed for the creation of an “uncertain” category, and partly limited false-positive predictions on the external dataset. The mean Dice score was 0.73 (0.60) per-image and 0.66 (0.48) per-LN for the validation (test) datasets. Our pipeline detects images with LNs more accurately than conventional methods, and high-throughput delineation of LNs can facilitate future LN content analyses of large datasets.

1. Introduction

Oeosphageal and gastric cancers (OeGC) were diagnosed more than 1.5 million times worldwide in 2020 and represented 13.2% of all cancer deaths (1). The treatment of OeGC patients depends on the disease stage, and patient performance status and preferences (2). For Western patients diagnosed with locally advanced resectable disease, the standard of care is neoadjuvant chemo(radio)therapy followed by surgery for oesophageal cancer and perioperative chemotherapy for gastric cancer according to the ESMO guideline (3).

The overall survival of Western OeGC patients is poor with a 3 year survival rate between 22.3% and 33.8% for gastric cancer and between 19.2% and 27.0% for oesophageal cancer (4).

Lymph node (LN) status (presence or absence of metastasis in regional LNs) is currently the strongest prognostic factors for OeGC patients irrespective of treatment modality, grade of primary tumour regression, or regression in LN (5,6). Our recent pilot study of digital haematoxylin and eosin (H&E) stained slides containing resection specimens from patients with oesophageal cancer from the OE02 trial (7) suggested that not only LN status but also the size of LNs might have prognostic value (8). Validation of these pilot study findings is needed in at least 1 independent large study assessing thousands of LNs before pathological LN size can be considered as a useful biomarker for routine use in OeGC patient management. This and possibly other imaging biomarkers could be useful to identify patients who will benefit most from (potentially) toxic adjuvant treatment.

However, manual review of digital H&E-stained slides to identify and delineate all LNs as previously performed in the pilot study is not feasible within a reasonable time frame in large datasets. Recent phase III trials in OeGC patients typically amount to 20 000 slides and 10 000 LNs per trial, as on average 30 slides are made per resection specimen and more than 15 LNs per patient are obtained. Thus, a toolbox for the automatic identification of image files containing LNs and their automatic delineation would be very desirable for a large-scale validation of our LN size findings and as a prerequisite for further characterisation of the LN architecture by quantitative image analysis. To the best of our knowledge, there are currently no fully automated solutions available for such a task.

We hypothesized that a computational pipeline using a deep learning (DL) model combined with imaging features extracted from the generated prediction map can: (1) identify which H&E-stained digitised slides from oesophagogastric cancer resection specimens contain LNs and (2) automatically delineate the LNs with higher accuracy than current stand-alone DL solutions.

The aim of the study was to develop, validate, and externally test a DL-based workflow to enable large-scale high throughput studies in digital H&E-stained LN tissue sections from resection specimens of oesophagogastric cancer patients.

2. Materials and methods

2.1. Hematoxylin & Eosin-stained digitised tissue section collection

H&E-stained slides were collected retrospectively from resection specimens from OeGC patients recruited into the phase III randomised con- trolled trial, UK MRC OE02 (7). Those samples were collected from 42 European centres. Whole slides were scanned using an Aperio XT Scanner. A total of 1695 scanned H&E slides from 493 resection specimens (on average 3,4 images per specimen) were manually reviewed and classified as containing one or more LNs (N=756 images) or no LN (N=939). All LNs were manually delineated by an expert pathologist using the Aperio ImageScope software (ground-truth delineations) and delineations were saved in an Aperio ImageScope XML annotation file format.

The image dataset was randomly split per patient, with 394 patients (∼80%) in the training dataset and 99 (∼20%) patients in the validation dataset. For the external dataset, 826 H&E slides were extracted from 33 resected specimens (on average 25 images per specimen) from the UK MRC OE05 trial (9). The OE05 dataset had 348 images with delineated LNs and 478 images identified without LNs. The study was approved by the South East Research Ethics Committee, London, UK, REC reference: 07/H1102/111.

2.2. Pre-processing of digital images for deep learning

Common pre-processing strategies for H&E-stained images as described by Li et al (10) were applied to the original images in our database to harmonise the dataset and remove noise: Python 3.7 was used and all packages/libraries used in this study are listed in supplementary material Table 1. As the resolution of slides scanned at 40× magnification can be up to 200 000×200 000 pixels, scanned images were extracted from the Aperio ScanScope files at a maximum size of 2048×2048 pixels, preserving the aspect ratio of the original image. To extract the image at a maximum resolution of 2048×2048 pixels, different downsample levels were tested until reaching the maximum resolution, at which point the downsampled image was extracted. Finally, the extracted images were converted into jpeg image file format to facilitate the use of standard python packages for pre-processing. One scanned image from the dataset was randomly selected to be the reference image for Macenko's colour normalisation strategy, which consists of colour deconvolution later matched to the colour characteristics extracted from the reference image (11). As the DL model required square images as input, scanned images with rectangular shapes (i.e., length > than 1.5 times the width) were split into 2 squares to avoid overstretching or compressing of the image. We also applied the Otsu thresholding method (12), a histogram-based filter able to generate a binary mask that separates the foreground (tissue) from the background (empty space) setting the background values to 255 to maintain a white background.

Subsequently, the scanned and cropped images were resized to 512x512 pixels by downsampling with bicubic interpolation (see Figure 1) to be suitable as input for U-Net.

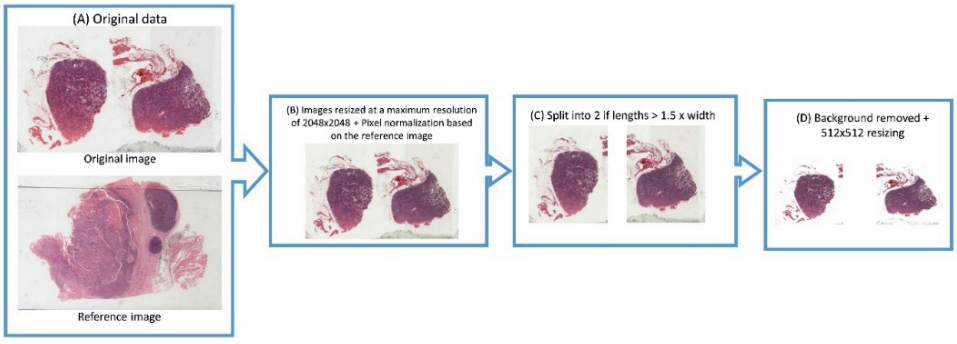


Figure 1: Pre-processing workflow: colour normalisation, resizing, splitting, and removal of background. (A) Digitised glass slide with 2 lymph nodes; bottom: randomly chosen reference image. (B) Image after colour normalisation per pixel and downsizing to 2048×2048. (C) Image split into 2 sub-images in cases where the original image was rectangular. (D) Removal of background and resizing to 512×512 pixels.

We named "pre-processed images" the resulting images. We also derived binary masks from the coordinates of the delineations saved in XML files indicating the pixels belonging to LN tissue. The number of samples used during the study before and after pre-processing can be seen in Table 1.

Table 1: Description of the 3 different datasets: number of patients with scanned H&E slides, number of scanned H&E slides, and number of images after pre-processing of the scanned H&E slide.

Data type/subset	Training dataset from OE02	Validation dataset from OE02	Test dataset from OE05
Number of patients	394	99	33
Number of images	1 340	355	826
Number of images after pre-processing	1 516	481	1 251

2.3. Deep learning model for automatic detection and delineation of lymph nodes

The DL model chosen to detect and delineate LNs was a U-Net (13) using ResNet-50 as backbone due to its proven good performance for histopathology whole slide image delineation (14). The loss function used was a combination of Dice loss weighted at 0.3 and binary cross entropy loss weighted at 0.7, which empirically gave the best Dice score on the validation dataset. We used Adam (adaptive moment estimation), an algorithm which optimises the model with a learning rate of 10^{-4} (15). The model was trained using 4 GPUs

(NVIDIA GeForce RTX 2080Ti) until overfitting was observed based on the surveillance of the mean Dice coefficient in the training and validation datasets calculated after each epoch, i.e., when the epoch just before the mean Dice continued increasing for the training dataset but stagnated or reduced for the validation dataset.

2.4. Identification of the images containing lymph nodes

The output of the DL model per pre-processed image was a probability map showing the predicted likelihood of a particular pixel being part of a LN. In order to convert the per-pixel prediction value into a binary classifier for the scanned image, we compared the results of 2 methods based on the probability maps. The first method was the current standard method which uses a simple threshold of the prediction map to obtain a binary mask as described by Ronneberger et al. (13), termed “conventional method” in this article. The per-pixel predictions were threshold at 0.5 probability, where every prediction higher than 0.5 was considered part of a LN, creating a binary mask. To remove potential artefacts, small areas (minimum area set at 5% of the smallest LN area found in the training set) were considered as potential outliers and excluded from further analyses. Any scanned image containing a region of interest greater than that size was labelled as potentially containing LN.

The second method used *a priori* knowledge of the LN shape (usually similar to a kidney bean) to analyse the predicted LN delineation and select the most likely correctly segmented ones. This selection allowed us to obtain a prediction score not just per pixel but per LN and quantify the results of our DL model. The following features were extracted from the prediction map of each candidate LN: descriptive statistics (geometric and harmonic means, standard deviation of prediction values, entropy, skewness, and kurtosis), and shape features (pixel count, number of delineations predicted, roundness, roundness disproportion, area, perimeter, centroid, orientation, major axis length, minor axis length, diameter, extent, solidity, eccentricity, elongation, perimeter/surface ratio). The features were normalised using z- score normalisation based on the mean and standard deviation derived from the training dataset and the correlation between features were tested using the Spearman rank correlation coefficient (16) on the training dataset. To remove redundant information, if a correlation coefficient was above 0.85 between 2 features, the feature with the highest correlation coefficient across the correlation table was deselected from the remaining feature set. The normalisation and the feature selection based on the training dataset was later applied to the validation and test datasets. Finally, recursive feature elimination with 10 cross-validation (RFECV) using default parameters was performed on the features extracted from the training set, optimising the area under the curve (AUC) of the receiver operating characteristic (ROC) score for an extreme gradient boosting (XGBoost) classifier. At every iteration of the RFECV model, the least predictive feature was removed from the dataset until only 1 remained. We visualised the RFECV curve (corresponding to the AUC against the number of features) and selected the number of features corresponding to the turning point of the curve, i.e., when no further increase in the AUC score was observed.

The selected features were used as input for an XGBoost classifier which was trained to give a prediction score between 1 and 0 whether a candidate delineation contained a LN or not. To fine-tune the parameters of this classifier, a grid-search with 10-fold cross-validation was performed on the training dataset. The parameters tested were maximum depth, minimum child weight, number of estimators, gamma, and the learning rate. The set parameters were the scoring system using the ROC AUC, the objective being binary logistic, and column sample by tree at 0.8. To determine whether a particular pre-processed image contained a LN or not, each potential LN within the pre-processed image was attributed a prediction score and the highest score was chosen to classify the pre-processed image. The workflow of both automatic classification strategies per pre-processed image (with LNs/without LNs) is shown Figure 2.

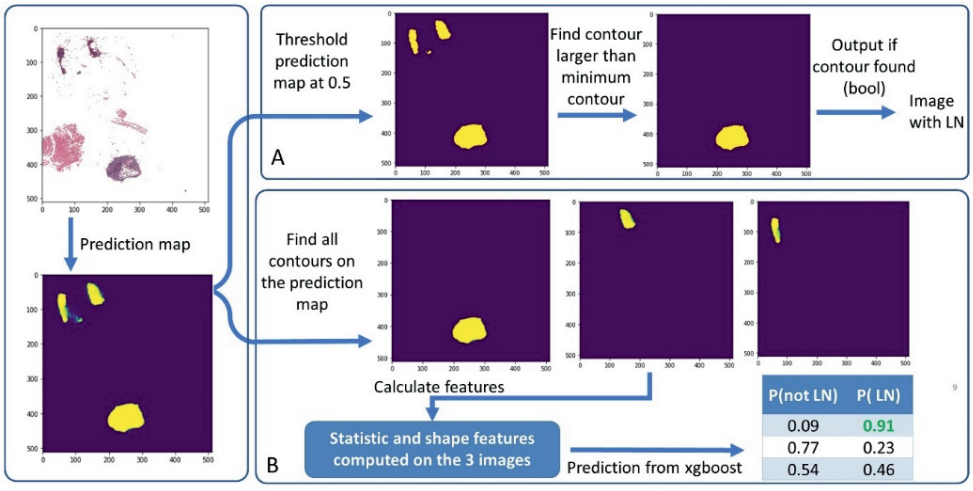


Figure 2: The 2 strategies for predicting whether an image contains a lymph node. (A) "conventional" method, (B) our prediction score method.

Reusing our prediction score computed per predicted delineations and in an attempt to make the model more robust, we empirically created a third "uncertain" category based on statistics from the validation dataset, for which the model could not predict with a high enough confidence whether the pre-processed image contained a LN or not. To define this new category, the lower bound corresponding to the lowest 5% of prediction scores in the distribution of pre-processed images with LNs and the upper bound corresponding to the highest 5% of scores in the distribution of pre-processed images without LNs in the validation dataset were extracted from the distribution of the confidence scores.

2.5. Analysis of the model's lymph node detection performance

We compared the performance of the 2 methods used for classification of pre-processed images on the validation, and external test datasets by comparing the normalised confusion

matrices (i.e., the rows are divided by the sum of the rows which then add up to 1). Performance metrics calculated on the validation and external datasets were balanced accuracy, sensitivity, specificity, and F1-score.

We reported the feature importance via the Gini index of the trained XGBoost model (17), the ROC curves of the candidate delineation classification prediction on the training, validation, and external test datasets with their confidence intervals at 95% calculated with 2000 bootstrapping of the results and the AUCs, along with the calibration curve based on the predictions obtained on the validation dataset. We also reported the results using the confusion matrices including the uncertain category on the validation and the external test datasets composed of the pre-processed images. To evaluate the added value of the uncertain category, we compared the false-negative and false-positive results on the external dataset with and without the uncertain category using 2 proportion z-test at a significance level of 0.05.

We used the maximum prediction score of a candidate delineation to establish the performance of the XGBoost model on the scanned images.

The scanned image predictions were used to calculate sensitivity and balanced accuracy per patient in the validation and external test datasets. We reported the mean sensitivity and the mean balanced accuracy per patient together with the violin plots of those metrics with a 2-sided Mann-Whitney-Wilcoxon test with Bonferroni correction to assess whether distributions of balanced accuracies and specificities were different between the validation and external test datasets.

Analysis of the false-negative results were performed by an expert pathologist. Observations were reported on the scanned images of the external test dataset containing LNs which remained undetected by the model to identify potential underlying causes or trends. The scanned images were considered false negatives when none of the pre-processed images belonging to those scanned images were falling in the category “uncertain” or “contain LN”.

2.6. Auto-delineation of the lymph nodes

Post-processing for auto-delineation extraction

If the pre-processed images had to be split in 2 during the pre-processing step, the delineations were performed on both pre-processed images and on a central image corresponding to both halves of the pre-processed images. Pixels with 2 probabilities were averaged. Finally, to evaluate the delineation results, we resized the pre-processed images to the original scanned image aspect ratio. Once the prediction was obtained from the scanned image, the delineations were automatically found as described in Suzuki and Be (18). Areas with LN candidates were removed if the area was smaller than the minimum area computed previously. Next, the delineations were made convex to roughly resemble the natural shape of LNs (19). From the scanned images, we filtered out the background to only delineate the tissue and increase the accuracy of the delineation, taking into consideration that potential concavity might not have been dealt with when making the delineations convex.

Analysis of lymph node delineation performance

The performance of the delineation model was reported using the average Dice coefficient (20) calculated from the original image, along with the average Dice per LNs from the validation and test dataset. Violin plots of the Dice coefficient per size category were reported. We calculated 4 size categories based on the distribution of areas of the ground-truth delineations in the original images in the training dataset, given in μm^2: (1) from the minimum area to the first quartile (Q1), (2) from Q1 to the median (M), (3) from M to the third quartile (Q3), and (4) from Q3 to the largest area. The distribution of the values within the violin plots were compared between the size categories using a 2-sided Mann-Whitney-Wilcoxon test with Bonferroni correction.

3. Results

After pre-processing of the datasets, the training dataset contained 1516 pre-processed images, the validation dataset 481, and the test dataset 1251 (see Table 1). Our U-Net model was trained for 28 epochs until the model began to overfit on the training dataset.

3.1. Evaluation of the lymph node detection performance

Comparison between conventional threshold method and newly developed prediction score method

Among the pre-processed images of the training dataset, 5% of the smallest LN area was equivalent to 167 pixels. The confusion matrices illustrating these results can be found in Figure 3 panels A and B. The detection performance was reported Table 2.

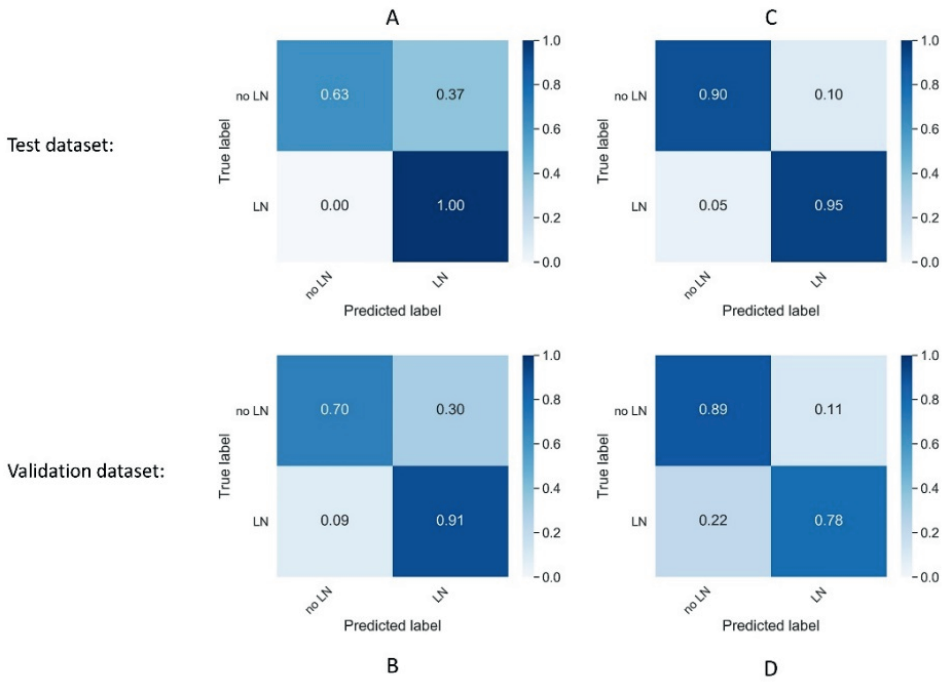


Figure 3: Comparison of the results obtained on the original images to detect LNs in the validation and external test dataset between the "conventional" method and our prediction score method. (A) Confusion matrix "conventional" method for the validation dataset, (B) confusion matrix "conventional" method for the external test dataset, (C) confusion matrix using the prediction score method on the validation dataset, (D) confusion matrix using the prediction score method on the external test dataset.

Table 2: Balanced accuracy, specificity, sensitivity and F1-score calculated on the validation and external test datasets for comparing the 2 classification methods. Bold indicates best performance on the external test dataset

method	dataset	balanced accuracy	specificity	sensitivity	F1-score
our method	validation	0.93 CI [0.90,0.95]	0.90 CI [0.87,0.94]	0.95 CI [0.92,0.98]	0.92 CI [0.89,0.94]
	test	**0.83 CI [0.81,0.86]**	**0.89 CI [0.86,0.91]**	0.78 CI [0.74,0.82]	**0.77 CI [0.74,0.81]**
conventional method	validation	0.81 CI [0.79,0.84]	0.63 CI [0.57,0.68]	1.00 CI [1.00,1.00]	0.81 CI [0.76,0.84]
	test	0.81 CI [0.78,0.83]	0.70 CI [0.67,0.73]	**0.91 CI [0.88,0.94]**	0.72 CI [0.69,0.75]

The optimal number of features calculated for our newly developed prediction score system was 6, namely perimeter surface ratio, standard devi- ation, roundness, harmonic mean,

number of contours, and centroid. The accuracy vs number of features curve calculated during the recursive feature elimination supporting the choice of number of features can be found in supplementary material Figure 1 A. The optimum hyperparameters for XGBoost were found to be gamma = 0.8, learning rate = 0.01, maximum depth = 3, number of estimators was 1000, and minimum child weight = 5. Feature importance can be found in supplementary material Figure 1B. The detection accuracies on the scanned images were 0.92 on the validation dataset and 0.85 on the test dataset. The confusion matrices illustrating these results can be found in Figure 3C and D.
The AUCs of the training, validation, and test datasets were 0.98, 0.94, and 0.90, respectively. The ROC curves obtained on the training, validation, and test datasets are illustrated in supplementary material Figure 2 and the calibration curve calculated on the validation dataset can be found in supplementary material Figure 3.
For the interval of uncertainty we found the following values on the validation dataset: The lower boundary of the distribution scores for the pre-processed images which contained LNs at a 5% cut-off was found to be 0.48; the upper boundary at 95% obtained on the score of the pre-processed images which didn't contain LNs had a prediction score of 0.72. Table 3 displays the results found on the validation and external test datasets, respectively.

Table 3: Predictions on the validation and external test datasets split into three categories according to the level of certainty. The ground truth (image with or without LN) was obtained from manual review by a pathologist.

	Pre-processed images predicted to contain one or more LN n (%)	Uncertain category n (%)	Pre-processed images predicted to contain no LN n (%)	Total
Dataset	Validation dataset			
Images with LN	181 (0.87)	17 (0.08)	**11 (0.05)**	209 (1)
Images without LN	**14 (0.05)**	11 (0.04)	247 (0.90)	272 (1)
Total	258 (0.53)	**28 (0.06)**	195 (0.41)	481 (1)
Dataset	External test dataset			
Images with LN	280 (0.69)	32 (0.08)	**92 (0.23)**	404 (1)
Images without LN	**61 (0.07)**	33 (0.04)	753 (0.89)	847 (1)
Total	341 (0.38)	**65 (0.06)**	845 (0.56)	1 251 (1)

The uncertain category, i.e., a category which would require manual rechecking of the original image by a pathologist, comprised 6% of the validation dataset. The same proportion was obtained on the external test dataset using the lower and upper bounds calculated on the validation dataset. Comparing the results obtained in the uncertain table score to the confusion matrix in Figure 3D, the false-negative rate was similar: 23 % in our uncertain table versus 22% in the confusion matrix (P = .73). However, we observed a significant decrease in false-positive findings: 7% in our uncertain table versus 11% in the confusion matrix (P < .05).

Accuracy and sensitivity distributions

The balanced accuracy and sensitivity distribution for the detection of LN per scanned image reported per patient in the validation and external test datasets is illustrated in supplementary material Figure 4. The mean sensitivities of the LN detection were 1 and 0.72 for the validation and test dataset, respectively. The mean balanced-accuracies for the LN detection were 0.58 for both the validation and test datasets.

False negative analysis on the test dataset

30 (9%) scans out of 348 were classified as "not containing LNs" although they contained a LN. Figure 4 summarizes the description of those images.

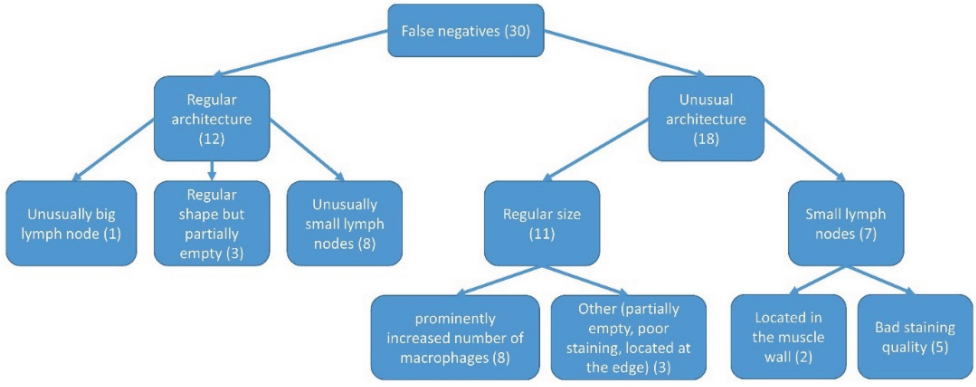


Figure 4: Analysis of the LNs architecture in the scanned images belonging to the external test dataset wrongly categorized without LNs

We observed that some of the LNs 'undetected' by the algorithm were (a) very small collections of lymphocytes which did not have a capsule or (b) did not display the usual LN microarchitecture with loss of lymphocytes and massive increase of macrophages occupying large part of the node, while others appeared 'empty', i.e. devoid of immune cells.

3.2. Evaluation of the model's lymph nodes delineation performance

The delineation performance was computed on the validation and external test datasets, comparing the ground truth delineated by a pathologist with the fully automatic delineation in original images containing LNs. The mean Dice score per original image was 0.73 and the mean Dice score per LNs was 0.66 for the validation dataset and 0.60 per original image and 0.48 per LNs for the external test dataset. The parameters used to create different intervals computed on the distribution of delineation areas in the train dataset were: Q1= 278 061.5, M= 737 090.6, and Q3= 1 707 021.1 (areas in μm^2). The violin plots of the Dice scores per interval for the validation and external test datasets are displayed in Figure 5. Examples of different quality of auto-delineations are displayed in supplementary material Figure 5.

Accurate delineation of small LNs (<Q1) seem to be significantly lower than the detection of the LNs at another size range, for both the validation and test datasets. However, the results are significantly different for the first and last categories (Figure 5).

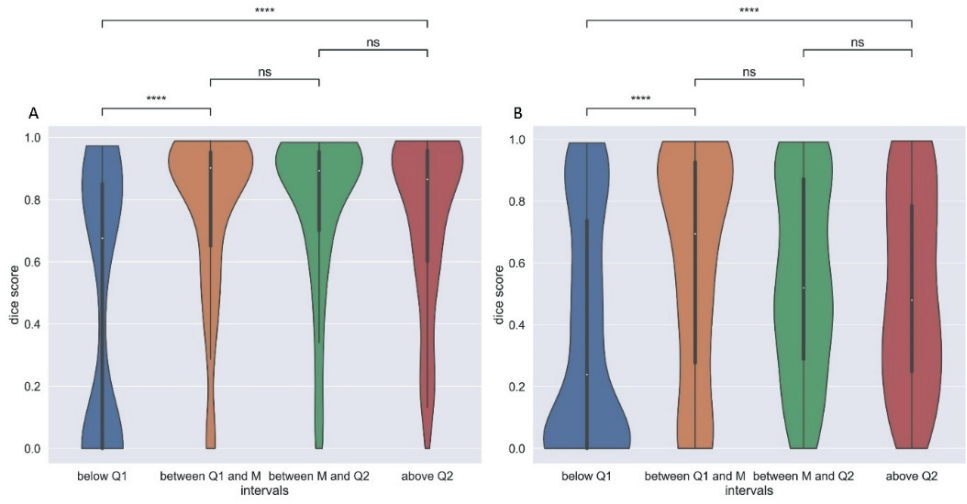


Figure 5: Violin plot of the dice score per LNs split into 4 intervals defined based on the quartiles found on the ground-truth delineations distribution of the training dataset in μm^2 (A) on the validation dataset and (B) the external test dataset. Legend of the annotations (P-values): ns = non-significant i.e. 5×10^{-2} < P-value <= 1; ****= P-value<= 10^{-4}

4. Discussion

In the current study, we developed a novel machine learning based pipeline to: (1) find and (2) delineate LNs in large collections of digitised H&E-stained slides from oesophagogastrectomy specimens and tested the performance on 1 independent dataset, while attempting to increase the explainability of the models. For finding the digital images containing LNs, we compared the performance of a conventional U-Net with thresholding method with our newly developed prediction score approach and observed a lower number of false-positives in both the validation and test datasets using our method. Furthermore, our approach had a higher accuracy in predicting whether a pre-processed image contains LN or not in the external test dataset (0.77 conventional method vs 0.81 our approach). Another study to delineate LNs in H&E-stained images for gastric cancer patients using a U-Net architecture and thresholding reports a Dice score of 0.986 on the validation dataset (21). The main difference to our study is that the training dataset of the U-Net model consisted exclusively of H&E- stained images containing LNs in every slide, meaning that the network would only have to exclude the background and small artefacts to allow LN delineation. Moreover, metrics per LN such as sensitivity or Dice score were not reported, leading to the performance of this model on small LNs to remain unknown.

When inspecting the feature importance within the XGBoost model ranked by the Gini coefficient, we observe that roundness and perimeter-to-surface ratio where among the 3 most important features. This correlates well with semantic knowledge that LNs often have an oval shape (19), leading to irregular shapes being filtered out by our model. The delineation results obtained on the test dataset were adequate (mean Dice of 0.60 per original image, 0.48 for per LNs). When looking at the results divided by LN size in Figure 5, the Dice scores for smaller LNs were significantly lower than for larger LNs. This is partially due to the penalisation of small structures by the Dice score, but might also be partially due to mislabelled data, where artefacts were wrongly attributed a label during pre-processing. Further analysis of the small LNs is being conducted. Although we attempted to limit the change of appearance of our data by splitting our images in 2 when the length/height difference was greater than 1.5, resizing the images into square images compresses the data and possibly negatively impacted our feature extractions and thus our results. Another pre-processing method such as tiling the extracted images instead of resizing them could alter the effect of the resizing.

Individual LNs can vary substantially in their microarchitecture (22) which can impact on the successful training of a DL model to identify LNs (23). Delineations of the lymph nodes could be impacted by tumour invasion, as this could change their structure and appearance. A follow-up study could evaluate the results on the positive and negative lymph nodes and show if a significant difference exists between the results in the 2 categories. Furthermore, H&E-stained tissue sections can vary in colour even if they originate from the same laboratory. Our analysis pipeline therefore included normalisation of the data, making the datasets less dependent from differences in staining. It is also possible that our normalisation method was not sufficient to prevent a domain shift in the external test dataset. Other method such as Fourier-based data augmentation as described in Wang et al. (24) could be adopted in a follow-up study to overcome this issue. We have chosen to train a U-Net model as this has been shown to be one of the most often-used models for automatic delineation in histopathology (14). Further, our attempt to make the results of a U-Net model trained on histopathology data more explainable (certainty score, uncertain class creation, and most important features extracted from the prediction map) could solve 2 major roadblocks to clinical implementation: DL models lack explainability (the “black-box problem”) and are incapable of assessing whether a new dataset is useable or should be rechecked by a pathologist (the “generalisability problem”). Our “uncertain” class could help solve this issue although our current results don’t generalise well on the external dataset, with almost a quarter of pre-processed images being classified as not containing LNs while containing LNs (23%).

The different results observed between the validation and test datasets could be due to the fact that the validation dataset is from the same source as the training dataset while the test dataset is from another cohort.

Looking towards clinical application, our model for the detection and delineation of LNs could be integrated into software used for reviewing H&E-stained slides in the diagnostic setting and tested prospectively on H&E data from UGI patients. To obtain better delineations and detection results, we suggest implementing a continual learning process

which would retrain the model with corrected delineations and detections predictions on the new dataset such as in Perkonigg et al. (25). A follow-up project will introduce analysis of handcrafted features extracted from the H&E-stained images (histomics analysis) to complete the work performed here and predict tumour infiltration within LNs.
In conclusion, we created a pipeline using deep learning for initial detection and handcrafted features to reinforce the predictions of a semantic delineation model which outperformed the conventional approach. Thanks to our scoring model, we could create an uncertain category for which the model is not confident to classify the image into with or without LNs that pathologists would have to review. Although good performance was obtained on the validation dataset, medium performance was obtained on the test dataset for both the classification and delineation tasks, which might be due to high heterogeneity in the external test dataset which might not have been there in the training dataset. The first part of our workflow could be used in a routine diagnostic setting for H&E-stained images of esophageal tissue after further prospective validation, and the second part could be useful for further work on measuring LN areas and characterising the structure of LNs, potentially useful for personal treatment planning for patients with UGI cancer.

Conflict of Interest

H.W. has minority shares in the company Radiomics SA. D.C. declares grants from Medimmune/AstraZeneca, Clovis, Eli Lilly, 4SC, Bayer, Celgene, Leap, and Roche, and Scientific Board Membership for OVIBIO. P.L. has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/059089), licensed to ptTheragnostic/DNAmito; one non-issued patent on LSRT (PCT/ P126537PC00), licensed to Varian; three non-patented inventions (softwares) licensed to ptTheragnostic/ DNAmito, Radiomics SA and Health Innovation Ventures and two non- issued, non-licensed patents on Deep Learning-Radiomics (N2024482, N2024889). He confirms that none of the above entities or funding sources were involved in the preparation of this paper. H.I.G. reports personal fees from Merck Sharp & Dohme, outside the submitted work.

Ethics Approval and Consent to Participate

The study was approved by the South East Research Ethics Committee, London, United Kingdom, REC reference: 07/H1102/111.

Author Contributions

M.B. performed all the analysis, analyzed the results of the classifiers and wrote the manuscript. D.M. and A.C. supervised the writing of this article and guarantees the integrity of the analysis and results presented. H.W. and P.L. supervised the progression of the project and the writing of this article. R.L., W.A. and M.N. collected and provided the

samples. H.I.G. devised the project's aim, supervised the project, annotated the images, and analyzed the results. All authors have participated in writing the manuscript. All authors read and approved the manuscript.

Funding

This work was made possible through the support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276). Authors furthermore acknowledge financial support from ERC-2020-PoC: 957565-AUTO.DISTINCT, CHAIMELEON no. 952172, EuCanImage no. 952103, the Dutch Cancer Society (KWF Kankerbestrijding) project number 12085/2018–2 and IMI-OPTIMA n° 101034347. The material collection was funded by Cancer Research UK (grant number C26441/A8944 to PI H.I.G.).

Data Availability Statement

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Supplementary Information

Table S1: python packages and their versions

purpose	packages	versions
pre-processing	histomicsTK	1.1.1
	os	n/a
	numpy	1.19.2
	pandas	0.24.2
	pillow	8.2.0
	py-wsi	2.1
	opencv	4.1.0
	scikit-image	0.18.1
	scipy	1.5.2
	simpleITK	1.2.0
deep learning	keras	2.2.4
	tensorflow-gpu	1.13.1
feature processing and calculation	pyradiomics	3.0.1
machine learning	xgboost	1.4.2
visualisation	matplotlib	3.0.3
results analysis	scikit-learn	0.20.3
	statannot	0.2.2

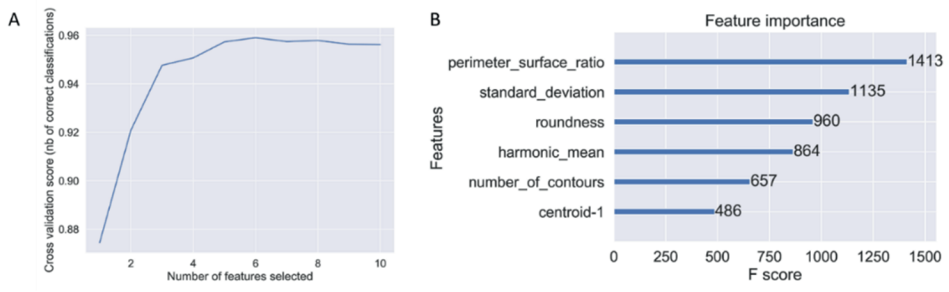


Figure S1: (A) feature selection with RFECV and (B) feature importance

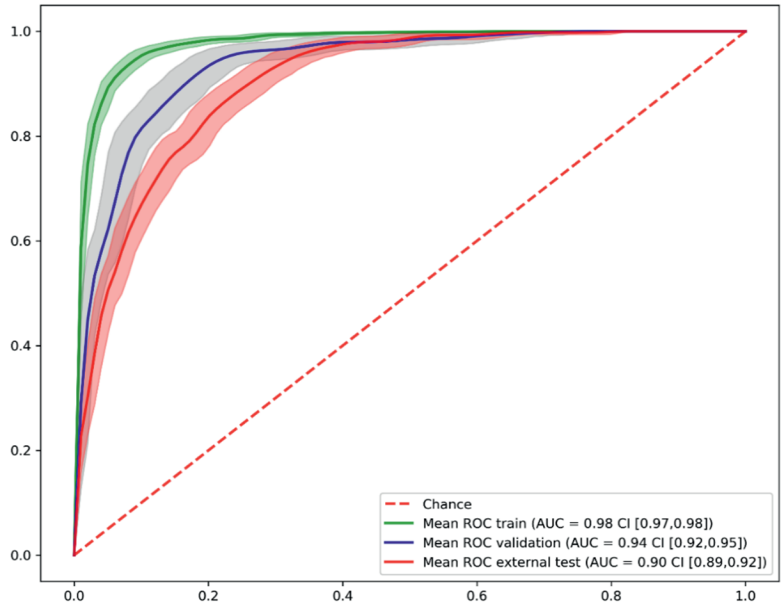


Figure S2: ROC curves of the train, test and external pre-processed datasets for the classification of LN candidates with our newly developed prediction score method. The DL model overfits slightly on the training dataset when looking at the ROC curves (figure 4) but still obtained good classification performance per candidate LNs on the external test dataset (AUC=0.90).

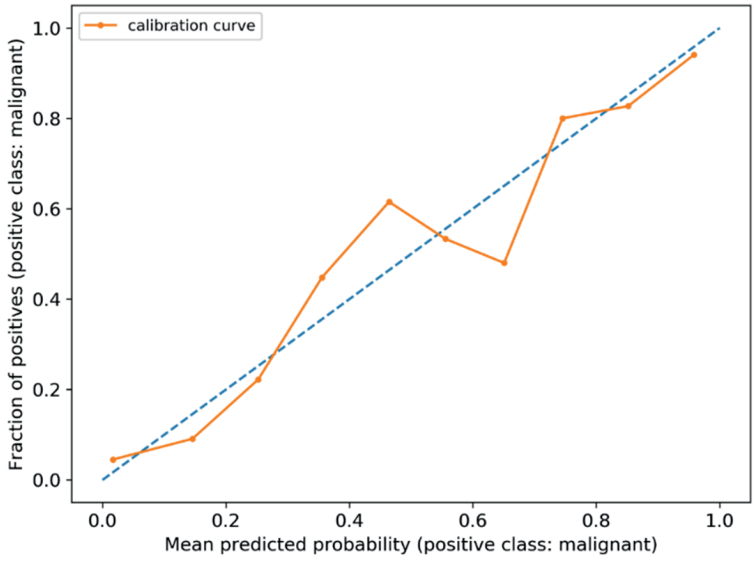


Figure S3: calibration curve on the validation dataset

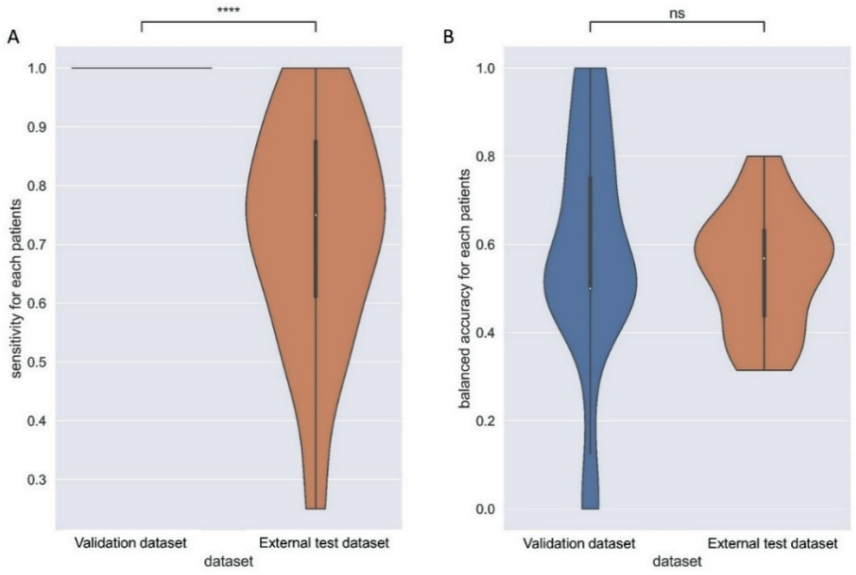


Figure S4: Report of the performance of the model to detect LNs with (A) violin plot of the sensitivity per patient on the validation dataset and on the external test dataset (B) violin plot of the balanced accuracy per patient on the validation dataset and on the external test dataset

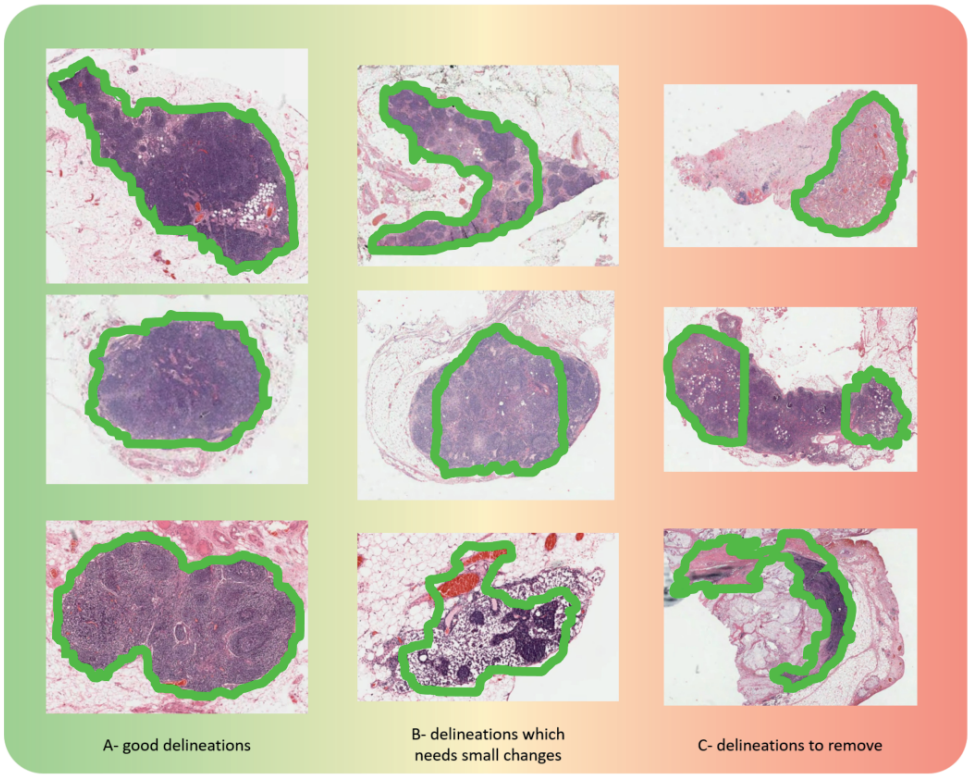


Figure S5: examples of autodelineations separated in three categories: (A) perfect delineations (B) delineations requiring small adjustments (C) unacceptable delineations

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249. https://doi.org/10.3322/caac.21660.

2. Lordick F, Mariette C, Haustermans K, Obermannová R, Arnold D, Committee EG. Oesophageal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2016;27:v50–v57. https://doi.org/10.1093/annonc/mdw329.

3. Smyth EC, Verheij M, Allum W, Cunningham D, Cervantes A, Arnold D. Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 2016;27:v38–v49. https://doi.org/10.1093/annonc/mdw350.

4. Arnold M, Morgan E, Bardot A, et al. International variation in oesophageal and gastric cancer survival 2012–2014: differences by histological subtype and stage at diagnosis (an ICBP SURVMARK-2 population-based study). Gut 2021. https://doi.org/10.1136/ gutjnl-2021-325266.gutjnl-2021-325266.

5. Smyth EC, Fassan M, Cunningham D, et al. Effect of pathologic tumor response and nodal status on survival in the medical research council adjuvant gastric infusional chemotherapy trial. J Clin Oncol 2016;34:2721–2727. https://doi.org/10.1200/jco.2015.65.7692.

6. Davarzani N, Hutchins GGA, West NP, et al. Prognostic value of pathological lymph node status and primary tumour regression grading following neoadjuvant chemotherapy-results from the MRC OE02 oesophageal cancer trial. Histopathology 2018; 72:1180–1188. https://doi.org/10.1111/his.13491.

7. Medical Research Council Oesophageal Cancer Working G. Surgical resection with or without preoperative chemotherapy in oesophageal cancer: a randomised controlled trial. Lancet 2002;359:1727–1733. https://doi.org/10.1016/S0140-6736(02)08651-8.

8. Kloft M, Ruisch JE, Raghuram G, et al. Prognostic significance of negative lymph node long axis in esophageal cancer: results from the randomized controlled UK MRC OE02 trial. Ann Surg 2021. https://doi.org/10.1097/sla.0000000000005214.

9. Alderson D, Cunningham D, Nankivell M, et al. Neoadjuvant cisplatin and fluorouracil versus epirubicin, cisplatin, and capecitabine followed by resection in patients with oesophageal adenocarcinoma (UK MRC OE05): an open-label, randomised phase 3 trial. Lancet Oncol 2017;18:1249–1260. https://doi.org/10.1016/S1470-2045(17)30447-3.

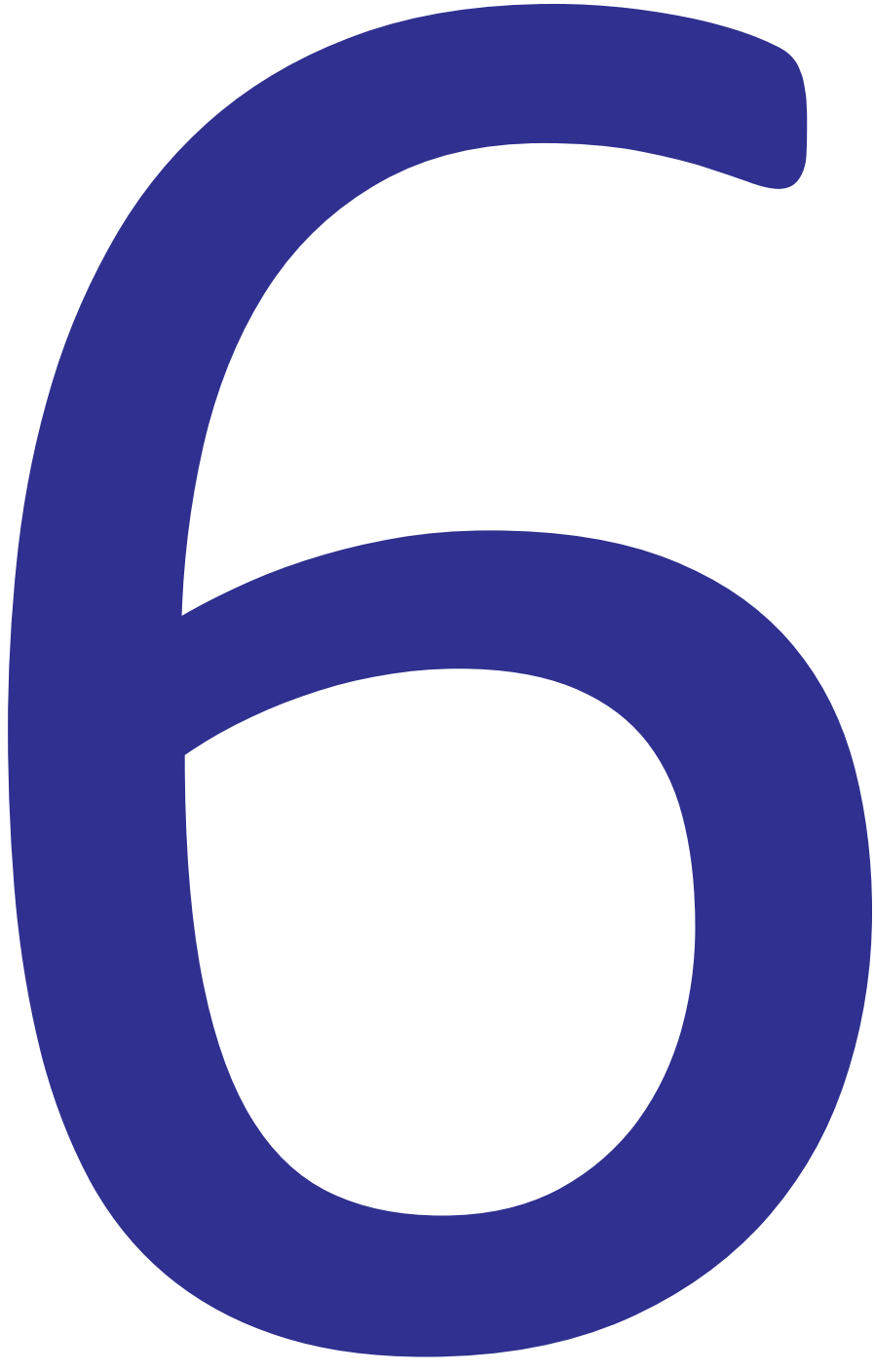
10. Li X, Li C, Rahaman MM, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artif Intel Rev 2022. https://doi.org/10.1007/s10462-021-10121-0.

11. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun G et al. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 1107-1110.

12. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 1979;9:62–66. https://doi.org/10.1109/TSMC.1979.4310076.

13. Ronneberger O, Fischer P, Brox T. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing. 2015:234–241.

14. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal 2021;67, 101813. https://doi.org/10.1016/j.media. 2020.101813.
15. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint 2014. arXiv:1412.6980.
16. Spearman C. The proof and measurement of association between two things. By C. Spearman, 1904. Am J Psychol 1987;100:441–471.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Routledge. 2017.
18. Suzuki S, Be K. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing, 30. 1985. p. 32–46. https://doi.org/10.1016/0734-189X(85)90016-7.
19. Ganeshalingam S, Koh D-M. Nodal staging. Cancer Imaging 2009;9:104–111. https://doi.org/10.1102/1470-7330.2009.0017.
20. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.
21. Wang X, Chen Y, Gao Y, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nat Commun 2021;12:1637. https://doi.org/10.1038/s41467-021-21674-7.
22. Elmore SA. Histopathology of the lymph nodes. Toxicol Pathol 2006;34:425–454. https://doi.org/10.1080/01926230600964722.
23. Wu Y, Cheng M, Huang S, et al. Recent advances of deep learning for computational histopathology: principles and applications. Cancers 2022;14:1199.
24. Wang X, Zhang J, Yang S, et al. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. Med Image Anal 2023;84, 102703. https://doi.org/10.1016/j.media.2022.102703.
25. Perkonigg M, Hofmanninger J, Herold CJ, et al. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat Commun 2021;12: 5678. https://doi.org/10.1038/s41467-021-25858-z.



Chapter 6

From identification to classification of lesions in contrast-enhanced mammography combining deep learning and handcrafted radiomics

Manon Beuque, Marc B.I. Lobbes, Yvonka van Wijk, Yousif Widaatalla, Sergey Primakov, Michael Majer, Corinne Balleyguier, Henry C. Woodruff[1], Philippe Lambin[1]

[1] shared senior authorship

Adapted from:
Manon P. L. Beuque, Marc B. I. Lobbes, Yvonka van Wijk, Yousif Widaatalla, Sergey Primakov, Michael Majer, Corinne Balleyguier, Henry C. Woodruff, Philippe Lambin.
Combining Deep Learning and Handcrafted Radiomics for Classification of Suspicious Lesions on Contrast-enhanced Mammograms. Radiology 2023; 307:5. doi: https://doi.org/10.1148/radiol.221843

Abstract

Background: Handcrafted radiomics (HR) and deep-learning (DL) models individually achieve good classification performance (benign/malignant) on contrast-enhanced mammography (CEM).

Purpose: We hypothesize that combining both allows for automated identification, delineation, and improved classification of lesions. Such system could potentially aid clinicians in their workflow and decision-making by highlighting lesions and making diagnostic suggestions.

Materials and Methods: Imaging and clinical data were retrospectively collected for 1,062 recall patients who underwent CEM acquired at the Maastricht University Medical Centre+, including diagnosis and breast cancer subtype, and 279 cases were acquired at Institute Gustave Roussy for external validation. Lesions with a known status (malignant/benign) were delineated by an expert radiologist. Pre-processed low-energy and recombined images were used to train a DL-model for automatic lesion identification, contouring, and classification. HR models were trained to classify lesions identified and contoured both by the radiologist and by the DL-model. Identification sensitivity and the area under the curve (AUC) for classification were compared between the different approaches at image and patient-level and both classification models were combined by averaging the predictions.

Results: On the external dataset, identification sensitivity of the lesions was 90% (99%) and mean Dice was 0.71 (0.80) on the image (patient) level. Using manual contours, the combined classification model achieved the highest sensitivity of 83% (95% confidence interval (CI) [79,87]%) on the image level, as well as the highest AUC of 0.88 (CI [0.86,0.91]). Using DL-generated contours, DL achieved the highest sensitivity of 90% (CI [87,93]%) while the highest AUC was reached with the combination model (0.95 (CI [0.94,0.96])). The two classification models agreed on 84% of the DL-generated contours and obtained an AUC of 0.96 (CI [0.95,0.97]) on this subset.

Conclusion: DL was able to accurately identify and delineate suspicious lesions and the combination and agreement of DL and HR achieved good diagnostic performance on CEM.

1. Introduction

Full-field digital mammography (FFDM) continues to be the primary breast imaging tool for the detection of breast cancer. However, diagnostic accuracy of FFDM is decreased in breasts with dense fibroglandular tissue (1), and FFDM specificity to detect cancer is moderate (2). Hence, there remains a clinical need to increase FFDM's diagnostic accuracy, either by using supplemental imaging modalities, such as ultrasound or breast MRI, or technically advanced mammography such as digital breast tomosynthesis or contrast-enhanced mammography (CEM).

CEM has a better diagnostic performance compared to FFDM, both in terms of sensitivity and specificity. Although CEM has a high sensitivity to identify breast cancer, specificity remains moderate (3). In addition, the currently described diagnostic performance of CEM is based on studies using visual assessment of the images by radiologists without aid of computerized techniques.

Studies suggest that diagnostic accuracy of FFDM might be improved with the help of machine learning (ML) based image analysis. McKinney et al. (4) showed that in some FFDM the expert radiologists were unable to provide a correct diagnosis, whereas the ML model did. However, the ML model would sometimes be unable to recognize "obvious" cases, i.e., those easily detected by expert radiologists. Many studies are already available using ML on FFDM, for example using handcrafted radiomics (HR) models (5, 6) to classify breast lesions (7) and deep learning (DL) to identify and segment lesions (8, 9), but their combination remains rarely reported. As an example, in computed tomography of the lung the combination of HR and DL was reported to show improved results for the diagnosis of idiopathic pulmonary fibrosis (10).

We hypothesized that the combination of HR and DL can match radiologists' performance in both identification and classification of suspicious breast lesions. For this purpose, we aimed to develop a comprehensive ML tool able to fully automatically identify, contour, and classify breast lesions based on CEM images of recall patients. In our approach, a DL-model was first trained to identify and segment suspicious lesions within CEM images, and to classify them into benign or malignant. Furthermore, HR classification models based both on manually and automatically delineated regions of interest and clinical parameters were trained, evaluated, and combined with the DL predictions.

2. Methods

2.1. Patient population

In this retrospective study a consecutive series of 1,601 patients who underwent CEM mostly for the assessment of breast lesions recalled from screening was assembled at the Maastricht University Medical Centre+, and their images and clinical data were collected. Other indications could be inconclusive findings on FFDM and/or ultrasound, suspicious (palpable) findings during physical examination, and as an alternative to breast MRI in cases where patients were unable to undergo MRI. Requirement for informed consent was

waived by the institutional review board (METC 2019-0995). Data were collected using the Picture Archiving and Communication System (PACS), and anonymized. 539 patients were excluded as they were deemed negative by an expert radiologist (i.e., no suspicious lesion was found) (Figure 1).

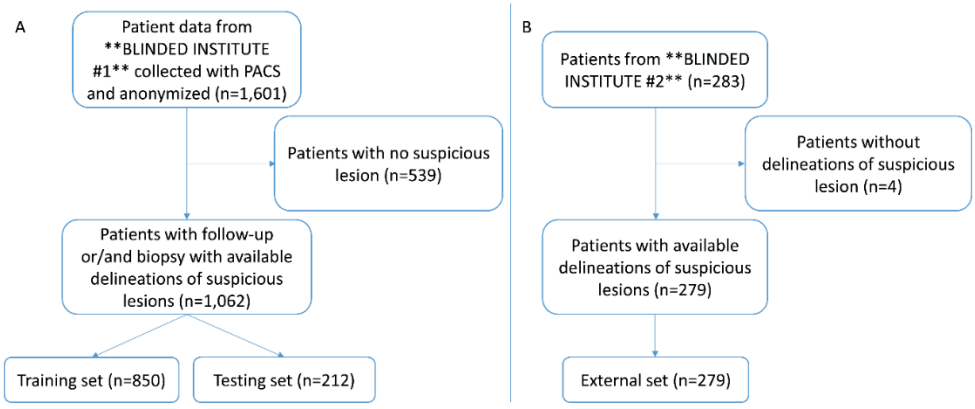


Figure 1: Flowcharts describing patient inclusion (A) used for training and testing (B) used for external validation of the machine learning models

279 patients with both images and clinical data were collected as an external validation dataset from Institute Gustave Roussy, for which the institutional review board waived informed consent (2022-140). The data from both institutes were never reported in a prior publication.

2.2. Imaging

The acquisition protocol of CEM was as described before (11, 12). In short, an iodinated contrast agent was intravenously administered two minutes before the acquisition of dual-energy mammography images of both breasts in the standard mediolateral-oblique (MLO) and craniocaudal (CC) views as minimum. A typical CEM acquisition results in a low-energy (LE) equivalent to FFDM (11) and a recombined image in which areas of contrast accumulation can be assessed (12). For analyses, we used LE and recombined images. All images included in the study were delineated by a research assistant Y.W. supervised by a certified breast radiologist M.L. with 13 years experience in CEM using MIM software v4.1. The delineations were made based on the information retrieved from the patient records and radiology reports. The final ground truth (diagnosis) of each delineation was assigned based on results obtained after reviewing of the pathology reports and/or two years follow-up. In our study, the term breast lesion is defined as architectural distortions, asymmetries, masses, and clusters of suspicious calcifications.

We reported patients' characteristics per dataset and their differences: For the numerical variables, we used Mann-Whitney U-test for two independent samples. For categorical

variables, if every category had more than ten samples we used chi-squared test, else we used a two proportions z-test. We considered the null hypothesis rejected if the p-values were smaller than 0.05.

2.3. Automatic detection and delineation of suspicious lesions using Mask R-CNN

Images were first preprocessed to filter out noise or irrelevant details reducing the size of the images, and to limit the computational cost (11). Since the pre-processing of CEM images for DL is not standardized, we suggested a series of pre-processing steps including histogram and intensity normalizations and combination of LE and recombined images into one image. We implemented a per-image normalization pipeline, starting by removing unwanted information from the image (supplemental Figure 1 A), such as background pixels and possible external objects present on the scan, following the recommendations of Perez-Garcia et al. (13). We applied Otsu's thresholding on the recombined image to find the region of interest of the breast and potential foreign objects (e.g. surgical clips, markers, etc.) within the image. We kept the mask of the largest continuous contour found on the threshold image, considering it to be the breast, removing possible external objects. The recombined and LE images were then cropped to keep only the largest object (i.e. the breast) and all background pixels were replaced with a pixel value of zero (supplemental Figure 1 B). The region of interest of the suspicious lesion was cropped using similar transformations. To obtain a smaller pixel range and avoid information loss during rebinning, instead of clipping at a set range of values such as in (12), we used the statistics of the pixel distribution in the image. We clipped the values higher than 99th percentile of the pixel values within the breast and the values lower than the 1st percentile of the pixel values within the breast for both LE and recombined images (supplemental Figure 1 C). We then normalized the images with minimum/maximum normalization and resampled the images to convert the pixels into 8-bit unsigned integers (supplemental Figure 1 D). The images were pre-processed to create 3 versions of the image to be combined into one image: We applied Contrast Limited Adaptive Histogram Equalization (CLAHE), a common filtering technique for pre-processing medical images (13) on the pre-processed LE and recombined images with a clip limit at 2.5, which redistributed the pixels in excess above the clip limit to limit the contrast in the image and a tile grid size of 16x16 pixels^2 (supplemental Figure 1 E). We also applied CLAHE with a lower clip limit of 1 to the recombined images and a tile grid size of 16x16 pixels^2. The three pre-processed images were merged to form a three-channel RGB image, using the LE in the red channel, recombined images in the green channel, and the second pre-processed recombined images in the blue channel.

The DL-model chosen to identify (i.e. generate a bounding-box around the lesion of interest), delineate, and classify lesions as either benign or malignant was Mask R-CNN with a resnet101-FPN backbone (14). This model has the advantage to preserve the scale ratio of the image and to be independent from the actual size of the lesions as predictions are made at different size resolution of the images. The weights pre-trained on the COCO

dataset (15) were used for transfer learning. Random transformation of the dataset was applied during training, setting minimum shearing at -0.1, maximum shearing at 0.1, minimum scaling at 0.9, maximum scaling at 1.1 and a 0.5 chance to flip the image along the y axis. The minimum size of the image was set to 533 pixels (minimum width of training dataset) and stay within memory constraints, the maximum size was set at 2561 pixels, the median value of the height distribution. The batch size was 1. The total loss was composed of three losses for the three different tasks: the regression loss for the bounding-boxes prediction was smooth L1 with sigma set at 3, classification loss was focal cross-entropy and the mask loss was based on binary cross-entropy. The optimizer chosen to train the DL model was Adam (16) with a learning rate starting at 10^{-5}. The learning rate was decreased with a factor 0.5 if the total loss didn't decrease for 2 epochs with a minimum learning rate at 10^{-7}. The optimal weights were chosen based on the best mean average precision results obtained on the test dataset.

Each suspicious lesion candidate outputted by the DL-model consisted into three predictions: a bounding-box containing a segmentation, a predicted label benign/malignant and an associated confidence score ranging from 0 to 1. The images were considered independent during training (CC and MLO views are both used as independent samples). We limited the output of the model to a maximum of five bounding-boxes per image as agreed with our expert radiologist, only highlighting potential lesions of interest. The bounding-boxes predictions were kept if their confidence score was higher than 0.1.

The results were reported on the test and external datasets for comparison. The lesion which had the highest confidence score in an image was considered the best candidate selected by the DL-model and was used to calculate the accuracy of the identification. The sensitivity was calculated based on all the bounding-boxes predicted per image or per patient. To calculate the sensitivity and accuracy of the identification, if the predicted bounding-box had an intersection over union (IoU) with the ground-truth bounding-box of more than 0.1, it was considered identified. To calculate the accuracy identification per patient, we used the maximum IoUs found across images from the same patient and evaluated if it was higher than 0.1, in which case the lesion was considered identified. Proportion z-tests (alpha=0.05) were calculated on the accuracies and the sensitivities to test for significant differences between the results from the test and external datasets. To obtain the performance of the segmentation task, if the manually delineated lesion was correctly identified, we reported the Dice coefficient of this lesion, if it was not identified, the score reported was 0. To give a performance on a patient-level, we used the highest Dice score available across the images corresponding to this patient. Significant differences between the Dice distributions were tested using Mann-Whitney U tests.

Identification sensitivity, accuracy, and mean Dice of the segmentations were reported per image and per patient on the test and external datasets. The Dice coefficient was computed per contour and reported in a violin plot.

In a retrospective sub-analysis of the cases where the presence of a lesion was not identified by the DL-model (i.e., 'false-negatives'), we calculated proportion z-tests (alpha=0.05) on the false-negative results and reviewed the images with our certified breast radiologist to

establish potential causes for these false-negative findings. The python packages used in our study are listed in supplemental Table 1.

2.4. Classification of lesions

HR features were computed on the manual contours and on the contours generated by our DL-model. We extracted these features from the recombined and LE images and combined them in the same table. In order to increase HR feature homogeneity and robustness, and to decrease sources of noise, grey-levels were discretized into a fixed width of 3 grey-levels. 660 HR features were extracted from the regions of interest using the package pyRadiomics (17) on the recombined images and 660 HR features on the LE images, giving a total of 1320 features extracted from the CEM images. Features describing first order statistics, 2D shape features, and texture were calculated for the unfiltered image and after applying 4 Laplacian of Gaussian filters with a kernel width of 0.1, 0.4, 0.7 and 1.0 mm. Feature selection and model building was performed on the training set and the final model was validated on the testing and external validation datasets. Given the large number of features extracted from each image and the relatively small size of the patient cohort, feature selection becomes a key step in the HR process. The first step was to apply z-score normalization on all features. Next, features were examined for pairwise feature correlations using Spearman's correlation coefficient r. Feature pairs with $r \geq 0.85$ were considered highly correlated and the feature with the highest average correlation to the remaining features was removed. After feature selection, we used recursive feature elimination with stratified 10-fold cross-validation to select the optimal number of features to use in our model, using an XGBoost classifier. To prevent the model to overfit due to imbalance, a higher weight was given to the minority class by using the XGBoost parameter scale_pos_weight, which was set to the ratio of the majority class over the minority class.

To train and evaluate the performance of the XGBoost model, the same split in training and testing used in the DL-model were used. We also compared the performance of the model with and without CF added to the HR based model. For the ground truth contours, the classifiers predicted benign versus malignant lesions. For the auto-predicted contours generated by our DL-model, the classifiers predicted cancerous lesion versus others (benign and false positive). For both datasets, we used grid search with stratified 10-fold cross-validation, optimizing XGBoost classifiers. The parameters tested were the maximum depth, gamma, the number of estimators, and the minimum of samples leaf to optimize the AUC score.

To test if the CF had an adding predictive value, we reused the pipeline described previously using this time radiomics and CF for feature selection. The CF included were pregnancies, number of children, family history, personal history, age, menopause status, medication, and cup size. To deal with missing CF, we used the missForest algorithm (18), an imputation method which can handle continuous and categorical data.

To understand the importance given to the features selected by the different models, SHapley Additive exPlanations (SHAP) values were calculated on the training dataset. Those values indicates how a certain feature influence the model to predict a positive or a negative

outcome: when a SHAP value is negative, the weight is toward a negative prediction. If a SHAP value is positive, the weight is toward a positive prediction. A higher absolute value of the SHAP values indicates a larger influence on the final prediction.

To combine the DL and HR models we averaged their classification probabilities to arrive at a single classification prediction. We repeated this process for the DL-model and the model based on HR and clinical-features (CF), and reported the results of these four models (two "ensembled" models just describes applied on two types of contour, manual and automatically generated) on the two available datasets via the receiver operating characteristic (ROC) curves. The calibrations curves obtained with the DL and HR methods on the test dataset were also joined. The area under the ROC curve (AUC), accuracy, sensitivity, specificity, and F1-score were computed on the external dataset at both the contour and the patient level. To obtain the results at patient-level, we averaged the probability scores given per contour for this patient when we computed the predictions on the ground truth contour. On the predicted contour, we reported the performance of the model per patient only when the bounding-box with the highest score corresponded to the ground truth contour and we attributed the maximum score found across the images to the patient to evaluate the classification prediction.

The thresholds used to obtain binary predictions were selected based on the statistics obtained on the training dataset with the Youden index (19). We listed the results obtained when the binary predictions of the best performing two models were in agreement and reported the percentage of cases for which the models agreed. The 95% confidence intervals (CIs) were computed for all the metrics using bootstrapping, resampling the datasets 2000 times and Tukey's tests were performed between the different metrics to assess significant differences for alpha=0.05. The complete workflow is presented Figure 2.

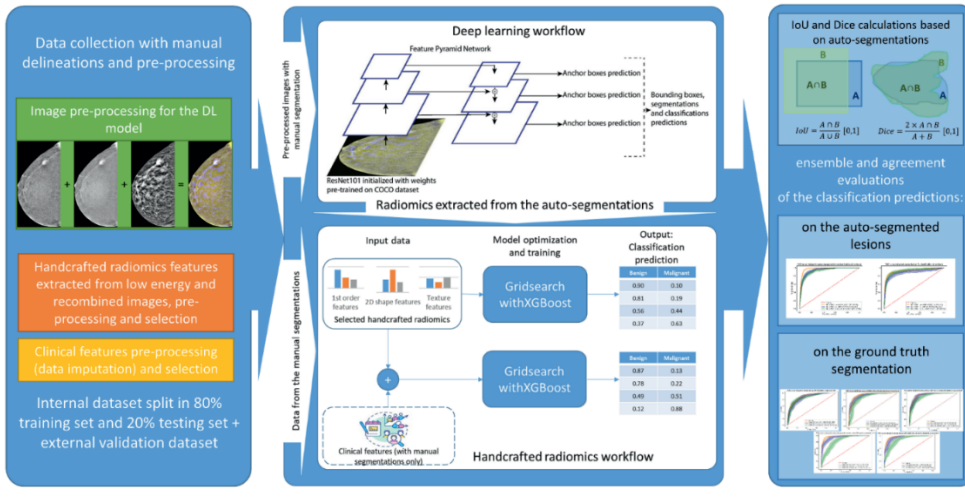


Figure 2: workflow of the detection, delineation, and classification predictions using handcrafted radiomics models and deep learning. Abbreviation: IoU = intersection over union

3. Results

3.1. Patient characteristics

The 1,062 patients were split randomly into a training dataset and a test dataset with a ratio 80/20. This split resulted in 850 patients for the training dataset and 212 patients for the test dataset. For the external validation, the data of 279 patients were used. The clinical characteristics of those patients are described in Table 1.

Table 1: Patient characteristics

Clinical characteristics	Training dataset	Test dataset	External dataset
number of patients (number of lesions)	850 (850)	212 (212)	279 (319)
mean age (±SD)	60.3 ±8.2***	60.2 ±8.1***	54.7 ±12.2*,**
Menopause			
pre	96 (0.11)	18 (0.08)	93 (0.33)
peri	65 (0.08)	14 (0.07)	25 (0.09)
post	504 (0.59)	136 (0.64)	156 (0.56)
Not reported	185 (0.21)	44 (0.21)	5 (0.02)
mean pregnancies	1.9 ±1.3***	2.0 ±1.2***	2.2 ±1.7*,**
mean children (± SD)	1.7 ±1.0***	1.8 ±1.1***	1.9 ±1.5 *,**
Medication			
None	426 (0.50)	93 (0.44)	215 (0.77)
OCP	226 (0.27)	70 (0.33)	37 (0.13)
HRT	17 (0.02)	5 (0.02)	17 (0.06)
Not reported	181 (0.21)	44 (0.21)	10 (0.03)
Family history positive for breast cancer (±SD)	18 ±39 %***	18 ±39 %***	43 ±59 %*,**
Personal history positive for breast cancer (±SD)	>2% ±14 %***	>1 ±8 %***	3.5 ±18 %*,**
Cup size			
A-C	418 (0.49)	104 (0.49)	164 (0.59)
D-F	241 (0.28)	62 (0.29)	72 (0.26)
>F	10 (0.01)	3 (0.01)	3 (0.01)
Not reported	181 (0.21)	43 (0.20)	40 (0.14)
Disease characteristics per lesion			
NST	227 (0.27) ***	58 (0.27) ***	163 (0.51) *,**
DCIS	63 (0.07) ***	19 (0.09) ***	9 (0.03) *,**
Other carcinoma	69 (0.08) ***	10 (0.05) ***	23 (0.07)
Cyst	310 (0.36) ***	80 (0.38) ***	75 (0.23) *,**
Fibroadenoma	68 (0.08)	17 (0.08)	29 (0.09)
Negative	6 (0.01)	1 (>0.01) ***	8 (0.03) **
Not reported	107 (0.13) ***	27 (0.13) ***	12 (0.04) *,**

Footnote: proportions are reported in parentheses; SD = standard deviation, OCP=oral contraceptive pill, HRT=hormone replacement therapy, NST=no special type, DCIS= ductal carcinoma in situ, *= null hypothesis rejected with training dataset (p-value<0.05), **= null hypothesis rejected with the test dataset (p-value<0.05), *** = null hypothesis rejected with the external dataset (p-value<0.05).

3.2. Identification and delineation of the lesions

After pre-processing of the image data, the mask R-CNN model was trained on 1,810 images from 850 patients in the training dataset, tested on 454 images from 212 patients, and validated on 590 images from 279 patients in the external dataset.

The DL-model was trained for 30 epochs and the best weights were obtained for epoch 13, at which point the model had the lowest total loss on the test dataset. The accuracy, sensitivity, and mean Dice are reported in Table 2 for the test and external datasets per contour and per patient.

The distribution of the Dice scores can be seen in supplemental Figure 2.

Table 2: detection and segmentation results of the deep learning model

datasets	accuracy delineations per contour	sensitivity delineations per contour	mean dice per contour	accuracy delineations per patient	sensitivity delineations per patient	mean dice per patient
test	0.64 (279/436)	0.85 (371/436)	0.65	0.80 (170/212)	0.94 (200/212)	0.75
external	0.73 (431/590)	0.90 (532/590)	0.71	0.88 (245/279)	0.99 (275/279)	0.80
p-values	< 0.01 (z-test)	0.01 (z-test)	< 0.01 (Mann-Whitney)	0.02 (z-test)	< 0.01 (z-test)	< 0.01 (Mann-Whitney)

3.3. Analysis of false negative

4/279 lesions were not identified by the DL-based model in the external validation, 12/212 in the test dataset and 43/850 in the training dataset. The proportion of false-negatives was significantly different ($p<0.05$) in the external dataset compared to the training and test datasets. We did not analyze the external validation cases, as privacy legislation did not permit us to retrospectively assess in-depth the data of these patients. The results of the analysis performed by our expert radiologist M.L. for the training and testing datasets based on the radiologist reports are displayed Figure 3.

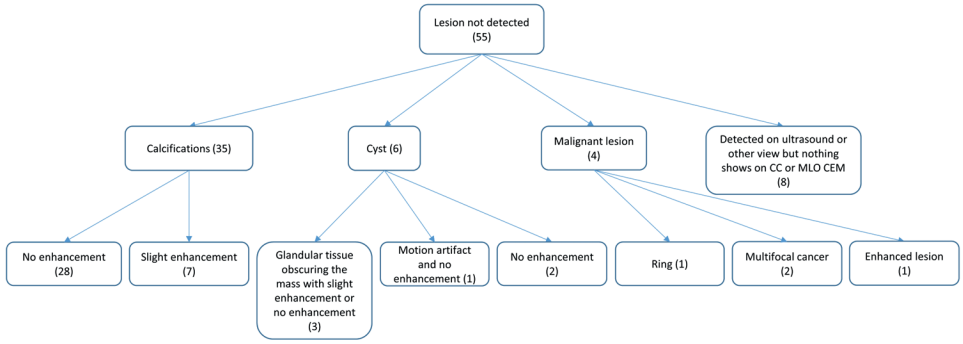


Figure 3: Analysis of the different type of lesions not detected by the DL model made by an expert radiologist. Abbreviations: CC: cranio-caudal, CEM: contrast-enhanced mammography, MLO: mediolateral oblique

The majority (7/12 in the test dataset, 28/43 in the train dataset) of false negatives were calcifications. Examples of unidentified and identified calcifications using the DL-model are displayed Figure 4.

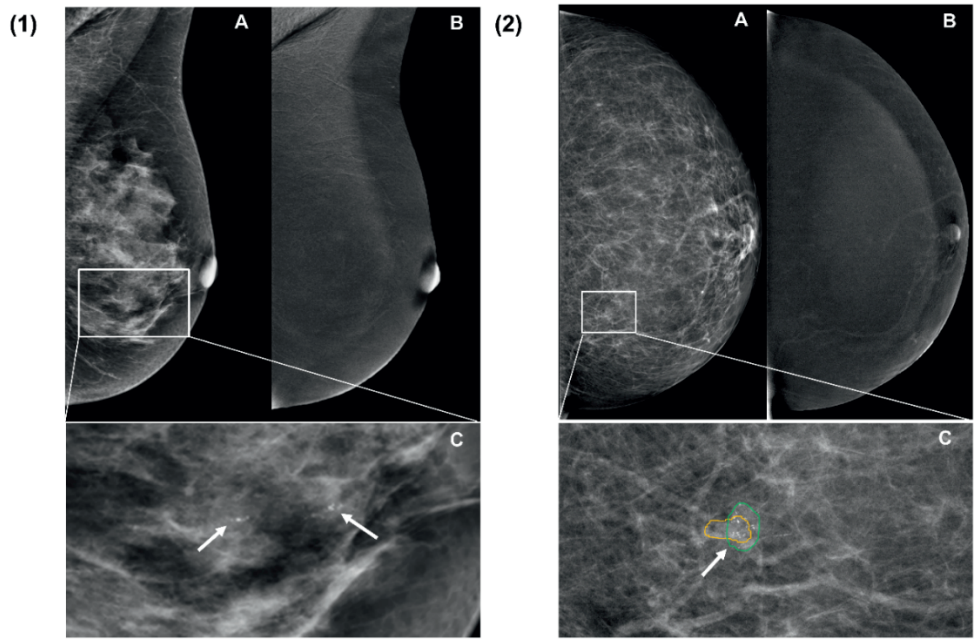


Figure 4: Example of a false negative finding (1) and a correct finding (2) by the deep learning model of suspicious calcifications. (1) On the low-energy images of the left breast (A, only mediolateral oblique view is shown) a cluster of fine linear and branching calcifications were observed (see outtake in C). On the recombined images (B), no enhancement was observed. The model did not provide any prompts. However, stereotactic vacuum-assisted core needle biopsy showed ductal carcinoma *in situ*. (2) Example of correctly detected calcification cluster.

3.4. Classification of the lesions

The optimal parameters found with grid-search for the HR models are available in supplemental Table 2 and the feature importance is available in supplemental Figure 3 with the summary plot of the SHapley Additive exPlanations values. The ROC curves are available in supplemental Figure 4 for the manual contours, and in supplemental Figure 5 for the automatic contours. For the predicted classification obtained on the manual contours, we observed based on the intersections of the CIs that the models seemed to overfit on the training datasets. However, the CIs of the ROC curves obtained with the predictions on the test dataset always overlapped with the CIs based on the predictions obtained on the training dataset. For the automatically generated contour, a similar phenomenon was observed: the DL-model did not overfit on the training dataset, but the HR model did.

The classification results on the external dataset obtained with the multiple approaches are provided in Table 3 (AUC, specificity, and sensitivity) and supplemental Table 3 (accuracy and F1-score). In the following text we report only on the best performing models for per-image and per-patient results for manual and automatically generated contours. For a classification per-image for the manual contours, a combination of HR and DL classification models yielded the highest AUC (0.88) and sensitivity (83%) and HR the highest specificity (80%). For a classification per patient based on the manual contours, the highest AUC (0.88) and sensitivity (89%) were found with DL and the highest specificity (83%) was found using HR. For the automatically generated contours, HR and DL obtained the highest AUC (0.95) and specificity (85%), while DL yielded the highest sensitivity (90%). For the classification per patient based on the automatically generated contours, HR yielded the highest specificity (74%) while the highest sensitivity was obtained with DL alone (100%), and the combination of DL and HR achieved the highest AUC (0.91). The calibration curves are shown in supplemental Figure 6.

Table 3: AUCs, specificities, and sensitivities on the external dataset with confidence intervals for the different model combinations on the manual contours and the automatically generated contours.

On the manual contours							
Per contour				Per patient			
approaches	**AUC**	**specificity**	**sensitivity**	**approaches**	**AUC**	**specificity**	**sensitivity**
radiomics	0.84 CI [0.81,0.88] *	182/227 (80 %) CI [75,85] *	242/363 (66 %) CI [62,71]	radiomics	0.83 CI [0.77,0.88] *	83/100 (83 %) CI [75,90]	113/179 (63 %) CI [56,70]
deep learning	0.86 CI [0.83,0.89]	170/227 (74 %) CI [69,81]	302/363 (83 %) CI [79,87] *	deep learning	0.88 CI [0.84,0.93]	73/100 (73 %) CI [64,81]	160/179 (89 %) CI [85,93]
radiomics + CF	0.84 CI [0.81,0.88] *	173/227 (76 %) CI [71,82]	269/363 (74 %) CI [69,79]	radiomics + CF	0.83 CI [0.77,0.88] *	79/100 (79 %) CI [71,86]	119/179 (66 %) CI [60,73]
deep learning + radiomics	0.88 CI [0.86,0.91] **	177/227 (77 %) CI [72,84]	302/363 (83 %) CI [79,87] *	deep learning + radiomics	0.88 CI [0.83,0.92] **	77/100 (77 %) CI [69,85] *	144/179 (80 %) CI [74,86]
deep learning + radiomics + CF	0.88 CI [0.86,0.91] **	179/227 (78 %) CI [73,84]	289/363 (79 %) CI [76,84]	deep learning + radiomics + CF	0.88 CI [0.83,0.92] **	78/100 (78 %) CI [70,86] *	140/179 (78 %) CI [72,84]
agreed labels	0.95 CI [0.92,0.97]	128/160 (80 %) CI [74,86] *	279/288 (96 %) CI [95,99]	agreed labels	0.93 CI [0.89,0.96]	71/91 (78 %) CI [69,86]	161/167 (96 %) CI [93,99]
On the predicted contours							
Per contour				Per patient			
approaches	**AUC**	**specificity**	**sensitivity**	**approaches**	**AUC**	**specificity**	**sensitivity**
radiomics	0.93 CI [0.92,0.94]	2027/ 2463 (82 %) CI [81,84]	315/353 (89 %) CI [86,92]	radiomics	0.89 CI [0.85,0.94]	55/74 (74 %) CI [64,84]	150/ 171 (87 %) CI [83,92]
deep learning	0.93 CI [0.92,0.95]	2043/ 2463 (82 %) CI [81,84]	319/353 (90 %) CI [87,93]	deep learning	0.87 CI [0.82,0.92]	33/74 (44 %) CI [34,57]	171/171 (100 %) CI [100,100]
deep learning + radiomics	0.95 CI [0.94,0.96]	2106/ 2463 (85 %) CI [84,87]	317/353 (89 %) CI [87,93]	deep learning + radiomics	0.91 CI [0.86,0.95] *	44/74 (59 %) CI [48,70] *	168/171 (98 %) CI [96,100] *
agreed labels	0.96 CI [0.95,0.97]	1828/ 2049 (89 %) CI [88,91]	297/313 (94 %) CI [92,97]	agreed labels	0.91 CI [0.86,0.95] *	44/74 (59 %) CI [48,70] *	168/ 171 (98 %) CI [96,100] *

Footnote: 95% CIs were reported within brackets; AUC = Area under the receiver operating characteristics curve; CF: clinical features; CI: confidence interval; * = not significantly different

The HR and DL-model predictions based on manual contours agreed for 76% of the lesions and AUC, specificity, and sensitivity within that subset were 0.95, 80%, and 94% respectively. For the per-patient predictions, the HR and DL-models agreed 92% of the time and the AUC, specificity, and sensitivity were 0.93, 78%, and 96% respectively on that subset. For the automated contours, the models agreed on 84% of the lesions and for all of the patients. The AUC, specificity, and sensitivity were 0.96, 89%, and 94% (0.91, 59%, 98%) per lesion (per patient), respectively. For results achieved through other combinations we refer to Table 3.

4. Discussion

In this study, we aimed to build and validate a workflow which would find suspicious lesions within CEM images and give a classification score using HR and DL-models. Additionally, we assessed the added value of CF and HR to classify the manual delineation and the auto-delineations. Our model found 90% of the lesions on the external validation dataset while correctly identifying 99% of patients with lesions. For the classification of lesions, and for most performance evaluation measures, the combination of HR and DL provided the best results on the manual delineations (AUC of 0.88), as well as on the DL-generated contours (AUC of 0.95). Hence, we concluded that our identification and classification model performed at the level to make it potentially generalizable.

This is to our knowledge the first study to provide a full workflow for identification, segmentation, and classification of suspicious lesions in CEM, and to compare the results between HR and DL-models. A similar study was done on FFDM using DL only, reporting a sensitivity of 90% and a false positive rate of 30% for identification of malignant lesions (20) which is similar to our study. It is important to note that imaging modalities (CEM and FFDM) are not directly comparable, and that our model also provided automatically generated delineations of the lesions.

Perek et al. proposed multi-model classification methods that combined a neural classifier with the Breast Imaging-Reporting And Data System classification on CEM, reaching a specificity of 66% for a sensitivity set at 100% (12). In the study by Wang et al. (21) the authors showed that a HR model extracted from the high-energy contrast image or the combinations of all CEM images obtained the highest performance with an AUC of 0.89 on the testing dataset, significantly better than using the LE contrast (which is generally accepted to be roughly equivalent to FFDM (22)) which achieved an AUC of 0.87. Although these studies showed promising results for automatically classifying benign and malignant lesions on CEM using ML approaches, they suffered from relatively small datasets and lacked external validation, which doesn't allow the reader to conclude that their model would obtain a similar performance on a dataset acquired externally. Our study is notable for its large training dataset, the validation of the model on an external dataset, and the combination of HR and DL.

Regarding the classification performance obtained by our best-performing model on manual contours, the AUC value is comparable to those obtained by radiologists across multiple studies. In the meta-analysis by Suter et al. (23), the (pooled) AUC for suspicious findings was 0.89 similar to the result obtained by our model based on the combination of DL and HR (AUC 0.88). In fact, for the cases in which our models are most certain, i.e. for which the DL and HR models agree, we can report an AUC of 0.95 (0.93) per-lesion (per patient) for the manual contours and 0.96 (0.91) per-contour (per patient) for the automated contours.

One notable limitation of this study was that the models were not optimized to identify calcifications, which do not always enhance on CEM. Most false negatives were a result of this limitation. In literature, contradictory results regarding the benefit of CEM compared to FFDM for the classification of calcifications by radiologists have been reported. It is also

possible that the resolution of the images is too low for our model to identify certain calcifications. A solution to this problem could be to combine the model for the identification and classification of lesion on CEM with a different model which would specifically target calcifications using FFDM (or the LE images) only. Moreover we only evaluated the delineations of lesions for which we have either a biopsy or prolonged follow-up, but it is theoretically possible that other benign lesions of no clinical importance were present on the image, potentially making false-positive identifications actually true-positives. The identification algorithm was not tested on images which didn't contain lesions, as the CEM scans were acquired to identify suspicious lesions already spotted during screening. A follow-up study in which FFDM is replaced by CEM might be interesting to conduct in order to compare the performance of those systems and to test our algorithm's capacity to detect lesions. Another limitation is that the contours and the evaluation of the models were made by a single certified radiologist with thirteen years of reviewing CEM experience. A consensus between breast radiologists would be preferable to limit bias.

As a future perspective, it could be interesting to test whether the performance of our model would vary in different breast density categories, as increased breast density is linked to poorer sensitivity within mammography (3). However, the systems used in this study are not yet equipped with automated breast density measurement software, nor are other available tools validated for use in CEM. Hence, we were not able to present our results per breast density category. To confirm our findings and support the utility of our model, a clinical trial should be established to evaluate results in daily clinical routine, such as described in (4) or in silico trial such as in (24). Data of every patient who underwent CEM could be collected and our model should serve as second reader, identifying potential lesions and giving a classification per-contour followed by a confidence score. The added value of the model could then be evaluated in different configuration: for triage, were the cases labeled as malignant with a low confidence score would be reavaluated first by radiologists; or as second reader were a first reader would evaluate the image and only if the model would disagree with the first reader, the case would be send to a second reader. Our current results need to be interpreted with study limitations mentioned above in mind. In conclusion, our automated identification and delineation tool was able to identify the vast majority of suspicious lesions seen on CEM, thus obtaining good performance for finding malignant lesions.

Supplemental Materials

Supplemental figures

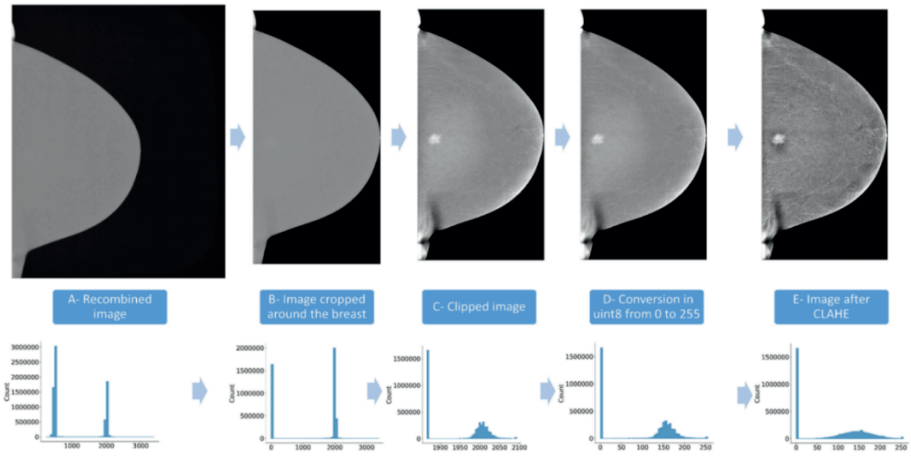


Figure 1: pre-processing steps illustrated with the example of a recombined scan left craniocaudal view from a standard CEM acquisition. The higher row displays the images after the pre-processing stepped indicated in the text below every images and the lower row represent the histograms of the pixel distribution: the y-axis is the pixel counts and the x-axis the pixel values (A) represent the information of the recombined image before pre-processing; the pixel value range is high, between 0 and 3000 (B) the image was cropped around the breast and the background pixels were set at 0 (C) the extreme pixel values within the breast tissue were removed: the values below the 1st percentile of the pixel distribution were set at the 1st percentile; the values above the 99th percentile of the pixel distribution were set at the 99th percentile, reducing the pixel value range drastically from 1800 to 2100 (D) the pixels were resampled to values ranging from 0 to 255 to convert the image to uint8 (E) a CLAHE filter was applied to the image.

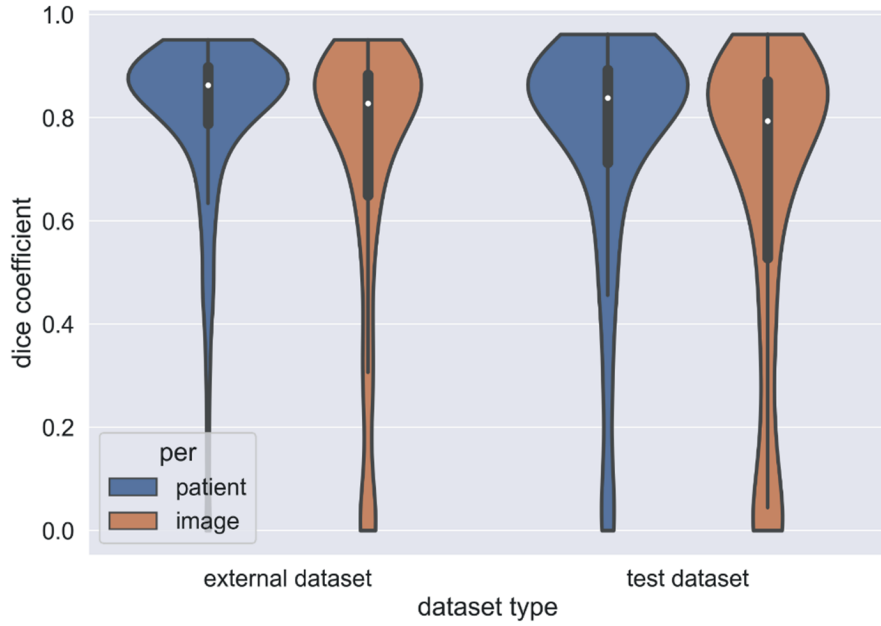


Figure 2: violin plots of the dice coefficients on the external and test datasets per image and per patient; y-axis displays the dice coefficient between 0 and 1 and x-axis represents the dataset type (external or test dataset)

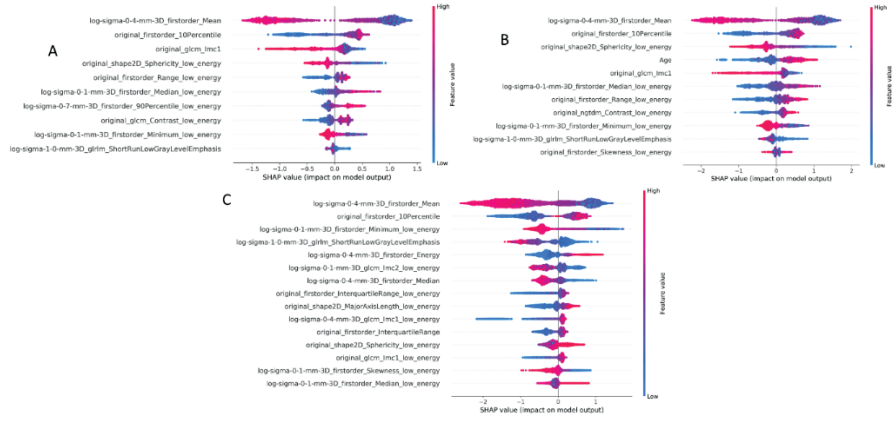


Figure 3: summary plot of the SHAP values calculated on the training dataset (A) for the classification benign versus malignant on the manual delineations with radiomics, (B) for the classification benign versus malignant on the manual delineations with radiomics and clinical features (C) for the classification malignant versus benign and false positive on the automatic delineations with radiomics

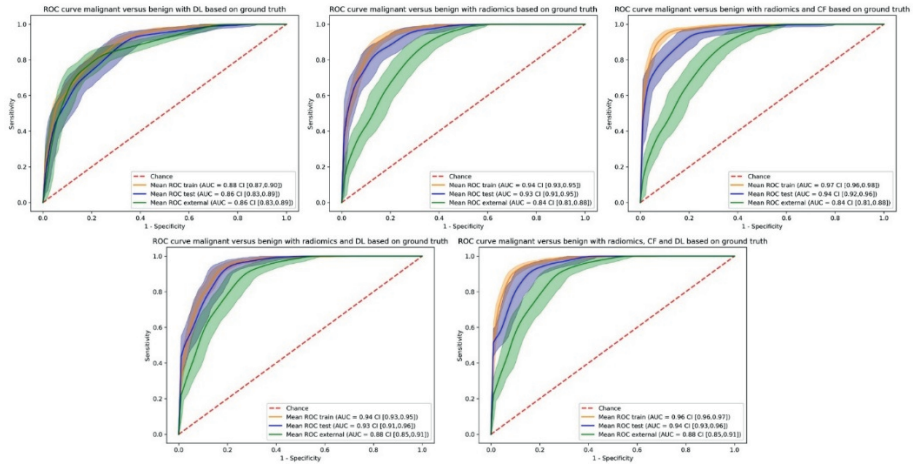


Figure 4: ROC curves for the predictions on the train/test/external datasets for prediction benign/malignant per image based on the ground truth with the five classifiers; abbreviations: AUC= area under the ROC curve, CF = clinical features, CI= confidence interval, DL=deep learning, ROC = receiver operating characteristic

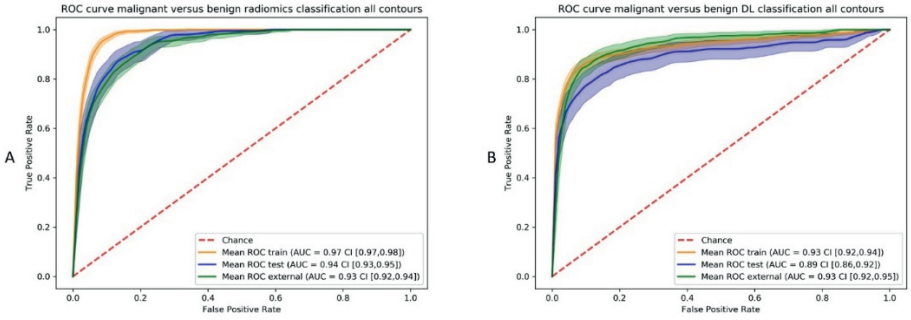


Figure 5: ROC curves for the predictions on the train/test/external datasets for prediction other/malignant per image based on the automatic contours with the two classifiers (A) radiomics based model, (B) deep learning. Abbreviations: AUC= area under the ROC curve, CI= confidence interval, DL=deep learning, ROC = receiver operating characteristic

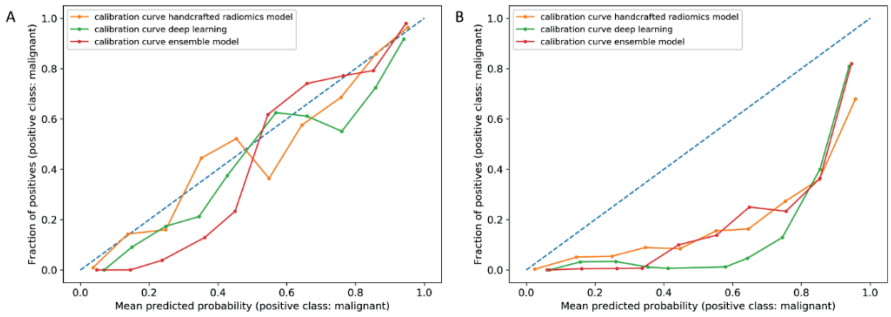


Figure 6: calibration curves performed on the test dataset for (A) the manual contours and (B) the automatically generated contours, drawn for 10 bins. We observe that the handcrafted radiomics (orange curve) predictions and the deep learning (green curve) predictions have similar calibration curves thus can be averaged to form a new ensemble model (red curves). The models based on the manual contours in (A) are well calibrated but we observe that the models based on the automatically generated contours in (B) are uncalibrated with over-confident calibration curves. This is due to the large imbalance in the datasets.

Supplemental tables

Table 1: python packages used and their versions

purpose	package	version
Pre-processing	os	-
	numpy	1.19.2
	pandas	0.24.2
	opencv	4.1.0.25
	scikit-image	0.18.1
	scipy	1.5.2
	simpleITK	1.2.0
Deep learning	keras	2.2.4
	keras-maskrcnn	0.2.2
	keras-retinanet	0.5.1
	tensorflow-gpu	1.13.1
Feature processing and calculation	pyradiomics	3.0.1
	missingpy	0.2.0
	scikit-learn	0.20.3
	statsmodels	0.9.0
	shap	0.39.0
Machine learning	xgboost	1.4.2
Visualisation	matplotlib	3.0.3
	seaborn	0.11.1

Table 2: XGBoost characteristics per dataset type

parameters/data	radiomics on manual delineation	radiomics and clinical features on manual delineations	radiomics on automated contours
gamma	0,5	0,5	0,6
learning rate	0,01	0,01	0,01
maximum depth	3	3	3
minimum child weight	1	5	1
number of estimators	340	890	1000
number of features selected	10	11	15

Table 3: accuracies and F1-scores on the external dataset with confidence intervals for the different model combinations on the manual contours and the automatically generated contours

on the manual contours					
per contour			per patient		
approaches	**accuracy**	**F1-score**	**approaches**	**accuracy**	**F1-score**
radiomics	424/590 (71 %) CI [68,75]	74 % CI [71,78]	radiomics	196/279 (70 %) CI [65,76]	73 % CI [67,79]
deep learning	472/590 (80 %) CI [77,83]	83 % CI [81,86]	deep learning	233/279 (83 %) CI [79,87]	87 % CI [83,91]
radiomics + CF	442/590 (74 %) CI [72,78]	78 % CI [75,82]	radiomics + CF	198/279 (70 %) CI [66,76]	74 % CI [69,80]
radiomics + deep learning	479/590 (81 %) CI [78,84]	84 % CI [81,87]	radiomics + deep learning	221/279 (79 %) CI [75,84]	83 % CI [79,87]
deep learning + radiomics + CF	468/590 (79 %) CI [76,83]	82 % CI [79,86]	deep learning + radiomics + CF	218/279 (78 %) CI [73,83]	82 % CI [78,86]
agreed labels	407/448 (90 %) CI [88,94]	93 % CI [91,95]	agreed labels	232/258 (89 %) CI [86,93]	92 % CI [89,95]
on the predicted contours					
per contour			per patient		
approaches	**accuracy**	**F1-score**	**approaches**	**accuracy**	**F1-score**
radiomics	2342/ 2816 (83 %) CI [82,85]	57 % CI [53,61]	radiomics	204/245 (83 %) CI [78,88]	88 % CI [84,91]
deep learning	2362/ 2816 (83 %) CI [83,85]	58 % CI [55,62]	deep learning	205/245 (83 %) CI [79,88]	89 % CI [86,92]
deep learning + radiomics	2423/ 2816 (86 %) CI [85,87]	61 % CI [58,65]	deep learning + radiomics	212/245 (86 %) CI [82,91] *	91 % CI [88,94] *
agreed labels	2125/ 2362 (89 %) CI [89,91]	71 % CI [68,75]	agreed labels	212/245 (86 %) CI [82,91] *	91 % CI [88,94] *

Footnote: 95% CIs were reported within brackets; CF: clinical features; CI: confidence interval; * = not significantly different

References

1. Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, Nakamura S. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts. Breast Cancer 2017;24(1):104-110. doi: 10.1007/s12282-016-0681-8

2. Zeeshan M, Salam B, Khalid QSB, Alam S, Sayani R. Diagnostic Accuracy of Digital Mammography in the Detection of Breast Cancer. Cureus 2018;10(4):e2448-e2448. doi: 10.7759/cureus.2448

3. Cozzi A, Magni V, Zanardo M, Schiaffino S, Sardanelli F. Contrast-enhanced Mammography: A Systematic Review and Meta-Analysis of Diagnostic Performance. Radiology 2022;302(3):568-581. doi: 10.1148/radiol.211412

4. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89-94. doi: 10.1038/s41586-019-1799-6

5. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, Aerts HJWL. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 2012;48(4):441-446. doi: https://doi.org/10.1016/j.ejca.2011.11.036

6. Zwanenburg A, Vallières M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit MC, Dinapoli N, Dinh CV, Echegaray S, El Naqa I, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Socarras Fernandez J, Spezi E, Steenbakkers R, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, van Dijk LV, van Griethuysen J, van Velden FHP, Whybra P, Richter C, Löck S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295(2):328-338. doi: 10.1148/radiol.2020191145

7. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. Seminars in Cancer Biology 2021;72:238-250. doi: https://doi.org/10.1016/j.semcancer.2020.04.002

8. Baccouche A, Garcia-Zapirain B, Castillo Olea C, Elmaghraby AS. Connected-UNets: a deep learning architecture for breast mass segmentation. npj Breast Cancer 2021;7(1):151. doi: 10.1038/s41523-021-00358-x

9. Ueda D, Yamamoto A, Onoda N, Takashima T, Noda S, Kashiwagi S, Morisaki T, Fukumoto S, Shiba M, Morimura M, Shimono T, Kageyama K, Tatekawa H, Murai K, Honjo T, Shimazaki A, Kabata D, Miki Y. Development and validation of a deep learning model for detection of breast cancers in mammography from multi-institutional datasets. PLOS ONE 2022;17(3):e0265751. doi: 10.1371/journal.pone.0265751

10. Refaee T, Salahuddin Z, Frix A-N, Yan C, Wu G, Woodruff H, Gietema H, Meunier P, Louis R, GUIOT J, Lambin P. Diagnosis of Idiopathic Pulmonary Fibrosis in HRCT Scans using a

combination of Handcrafted Radiomics and Deep Learning. Frontiers in Medicine 2022. doi: https://doi.org/10.3389/fmed.2022.915243
11. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv preprint arXiv:200304696 2020.
12. Perek S, Kiryati N, Zimmerman-Moreno G, Sklair-Levy M, Konen E, Mayer A. Classification of contrast-enhanced spectral mammography (CESM) images. International Journal of Computer Assisted Radiology and Surgery 2018. doi: 10.1007/s11548-018-1876-6
13. Reza AM. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. Journal of VLSI signal processing systems for signal, image and video technology 2004;38(1):35-44. doi: 10.1023/B:VLSI.0000028532.53893.82
14. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision2017; p. 2980-2988.
15. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. European conference on computer vision: Springer, 2014; p. 740-755.
16. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014.
17. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin J-C, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research 2017;77(21):e104-e107. doi: 10.1158/0008-5472.Can-17-0339
18. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2011;28(1):112-118. doi: 10.1093/bioinformatics/btr597
19. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32-35. doi: https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
20. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports 2018;8(1):4165. doi: 10.1038/s41598-018-22437-z
21. Wang S, Mao N, Duan S, Li Q, Li R, Jiang T, Wang Z, Xie H, Gu Y. Radiomic Analysis of Contrast-Enhanced Mammography With Different Image Types: Classification of Breast Lesions. Frontiers in oncology 2021;11:600546-600546. doi: 10.3389/fonc.2021.600546
22. Lalji UC, Jeukens CRLPN, Houben I, Nelemans PJ, van Engen RE, van Wylick E, Beets-Tan RGH, Wildberger JE, Paulis LE, Lobbes MBI. Evaluation of low-energy contrast-enhanced spectral mammography images by comparing them to full-field digital mammography using EUREF image quality criteria. European Radiology 2015;25(10):2813-2820. doi: 10.1007/s00330-015-3695-2
23. Suter MB, Pesapane F, Agazzi GM, Gagliardi T, Nigro O, Bozzini A, Priolo F, Penco S, Cassano E, Chini C, Squizzato A. Diagnostic accuracy of contrast-enhanced spectral mammography for breast lesions: A systematic review and meta-analysis. The Breast 2020;53:8-17. doi: https://doi.org/10.1016/j.breast.2020.06.005
24. Primakov SP, Ibrahim A, van Timmeren JE, Wu G, Keek SA, Beuque M, Granzier RWY, Lavrova E, Scrivener M, Sanduleanu S, Kayan E, Halilaj I, Lenaers A, Wu J, Monshouwer R, Geets X, Gietema HA, Hendriks LEL, Morin O, Jochems A, Woodruff HC, Lambin P. Automated detection and segmentation of non-small cell lung cancer computed

tomography images. Nature Communications 2022;13(1):3423. doi: 10.1038/s41467-022-30841-3

7

Chapter 7

Discussion

Handcrafted feature-based machine learning (ML) models and deep learning (DL) models that automatically classify medical imaging datasets are increasingly published in the literature in recent years. Few studies have been conducted to compare and/or combine these methods, both of which have different strengths and weaknesses. The over-arching goal of the research presented in this thesis was to assess whether the combined use of handcrafted feature-based ML and DL classification models results in better performance compared to the use of each method alone. In this thesis, we hypothesised that **feature-based ML models and DL models capture information from medical imaging datasets which is complementary and that their combined use results in more accurate classification predictions.** In this Chapter, the individual studies presented in this thesis are summarized and discussed. First, we discuss the comparison and combination of deep learning and feature-based models. Second, we discuss the evaluation of feature-based models used to augment deep learning predictions. Third, we highlight the necessary steps to introduce ML studies into clinical workflows comparing our results to the current state of the art. Finally, we discuss future prospects suggesting what progress still needs to happen in the field of precision medicine to have a complete, sustainable, and generalizable workflow.

In **Chapter 2**, we analysed the use of feature-based ML models (also referred to as radiomics in our context) as well as DL methods for quantitative image analyses, highlighting their strengths and weaknesses. We hypothesised that DL has the potential to complement a radiomic workflow by detecting and segmenting region of interests (ROIs) within medical images and that both, radiomics and DL, could be used individually or in combination to classify these ROIs. However, we observed that the analysis with both techniques was suffering from a lack of stability and reproducibility, which could be solved through image homogenization of the datasets, removing differences due to different acquisition parameters and vendors/scanners for radiology images (1) and for histopathology data (2). Such harmonization should ultimately lead to more robust and generalizable models. Furthermore, for radiomics in particular, we realised that a robust feature selection method, standardization and harmonization of the features (as suggested by the image biomarker standardisation initiative (IBSI) (3) and ComBat (4)) are important areas for further studies to improve the stability and reproducibility of those features.

Comparing and combining deep learning and feature-based models

In **Chapter 3**, we analysed the capability of mass spectrometry imaging (MSI) data and data obtained from haematoxylin & eosin (H&E) stained histological slides to automatically identify patients with Barrett's oesophagus and to predict disease progression in patients with low grade dysplasia. The datasets were co-registered and the H&E data corresponding to the resolution of the MSI acquisition was extracted at a resolution of 96x96 pixels, resulting in a new dataset consisting of 144 823 tiles containing H&E data and MSI data. To our knowledge, this is the first study that compared the predictive value of H&E and MSI in parallel. This study showed that the model based on H&E data was better at differentiating

epithelial tissue from stroma than MSI data, whereas MSI data was more suitable for predicting the grade of dysplasia per tile and predicting disease progression for patients with low grade dysplasia in Barrett's oesophagus. Combining the predictions of the models based on the two datasets and obtained for the same tasks didn't improve the results per task. Thus, the two datasets might contain complementary information which would be better suitable for different but complementary tasks.
Stroma and epithelial tissue have very different histological appearances and are relatively easy to distinguish for a trained non-expert (e.g. non-pathologist) and therefore, probably also relatively easy to be learned by a DL model from H&E. However, grading dysplasia on H&E stained tissue sections can be very difficult for a pathologist looking at the whole slide and is even more difficult if only a small fraction of the images (H&E tile) is available for review due to the lack of surrounding context which pathologists regularly use to make a decision. On the other hand, the MSI data might contain molecular information specific to the respective dysplasia grade. A limitation of this study was that in order to superimpose and compare the predictions of the two datasets, the size of the H&E images had to be adapted to match the raster size of the MSI acquisitions. There is literature which shows that model prediction performance may depend greatly on the zoom level of the H&E images and the spatial resolution used (5). Thus, we could potentially have obtained better results by implementing a different pre-processing strategy which would have allowed us to use H&E images at different spatial resolution. To do this we could have, for example, averaged neighbouring MSI tiles and try to train models on larger H&E images. However, this would have been beyond the aim of the current study which was a proof of concept study to demonstrate that these two very different datasets can be analysed jointly using ML and DL methodology. In a recent study, Faghani et al. were able to predict the grade of dysplasia on H&E with high performance choosing a larger tile size in a dataset from 542 patients with Barrett's oesophagus (6). The relatively small number of patients was another limitation of our study. Although we had more than 100,000 tiles available for training and testing, this data originated from a total of 57 patients only, which limited our conclusions for disease prediction at patient level to "potentially predictive" or "random predictions". Furthermore, data from an independent study cohort to validate the results was not available to us. Identification of the actual peptide/protein from the mass-to-charge ratio values of the MSI analysis was not performed in the current study, but would be of clinical interest. Such identification may allow for a more specific and precise test to be developed to replace MS acquisition which is costly and time consuming.

In **Chapter 4**, we compared and combined a radiomics based model and a DL model for predicting adverse radiation effects (ARE) using a pre-treatment brain magnetic resonance imaging (MRI) dataset from patients with brain metastasis who underwent radiotherapy. Different image pre-processing techniques were studied to improve model performance. For this, we tested white-stripe, z-score and CLAHE normalizations, inspired by the guidelines provided in (1). Our results showed that the predictions obtained with a combined radiomics based and DL based model had a better performance than predictions based on each model alone. To the best of our knowledge, this was the first study using pre-

treatment brain MRI images to predict the risk of ARE suggesting that the use of radiomics together with DL predictions may provide better and more stable results. Interestingly, the performance obtained in this study in pre-treatment MRI images was similar to that obtained from classifying ARE versus tumour in MRI images after stereotactic radiotherapy: in our study we obtained an AUC of 0.71 CI [0.60,0.82] compared to AUC of 0.73 obtained by Zhang et al. using delta radiomics on T1 and T2 MRIs after stereotactic radiotherapy (7). However, the prediction score reached in our study is still too low to be used confidently for treatment planning and requires further study with a dataset containing more positive examples as one of the limitation of our study was that the dataset was highly imbalanced (about 10% of the scans were from patients with ARE).

Evaluation of feature-based models to augment deep learning predictions

In **Chapter 5**, we evaluated whether the use of a ML based model built on the predictions of a DL model can improve performance when trying to identify digital H&E stained slides that contain lymph nodes (LN) and subsequently segment them within a large H&E dataset from resection specimens from patients with oesophageal cancer. We compared our results to the conventional U-Net model approach and found that the accuracy of our model was better than that using conventional U-net model approaches. Moreover, our method allowed us to obtain a probability score per potential LN found, allowing the creation of an "uncertain" class for predicted contours, where the model cannot give a prediction whether the candidate contour is a LN or not. This addition of an "uncertain" class allowed us to reduce the false positive findings, only 6% of the dataset would be classified as "uncertain" requiring checking by the pathologist. Our results suggest that there is utility in combining a ML model with a DL model (a) to increase model performance and (b) to enable the addition of an uncertainty score, allowing the pathologist to decide which contours to review as part of the quality control measures based on a pre-defined range of certainty values.

Although we applied extensive pre-processing to the H&E data to reduce or eliminate the inherent variability of the H&E staining, our results indicate that further data homogenizing is necessary to obtain more consistent results between datasets originating from different centres. In our case, the differences in prediction accuracy might also be related to the fact that patients from our external dataset received different treatment (some received pre-operative chemotherapy whereas others did not), or to the imbalance between the number of images containing LNs and the number of images not containing LNs being different in the two datasets. Thus, more training data from other centres might improve the detection and segmentation performance of our model for future use and to validate the findings presented in the study of Kloft et al., which suggested that LN size correlates with prognosis (8).

In **Chapter 6**, we implemented a DL model which takes as input pre-processed contrast enhanced mammograms (CEM) containing a suspicious mass, returning as output predictions on lesion location, lesion contour and a label differentiating between "benign" and "malignant". In parallel, we implemented a radiomics-based model using the contours made by a radiologist to predict malignancy of the lesions, comparing and combining the predictions obtained from different models. We also implemented a radiomics-based model based on the predicted contours obtained by the DL model and compared and combined the scores obtained. This is to our knowledge the first study to provide such workflow for CEM. We observed that for classifications made on the ground truth contours from the radiologist or on the predicted contours, a combination of radiomics and DL results obtained the best performance. Our results confirm those from a previous study, where the combination of radiomics and DL increased the performance of a classifier in full-field digital mammography (FFDM) images (9). However, the previous study used a different approach extracting DL features from one of the fully connected layer and combining it later with clinical and radiomics features to train a ML model which classified malignant and benign regions. Compared to our study, the previous study did not implement a model to detect and segment the lesions within the images, leaving this part of the work to radiologists. In our study, we also identified the location of the suspicious lesion in the CEM scan. Our approach of identifying the suspicious lesion within the image, segmenting it and classifying it into benign/malignant could be transferred to any medical imaging data.
To better interpret the results of the model, we would propose a follow-up study in which we would create an uncertain class such as introduced in Chapter 5 and thus limit the number of (presumed) false positive findings while maximising the identification of malignant lesions. Interestingly, we noticed that our model had problems with detecting calcifications in particular, most likely due to their size and difference in shape compared to other lesions. Thus, it could be clinically helpful to train a specific model to recognize calcifications which could later be combined with our model.

Steps necessary to translate research into practice

Removing heterogeneity of the datasets

Pre-processing of the images appears sometimes insufficient to harmonize datasets and to obtain the same performance on an external validation dataset than on the dataset used to train a model. We only partially worked on this issue due to time constraint as such a study would have required the setting up of a multi-centre study, data collection after acquisition of the same image modality used to train a model, and re-evaluation of the model on all the collected datasets to assess whether the harmonization method was generalizable. This topic was explored in depth by Castro et al. (10), highlighting why models trained on certain dataset might have disappointing performance on a new dataset. One of the proposed reasons was 'dataset shift' (e.g. changes in data distribution) which might explain why the model we developed in Chapter 5 underperformed on the external dataset. To overcome this potential issue, one solution could be to remove the scanner or acquisition information

from the dataset by training a model to unlearn confounding parameters as described in (11). In this article, the authors used brain MRI data and implemented a method to make feature space invariant on three different tasks: regression, classification, and segmentation. They trained their models to remove scanner bias and found that unlearning those confounders lead to better results. Whilst we considered utilizing the scan information to standardize directly MRI scans in Chapter 4 with Piecewise Linear Histogram Matching or Nyul normalization (12), unfortunately, the values of the scanner parameters were not available to us. Unlearning confounders with a DL model has similar limitations: it also requires scanner information which is not always available in the datasets and the larger the datasets, the higher the chance that there will be missing values. Moreover, this approach might also lead to cancelling biological information which might have a predictive value for the model. Thus, further work on data collection from multiple centres is necessary to analyse and remove confounders in medical imaging datasets to avoid overfitting of ML models and allow reproducibility of the results on data from different origins.

Increasing interpretability towards trustworthy AI to facilitate clinical adoption

It can be difficult to understand why a machine-learning model made a particular decision on certain input data, specifically to understand when the model makes false predictions or classifications. To facilitate the adoption of ML models in clinical practice, there is a need to explain how the model works to allow verification that a decision was taken based on logical parameters, as we discussed in Chapter 2. In this context, Salahuddin et al. (13) distinguishes between global and local interpretability methods to potentially help clinicians understand and accept the output of a ML model. In this thesis, we tried to provide global interpretability information for all our studies. To this effect, we provided feature importances when training a ML model, to help the reader understand which features were most relevant for the task in Chapter 3, 4 and 5. In Chapter 6, we implemented summary plots of the SHapley Additive exPlanation (SHAP) values to increase understanding on how the different values affect the decision taken by the algorithm. We didn't present interpretability methods on individual outcome in this thesis, it would be another research topic for clinical implementation, but ideally, local interpretability should be implemented for all classification predictions, to explain the output of a model to a clinician, such as presented in the work of Barnett et al. (14).

Uncertainty predictions and outlier detection

We believe that uncertainty predictions are necessary to interpret ML model results, as described by Kompa et al. (15). The authors strongly encourage researchers to always include uncertainty predictions when realizing a new AI model based on medical imaging. This would make the model more trustworthy and help spot dataset shifts. Moreover, uncertainty prediction could be used for triage, as the clinicians could focus on the patient data with strong positive certainty, to rapidly make a decision while patient predictions with

uncertain label would need to be further evaluated, potentially initiating further tests to establish a diagnosis/prognosis. Chapter 5 demonstrates a possible approach to uncertainty prediction, using handcrafted features on U-Net output to establish a confidence score and determining whether the predicted contours were correct. The score allowed us to create an uncertain category needed to be reviewed by pathologists, which would reduce the amount of false positive detections. Another possible approach would have been to train a model to output uncertainty predictions in parallel with classification/detection predictions such as presented in (16).

Clinical trial and real world implementation

As discussed in this thesis, generalizability of ML models is an issue faced by many studies. Performance during training might outperform clinicians or achieve similar results but often the results are worse when used in a clinical setting. In this thesis, we tested the performance of our models in an unbiased fashion using leave-one-out cross validation (see Chapter 2) or validating the model on retrospectively collected data from external centres (see Chapters 3, 4, 5 and 6). This still may not be sufficient to predict results in a prospective setting where data can be more heterogeneous, substantially more imbalanced and where unknown implementation challenges might arise. Thus implementation of a clinical trial before integration of the models in the clinic is recommended. Guidelines have been established for clinical trials specifically for AI applications such as SPIRIT-AI (for writing clinical trial protocols) (17) and CONSORT-AI (for randomized controlled trials) (18). Moreover, software/models are considered medical devices and therefore require CE-marking and FDA or EMA approval, sometime classified as high risk (MDR class III) (19).

Sometimes, research seems to focus on the technical challenge of a problem and fails to report on clinical usefulness. For example, reporting results per patient and not per data point or misusing Dice metrics by reporting the overall result per image while the image contained multiple objects which needed to be segmented. The Metrics Reloaded Delphi Consortium introduced a framework to help researchers defining the problem they want to address, called “problem fingerprint” (20). After defining the problem and its scope, the correct corresponding metric and its application can be chosen from their metric proposal based on how the performance of those models should be reported.

We decided to follow this framework in this thesis and aimed to not only report the results per sample, but more generally report usable outcome variables. For example, in Chapter 3 we reported the grade of dysplasia prediction on a tile level, we showed the classification predictions on a full H&E image and we predicted progression of low grade dysplasia at patient level. In Chapter 4, we used the maximum risk predicted of developing ARE within the patient scan to determine the risk per patient to enable patient treatment decisions. In Chapter 5, we reported the Dice coefficient not only per image but also per single contour found within the H&E image, making the results more transparents, and also per patient because we needed to know if at least one image with LNs was correctly found per patient to make it clinically useful. In Chapter 6, we reported the results per region per image and

per patient which is the more relevant as a mass might not be found on one of the image from one patient, but might be found in another image of the same patient.
Not only the problem-statement and metrics need to be correctly defined, but the solution needs to be usable in a research setting and/or in the clinic. To respond to this challenge, (21) and (22) strongly recommend to involve the potential users (i.e. the clinicians) in all phases of the development of the model, making the AI applications user-centred. This is why in all of our studies we involved clinicians and made sure that our studies answered unmet clinical needs. This means to first ask the clinicians what they need, then constantly reassess usefulness of the model developed during the development phase, and ask how the clinicians would like to interact with the application to implement it efficiently. Later on, clear documentation needs to be available at all times for the clinicians to understand how and for what purpose the model can and cannot be used.

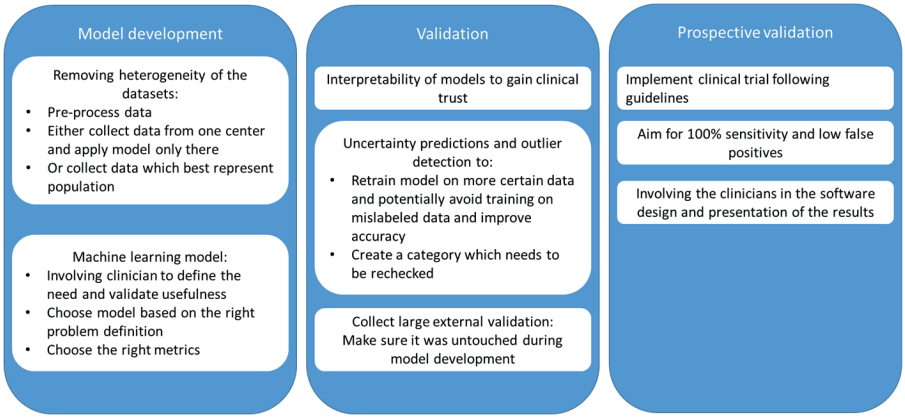


Figure 1: lessons learned from our studies - workflow to bring a ML model to the clinic

Future prospects: improving the predictions, replacing the empiric tools and predicting the future

Personalized medicine based on multi-modal datasets

The aim of personalized medicine is to provide the optimal treatment for the individual patient based on all available data. Lot of models were proposed in the literature to stratify risk, predict survival or predict treatment response based on feature-based models or DL models. Due to insufficient performance or lack of validation, only few of those models were implemented in the clinic (23). Models developed on multimodal datasets might be a solution to increase translation into the clinic. Data collected per patient may not be limited to one modality as we saw in Chapter 1 and this collection of data per patient should be considered as a whole for ML studies, although currently those datasets are most often used independently in unimodal studies (24). We were able to integrate patient characteristics in our models in Chapters 4 and 6 and saw that the performance of the model improved when using this data for patient classification in Chapter 4, although it didn't help

our model in Chapter 6, probably because the patient characteristics available in our study were not relevant for the question e.g. to predict benign or malignant status of a suspicious lesion. Moreover, we saw in Chapter 3 that multimodal datasets could be useful for different prediction tasks into one workflow but it is also possible that the combination of predictions for one task obtained on multimodal datasets would withhold better, more robust predictive power. Indeed in the work of (25), the authors found that merging histopathology, radiology and clinicogenomic models in one model was better in stratifying risk in high-grade serous ovarian cancer patients than each model by itself. We know that for CEM, a prior FFDM was acquired. Although FFDM data was not collected in our study, a follow-up work could include it and have a model similar to the one developed Chapter 6 applied to the FFDM and which could be combined with the results obtained on both FFDM and CEM data.

Training update strategies

We noticed in Chapter 5 that changes in data distribution can be an issue for keeping the model at a constantly good performance. As mentioned in Chapter 5, strategies need to be implemented to retrain the models and keep them up to date, without compromising patients' privacy rights. One way to achieve this would be to use federated learning, which would allow the models to be retrained without a company/researcher having access to the data directly e.g. the patient data never leave the hospital environment (26). Distributed learning is also a possible option and differ from federated learning by the fact that only one model is trained rather than assembling multiple models trained for each different location. Once the models would be validated and implemented in the clinic, they would still need to be updated to preserve performance when there is a change of acquisition parameters for the imaging data. In the article of Perkonigg et al. (27), the authors suggests to adopt a continual learning approach with dynamic memory to preserve the performance of the model trained on previous dataset while adapting to new dataset which might be acquired with different acquisition parameter or are coming from new scanners.

Universal models

Ideally, ML models need to be trained on a large database (as a rule of thumb, more than hundred samples) composed of data from multiple centre worldwide. To improve the results obtained in Chapter 6, we could imagine implementing a workflow such as in (28), where the authors pre-trained classifiers on ImageNet (29) or JFT-300M (30), and then retrained those models on a large chest radiographs database formed of 5 datasets with a total of 821 544 chest radiographs using supervised contrastive learning. Then the authors reused the models to perform different classification tasks, freezing the models and training only the last layers. The authors obtained good classification results using small dataset and when using large dataset the models outperformed state of the art. Instead of chest X-rays, we could pre-train a model with large FFDM dataset such as the one described in (31) where the authors collected 1 001 093 images. For histopathology, the Bigpicture project

(https://bigpicture.eu) aims to collect large amount of data to develop AI tools, which could be used similarly to improve the performance of our models in Chapter 3 and 5.

Conclusion

In the studies presented in this thesis, we implemented different workflows using different imaging modalities, clinical questions, and different combination of ML and DL models. Those studies aimed to be reproducible by establishing a common pre-processing strategy and to be validated either using cross-validation or an external validation dataset provided by another institute. We demonstrated that using feature based models in combination with DL models can make stronger predictions, with different but complementary information being extracted with two methods and combining the information of both models to yield the best results. Challenges still remain before sustainable implementation in the clinic: homogenisation of datasets from different source needs to be improved and the limits of the models' applications requires to be well defined and explained. Moreover, we should keep the clinical application of those models in sight and evaluate their impact once implemented in the clinic, adapting and updating them according to the needs of the clinicians.

References

1. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, Choyke PL, Turkbey B. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. J Med Imaging (Bellingham) 2021;8(1):010901. doi: 10.1117/1.Jmi.8.1.010901

2. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst J-M, Ciompi F, van der Laak J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis 2019;58:101544. doi: https://doi.org/10.1016/j.media.2019.101544

3. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit M-C, Dinapoli N, Dinh CV, Echegaray S, Naqa IE, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Fernandez JS, Spezi E, Steenbakkers RJHM, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, Dijk LVv, Griethuysen Jv, Velden FHPv, Whybra P, Richter C, Löck S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295(2):328-338. doi: 10.1148/radiol.2020191145

4. Orlhac F, Eertink JJ, Cottereau A-S, Zijlstra JM, Thieblemont C, Meignan M, Boellaard R, Buvat I. A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. J Nucl Med 2022;63(2):172-179. doi: 10.2967/jnumed.121.262464

5. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal 2018;16:34-42. doi: https://doi.org/10.1016/j.csbj.2018.01.001

6. Faghani S, Codipilly DC, Vogelsang D, Moassefi M, Rouzrokh P, Khosravi B, Agarwal S, Dhaliwal L, Katzka DA, Hagen C, Lewis J, Leggett CL, Erickson BJ, Iyer PG. Development of a Deep Learning Model for the Histological Diagnosis of Dysplasia in Barrett's Esophagus. Gastrointestinal Endoscopy 2022. doi: https://doi.org/10.1016/j.gie.2022.06.013

7. Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, Brown PD, McGovern SL, Guha-Thakurta N, Ferguson SD, Fave X, Zhang L, Mackin D, Court LE, Li J. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. European Radiology 2018;28(6):2255-2263. doi: 10.1007/s00330-017-5154-8

8. Kloft M, Ruisch JE, Raghuram G, Emmerson J, Nankivell M, Cunningham D, Allum WH, Langley RE, Grabsch HI. Prognostic Significance of Negative Lymph Node Long Axis in Esophageal Cancer: Results From the Randomized Controlled UK MRC OE02 Trial. Annals of surgery 2021. doi: 10.1097/sla.0000000000005214

9. Cui Y, Li Y, Xing D, Bai T, Dong J, Zhu J. Improving the prediction of benign or malignant breast masses using a combination of image biomarkers and clinical parameters. Frontiers in Oncology 2021;11:629321.

10. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nature Communications 2020;11(1):3673. doi: 10.1038/s41467-020-17478-w

11. Dinsdale NK, Jenkinson M, Namburete AIL. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. NeuroImage 2021;228:117689. doi: https://doi.org/10.1016/j.neuroimage.2020.117689
12. Nyul LG, Udupa JK, Xuan Z. New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 2000;19(2):143-150. doi: 10.1109/42.836373
13. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine 2022;140:105111. doi: https://doi.org/10.1016/j.compbiomed.2021.105111
14. Barnett A, Schwartz F, Tao C, Chen C, Ren Y, Lo J, Rudin C. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence 2021;3. doi: 10.1038/s42256-021-00423-x
15. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine 2021;4(1):4. doi: 10.1038/s41746-020-00367-3
16. Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Gibson E, Vishwanath RS, Balachandran A, Balter JM, Cao Y, Singh R, Digumarthy SR, Kalra MK, Grbic S, Comaniciu D. Quantifying and leveraging predictive uncertainty for medical image assessment. Medical Image Analysis 2021;68:101855. doi: https://doi.org/10.1016/j.media.2020.101855
17. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, Darzi A, Holmes C, Yau C, Moher D, Ashrafian H, Deeks JJ, Ferrante di Ruffano L, Faes L, Keane PA, Vollmer SJ, Lee AY, Jonas A, Esteva A, Beam AL, Panico MB, Lee CS, Haug C, Kelly CJ, Yau C, Mulrow C, Espinoza C, Fletcher J, Moher D, Paltoo D, Manna E, Price G, Collins GS, Harvey H, Matcham J, Monteiro J, ElZarrad MK, Ferrante di Ruffano L, Oakden-Rayner L, McCradden M, Keane PA, Savage R, Golub R, Sarkar R, Rowley S, The S-A, Group C-AW, Spirit AI, Group C-AS, Spirit AI, Group C-AC. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nature Medicine 2020;26(9):1351-1363. doi: 10.1038/s41591-020-1037-7
18. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, Darzi A, Holmes C, Yau C, Ashrafian H, Deeks JJ, Ferrante di Ruffano L, Faes L, Keane PA, Vollmer SJ, Lee AY, Jonas A, Esteva A, Beam AL, Chan A-W, Panico MB, Lee CS, Haug C, Kelly CJ, Yau C, Mulrow C, Espinoza C, Fletcher J, Paltoo D, Manna E, Price G, Collins GS, Harvey H, Matcham J, Monteiro J, ElZarrad MK, Ferrante di Ruffano L, Oakden-Rayner L, McCradden M, Keane PA, Savage R, Golub R, Sarkar R, Rowley S, The S-A, Group C-AW, Spirit AI, Group C-AS, Spirit AI, Group C-AC. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature Medicine 2020;26(9):1364-1374. doi: 10.1038/s41591-020-1034-x
19. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health 2021;3(3):e195-e203. doi: https://doi.org/10.1016/S2589-7500(20)30292-2
20. Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv preprint arXiv:220601653 2022.

21. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, Aussó S, Alberich LC, Marias K, Tsiknakis M. FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. arXiv preprint arXiv:210909658 2021.
22. Filice RW, Ratwani RM. The Case for User-Centered Artificial Intelligence in Radiology. Radiology: Artificial Intelligence 2020;2(3):e190095. doi: 10.1148/ryai.2020190095
23. Zhu W, Xie L, Han J, Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. Cancers 2020;12(3):603.
24. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer 2022;22(2):114-126. doi: 10.1038/s41568-021-00408-3
25. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, Zamarin D, Roche KL, Liu Y, Patel D, Aukerman A, Pasha A, Rose D, Selenica P, Causa Andrieu PI, Fong C, Capanu M, Reis-Filho JS, Vanguri R, Veeraraghavan H, Gangai N, Sosa R, Leung S, McPherson A, Gao J, Lakhman Y, Shah SP, Consortium MM. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature Cancer 2022;3(6):723-733. doi: 10.1038/s43018-022-00388-9
26. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ. The future of digital health with federated learning. npj Digital Medicine 2020;3(1):119. doi: 10.1038/s41746-020-00323-1
27. Perkonigg M, Hofmanninger J, Herold CJ, Brink JA, Pianykh O, Prosch H, Langs G. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature Communications 2021;12(1):5678. doi: 10.1038/s41467-021-25858-z
28. Sellergren AB, Chen C, Nabulsi Z, Li Y, Maschinot A, Sarna A, Huang J, Lau C, Kalidindi SR, Etemadi M, Garcia-Vicente F, Melnick D, Liu Y, Eswaran K, Tse D, Beladia N, Krishnan D, Shetty S. Simplified Transfer Learning for Chest Radiography Models Using Less Data. Radiology;0(0):212482. doi: 10.1148/radiol.212482
29. Deng J, Dong W, Socher R, Li LJ, Kai L, Li F-F. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition2009; p. 248-255.
30. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. 2017 IEEE International Conference on Computer Vision (ICCV)2017; p. 843-852.
31. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, Jastrzębski S, Févry T, Katsnelson J, Kim E, Wolfson S, Parikh U, Gaddam S, Lin LLY, Ho K, Weinstein JD, Reig B, Gao Y, Toth H, Pysarenko K, Lewin A, Lee J, Airola K, Mema E, Chung S, Hwang E, Samreen N, Kim SG, Heacock L, Moy L, Cho K, Geras KJ. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging 2020;39(4):1184-1194. doi: 10.1109/TMI.2019.2945514

Summary

In 2024 the estimation of patients with cancer reaches 28 million people, an increase of almost 50% compared to the figures from 2020 (GLOBACAN 2020). This will make cancer even more of a burden for society and for healthcare. Moreover, lack of clinicians is already a worldwide issue, making the demand for tools to reduce their workload very high; hence the need to keep improving and developing clinical decision support systems, which is the focus of this thesis.

This thesis consists of two parts, both aiming to explore the combine value of feature-based models and deep learning models for medical image analysis in cancer. The first part investigated the combination of the predictions obtained with feature-based and deep learning models, potentially leading to more accurate and robust frameworks. The second part of this thesis explored the use of feature-based models to augment the predictions of deep-learning models.

Part 1: Comparing and combining deep learning and feature-based machine learning

Radiomics and deep learning are two machine-learning methods which can be used to classify medical images and promising results using these methods have been reported in the literature. In **Chapter 2** we explored the pros and cons of those two methods: compared to deep learning, radiomics can perform better on small datasets (which is usually the case when analyzing medical imaging datasets), but cannot be used to segment data. Furthermore, radiomics requires input from clinicians/experts who need to identify and delineate the data/region of interest by hand and from data scientists who select features based on a preset list. We hypothesized in this thesis that both methods yield complementary information and using them in combination improves the results.

In **Chapter 3**, we analyzed the capability of mass spectrometry imaging (MSI) data and data from haematoxylin & eosin (H&E) stained tissue sections for automatic identification of patients with Barrett's oesophagus and prediction of progression in patients with low grade dysplasia. Due to the differences in the acquisition method of these two datasets, the datasets needed to be homogenized. This study showed that the model based on H&E data alone was better at identifying tissue type and the models based on MSI data alone were more suitable for predicting dysplasia grade for all patients and disease progression in patients with low grade dysplasia.

In **Chapter 4**, we compared and combined a radiomics-based model and a deep learning based model aiming to predict adverse radiation effects (ARE) in a dataset of pre-treatment brain MRI images containing metastasis. In this case, the most efficient pre-processing method was selected independently for the two different models. We observed that the best results on the external dataset were obtained when combining the predictions of the radiomics-based model and the deep learning model. This study suggests that the

predictions of the two models could be used in combination and improve the classification of ARE/none-ARE lesions within brain-MRI compared to using those models independently.

Part 2: Using feature-based models to augment deep learning predictions

In **Chapter 5**, we evaluated wether the use of a machine-learning model based on handcrafted-features in addition to a conventional U-Net model can improve performance when trying to find digital H&E stained slides containing lymph node (LN) and subsequently segment them using a large H&E dataset from patients with oesophageal cancer. We compared our results to the conventional U-Net model approach and found that the accuracy of our model was better than conventional U-net model approaches. Moreover, our method allowed us to obtain a likelihood score per potential LN found, allowing the creation of an "uncertain" class, for which the model cannot provide a prediction whether the candidate contour is a LN or not. The addition of an "uncertain" class allowed us to identify slides, which (for sure) require manual quality control more specifically instead of quality controlling of a random set of few slides. Six percent of the images from the external dataset were classified in the uncertain category and thus would need to be quality checked by a pathologist/expert.

Finally, in **Chapter 6**, we implemented a deep learning model which uses pre-processed data of contrast enhanced mammograms containing a suspicious mass, returning predictions on mass location, contour and a label differentiating between "benign" and "malignant". In parallel, we implemented a radiomics-based model on the contours made by the radiologist to predict the malignancy of the masses, comparing and combining the predictions obtained with the different models. We also implemented a radiomics model based on the predicted contours obtained after prediction by the deep learning model and compared and combined the score obtained there. We observed that for both scenarios (ground truth contours and predicted contours), the combination of radiomics and deep learning results obtained the best performance.

In every study presented in this thesis, we implemented a reproducible workflow by establishing a common pre-processing strategy and validating the models either with cross-validation or with external validation datasets provided by another institute. We conclude from the studies presented in this thesis that using feature-based models (e.g. radiomics) in combination with deep learning models leads to stronger predictions as different but complementary information is extracted and processed by the two methods. Future studies are needed whether decision support models can be further improved by also including patient characteristics as a non-image based dataset in the final model.

Impact paragraph

This thesis investigated the individual and combined use of deep learning and handcrafted radiomics to improve the machine learning model predictions on different types of imaging data. Our first study was a systematic review about the strengths and weaknesses of handcrafted radiomics and deep learning methodologies, which informed the work conducted in this thesis. We were able to draw a number of important conclusions from this review:
(1) Researchers seems to either use handcrafted radiomics or deep learning based models, we were unable to identify a study which used these methods in combination.
(2) Having homogeneous datasets, which are independent from the machine they were acquired on, is very important for reproducibility and successful validation of the results.
(3) In order to increase usuability of our models, assessing performance on at least one external validation dataset and using cross-validation on a test dataset are key to success.

Scientific impacts

In the first part of this thesis, we compared and combined deep learning and feature-based machine learning. From our study, we concluded that the use of two different datasets, one containing histological information and the other one containing molecular information, might have complementary value for the prediction of dysplasia grades in Barrett's oesophagus. The analysis of the images with the molecular information was better in predicting dysplasia grade per image and progression of dysplasia to cancer whereas the images with the histological information were best to classify tissue type. This discovery might help clinicians to improve prediction of progression of dysplasia and hence could improve management of patients with Barrett's oesophagus. In another study on brain MRI, we discovered that combining the predictions of adverse radiation effect obtained with handcrafted features machine learning and deep learning method lead to better predictions than using one model alone. This important information might influence how MRI images are analysed in the future when trying to predict adverse effects of radiation therapy in patients with brain metastases.

In the second part of this thesis, we used feature-based models to augment deep learning predictions. We created a novel workflow based on histology images of lymph nodes and other tissues in order to detect and segment lymph nodes. We used first a deep learning model to segment the lymph nodes and then created a machine-learning model which takes those predicted segmentations as input and outputs a score from 0 to 1 quantifying the likelihood that the segmentation is indeed a lymph node which has been correctly segmented. Adding this step to a regular deep learning approach reduced the false positive results significantly as part of the false positive findings were reclassified in the uncertain category. This work could be used to detect and segment lymph nodes on histology images in a research setting as a prerequisite of performing detailed AI based analyses of the lymph node architecture (data of this study was used to successfully obtain funding for further investigations in lymph nodes). Our last study was performed on contrast enhanced mammography images. We trained a model to detect, segment and classify suspicious

lesions. Interestingly, also in this study, the best results were obtained using a combination of deep learning and handcrafted radiomics features.

We have shown in this thesis for the first time the potential of utilizing the strengths of handcrafted features based models in combination with the strengths of deep learning models for multiple medical imaging datasets and diverse tasks. Although our results are promising and offer insights into new avenues of research, validation of our findings in independent series is required.

Social impacts and knowledge transfer

According to the GLOBACAN 2020, it is expected that 28 million people will be diagnosed with cancer in 2040, which represents an increase of almost 50% compared to the figures from 2020. Cancer is already a leading cause of death in 2022 and will become even more of a burden on society.

Although lot of published studies are presenting tissue-based biomarkers based on medical imaging data, none was introduced into the routine practice. Clinical decisions are still based on a clinician assessing medical images (CT, MRI, X-rays, H&E stained tissue, etc.). This approach is subjective as it depends on the experience of the observer and experts are not always available to review those images. Thus, there is a need for accurate and fast tools to assist and support decisions of the clinicians objectively. We contributed to advance the field of personalized medicine by showing that using handcrafted features in combination with deep learning helps improve predictions eventually for detections, delineation and diagnosis tasks, making another step towards clinical implementation.

We communicated all the results of our research: all our studies are either published or submitted to peer-reviewed journals and made open access. The work presented in this thesis has been presented and discussed at multiple national and international conferences to disseminate our findings with medical imaging experts: Presentations were given at the GROW science day of Maastricht University (2019, 2020), the European Congress on Digital Pathology (2020, 2021), the European Congress of Radiology (2020, 2022) and the conference of the Pathological Society of Great Britain and Ireland (Manchester Pathology) (2021). Moreover, our work presented Chapter 5 is planned to be implemented by our department first for research purpose and if the results show to be consistent, then it will be implemented in the clinic. The model presented Chapter 6 will be made available for use by a company in the next years and could have a positive impact on the clinical workflow. Research is currently being done to improve the detection and diagnosis of micro-calcifications within the breast on contrast-enhanced mammography.

Addendum

Supplementary Table for Chapter 3

Supplementary Table 3. Features importance of random forest, xgboost and the average score based on MSI data for tissue type classification and grading.

Tissue Type Classification				Grade Classification			
m/z	**importances_rf**	**importances_xgboost**	**importances_mean**	**m/z**	**importances_rf**	**importances_xgboost**	**importances_mean**
1139.5610000000001	0,009018	0,028368	0,018693	862.459	0,023959	0,033814	0,028886
1155.5810000000001	0,011386	0,023511	0,017448	948.4939999999999	0,011251	0,026825	0,019038
814.4580000000001	0,017142	0,017239	0,017191	1034.528	0,008707	0,023367	0,016037
1077.565	0,005157	0,010287	0,007722	1303.629	0,005147	0,019402	0,012274
840.443	0,006317	0,008599	0,007458	947.4910000000001	0,009559	0,014432	0,011995
958.57	0,007067	0,008424	0,007745	912.447	0,009235	0,012216	0,010726
924.451	0,01044	0,007488	0,008964	1359.6879999999999	0,007489	0,011516	0,009502
1052.538	0,011357	0,007139	0,009248	2982.907	0,005709	0,011671	0,00869
1099.557	0,008229	0,006933	0,007581	964.497	0,006348	0,008988	0,007668
1115.5819999999999	0,009927	0,006846	0,008386	926.4860000000001	0,005446	0,009671	0,007558
1093.585	0,005547	0,006673	0,00611	1229.58	0,007447	0,007667	0,007557
1080.548	0,007526	0,006512	0,007019	874.438	0,007729	0,006638	0,007184
1342.6670000000001	0,006399	0,006494	0,006446	825.4110000000001	0,008719	0,005373	0,007046
1048.53	0,004038	0,006417	0,005227	1099.557	0,005636	0,008079	0,006858
858.448	0,004643	0,00637	0,005507	950.4839999999999	0,00433	0,009293	0,006812
1111.603	0,008633	0,006368	0,0075	1607.8010000000002	0,002663	0,010934	0,006798
990.4960000000001	0,00281	0,005801	0,004306	1119.5819999999999	0,007129	0,00614	0,006635
917.471	0,004379	0,0058	0,00509	955.5139999999999	0,007676	0,005399	0,006538
1113.594	0,003726	0,005628	0,004677	988.528	0,007269	0,005414	0,006341
1133.5919999999999	0,009468	0,005208	0,007338	1564.8	0,003894	0,008411	0,006152
969.51	0,006424	0,00512	0,005772	979.554	0,006556	0,005002	0,005779
1039.546	0,006188	0,005041	0,005615	924.451	0,006697	0,004764	0,00573
1021.528	0,003937	0,004945	0,004441	1249.6689999999999	0,006841	0,004184	0,005513

1002.5	0,004514	0,004756	0,004635	828.437	0,00377	0,00708	0,005425
947.4910000000001	0,00382	0,004737	0,004278	1279.62	0,003992	0,006834	0,005413
868.4630000000001	0,003349	0,004722	0,004036	906.461	0,006218	0,004608	0,005413
936.503	0,003535	0,004711	0,004123	801.433	0,006853	0,003961	0,005407
2181.185	0,006415	0,004651	0,005533	868.4630000000001	0,005502	0,005069	0,005285
809.429	0,002971	0,004649	0,00381	1141.5810000000001	0,005462	0,004957	0,005209
964.497	0,00332	0,00461	0,003965	809.429	0,007238	0,003123	0,00518
1066.5330000000001	0,005865	0,004565	0,005215	1184.572	0,004861	0,005317	0,005089
1184.572	0,002475	0,004439	0,003457	981.525	0,003618	0,006495	0,005057
971.573	0,004421	0,004435	0,004428	1022.53	0,00506	0,004871	0,004966
1027.565	0,003262	0,004346	0,003804	834.443	0,006772	0,002832	0,004802
890.453	0,003678	0,004296	0,003987	2750.663	0,004227	0,005361	0,004794
988.528	0,003007	0,00428	0,003644	856.4730000000001	0,005269	0,004265	0,004767
2690.586	0,003987	0,004243	0,004115	1066.5330000000001	0,005416	0,00395	0,004683
994.5139999999999	0,004056	0,00419	0,004123	1196.607	0,004275	0,005051	0,004663
1149.559	0,004507	0,004046	0,004276	1127.572	0,005085	0,004212	0,004649
957.574	0,004742	0,003988	0,004365	815.442	0,005768	0,003494	0,004631
1018.5010000000001	0,002701	0,003977	0,003339	1080.548	0,005106	0,004027	0,004566
1364.6589999999999	0,00493	0,003948	0,004439	867.4639999999999	0,005964	0,003124	0,004544
882.477	0,003523	0,003921	0,003722	1115.5819999999999	0,004812	0,004182	0,004497
918.465	0,002506	0,003866	0,003186	866.472	0,006199	0,002718	0,004459
1035.535	0,003835	0,003849	0,003842	911.4760000000001	0,004184	0,004655	0,00442
1074.55	0,00562	0,003839	0,004729	983.5169999999999	0,003818	0,004924	0,004371
856.4730000000001	0,003318	0,003734	0,003526	1116.576	0,004943	0,003752	0,004348
950.4839999999999	0,004743	0,003713	0,004228	1320.6760000000002	0,002595	0,00602	0,004307
829.429	0,002839	0,003709	0,003274	1117.57	0,003346	0,005157	0,004252
1117.57	0,004901	0,003708	0,004304	814.4580000000001	0,003849	0,004617	0,004233
912.447	0,003366	0,003695	0,00353	994.5139999999999	0,003511	0,004753	0,004132
942.4960000000001	0,00318	0,003684	0,003432	1012.5139999999999	0,00309	0,005148	0,004119

1127.572	0,00251	0,00367	0,00309	879.4830000000001	0,004941	0,003276	0,004108
859.447	0,002765	0,003615	0,00319	2115.2110000000002	0,004153	0,00401	0,004082
837.448	0,003244	0,003582	0,003413	995.512	0,004087	0,004035	0,004061
931.4860000000001	0,002691	0,003574	0,003132	1120.57	0,002872	0,00513	0,004001
1564.8	0,002867	0,003565	0,003216	980.549	0,004731	0,003235	0,003983
817.416	0,003727	0,003542	0,003634	1131.566	0,005578	0,002346	0,003962
878.4760000000001	0,003616	0,003531	0,003573	949.4889999999999	0,002925	0,004951	0,003938
944.551	0,003231	0,0035	0,003365	823.449	0,003776	0,004073	0,003924
874.438	0,002892	0,003497	0,003195	840.443	0,005381	0,002455	0,003918
1116.576	0,005111	0,003445	0,004278	1024.517	0,004901	0,002904	0,003902
1059.562	0,003373	0,003416	0,003395	934.475	0,005047	0,002644	0,003845
1044.537	0,003518	0,003413	0,003465	908.4630000000001	0,003598	0,004037	0,003817
1012.5139999999999	0,002889	0,003345	0,003117	1079.557	0,003143	0,004454	0,003798
1094.598	0,003996	0,003338	0,003667	859.447	0,003937	0,003641	0,003789
972.534	0,002955	0,003335	0,003145	1267.681	0,004142	0,003316	0,003729
884.467	0,00274	0,003335	0,003037	886.4639999999999	0,005355	0,002071	0,003713
981.525	0,006133	0,003329	0,004731	936.503	0,002864	0,004388	0,003626
1007.551	0,003157	0,003288	0,003222	1044.537	0,004438	0,002772	0,003605
1251.598	0,004159	0,003284	0,003721	822.444	0,003891	0,003294	0,003593
1297.634	0,002954	0,00324	0,003097	1093.585	0,004466	0,002711	0,003589
1081.565	0,00395	0,003226	0,003588	1077.565	0,003514	0,003655	0,003585
1062.552	0,002193	0,003209	0,002701	839.4	0,004278	0,002814	0,003546
1269.687	0,003838	0,003207	0,003523	836.449	0,003283	0,003775	0,003529
1173.563	0,002816	0,003203	0,00301	817.416	0,003016	0,004031	0,003523
904.4789999999999	0,002465	0,003199	0,002832	1095.5919999999999	0,002802	0,004222	0,003512
1024.517	0,003078	0,003166	0,003122	966.538	0,004127	0,00284	0,003484
1562.8139999999999	0,002467	0,003165	0,002816	990.4960000000001	0,004024	0,002906	0,003465
1022.53	0,004376	0,003162	0,003769	1302.6480000000001	0,00296	0,003947	0,003454
1546.7979999999998	0,002021	0,003162	0,002591	878.4760000000001	0,004577	0,002202	0,00339

898.4960000000001	0,002885	0,003161	0,003023	1584.805	0,001909	0,004869	0,003389
923.4889999999999	0,004529	0,003156	0,003842	927.4989999999999	0,003168	0,003538	0,003353
1542.783	0,004385	0,003155	0,00377	1173.563	0,001923	0,004705	0,003314
1280.624	0,00246	0,003126	0,002793	1087.5739999999998	0,003519	0,003105	0,003312
815.442	0,002269	0,003123	0,002696	1366.6629999999998	0,004577	0,001959	0,003268
1137.575	0,003215	0,003116	0,003165	1098.5710000000001	0,002964	0,003543	0,003253
1004.51	0,002876	0,003112	0,002994	1107.575	0,002379	0,00412	0,00325
1136.589	0,002311	0,003097	0,002704	967.5369999999999	0,003385	0,003108	0,003246
2695.6059999999998	0,002784	0,003095	0,002939	914.4810000000001	0,002799	0,003656	0,003228
800.422	0,003448	0,003078	0,003263	965.5310000000001	0,004409	0,00204	0,003225
862.459	0,003037	0,003058	0,003048	852.443	0,003187	0,003243	0,003215
879.4830000000001	0,002921	0,003057	0,002989	845.447	0,003742	0,002646	0,003194
2159.197	0,005334	0,003046	0,00419	909.472	0,0023	0,004033	0,003166
965.5310000000001	0,002138	0,003045	0,002592	963.5039999999999	0,003366	0,002948	0,003157
1138.573	0,005337	0,003039	0,004188	872.45	0,003021	0,003265	0,003143
1069.575	0,002402	0,003028	0,002715	1004.51	0,003339	0,002881	0,00311
839.4	0,003858	0,003027	0,003443	1006.507	0,002846	0,003327	0,003087
1160.569	0,003077	0,003003	0,00304	1337.6979999999999	0,00257	0,003598	0,003084
986.545	0,002698	0,002992	0,002845	954.4830000000001	0,003092	0,003026	0,003059
1095.5919999999999	0,003763	0,002988	0,003376	1007.551	0,002238	0,00387	0,003054
983.5169999999999	0,00333	0,002987	0,003158	969.51	0,003935	0,002165	0,00305
836.449	0,003176	0,002965	0,00307	957.574	0,003348	0,002732	0,00304
953.5010000000001	0,002878	0,002947	0,002913	982.5110000000001	0,003381	0,002673	0,003027
1366.6629999999998	0,002636	0,002941	0,002789	961.4910000000001	0,003326	0,00271	0,003018
987.5269999999999	0,00263	0,002939	0,002784	1339.682	0,002756	0,003253	0,003005
852.443	0,003224	0,002937	0,00308	889.465	0,002931	0,003071	0,003001
933.472	0,00242	0,002931	0,002676	806.42	0,003414	0,002554	0,002984
823.449	0,002311	0,00293	0,00262	896.422	0,003065	0,002898	0,002982
982.5110000000001	0,002872	0,002927	0,0029	1000.5219999999999	0,003409	0,002553	0,002981

985.572	0,004069	0,002924	0,003497	1052.538	0,003435	0,002518	0,002976
2084.03	0,002557	0,002915	0,002736	976.482	0,004011	0,001885	0,002948
1076.577	0,002542	0,002909	0,002725	1542.783	0,003327	0,002555	0,002941
963.5039999999999	0,002789	0,002875	0,002832	917.471	0,003251	0,002628	0,00294
830.45	0,001985	0,002875	0,00243	1110.556	0,003315	0,002556	0,002936
1119.5819999999999	0,002468	0,002873	0,002671	1198.7160000000001	0,003222	0,002615	0,002919
993.564	0,002967	0,00287	0,002919	987.5269999999999	0,002141	0,003677	0,002909
1090.562	0,002741	0,00287	0,002805	999.528	0,002314	0,003494	0,002904
948.4939999999999	0,002828	0,002866	0,002847	929.5219999999999	0,003213	0,002582	0,002897
2208.131	0,002807	0,002852	0,002829	2208.131	0,001602	0,004191	0,002897
1850.92	0,002337	0,002848	0,002592	1076.577	0,003177	0,002611	0,002894
1833.9470000000001	0,002936	0,002818	0,002877	933.472	0,003529	0,00223	0,00288
1105.579	0,002324	0,002808	0,002566	1020.513	0,003109	0,002632	0,002871
1040.525	0,002476	0,002795	0,002635	846.447	0,003708	0,001999	0,002854
1126.586	0,002102	0,002791	0,002447	898.4960000000001	0,003674	0,002029	0,002852
911.4760000000001	0,003075	0,002764	0,00292	1655.8239999999998	0,003477	0,00221	0,002844
1530.7320000000002	0,002783	0,002754	0,002769	819.427	0,002779	0,002902	0,002841
1303.629	0,002539	0,002753	0,002646	1111.603	0,001863	0,003777	0,00282
900.505	0,002771	0,00275	0,00276	922.5	0,002617	0,003002	0,00281
819.427	0,002304	0,002744	0,002524	1138.573	0,003682	0,001926	0,002804
914.4810000000001	0,002208	0,00274	0,002474	1136.589	0,002523	0,003057	0,00279
1242.692	0,002977	0,002737	0,002857	993.564	0,003026	0,002552	0,002789
834.443	0,002478	0,00273	0,002604	918.465	0,002842	0,002706	0,002774
1237.627	0,002216	0,002717	0,002467	1143.577	0,00243	0,0031	0,002765
872.45	0,00328	0,002712	0,002996	876.454	0,003516	0,002014	0,002765
1000.5219999999999	0,00236	0,002705	0,002532	942.4960000000001	0,00291	0,002527	0,002718
807.413	0,002771	0,002702	0,002737	1126.586	0,003497	0,001928	0,002712
1584.805	0,002356	0,002702	0,002529	971.573	0,0025	0,002924	0,002712
845.447	0,002178	0,002701	0,00244	1088.5539999999999	0,002949	0,002437	0,002693

876.454	0,002339	0,002688	0,002514	851.453	0,002617	0,002765	0,002691
1302.6480000000001	0,002437	0,002662	0,002549	1214.623	0,002491	0,002868	0,002679
1129.586	0,002252	0,00266	0,002456	892.4680000000001	0,002195	0,003144	0,00267
883.475	0,002198	0,002659	0,002428	915.492	0,002887	0,002421	0,002654
902.482	0,002667	0,002656	0,002662	882.477	0,003509	0,001791	0,00265
1196.607	0,002313	0,002653	0,002483	923.4889999999999	0,002609	0,002676	0,002643
831.446	0,002227	0,002648	0,002437	1239.618	0,002644	0,002622	0,002633
976.482	0,002986	0,002647	0,002816	998.503	0,002997	0,002262	0,002629
1198.7160000000001	0,002972	0,002647	0,002809	1147.5910000000001	0,003316	0,001937	0,002626
1107.575	0,002303	0,002638	0,002471	938.4889999999999	0,003745	0,001503	0,002624
1289.671	0,002296	0,002632	0,002464	890.453	0,002814	0,002401	0,002607
1028.599	0,002612	0,002631	0,002621	810.436	0,00212	0,003085	0,002602
1131.566	0,003769	0,002625	0,003197	952.4889999999999	0,003075	0,002071	0,002573
934.475	0,002697	0,002619	0,002658	1065.553	0,002368	0,002764	0,002566
980.549	0,004159	0,002616	0,003387	1094.598	0,002877	0,002252	0,002565
979.554	0,004023	0,002603	0,003313	857.4739999999999	0,002677	0,002416	0,002547
920.4910000000001	0,002418	0,002578	0,002498	1487.711	0,002892	0,002199	0,002546
984.508	0,002347	0,002577	0,002462	1010.5369999999999	0,003309	0,001777	0,002543
1176.5819999999999	0,002644	0,002572	0,002608	1381.7020000000002	0,002888	0,002198	0,002543
962.4960000000001	0,002183	0,002568	0,002375	807.413	0,002565	0,002514	0,00254
955.5139999999999	0,002451	0,002565	0,002508	2084.03	0,003247	0,001828	0,002538
1054.5639999999999	0,003511	0,002559	0,003035	984.508	0,002193	0,002824	0,002508
997.523	0,002353	0,002559	0,002456	962.4960000000001	0,002825	0,002177	0,002501
825.4110000000001	0,003612	0,002559	0,003085	883.475	0,002388	0,002606	0,002497
851.453	0,002074	0,002551	0,002313	1220.702	0,002952	0,002015	0,002484
943.54	0,00236	0,002548	0,002454	920.4910000000001	0,002715	0,002237	0,002476
1154.5810000000001	0,002198	0,002535	0,002366	858.448	0,002578	0,002372	0,002475
1267.681	0,002298	0,002533	0,002415	1072.566	0,003109	0,001825	0,002467
857.4739999999999	0,002471	0,002518	0,002495	1032.56	0,002171	0,002737	0,002454

996.5160000000001	0,002659	0,002514	0,002586	1074.55	0,002644	0,002251	0,002448
892.4680000000001	0,002627	0,002512	0,002569	1297.634	0,002734	0,002155	0,002445
1092.553	0,002607	0,002512	0,00256	1090.562	0,002292	0,002593	0,002443
974.5060000000001	0,002378	0,002511	0,002444	900.505	0,0028	0,002081	0,00244
886.4639999999999	0,002394	0,002507	0,002451	939.49	0,002554	0,002322	0,002438
1706.817	0,002638	0,002501	0,00257	1325.651	0,002079	0,002762	0,002421
915.492	0,002462	0,0025	0,002481	1158.583	0,002322	0,002513	0,002417
828.437	0,002343	0,002499	0,002421	2705.591	0,001769	0,003058	0,002413
1006.507	0,002677	0,00249	0,002584	1289.671	0,001869	0,00295	0,00241
956.5139999999999	0,002484	0,00248	0,002482	837.448	0,002803	0,002012	0,002408
1381.7020000000002	0,002527	0,002477	0,002502	1510.736	0,002157	0,002648	0,002403
1010.5369999999999	0,002109	0,00247	0,00229	978.518	0,0023	0,002484	0,002392
1220.702	0,002601	0,002466	0,002533	1149.559	0,001815	0,002957	0,002386
1036.55	0,003189	0,00245	0,002819	1105.579	0,003036	0,00173	0,002383
850.461	0,002151	0,002444	0,002298	1018.5010000000001	0,002105	0,002652	0,002379
1235.638	0,002565	0,002435	0,0025	854.451	0,003332	0,001418	0,002375
822.444	0,002314	0,002427	0,002371	2690.586	0,002261	0,002479	0,00237
1034.528	0,002346	0,002423	0,002385	986.545	0,002368	0,00237	0,002369
1151.5739999999998	0,002167	0,002421	0,002294	1324.644	0,00215	0,002583	0,002367
938.4889999999999	0,003716	0,00242	0,003068	1580.7920000000001	0,002634	0,002093	0,002363
1481.721	0,002249	0,002419	0,002334	1530.7320000000002	0,002958	0,001745	0,002352
975.5139999999999	0,002562	0,002412	0,002487	1002.5	0,003445	0,001242	0,002344
909.472	0,002379	0,00241	0,002395	928.482	0,001832	0,002854	0,002343
1257.627	0,002221	0,002409	0,002315	904.4789999999999	0,002274	0,002384	0,002329
1015.5360000000001	0,002167	0,002407	0,002287	921.4989999999999	0,002238	0,002411	0,002324
806.42	0,002737	0,002407	0,002572	1109.568	0,003424	0,001211	0,002317
1552.727	0,003066	0,002401	0,002733	1155.5810000000001	0,002165	0,002457	0,002311
2321.25	0,00235	0,002384	0,002367	1340.65	0,002832	0,00177	0,002301
954.4830000000001	0,0025	0,002383	0,002442	808.435	0,002027	0,002575	0,002301

1629.7910000000002	0,002147	0,002382	0,002265	930.4830000000001	0,002491	0,002043	0,002267
1359.6879999999999	0,002238	0,002381	0,002309	1342.6670000000001	0,002169	0,00235	0,00226
846.447	0,002133	0,002379	0,002256	1706.817	0,002567	0,001895	0,002231
905.472	0,002048	0,002372	0,00221	1275.644	0,001873	0,002587	0,00223
1042.5510000000002	0,002269	0,002363	0,002316	1021.528	0,002973	0,001455	0,002214
928.482	0,002169	0,002363	0,002266	931.4860000000001	0,00287	0,001551	0,00221
1347.641	0,003057	0,002362	0,002709	1129.586	0,001888	0,002471	0,002179
926.4860000000001	0,002408	0,002361	0,002385	944.551	0,002332	0,001997	0,002164
1110.556	0,00315	0,002356	0,002753	888.456	0,002274	0,002045	0,002159
929.5219999999999	0,002797	0,002356	0,002576	1364.6589999999999	0,002389	0,001924	0,002157
1147.5910000000001	0,002347	0,00235	0,002349	1459.715	0,002383	0,001926	0,002155
998.503	0,002944	0,002344	0,002644	1125.586	0,002313	0,001942	0,002128
966.538	0,002941	0,002342	0,002642	997.523	0,002086	0,002154	0,00212
1229.58	0,00319	0,002339	0,002765	905.472	0,002456	0,001784	0,00212
894.47	0,002608	0,002328	0,002468	1135.5910000000001	0,001812	0,002391	0,002101
921.4989999999999	0,001936	0,002324	0,00213	937.508	0,002281	0,001922	0,002101
1320.6760000000002	0,004195	0,002323	0,003259	901.5089999999999	0,002186	0,002005	0,002095
871.482	0,002914	0,002318	0,002616	1036.55	0,001985	0,002204	0,002095
1088.5539999999999	0,002428	0,002314	0,002371	2305.245	0,002285	0,001887	0,002086
844.4910000000001	0,002203	0,002309	0,002256	1280.624	0,002507	0,001635	0,002071
952.4889999999999	0,002196	0,002307	0,002251	2126.1710000000003	0,002208	0,001926	0,002067
1607.8010000000002	0,002454	0,002303	0,002379	816.449	0,002139	0,001994	0,002066
1279.62	0,002202	0,0023	0,002251	844.4910000000001	0,002766	0,001361	0,002064
927.4989999999999	0,002524	0,002295	0,00241	977.505	0,002064	0,002055	0,002059
1171.586	0,002834	0,00229	0,002562	1237.627	0,00208	0,002024	0,002052
1125.586	0,002255	0,002284	0,00227	829.429	0,002857	0,001238	0,002048
896.422	0,006875	0,002284	0,004579	1015.5360000000001	0,002405	0,001688	0,002047
889.465	0,002311	0,002278	0,002295	1176.5819999999999	0,002464	0,001629	0,002047
908.4630000000001	0,002262	0,002263	0,002263	1251.598	0,002516	0,001563	0,002039

903.482	0,002084	0,00226	0,002172	1508.726	0,00237	0,001709	0,002039
816.449	0,002002	0,002256	0,002129	1242.692	0,002355	0,001697	0,002026
939.49	0,002237	0,002255	0,002246	1050.569	0,001927	0,002117	0,002022
1275.644	0,002367	0,002254	0,002311	1562.8139999999999	0,00229	0,001723	0,002006
937.508	0,002299	0,002247	0,002273	1257.627	0,002437	0,001556	0,001996
967.5369999999999	0,002806	0,002246	0,002526	1081.565	0,002361	0,001621	0,001991
1339.682	0,002397	0,002231	0,002314	1465.713	0,002094	0,001844	0,001969
2303.23	0,002358	0,002225	0,002291	1833.9470000000001	0,002154	0,001783	0,001968
1014.5360000000001	0,00209	0,002222	0,002156	1040.525	0,002321	0,001579	0,00195
906.461	0,002736	0,002212	0,002474	1269.687	0,002381	0,001507	0,001944
805.416	0,002797	0,002209	0,002503	800.422	0,002525	0,001351	0,001938
1580.7920000000001	0,002077	0,002209	0,002143	1014.5360000000001	0,00198	0,001884	0,001932
1214.623	0,002358	0,002206	0,002282	956.5139999999999	0,002048	0,001785	0,001917
1655.8239999999998	0,002464	0,002205	0,002334	960.502	0,002112	0,001683	0,001897
1001.5310000000001	0,002039	0,002195	0,002117	1235.638	0,002173	0,00162	0,001897
1087.5739999999998	0,002574	0,002195	0,002384	1177.585	0,002665	0,001127	0,001896
1653.8129999999999	0,004325	0,002186	0,003256	894.47	0,002137	0,001652	0,001894
977.505	0,002561	0,002186	0,002374	830.45	0,002168	0,001596	0,001882
1098.5710000000001	0,002955	0,002184	0,00257	884.467	0,001783	0,001974	0,001879
1508.726	0,002468	0,002181	0,002325	1751.8629999999998	0,00216	0,001596	0,001878
1855.942	0,003823	0,002179	0,003001	903.482	0,002144	0,001592	0,001868
930.4830000000001	0,002362	0,002178	0,00227	1568.7910000000002	0,002388	0,001321	0,001855
1319.6529999999998	0,002377	0,002171	0,002274	1054.5639999999999	0,001642	0,002067	0,001854
1465.713	0,002326	0,002166	0,002246	945.533	0,002251	0,001451	0,001851
888.456	0,00214	0,002164	0,002152	1271.672	0,001744	0,001941	0,001843
1324.644	0,002361	0,002158	0,00226	1869.9229999999998	0,001974	0,001711	0,001843
1143.577	0,002293	0,002156	0,002224	1151.5739999999998	0,001657	0,002025	0,001841
2216.1620000000003	0,002402	0,002155	0,002278	1101.563	0,00239	0,001275	0,001833
1325.651	0,002253	0,002153	0,002203	1139.5610000000001	0,002026	0,001612	0,001819

1079.557	0,002307	0,002153	0,00223	972.534	0,001679	0,001951	0,001815
2126.1710000000003	0,002113	0,002139	0,002126	1039.546	0,002066	0,001537	0,001801
1020.513	0,002013	0,002136	0,002075	1062.552	0,001753	0,001821	0,001787
1065.553	0,002521	0,002134	0,002327	1154.5810000000001	0,001754	0,001818	0,001786
1487.711	0,001981	0,002129	0,002055	850.461	0,001698	0,001848	0,001773
1751.8629999999998	0,002165	0,002113	0,002139	1070.572	0,001781	0,001762	0,001772
999.528	0,002354	0,002112	0,002233	902.482	0,001857	0,001663	0,00176
945.533	0,002108	0,002112	0,00211	805.416	0,001874	0,001631	0,001752
1009.538	0,002062	0,002111	0,002086	2216.1620000000003	0,001926	0,001577	0,001751
808.435	0,001956	0,002111	0,002033	996.5160000000001	0,001998	0,001491	0,001745
1135.5910000000001	0,002418	0,00211	0,002264	1653.8129999999999	0,001948	0,001508	0,001728
1050.569	0,002096	0,00211	0,002103	1092.553	0,001721	0,001709	0,001715
961.4910000000001	0,002219	0,002108	0,002163	2321.25	0,001396	0,002032	0,001714
854.451	0,002149	0,002104	0,002126	1133.5919999999999	0,001762	0,001649	0,001706
1239.618	0,002088	0,002104	0,002096	1347.641	0,002045	0,001333	0,001689
1177.585	0,003289	0,002101	0,002695	1546.7979999999998	0,001678	0,00169	0,001684
960.502	0,002247	0,0021	0,002174	953.5010000000001	0,001725	0,001639	0,001682
1340.65	0,00213	0,002092	0,002111	943.54	0,001757	0,001605	0,001681
810.436	0,002255	0,00209	0,002173	1307.62	0,001841	0,001514	0,001677
901.5089999999999	0,003745	0,002084	0,002914	2104.1620000000003	0,001577	0,001766	0,001672
949.4889999999999	0,002345	0,002083	0,002214	1160.569	0,001406	0,001932	0,001669
1141.5810000000001	0,002387	0,002078	0,002232	1069.575	0,001622	0,001711	0,001667
1234.6760000000002	0,002488	0,002077	0,002282	1048.53	0,002105	0,001194	0,001649
1030.5410000000002	0,002135	0,002074	0,002104	1234.6760000000002	0,001701	0,001585	0,001643
801.433	0,002297	0,002073	0,002185	1356.658	0,00187	0,001413	0,001642
866.472	0,002116	0,00207	0,002093	985.572	0,001764	0,00151	0,001637
922.5	0,00239	0,002066	0,002228	2159.197	0,001662	0,001604	0,001633
1109.568	0,00247	0,002066	0,002268	2958.9759999999997	0,001613	0,00164	0,001627
1072.566	0,002948	0,002064	0,002506	974.5060000000001	0,001647	0,001605	0,001626

1307.62	0,003788	0,002059	0,002924	2137.188	0,001713	0,001535	0,001624
2137.188	0,002189	0,002059	0,002124	1035.535	0,001546	0,001682	0,001614
1032.56	0,002061	0,002057	0,002059	975.5139999999999	0,001838	0,001364	0,001601
1869.9229999999998	0,002376	0,002057	0,002216	1481.721	0,001835	0,001365	0,0016
2318.2329999999997	0,002238	0,002052	0,002145	1319.6529999999998	0,001987	0,001174	0,00158
978.518	0,002245	0,002049	0,002147	1001.5310000000001	0,001634	0,001509	0,001571
1249.6689999999999	0,002228	0,002036	0,002132	1217.63	0,001727	0,001415	0,001571
867.4639999999999	0,00229	0,002024	0,002157	1059.562	0,001685	0,001441	0,001563
1271.672	0,00215	0,002022	0,002086	1009.538	0,001758	0,00135	0,001554
2705.591	0,002219	0,00202	0,002119	1171.586	0,001604	0,001491	0,001547
2461.343	0,002274	0,002015	0,002144	958.57	0,001483	0,001607	0,001545
1101.563	0,002324	0,002013	0,002168	2219.113	0,001575	0,001503	0,001539
1217.63	0,002134	0,00201	0,002072	1027.565	0,001598	0,00146	0,001529
2104.1620000000003	0,002208	0,002004	0,002106	2695.6059999999998	0,001604	0,001453	0,001528
1158.583	0,002334	0,001982	0,002158	1137.575	0,001578	0,00142	0,001499
1612.819	0,002136	0,001976	0,002056	1028.599	0,001457	0,001524	0,001491
1356.658	0,002156	0,001962	0,002059	871.482	0,00136	0,001618	0,001489
2219.113	0,002449	0,001962	0,002205	1850.92	0,001415	0,001545	0,00148
1120.57	0,002461	0,001955	0,002208	1629.7910000000002	0,001424	0,001534	0,001479
1510.736	0,002045	0,001955	0,002	1612.819	0,001409	0,001542	0,001476
1867.932	0,002053	0,001944	0,001999	1113.594	0,001605	0,001311	0,001458
1459.715	0,00215	0,001935	0,002043	831.446	0,001615	0,001262	0,001438
2115.2110000000002	0,002042	0,001931	0,001987	2727.6059999999998	0,001638	0,001172	0,001405
2958.9759999999997	0,002357	0,001928	0,002142	1867.932	0,001347	0,001462	0,001404
2305.245	0,002294	0,001928	0,002111	940.507	0,001544	0,001142	0,001343
995.512	0,002313	0,001923	0,002118	1677.8239999999998	0,001374	0,001309	0,001342
1070.572	0,001947	0,001916	0,001931	1585.8	0,001497	0,001186	0,001341
1568.7910000000002	0,002286	0,001908	0,002097	1855.942	0,001456	0,001225	0,00134
940.507	0,002366	0,001898	0,002132	2318.2329999999997	0,001223	0,001291	0,001257

Appendices

Acknowledgement
Curriculum Vitae
List of publications

Acknowledgement

I would like to thank my promoter Prof. Dr. Philippe Lambin, for giving me the opportunity to do a PhD at the Precision Medicine department. He motivated me to consistently provide comprehensive arguments and delve deeper into study cases, promoting academic excellence and challenging the applicability of my research.

To my promoter Prof. Dr. Heike Grabsch, thank you for your support and for trusting me with your project. Our exchanges were very insightful and opened the door to new possibilities regarding the application of my research. I couldn't have completed my PhD without your mentoring.

I would like to express my gratitude to Dr. Henry Woodruff for his role as co-supervisor and giving me the autonomy to manage my projects as I deemed appropriate.

Prof. Dr. D. Keszthelyi, Prof. Dr. M.L. Smidt, Prof. Dr. J.N. Kather, Prof. Dr. W.J. Niessen, I would like to thank you for taking the time to review my thesis. Marta and Benjamin, my colleagues from the M4I, thank you for co-authoring my first paper. It was challenging but I am proud of what we achieved together. I would also like to thank all my co-authors for the time they invested to review our manuscripts and take them to a higher level.

I am grateful for the friends I made along the way thanks to the grant I was part of: Michael, Patrick, and Akshayaa, it was a pleasure to work with you. To my colleagues of the D-Lab, thank you for your camaraderie and for sharing your knowledge and ideas with me. Your insights have been invaluable, and I have learned a lot from our conversations. To Anke, I am grateful for your guidance and mentorship throughout my PhD journey. Your advice and support have been crucial in shaping my research and helping me navigate the challenges along the way. To Natacha, Anouk, Dasha, Lars, Inez, and Kosta, I am honored to have had the opportunity to supervise such brilliant students. Your dedication, hard work, and creativity have been truly inspiring, and I am proud of all that you have accomplished.

Abdalla, Lisa, Sergey, Simon, and Will, I want to express my appreciation to all of you. The best memories I have from my time in Maastricht were with you. Sergey, you are the first person I met in the lab, and you always made it seem like any challenge coming your way was so easy to handle. I am very impressed by your optimism and I am proud to be your friend. I am looking forward to our future adventures with you and Kate! Abdalla, thank you for your kindness and your friendship. I really enjoyed having you as neighbor, in our office and in the building where we lived. Your commitment to making great science is something I really admire. Simon, thank you for your friendship, it was a true lifeline. I am deeply inspired by your ability to remain patient and calm in any situation. We had a great time in Maastricht, and I'm glad we get to enjoy the North together. Lisa, I loved the time we shared and listening to your stories. Your courage is truly admirable and serves as a powerful inspiration for me to step out of my comfort zone and embrace fearlessness. It was a pleasure to have you as movie companion and I hope we get to meet again all over Europe.

Will, you were a great friend and a good colleague, Maastricht will never be the same without you.

My friends from Maastricht: Niloo, Eva and Alex, I really enjoyed your company during our nice diners and our long conversations. I hope you will come visit me often to continue those activities!

Laetitia, Lise, Gianna, merci de m'avoir rendu visite aux Pays-Bas et d'avoir toujours été là pour moi. Vous me faites sentir appréciée, entendue et reconnue. Nos weekends et vacances ensemble ont été mémorables, j'ai hâte d'en partager des milliers d'autre avec vous. Vous êtes des amies exceptionnelles. Merci aussi à Julien, Denis, Clément et tous les autres metteurs d'ambiance. Merci de m'avoir soutenue tout au long de mon parcours. Vous êtes ma deuxième famille et malgré nos rencontres espacées, je me sens chez moi avec vous.

Merci à ma famille pour votre soutien indéfectible, vos encouragements et vos conseils. Je suis infiniment heureuse de vous avoir dans ma vie, malgré la distance qui nous sépare. Je vous aime tous tellement et je vous suis reconnaissante pour tout.

I am also grateful for my second family in the Netherlands, Gerda, Marcel, Lucas and Alyona, thank you for being so supportive and open-minded. Our conversations really helped me through my PhD.

Finally, Alex, thank you for your infinite support and your help. I feel incredibly lucky to have found such a smart, compassionate and loving partner. I am looking forward to the rest of our journey together.

Curriculum Vitae

Manon was born on the 14th of November 1995 in Saint-Cloud, France. She successfully completed a preparation program for admission to Engineering schools at Lycée Chaptal in Paris in 2015. She then completed her Master of Science in Biomedical Engineering at the University of Technology of Compiègne in France in October 2018.

Before diving into her PhD studies, Manon gained valuable experience as a Software Engineer in Medical Imaging during her internship at Imabiotech in Lille, France. This opportunity allowed her to apply her technical skills in a practical setting and gain hands-on experience in the medical imaging field.

Manon started her PhD in October 2018, implementing Computer Vision solutions for Medical Imaging at the Precision Medicine department of Maastricht University. She was involved in every aspect of the research process, from data collection and preprocessing of high-resolution medical images to developing AI models for specific tasks, testing and validating the models, and generating results.
During her PhD, Manon worked on several projects, including the detection and delineation of suspicious masses in contrast-enhanced mammography, predicting malignancy, and assessing radio-necrosis in pre-treatment MRIs. She also focused on detecting and delineating lymph nodes in large H&E datasets, while exploring the comparative predictive values of H&E and MSI data.
She designed a deep learning exercise for the AI4Imaging workshop in 2019, mentored second-year bachelor's students enrolled in the course "Non-Invasive Techniques in Biomedical Research" in 2020 and co-supervised a research project for a honours group in 2021-2022. She has also collaborated with leading organizations, such as Mirada and DKFZ, to explore automated delineation of the diaphragm in CT scans and to study stained tissues from patients with glioma, respectively. Throughout her PhD training, Manon has had the privilege of working under the guidance of distinguished supervisors: Prof. Dr. Philippe Lambin, Prof. Dr. Heike Grabsch, and Dr. Henry Woodruff. Their expertise and mentorship have played a pivotal role in shaping Manon's professional growth and accomplishments.

Since November 2022, Manon has been working as a Deep Learning Scientist at Agendia in Amsterdam.

List of publications

Beuque M., M. B. I. Lobbes, Y. v. Wijk, Y. Widaatalla, S. Primakov, M. Majer, C. Balleyguier, H. C. Woodruff and P. Lambin (2023). "Combining Deep Learning and Handcrafted Radiomics for Classification of Suspicious Lesions on Contrast-enhanced Mammograms." Radiology 307(5): e221843.

Lavrova, E., S. Primakov, Z. Salahuddin, **M. Beuque**, D. Verstappen, H. C. Woodruff and P. Lambin (2023). "Precision-medicine-toolbox: An open-source python package for the quantitative medical image analysis." Software Impacts: 100508.

Rogers, W., S. A. Keek, **M. Beuque**, E. Lavrova, S. Primakov, G. Wu, C. Yan, S. Sanduleanu, H. A. Gietema, R. Casale, M. Occhipinti, H. C. Woodruff, A. Jochems and P. Lambin (2023). "Towards texture accurate slice interpolation of medical images using PixelMiner." Computers in Biology and Medicine 161: 106701.

Beuque M., D. R. Magee, A. Chatterjee, H. C. Woodruff, R. E. Langley, W. Allum, M. G. Nankivell, D. Cunningham, P. Lambin and H. I. Grabsch (2023). "Automated detection and delineation of lymph nodes in haematoxylin & eosin stained digitised slides." Journal of Pathology Informatics 14: 100192.

Keek, S. A., **M. Beuque**, S. Primakov, H. C. Woodruff, A. Chatterjee, J. E. van Timmeren, M. Vallières, L. E. Hendriks, J. Kraft and N. Andratschke (2022). "Predicting adverse radiation effects in brain tumors after stereotactic radiotherapy with deep learning and radiomics." Frontiers in Oncology: 3341.

Van Camp, A., **M. Beuque**, L. Cockmartin, H. C. Woodruff, N. W. Marshall, M. Lobbes, P. Lambin and H. Bosmans (2022). Synthetic data of simulated microcalcification clusters to train and explain deep learning detection models in contrast-enhanced mammography. 16th International Workshop on Breast Imaging (IWBI2022), SPIE.

Van Camp, A., L. Cockmartin, **M. Beuque**, H. Woodruff, N. Marshall, P. Lambin and H. Bosmans (2022). The creation of a large set of realistic synthetic microcalcification clusters for simulation in (contrast-enhanced) mammography images. Medical Imaging 2022: Physics of Medical Imaging, SPIE.

Primakov, S. P., A. Ibrahim, J. E. van Timmeren, G. Wu, S. A. Keek, **M. Beuque**, R. W. Granzier, E. Lavrova, M. Scrivener and S. Sanduleanu (2022). "Automated detection and segmentation of non-small cell lung cancer computed tomography images." Nature communications 13(1): 3423.

Ibrahim, A., S. Primakov, **M. Beuque**, H. Woodruff, I. Halilaj, G. Wu, T. Refaee, R. Granzier, Y. Widaatalla and R. Hustinx (2021). "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework." Methods 188: 20-29.

Beuque, M., M. Martin-Lorenzo, B. Balluff, H. C. Woodruff, M. Lucas, D. M. de Bruin, J. E. van Timmeren, O. J. de Boer, R. M. Heeren and S. L. Meijer (2021). "Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging." Computers in Biology and Medicine 138: 104918.

Chatterjee, A., H. Woodruff, M. Lobbes, Y. van Wijk, **M. Beuque** and P. Lambin (2021). "Machine learning with imbalanced clinical data: does synthetic minority oversampling help?" Physica Medica: European Journal of Medical Physics 92: S7.

Ibrahim, A., S. Primakov, **M. Beuque**, H. Woodruff, I. Halilaj, G. Wu, T. Refaee, R. Granzier, Y. Widaatalla and R. Hustinx (2021). "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework." Methods 188: 20-29.

Chatterjee, A., H. Woodruff, M. Lobbes, Y. van Wijk, **M. Beuque**, J. Seuntjens and P. Lambin (2020). "Altering the decision threshold as a simple and effective method for machine learning-based classification of imbalanced radiation oncology data." International Journal of Radiation Oncology, Biology, Physics 108(3): e332.

Chatterjee, A., H. Woodruff, M. Lobbes, Y. van Wijk, **M. Beuque**, J. Seuntjens and P. Lambin (2020). Effectiveness of Simple Data Imputation for Missing Feature Values in Binary Classification. MEDICAL PHYSICS, WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

Rogers, W., S. Thulasi Seetha, T. A. Refaee, R. I. Lieverse, R. W. Granzier, A. Ibrahim, S. A. Keek, S. Sanduleanu, S. P. Primakov and **M. Beuque** (2020). "Radiomics: from qualitative to quantitative imaging." The British journal of radiology 93(1108): 20190948.

Drochon, A., **M. Beuque** and D. D. R. A.-A. Rodriguez (2018). "A review of some reference analytic solutions for the magnetohydrodynamic flow of blood." Applied Mathematics 9(10): 1179-1192.

Drochon, A., **M. Beuque** and A.-A. R. Dima (2017). "Impact of an External Magnetic Field on the Shear Stresses Exerted by Blood Flowing in a Large Vessel." Journal of Applied Mathematics and Physics (JAMP) 5(7): 1493-1502.