AI applications in routine clinical imaging

Citation for published version (APA):

Vaidyanathan, A. (2023). *AI applications in routine clinical imaging: detection, segmentation, diagnosis, and prognosis*. [Doctoral Thesis, Maastricht University]. Maastricht University. https://doi.org/10.26481/dis.20230228av

Document status and date:
Published: 01/01/2023

DOI:
10.26481/dis.20230228av

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 18 Apr. 2024

AI applications in routine clinical imaging: detection, segmentation, diagnosis, and prognosis

Akshayaa Vaidyanathan

The work presented in this thesis was made possible by the financial support of the European Program PREDICT - ITN - n° 766276).
Cover design: Dennis Hendriks || ProefschriftMaken.nl
Layout: ProefschriftMaken.nl
ISBN: 978-94-6469-195-5
© 2023 Akshayaa Vaidyanathan
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author or the copyright-owning journals for previous published chapters.

AI applications in routine clinical imaging: detection, segmentation, diagnosis, and prognosis

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 Tuesday February 28th, 2023, at 16:00 hours

by Akshayaa Vaidyanathan

Supervisor:
Prof. P. Lambin

Co-supervisors:
Dr. H. C. Woodruff
Dr. S. Walsh, Radiomics, Belgium

Assessment Committee:
Prof. dr. Manon van Engeland (Chair)
Prof. dr. Rhodri Cusack**,** Trinity College Dublin, Ireland
Prof. dr. Eline Kooi
Prof. dr. Greetje Vande Velde, Catholic University Leuven, Belgium

Contents

Chapter 1

Introduction ……………………………………………………………………………………………5

PART 1 - AI-BASED DIAGNOSTIC MODELS FOR RESPIRATORY DISEASES

Chapter 2

Development and validation of an automated radiomic CT signature for
detecting COVID 19……………………………………………………………………………23

Chapter 3

Deep learning algorithm to classify COVID-19 and other pneumonias on chest CT
based on automatic segmentation of lung abnormalities ……………………………41

Chapter 4

Externally validated deep learning model for the diagnosis and detection of pulmonary
embolism on chest CTPA images ……………………………………………………………61

PART 2 – AI-BASED MODELS FOR TREATMENT OUTCOME PREDICTION AND DETECTION OF DISEASE IN ONCOLOGICAL USE CASES

Chapter 5

Prospective validation of a prognostic radiomics signature for patients with non-metastatic NSCLC
treated with standard of care non-surgical therapy……………………………………77

Chapter 6

Deep learning-based identification of bone scintigraphies containing
metastatic bone disease foci……………………………………………………………………95

PART 3 – AI BASED MODEL FOR DIAGNOSIS OF A DISORDER IN THE INNER EAR

Chapter 7

A non-invasive, automated diagnosis of Menière’s disease using radiomics and machine learning
on conventional magnetic resonance imaging: A multicentric, case-controlled feasibility study……111

Chapter 8

Deep learning for the fully automated segmentation of the inner ear on MRI ……………………125

PART 4 – SUMMARY AND FUTURE PERSPECTIVES

Chapter 9

Discussion ……145

English & Dutch Summary……………………………………………………………………………………………159

Appendix…….163

Impact addendum……………………………………………………………………………………………………….199

Acknowledgments………………………………………………………………………………………………………201

Curriculum vitae……202

List of manuscripts……………………………………………………………………………………………………203

Chapter 1

Introduction

1 RADIOMICS APPLICATIONS IN MEDICINE

Imaging is a fundamental technology in medicine and is used in clinical practice to aid decision-making for screening, diagnostic [1] therapeutic [2] and follow-up purposes. Radiomics was born in 2012 as an innovative approach to image analysis, that focuses on augmentation of traditional quantitative image analysis [3], [4] using automated high-throughput extraction of large amounts (200+) of quantitative features from medical images, such as CT, MRI and PET scans. To extract relevant features from the images, it is very important that the region of interest (ROI) is appropriately selected and delineated. Below we report a typical step-by-step radiomics workflow (Figure 1.1)

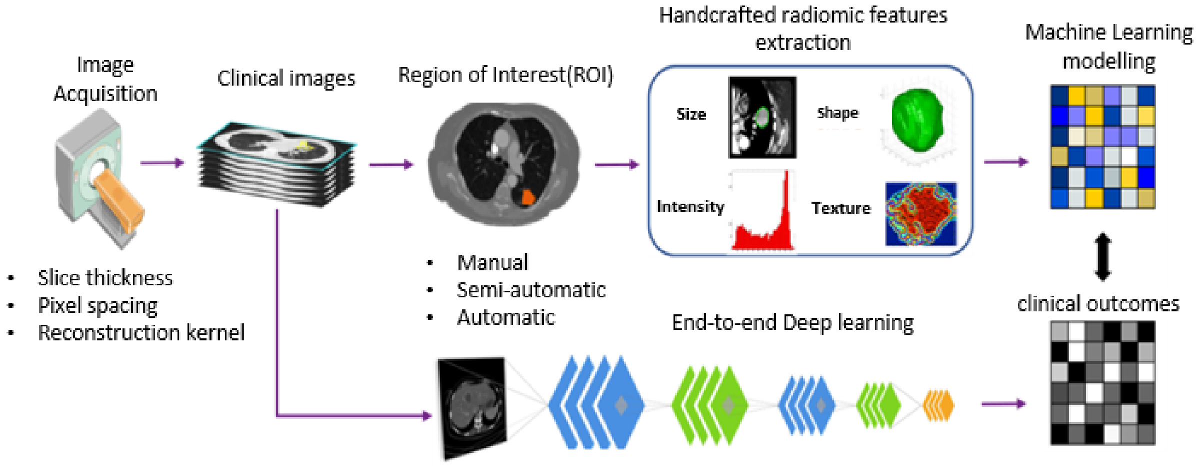


Figure 1.1 Scheme of the radiomics workflow for hand-crafted features (top) and deep learning (bottom)

The hypothesis is that quantitative analysis of medical image data can provide complementary information to aid physicians in the decision-making process, aided by automatic or semi-automatic software, in a fast and reproducible way [5]. Radiomics is the result of several decades of computer-aided diagnosis, prognosis, and therapeutics research [6], [7]. A robust radiomics approach consists of the identification of a wide variety of quantitative features from medical images, the storage of such data in several independent databases functioning as a single entity (federated databases) [8] and the subsequent data mining to obtain clinically relevant outcomes [9]. Medical images such as CT, MR, and/or PET scans can be analyzed and processed to extract relevant radiomics features which can be used for screening, diagnostic [10], follow-up, and prognostic [11] purposes as well as for pharmacokinetic and pharmacodynamic studies [12]–[14]. Databases that collect and cross-reference vast amounts of radiomics data along with other relevant patient information from millions of cases are already a reality, but still present considerable management problems [15]–[18]. Since radiomics' inception in 2012, the number of radiomics publications has grown exponentially (See Figure 1.2) as well as its detractors and disbelievers. The proven efficacy of radiomics approaches and the enthusiasm around this new method has to be tempered by its informed application and the careful evaluation of its real potential.

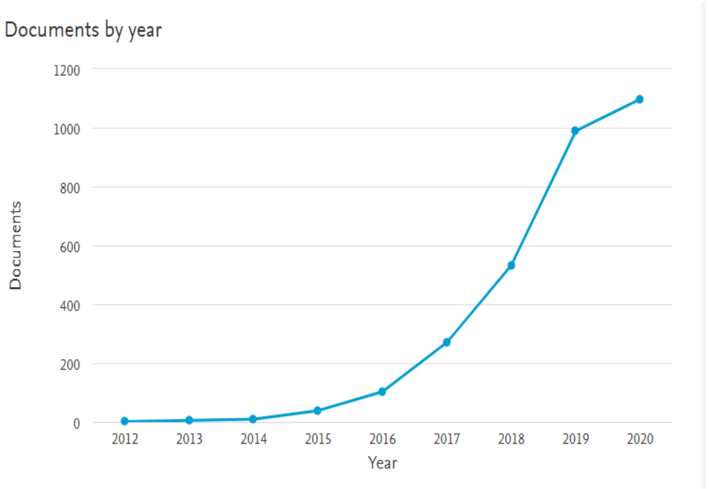


Figure 1.2 Number of "radiomics" publications per year (2012 -2020). Data obtained from Scopus (09/09/2020)

Two main approaches are used for radiomics analysis, hand-crafted features and deep learning (DL). Radiomic hand-crafted features (such as intensity, shape, texture, or wavelet) offer information on the specific area of the imaging scan one wishes to investigate, which might be a tumor region or a whole organ. These features are distinct yet interconnected to other data sources (such as clinical, treatment, or genomic data) [19]. The main challenge lies in the collection and integration of multimodal data sources quantitatively, delivering unambiguous clinical information and in turn allowing accurate and robust outcome prediction [20]. DL methods instead use a data-driven approach for model creation, mimicking simplified brain neuron interactions. DL has the advantage of not needing prior segmentation of the imaging scan. However, the "black box" approach of DL and the lack of interpretability of the models are seen as the main limitation of clinical applicability. Moreover, DL approaches need a large amount of data to truly express their potential, and sometimes the patient cohorts available, for example in case of rare diseases, are not enough to leverage a DL architecture effectively. For as much as this scenario seems straightforward and most alluring for clinicians, there are still too many published prediction models which lack standardized evaluation of their performance, reproducibility, and/or clinical utility [21], [22].

2 ARTIFICIAL INTELLIGENCE IN HEALTHCARE

Recent years have seen significant advances in the capacity of Artificial Intelligence (AI), which is growing in sophistication, complexity, and autonomy. AI attempts to emulate the neural processes of humans, and it introduces a paradigm change to healthcare, driven by growing healthcare data access and rapid development in analytical techniques. The rapid explosion of AI has given rise to the possibilities of using aggregated health data to generate powerful models that can automate diagnosis and also allow an increasingly precise approach to medicine by tailoring therapies and targeting services with optimal efficacy in a timely and dynamic manner. Despite the remarkable advances in DL-based approaches, the notion that AI technologies will swiftly usher in a new utopian era of digital healthcare is visionary at best and delusionary at worst, real and substantial challenges still exist. Explainable AI can play a critical role in identifying radiomic features that are clinically meaningful. Moreover, many high-performance DL models produce findings that are next to impossible for non-AI-expert humans to comprehend. While these models can produce better-than-human efficiency, it is not easy to express intuitive interpretations that can justify model findings, define model uncertainties, or derive additional clinical insights from these computational ’black-boxes.’ The ideal solution should have both high explainability and high performance. However, existing linear models, rule-based models, and decision trees are more transparent, but with lower performance in general. In contrast, complex models, e.g., DL and ensembles, manifest higher performance while less explainability can be obtained.

2.1 POST-HOC EXPLAINABILITY

Post-hoc explainability targets models that are not readily interpretable by design, by resorting to diverse means to enhance their interpretability, such as visual explanations, local explanations, and feature relevance explanations techniques. Each of these techniques covers one of the most common ways humans explain systems and processes by themselves. Figure 1.3 shows a conceptual diagram of the most common post-hoc explainability approaches available for AI models applied in healthcare.

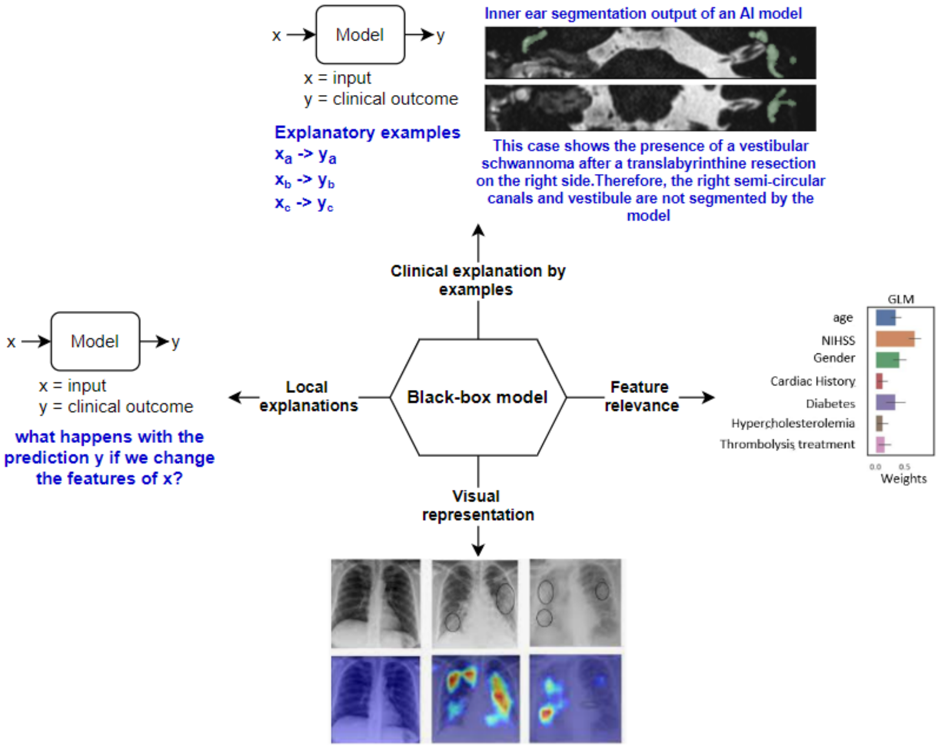

x
Model
y
x = input
y = clinical outcome
Explanatory examples
x_a -> y_a
x_b -> y_b
x_c -> y_c
Inner ear segmentation output of an AI model
This case shows the presence of a vestibular schwannoma after a translabyrinthine resection on the right side.Therefore, the right semi-circular canals and vestibule are not segmented by the model
Clinical explanation by examples
x
Model
y
x = input
y = clinical outcome
what happens with the prediction y if we change the features of x?
Local explanations
Black-box model
Feature relevance
GLM
age
NIHSS
Gender
Cardiac History
Diabetes
Hypercholesterolemia
Thrombolysis treatment
0.0
0.5
Weights
Visual representation

Figure 1.3: Conceptual diagram showing the different post-hoc explainability approaches available for AI models applied in healthcare

3 GOOD PRACTICES IN RADIOMICS STUDIES

Radiomics can be defined as a collection of methods (algorithms) that produces useful insight based upon a large number of extracted features from radiographic medical images [23]. Radiomics emerged originally in the field of oncology [24], [25]; however, it can be applied to any medical study where a disease or a condition can be imaged [26]–[29]. A radiomics study can be divided into four main phases: data selection and curation, features extraction, exploratory analysis, and modeling. Below we report a typical step-by-step radiomics workflow (Figure. 1.4) and the application of an evaluation protocol called Radiomics Quality Score (RQS). In 2017 the Radiomics Quality Score (RQS) was proposed and defined to help the scientific community assess the quality and scientific/clinical value of a radiomics study at a glance [30]. A similar example is the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative [31]. The RQS is determined by 16 key criteria which are assigned to a point value for a maximum of 36 points (100%). These criteria cover image acquisition protocols, statistical data treatment, cohort provenance, and open science policies, encompassing all the relevant aspects that a reliable radiomics publication should present. The latest version of the RQS questionnaire (RQS 2.0) which is under development is aimed to cover criteria like interpretability and explainability which are more specific and relevant to DL-based radiomics

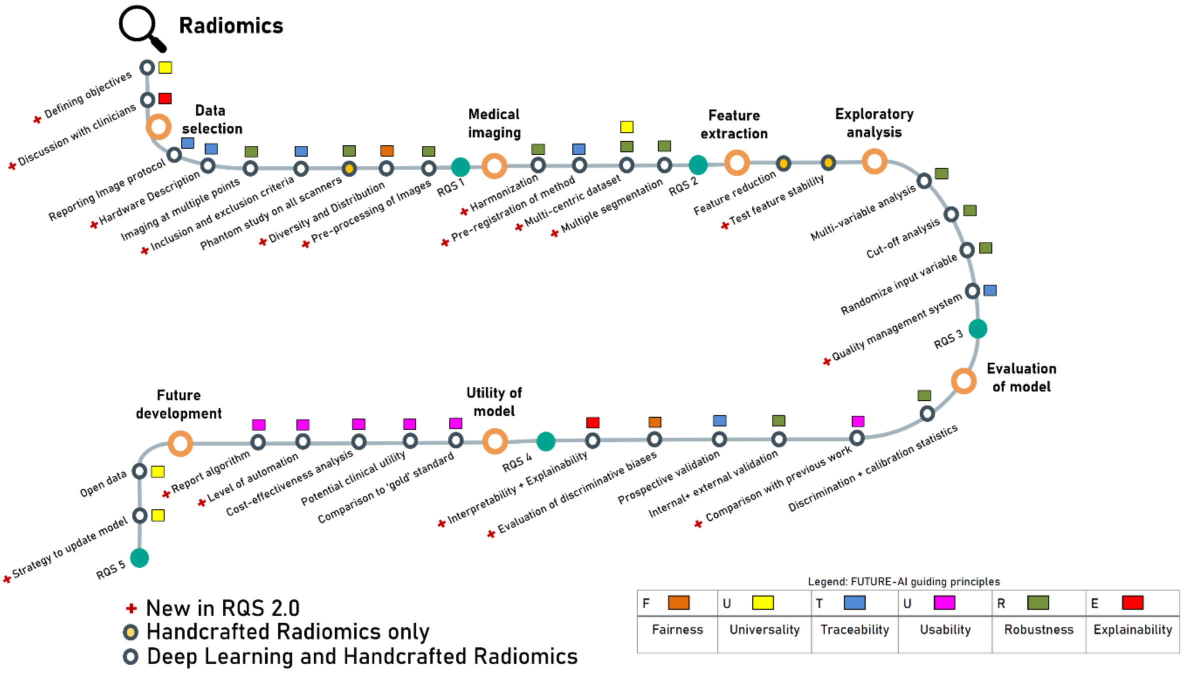

Radiomics
* Defining objectives
* Discussion with clinicians
Data selection
Reporting Image protocol
* Hardware Description
Imaging at multiple points
* Inclusion and exclusion criteria
Phantom study on all scanners
* Diversity and Distribution
* Pre-processing of Images
RQS 1
Medical imaging
* Harmonization
* Pre-registration of method
* Multi-centric dataset
* Multiple segmentation
RQS 2
Feature extraction
Feature reduction
* Test feature stability
Exploratory analysis
Multi-variable analysis
Cut-off analysis
Randomize input variable
* Quality management system
RQS 3
Evaluation of model
Discrimination + calibration statistics
* Comparison with previous work
Internal+ external validation
Prospective validation
* Evaluation of discriminative biases
* Interpretability + Explainability
RQS 4
Utility of model
Comparison to 'gold' standard
Potential clinical utility
Cost-effectiveness analysis
* Level of automation
* Report algorithm
Future development
Open data
* Strategy to update model
RQS 5
+ New in RQS 2.0
Handcrafted Radiomics only
Deep Learning and Handcrafted Radiomics
Legend: FUTURE-AI guiding principles
F Fairness
U Universality
T Traceability
U Usability
R Robustness
E Explainability

Figure.1.4. Scheme of the radiomics workflow for hand-crafted features (top) and DL (bottom)

3.1 DATA SELECTION AND CURATION

The starting point of the Radiomic analysis is the selection of an unmet clinical need, the appropriate imaging technique (CT, MRI, PET, etc.), the identification of the volume of interest (VOI), and the choice of a specific prediction target — the relevant clinical question that the radiomics analysis aims to answer. For example, in a typical oncological study, the entire primary tumor is analyzed and linked to available data on treatment outcomes, such as survival rate. Radiomic analyses can be performed on subregions of the tumor (habitats), metastatic lesions, as well as in normal tissues. Radiomics analysis, however, is not restricted to oncology and can be applied to any image generated in the clinical setting [32]–[34]. The use of standardized imaging protocols to eliminate unnecessary confounding variability is of paramount importance [23], [35] and has been recognized through the years as one of the main factors leading to low-quality radiomics analysis [36]. Still nowadays, however, non-standardized imaging protocols are commonplace. Reproducibility and comparability of radiomic studies can be achieved only by extensive disclosure of imaging protocols along with clear guidelines on how such protocols should be applied and reported. To overcome at least partially these issues, it is important to introduce standardization techniques for data storage and usage, such as 1) the digital imaging and communications in medicine (DICOM) standard [37], 2) the clinical data interchange standards consortium (CDISC) [38], 3) the health level seven (HL7) standards [39] to guide data transfer and sharing, and FAIR (findable, accessible, interoperable, reusable) data principles, as a routine practice, despite the type of the analysis that will be applied for the data. This standardization is key to effective manipulation of the data and saves time and expenses in the long term.

3.2 MEDICAL IMAGING

3.2.1 Segmentation

Segmentation is the first fundamental step in radiomics analysis and can be performed manually by expert radiologists/clinicians or (semi-) automatically [40]. Both approaches have their pros and cons and the most suited one varies on a case-by-case basis [41], [42]. In general, automatic segmentation is more reproducible and faster than hand-made segmentation. The segmentation step determines which voxels within an image are analyzed: it is easy to see that the variability in segmentation (both human- and machine-driven) can introduce bias in the evaluation of the derived radiomic features [43]. For example, a semi-automatic segmentation method can result in different radiomic feature values than a manual delineation, as well as segmentation performed by two different physicians. Multiple segmentation is a method to limit the extent of this bias [44] including evaluation by multiple clinicians, perturbation of the segmentations with noise [45], and combination of diverse algorithms [46], [47]. Also, the segmentation models' performances are usually evaluated by comparing them against radiologists' segmentation. However, several studies have mentioned inter and intra-reader variability in segmentation tasks [48] [49]. Another study explores several factors that influence readers' concordance [50]. Hence, an algorithm's performance should be considered acceptable if it is within the range of intra and/or inter-reader variability.

3.2.2 Phantom studies and feature stability

Another source of variability in the preliminary radiomics phase is the inter-machine and inter-vendor differences between the scanners employed [51]. In most real-life situations, the radiomics study must rely on data acquired on different scanners from different producers thus, not taking into account this systematic source of uncertainty might jeopardize the radiomics model prediction capabilities. To overcome at least part of this intrinsic limitation, the use of phantoms (i.e. an object built in shape and materials as close as possible to human tissue and organs) is a suitable means to assess and account for the possible similarities and differences [52]. Radiomics features need also to be robust concerning other possible sources of variability such as target volume motion (expansion or shrinkage). To probe the feature resilience, test-retest approaches [53], [54] can be exploited to measure feature stability: for example, two datasets of images acquired within a small period from the same patient cohort or the use of cohorts from multiple sources [55], [56]. Volatile or robust features can be identified and excluded from model development. For example, a feature that is robust for the prediction of overall survival for lung cancer for a given dataset could be volatile for the prediction of pneumonitis in lung cancer (imaged and segmented in an alternative way).

To ameliorate the reproducibility of radiomics features, several methods of harmonization have been proposed in the literature. The ComBat method initially developed for genomics aims to remove non-biological differences related to scanner type to combine radiomics features extracted from data coming from different centers [36], [57], [58]. Other methods include training Neural

Network to standardize radiomics feature [59], intensity and diffusion maps harmonization [56], [60] and data augmentation with generative adversarial networks (GAN) [61]. For a complete overview see [62].

3.3 FEATURE EXTRACTION

The essence of radiomics is the extraction of quantitative image features to characterize ROIs. Hand-crafted radiomic features can be divided into five groups;

- size and shape based–features,
- descriptors of the image intensity histogram,
- descriptors of the (spatial) relationships between image voxels,
- features extracted from filtered images, and fractal features [63], [64].

Feature values are dependent on pre-processing methods applied to the images such as filtering, or intensity discretization and reconstruction. Furthermore, variation exists in feature nomenclature, mathematical definition, extraction methodology, and software implementation of the applied extraction algorithms [65]–[67]. To harmonize radiomic features and model reports, all these differences have to be taken into account and clear specifications are to be included with each model [68].

3.4 EXPLORATORY ANALYSIS

The true potential of radiomics approaches lies in the possibility to combine radiomic and non-radiomic features with the prediction target to create a single dataset. This approach allows the evaluation of possible correlations between features. However, some radiomics features that are highly correlated with other routine clinical features (such as tumor stage) might not provide additional meaningful information. Approaches such as (unsupervised) clustering, PCA (Principal component analysis) [69] or MRMR (maximum relevance minimum redundancy) [70] permit identification and eliminate redundancy, for instance, by reducing highly correlated features to a single representative archetypical feature. This is a fundamental step to avoid overfitting [71], [72]. On the other hand, additional data collected, for example, from multiple segmentations or phantom studies can be used to test the feature robustness [73], [74]. This process of reduction and/or exclusion should be described clearly, to avoid misinterpretation and help in the unambiguous identification of relevant features. Also, univariate correlations of single radiomics features with clinical outcome is part of the exploratory analysis and could inform the subsequent modeling step, underlining relations between single radiomics features with clinical covariates of interest.

3.5 MODELING

After feature extraction and possible reduction, the creation of the radiomic model encompasses three major steps: feature selection, modeling methodology, and validation. Regarding the choice of modeling methodology, the identification of the optimal machine-learning method is a crucial step; thus, in an ideal scenario, multiple methods should be utilized and compared [75] and their implementation should be comprehensively documented. Another fundamental point in the modeling phase is the validation, which has to be performed to verify the applicability of the model in a real-world situation. Ideally, the model should be internally and externally prospectively validated, using real world evidence or controlled trial datasets and the performance compared and reported [76]–[78].

3.5.1 Feature selection

The number of radiomic features that can be extracted from images is technically unlimited. Several different filters, feature categories, and other parameters can be used to mine the information hidden inside an imaging scan. Including all the possible features, even if practically possible, would result in overfitting which in turn renders the model useless for patients not previously evaluated (the so-called curse of dimensionality) [79], [80]. The most used approach is the reduction to archetypal features representing a group or class of features, identified by dimensionality reduction techniques. Several different kinds of clustering algorithms and PCA are available and also this choice has to be justified and reported in detail, to promote transparency and replicability. Again, the same feature might be relevant for a given dataset, segmented in a certain way for a specific end-point prediction but not important whit a different segmentation routine or a different cohort of patients.

3.5.2 Modeling methodology

The choice of modeling technique has been proven to affect prediction performance in radiomics [75]. Ideally, multiple modeling methodologies should be tested to select the best approach for the given data set and the other parameters involved in the creation of the model. Comparisons between Machine Learning (ML) and DL approaches are common [81], [82] and the final choice must be consider the performance of the model, also the applicability of the proposed strategy in a real-world situation, considering for example computational burden or explicability of the resulting predictions [83], [84]. Another key point in the selection of modeling methodology is replicability by other researchers, in the light of responsible and transparent research and innovation. This can be achieved, for example, by making the software code available in public repositories such as GitHub [85], Gitlab [86] and OpenML [87]. Also, many scientific journals put in place, in the last years, tools to help data and algorithms sharing, making these available to the scientific community.

3.5.3 Validation

Validation techniques are needed to assess whether the model is predictive for the target patient population or just for a particular subset of samples analyzed. Model performances are typically measured in terms of *discrimination* and *calibration*. Discrimination is represented by concordance statistics. For example, the discrimination metric for a binary outcome is the receiver operating characteristic (ROC) curve, or area under the ROC curve (AUC) [88]. The AUC relates to the sensitivity and specificity of the model and represents the probability that a random patient matching an outcome is assigned in the class-specific for that outcome with a larger probability than another random patient who does not match the outcome. The calibration, instead, is a measure of the agreement between observed outcomes and model predictions [89]. Calibration can be reported using a calibration plot and calibration-in-the-large/slope, with the Brier score, the mean squared prediction error, as a measure of overall performance.

The statistical methods used on both training and validation data sets need to be reported in detail. A valid model must exhibit statistical consistency between the training and validation sets. In terms of validation set selection, an externally validated model has more credibility than an internally validated one because validation with independent data sets is considered more robust [77], [89]–[91]. For "good radiomics practice", the reproducibility and replicability of the model should also be included in the validation step. Reproducibility relates to the verification of the result by independent researchers using the same methodology and data set, to verify the absence of errors, while replicability means the possibility of replicating the radiomics analysis with the same methodology but different appropriate datasets, to generalize the original findings [92]–[96]. Reproducibility and replicability in radiomics are, however, not possible if researchers do not disclose all the details of the analysis performed. Each radiomics model must be accompanied by the disclosure of imaging protocols, analyzed scans, segmentations of VOIs, detailed accounts of how features were extracted (including the formulae), and of the modeling methodology used (ideally, the code) [97].

4 ARTIFICIAL INTELLIGENCE FOR AUTOMATIC DELINEATIONS

A critical aspect of the Radiomics workflow is the need for segmentation of the region of interest from which relevant quantitative features are to be extracted. Manual delineations of anatomical structures and abnormal areas in different organs could be time-consuming and result in variability across annotators. AI-based auto-segmentation models can complement radiomic feature extraction, delivering a fully automated framework for further analysis of the radiomic features. In the past years, several researchers have proved the efficiency of AI in medical imaging for the segmentation task. A recent article [98] provides a comprehensive overview of various architectural types of deep learning models being researched for the segmentation of various anatomical structures. Of all the different architectures, U-net [99], a type of convolutional neural network designed specifically for biomedical image segmentation tasks has been researched and used successfully in different flavors for various medical image segmentation tasks. Another research article [100] presents a literature review of medical image segmentation based on U-net, focusing on the successful segmentation experience of U-net for different lesion regions in six medical imaging systems.

AI-based segmentation models can also be used to improve classification models to select appropriate slices from 3D images to make the model focus on the region of interest and to prevent overfitting by learning the irrelevant patterns existing outside the ROI. Deep learning models have been proven the most successful approach to auto-segmentation of medical images and the output can be easily integrated into a full radiomics pipeline. Also, the radiomics model itself, the so-called radiomics signature, can be based on hand-crafted features which can be used in traditional Machine Learning models and “deep” features for an end-to-end deep learning approach to the classification or prediction task at hand.

In the present thesis, I explored the potential of applying Artificial intelligence approaches in combination with Radiomics for several disease detection and classification pertaining to unmet clinical needs.
In the development of the research trajectory, I have taken into account three main topics:

Robustness and generalizability: Generalizability and robustness are essential for real clinical applications. Generalization refers to a model’s ability to adapt properly to new, previously unseen data. Each model was externally validated on data coming from different sources compared to the training dataset to verify generalizability, thereby ensuring the reliability of a model’s performance on a dataset presenting from sources employing varying imaging parameters, from a population presenting ‘unseen’ abnormalities in the region of interest.

Explainability of the model decision: Several explainability methods, both post-hoc and ante-hoc, have been used to assure the transparency of the model decision, linking the output or the input to a tangible clinical and radiological characteristic of the medical images analyzed. The Explainability of AI models promotes trust among non-specialists and the acceptance of these methods in the clinic.

Clinical relevance: Unmet clinical need for a particular tool was researched based on the literature and interactions with clinicians at Maastricht university medical center and the developed models were evaluated on external validation datasets which represent populations in real-world clinical settings.

5 RESEARCH HYPOTHESIS

The research work presented in this thesis focuses **on the overall hypothesis that (semi) automated Radiomics and AI based methodologies can produce generalizable performance equivalent/non-inferior/superior to that of an expert human charged with the same tasks and is exemplified in detection, diagnosis, and treatment response prediction use cases. The (semi) automated radiomics and AI based methodologies throughout my research were explored specifically on the following clinical problem statements,**

- Covid-19 diagnosis and differential diagnosis using CT imaging
- Pulmonary embolism detection and diagnosis using CTPA imaging
- Treatment response prediction in patients with non-metastatic non-small cell lung cancer using CT imaging
- Menière's disease diagnosis using MRI imaging

The thesis is organized around the following chapters:

Chapter 2 focuses on the combination of an automatic deep learning segmentation model for lungs to expedite and automatize the application of radiomics signature to detect COVID-19 infected patients from chest CT scans.

Chapter 3, upon building on the expertise acquired in the previous approach, reports the development of an end-to-end deep learning framework for the differential diagnosis of pneumonia patients infected with COVID-19, compared to Influenza/CAP patients and no infection patients. The automatic lung segmentation model was combined with a lung abnormalities segmentation model which, filtering the lung slices with no abnormal radiological features, assures that the classification model focuses its decision only on the relevant slices from each chest CT scan volume. The explainability is assured by an automatic clinical summary report, which allows the clinician to review all the slices with the segmented abnormalities and the associated classification probabilities for each class produced by the model.

In Chapter 4, the detection of pulmonary embolism from chest CT angiography was tackled with a fully automated deep learning framework, training the model on a slice level and aggregating the results to obtain patient-level prediction, after having chosen the optimal classification threshold on the internal validation set. The Grad-CAM method was used to assess model explainability, which was quantified by comparing the activations maps with the radiologist's manual delineation of emboli.

Chapter 5 focuses on the application of a radiomic signature to predict survival as the primary outcome in patients with non-metastatic non-small cell lung cancer (NSCLC) for chemotherapy. We have compared the performances of the models trained on the features extracted from the Gross Tumour Volume (GTV) segmented manually and by using an AI-based GTV segmentation model.

Chapter 6 presents the development and validation of a deep learning framework for the identification of metastatic bone foci on bone scintigraphy scans. The model can distinguish between metastatic and non-metastatic disease and the model performances were compared with radiation oncologists in an *in-silico* clinical study, also in terms of speed of classification. Also, in this case, Grad-Cam explanation maps were used to assess model explainability.

In Chapter 7, I have investigated the application of radiomic signature for the diagnosis of Menière's disease from the features extracted from the inner ear region on conventional T2- weighted MRI scans

In Chapter 8, as an attempt to fully automate the radiomics-based diagnostic tool for Menière's disease, an AI-based auto-segmentation model was explored for segmentation of the inner ear on MRI scans. The trained model was validated for generalizability on multi-vendor, multi-centric data with diverse abnormalities presenting in the inner ear.

To conclude, chapter 9 provides a general discussion and future perspectives for AI-based diagnostic, prognostic, and treatment outcome applications.

6 REFERENCES

[1] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.," *Nat Commun*, vol. 5, p. 4006, Jun. 2014, doi: 10.1038/ncomms5006.

[2] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nat Rev Clin Oncol*, vol. 8, no. 3, pp. 184–187, 2011, doi: 10.1038/nrclinonc.2010.227.

[3] P. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur J Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, doi: 10.1016/j.ejca.2011.11.036.

[4] P. Lambin *et al.*, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12. Nature Publishing Group, pp. 749–762, Dec. 01, 2017. doi: 10.1038/nrclinonc.2017.141.

[5] V. Kumar *et al.*, "Radiomics: The process and the challenges," *Magn Reson Imaging*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012, doi: 10.1016/j.mri.2012.06.010.

[6] AA.VV., "Medicine: Computers by the Bedside," *Nature*, vol. 224, no. 5220, pp. 636–637, 1969, doi: 10.1038/224636b0.

[7] H. M. Schoolman and L. M. Bernstein, "Computer use in diagnosis, prognosis, and therapy," *Science (1979)*, vol. 200, no. 4344, pp. 926 LP – 931, May 1978, doi: 10.1126/science.347580.

[8] F. Zerka *et al.*, "Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care," *JCO Clin Cancer Inform*, no. 4, pp. 184–200, 2020, doi: 10.1200/cci.19.00047.

[9] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data.," *Radiology*, 2016, doi: 10.1148/radiol.2015151169.

[10] Z. Liu *et al.*, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges," *Theranostics*, vol. 9, no. 5, pp. 1303–1322, 2019, doi: 10.7150/thno.30309.

[11] B.-H. Zheng *et al.*, “Radiomics score: a potential prognostic imaging feature for postoperative survival of solitary HCC patients,” *BMC Cancer*, vol. 18, no. 1, p. 1148, 2018, doi: 10.1186/s12885-018-5024-z.

[12] S. Monti *et al.*, “DCE-MRI pharmacokinetic-based phenotyping of invasive ductal carcinoma: A radiomic study for prediction of histological outcomes,” *Contrast Media Mol Imaging*, vol. 2018, 2018, doi: 10.1155/2018/5076269.

[13] J. E. Bibault *et al.*, “Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer,” *Sci Rep*, vol. 8, no. 1, pp. 1–8, 2018, doi: 10.1038/s41598-018-30657-6.

[14] X. li Song, J. L. Ren, D. Zhao, L. Wang, H. Ren, and J. Niu, “Radiomics derived from dynamic contrast-enhanced MRI pharmacokinetic protocol features: the value of precision diagnosis ovarian neoplasms,” *Eur Radiol*, vol. I, 2020, doi: 10.1007/s00330-020-07112-0.

[15] E. Roelofs, A. Dekker, E. Meldolesi, R. G. P. M. van Stiphout, V. Valentini, and P. Lambin, “International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining,” *Radiotherapy and Oncology*, vol. 110, no. 2, pp. 370–374, Feb. 2014, doi: 10.1016/j.radonc.2013.11.001.

[16] E. Roelofs, L. Persoon, S. Nijsten, W. Wiessler, A. Dekker, and P. Lambin, “Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial,” *Radiotherapy and Oncology*, vol. 108, no. 1, pp. 174–179, Jul. 2013, doi: 10.1016/j.radonc.2012.09.019.

[17] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records,” *Sci Rep*, vol. 6, no. 1, p. 26094, 2016, doi: 10.1038/srep26094.

[18] K. T. Nead *et al.*, “Androgen Deprivation Therapy and Future Alzheimer’s Disease Risk,” *Journal of Clinical Oncology*, vol. 34, no. 6, pp. 566–571, Dec. 2015, doi: 10.1200/JCO.2015.63.6266.

[19] R. A. Gatenby, O. Grove, and R. J. Gillies, “Quantitative Imaging in Cancer Evolution and Ecology,” *Radiology*, vol. 269, no. 1, pp. 8–14, Oct. 2013, doi: 10.1148/radiol.13122697.

[20] P. Lambin *et al.*, “Decision support systems for personalized and participative radiation oncology,” *Adv Drug Deliv Rev*, vol. 109, pp. 131–153, 2017, doi: https://doi.org/10.1016/j.addr.2016.01.006.

[21] A. J. Vickers, “Prediction Models: Revolutionary in Principle, But Do They Do More Good Than Harm?,” *Journal of Clinical Oncology*, vol. 29, no. 22, pp. 2951–2952, Jun. 2011, doi: 10.1200/JCO.2011.36.1329.

[22] E. J. Limkin *et al.*, “Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology,” *Annals of Oncology*, vol. 28, no. 6, pp. 1191–1206, Jun. 2017, doi: 10.1093/annonc/mdx034.

[23] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images Are More than Pictures, They Are Data,” *Radiology*, vol. 278, no. 2, pp. 563–577, Nov. 2015, doi: 10.1148/radiol.2015151169.

[24] H. J. W. L. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.,” *Nat Commun*, vol. 5, p. 4006, Jun. 2014, doi: 10.1038/ncomms5006.

[25] R. T. H. Leijenaar *et al.*, “Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: A multicenter study,” *British Journal of Radiology*, vol. 91, no. 1086, p. 20170498, Jun. 2018, doi: 10.1259/bjr.20170498.

[26] A. F. Leite, K. de F. Vasconcelos, H. Willems, and R. Jacobs, “Radiomics and Machine Learning in Oral Healthcare,” *Proteomics Clin Appl*, vol. 14, no. 3, 2020, doi: 10.1002/prca.201900040.

[27] H. Sun *et al.*, “Psychoradiologic utility of MR imaging for diagnosis of attention deficit hyperactivity disorder: A radiomics analysis,” *Radiology*, vol. 287, no. 2, pp. 620–630, 2018, doi: 10.1148/radiol.2017170226.

[28] P. Lovinfosse, D. Visvikis, R. Hustinx, and M. Hatt, “FDG PET radiomics: a review of the methodological aspects,” *Clin Transl Imaging*, vol. 6, no. 5, pp. 379–391, 2018, doi: 10.1007/s40336-018-0292-9.

[29] L. Sibille *et al.*, “18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks,” *Radiology*, vol. 294, no. 2, pp. 445–452, Dec. 2019, doi: 10.1148/radiol.2019191114.

[30] P. Lambin *et al.*, “Radiomics: the bridge between medical imaging and personalized medicine.,” *Nat Rev Clin Oncol*, vol. 14, no. 12, pp. 749–762, Oct. 2017, doi: 10.1038/NRCLINONC.2017.141.

[31] J. Zhong *et al.*, “An updated systematic review of radiomics in osteosarcoma: utilizing CLAIM to adapt the increasing trend of deep learning application in radiomics,” *Insights Imaging*, vol. 13, no. 1, pp. 1–15, Aug. 2022, doi: 10.1186/S13244-022-01277-6/TABLES/5.

[32] S. Röhrich, J. Hofmanninger, F. Prayer, H. Müller, H. Prosch, and G. Langs, “Prospects and Challenges of Radiomics by Using Nononcologic Routine Chest CT,” *Radiol Cardiothorac Imaging*, vol. 2, no. 4, p. e190190, Aug. 2020, doi: 10.1148/ryct.2020190190.

[33] M. Bogowicz *et al.*, “Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models,” *Radiotherapy and Oncology*, vol. 125, no. 3, pp. 385–391, 2017, doi: 10.1016/j.radonc.2017.10.023.

[34] J. E. van Timmeren *et al.*, “Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images,” *Radiotherapy and Oncology*, vol. 123, no. 3, pp. 363–369, 2017, doi: 10.1016/j.radonc.2017.04.016.

[35] S. S. F. Yip and H. J. W. L. Aerts, “Applications and limitations of radiomics,” *Phys Med Biol*, vol. 61, no. 13, pp. R150–R166, 2016, doi: 10.1088/0031-9155/61/13/r150.

[36] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, “Validation of a method to compensate multicenter effects affecting CT radiomics,” *Radiology*, vol. 291, no. 1, pp. 53–59, 2019, doi: 10.1148/radiol.2019182023.

[37] B. Gibaud, “The DICOM standard: A brief overview,” *NATO Science for Peace and Security Series B: Physics and Biophysics*, pp. 229–238, 2008, doi: 10.1007/978-1-4020-8752-3_13/COVER.

[38] R. Facile *et al.*, “Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey,” *JMIR Med Inform*, vol. 10, no. 1, Jan. 2022, doi: 10.2196/30363.

[39] “HL7 | Digital Healthcare Research.” https://digital.ahrq.gov/hl7 (accessed Sep. 12, 2022).

[40] D. F. Polan, S. L. Brady, and R. A. Kaufman, “Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study,” *Phys Med Biol*, vol. 61, no. 17, pp. 6553–6569, 2016, doi: 10.1088/0031-9155/61/17/6553.

[41] N. Porz *et al.*, “Multi-modal glioblastoma segmentation: Man versus machine,” *PLoS One*, vol. 9, no. 5, pp. 1–9, 2014, doi: 10.1371/journal.pone.0096873.

[42] N. Porz *et al.*, “Fully automated enhanced tumor compartmentalization: Man vs. Machine reloaded,” *PLoS One*, vol. 11, no. 11, pp. 1–16, 2016, doi: 10.1371/journal.pone.0165302.

[43] Y. Balagurunathan *et al.*, “Reproducibility and Prognosis of Quantitative Features Extracted from CT Images,” *Transl Oncol*, vol. 7, no. 1, pp. 72–87, 2014, doi: https://doi.org/10.1593/tlo.13844.

[44] W. Grootjans *et al.*, “The Impact of Optimal Respiratory Gating and Image Noise on Evaluation of Intratumor Heterogeneity on 18F-FDG PET Imaging of Lung Cancer.,” *J Nucl Med*, vol. 57, no. 11, pp. 1692–1698, Nov. 2016, doi: 10.2967/jnumed.116.173112.

[45] A. Zwanenburg *et al.*, “Assessing robustness of radiomic features by image perturbation,” *Sci Rep*, vol. 9, no. 1, p. 614, 2019, doi: 10.1038/s41598-018-36938-4.

[46] A. A. Farag, H. E. A. E. Munim, J. H. Graham, and A. A. Farag, “A Novel Approach for Lung Nodules Segmentation in Chest CT Using Level Sets,” *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5202–5213, 2013, doi: 10.1109/TIP.2013.2282899.

[47] J. Guiot *et al.*, “Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19,” *Diagnostics* , vol. 11, no. 1. 2021. doi: 10.3390/diagnostics11010041.

[48] C. C. Willers *et al.*, “Inter-reader variation in lung segmentation of functional lung MRI quantification,” *European Respiratory Journal*, vol. 54, no. suppl 63, p. PA333, Sep. 2019, doi: 10.1183/13993003.CONGRESS-2019.PA333.

[49] F. Rizzetto *et al.*, “Impact of inter-reader contouring variability on textural radiomics of colorectal liver metastases,” *Eur Radiol Exp*, vol. 4, no. 1, Dec. 2020, doi: 10.1186/S41747-020-00189-8.

[50] A. M. Schmid *et al.*, “Radiologists and Clinical Trials: Part 1 The Truth About Reader Disagreements,” *Ther Innov Regul Sci*, vol. 55, no. 6, pp. 1111–1121, Nov. 2021, doi: 10.1007/S43441-021-00316-6/TABLES/1.

[51] D. Mackin *et al.*, “Measuring Computed Tomography Scanner Variability of Radiomics Features,” *Invest Radiol*, vol. 50, no. 11, 2015.

[52] R. Berenguer *et al.*, “Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters,” *Radiology*, vol. 288, no. 2, pp. 407–415, 2018, doi: 10.1148/radiol.2018172361.

[53] B. Zhao *et al.*, “Reproducibility of radiomics for deciphering tumor phenotype with imaging,” *Sci Rep*, vol. 6, no. 1, p. 23428, 2016, doi: 10.1038/srep23428.

[54] C. Haarburger, G. Müller-Franzes, L. Weninger, C. Kuhl, D. Truhn, and D. Merhof, “Radiomics feature reproducibility under inter-rater variability in segmentations of CT images,” *Sci Rep*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-69534-6.

[55] J. E. van Timmeren *et al.*, “Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific?,” *Tomography*, vol. 2, no. 4, pp. 361–365, Dec. 2016, doi: 10.18383/j.tom.2016.00208.

[56] J. Peerlings *et al.*, “Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial,” *Sci Rep*, vol. 9, no. 1, p. 4800, 2019, doi: 10.1038/s41598-019-41344-5.

[57] R. Da-ano *et al.*, “Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies,” *Sci Rep*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-66110-w.

[58] F. Orlhac *et al.*, “A postreconstruction harmonization method for multicenter radiomic studies in PET,” *Journal of Nuclear Medicine*, vol. 59, no. 8, pp. 1321–1328, 2018, doi: 10.2967/jnumed.117.199935.

[59] V. Andrearczyk, A. Depeursinge, and H. Müller, “Neural network training for cross-protocol radiomic feature standardization in computed tomography,” *Journal of Medical Imaging*, vol. 6, no. 02, p. 1, 2019, doi: 10.1117/1.jmi.6.2.024008.

[60] A. Crombé *et al.*, “Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients,” *Sci Rep*, vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-72535-0.

[61] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” *Sci Rep*, vol. 9, no. 1, pp. 1–9, 2019, doi: 10.1038/s41598-019-52737-x.

[62] R. Da-Ano, D. Visvikis, and M. Hatt, “Harmonization strategies for multicenter radiomics investigations,” *Phys Med Biol*, vol. 65, no. 24, p. 24TR02, 2020, doi: 10.1088/1361-6560/aba798.

[63] S. Ranjbar and J. Ross Mitchell, “Chapter 8 - An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine,” A. Depeursinge, O. S. Al-Kadi, and J. R. B. T.-B. T. A. Mitchell, Eds. Academic Press, 2017, pp. 223–245. doi: https://doi.org/10.1016/B978-0-12-812133-7.00008-9.

[64] L. Antonelli, M. R. Guarracino, L. Maddalena, and M. Sangiovanni, “Integrating imaging and omics data: A review,” *Biomed Signal Process Control*, vol. 52, pp. 264–280, 2019, doi: https://doi.org/10.1016/j.bspc.2019.04.032.

[65] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis, “Characterization of PET/CT images using texture analysis: the past, the present… any future?,” *Eur J Nucl Med Mol Imaging*, vol. 44, no. 1, pp. 151–165, 2017, doi: 10.1007/s00259-016-3427-0.

[66] Y.-H. D. Fang *et al.*, “Development and evaluation of an open-source software package ‘CGITA’ for quantifying tumor heterogeneity with molecular images.,” *Biomed Res Int*, vol. 2014, p. 248505, 2014, doi: 10.1155/2014/248505.

[67] L. Zhang, D. V Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court, “IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics.,” *Med Phys*, vol. 42, no. 3, pp. 1341–1353, Mar. 2015, doi: 10.1118/1.4908210.

[68] A. Zwanenburg *et al.*, “The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping,” *Radiology*, vol. 295, no. 2, pp. 328–338, 2020, doi: 10.1148/radiol.2020191145.

[69] D. A. P. Delzell, S. Magnuson, T. Peter, M. Smith, and B. J. Smith, “Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data ,” *Frontiers in Oncology* , vol. 9. p. 1393, 2019.

[70] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data,” *BMC Bioinformatics*, vol. 18, no. 1, p. 9, 2017, doi: 10.1186/s12859-016-1423-9.

[71] X. Ying, “An Overview of Overfitting and its Solutions,” p. 22022, 2019, doi: 10.1088/1742-6596/1168/2/022022.

[72] H. Song, M. Kim, D. Park, and J.-G. Lee, “PRESTOPPING: HOW DOES EARLY STOPPING HELP GENERALIZATION AGAINST LABEL NOISE?”

[73] S.-H. Lee, H. Cho, H. Y. Lee, and H. Park, “Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer,” *Cancer Imaging*, vol. 19, no. 1, p. 54, 2019, doi: 10.1186/s40644-019-0239-z.

[74] Q. Qiu *et al.*, “Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability,” *Quantitative Imaging in Medicine and Surgery; Vol 9, No 3 (March 2019): Quantitative Imaging in Medicine and Surgery*, 2019.

[75] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, “Machine Learning methods for Quantitative Radiomic Biomarkers,” *Sci Rep*, vol. 5, no. 1, p. 13087, 2015, doi: 10.1038/srep13087.

[76] N. Garau *et al.*, “External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis,” *Med Phys*, vol. n/a, no. n/a, Jun. 2020, doi: 10.1002/mp.14308.

[77] T. P. A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. M. Moons, “A new framework to enhance the interpretation of external validation studies of clinical prediction models,” *J Clin Epidemiol*, vol. 68, no. 3, pp. 279–289, Mar. 2015, doi: 10.1016/j.jclinepi.2014.06.018.

[78] P. Lambin *et al.*, “Decision support systems for personalized and participative radiation oncology,” *Adv Drug Deliv Rev*, vol. 109, pp. 131–153, 2017, doi: https://doi.org/10.1016/j.addr.2016.01.006.

[79] J. E. Park, S. Y. Park, H. J. Kim, and H. S. Kim, “Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives.,” *Korean J Radiol*, vol. 20, no. 7, pp. 1124–1137, Jul. 2019, doi: 10.3348/kjr.2018.0070.

[80] N. Altman and M. Krzywinski, “The curse(s) of dimensionality,” *Nat Methods*, vol. 15, no. 6, pp. 399–400, 2018, doi: 10.1038/s41592-018-0019-x.

[81] E. Capobianco and J. Deng, “Radiomics at a Glance: A Few Lessons Learned from Learning Approaches,” *Cancers* , vol. 12, no. 9. 2020. doi: 10.3390/cancers12092453.

[82] M. Avanzo *et al.*, “Machine and deep learning methods for radiomics,” *Med Phys*, vol. 47, no. 5, pp. e185–e202, May 2020, doi: 10.1002/mp.13678.

[83] P. Giraud *et al.*, “Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers ,” *Frontiers in Oncology* , vol. 9. p. 174, 2019.

[84] G. Choy *et al.*, “Current Applications and Future Impact of Machine Learning in Radiology,” *Radiology*, vol. 288, no. 2, pp. 318–328, Jun. 2018, doi: 10.1148/radiol.2018171820.

[85] “https://github.com/.”

[86] “https://about.gitlab.com/.”

[87] “https://www.openml.org/.”

[88] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988, doi: 10.2307/2531595.

[89] E. W. Steyerberg *et al.*, “Assessing the performance of prediction models: a framework for traditional and novel measures.,” *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.

[90] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement,” *BMC Med*, vol. 13, no. 1, p. 1, 2015, doi: 10.1186/s12916-014-0241-z.

[91] S. Lemeshow and D. W. Hosmer Jr., “A review of goodness of fit statistics for use in the development of logistic regression models1,” *Am J Epidemiol*, vol. 115, no. 1, pp. 92–106, Jan. 1982, doi: 10.1093/oxfordjournals.aje.a113284.

[92] J. T. Leek and R. D. Peng, “Statistics: P values are just the tip of the iceberg.,” *Nature*, vol. 520, no. 7549, p. 612, Apr. 2015, doi: 10.1038/520612a.

[93] C. Drummond, “Replicability is not reproducibility: nor is it good science,” *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada,2009*. 2009.

[94] R. D. Peng, “Reproducible Research in Computational Science,” *Science (1979)*, vol. 334, no. 6060, pp. 1226 LP – 1227, Dec. 2011, doi: 10.1126/science.1213847.

[95] R. D. Peng, F. Dominici, and S. L. Zeger, “Reproducible Epidemiologic Research,” *Am J Epidemiol*, vol. 163, no. 9, pp. 783–789, Mar. 2006, doi: 10.1093/aje/kwj093.

[96] S. Fiset *et al.*, “Repeatability and reproducibility of MRI-based radiomic features in cervical cancer,” *Radiotherapy and Oncology*, vol. 135, pp. 107–114, Jun. 2019, doi: 10.1016/j.radonc.2019.03.001.

[97] P. Kalendralis *et al.*, “FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections,” *Med Phys*, vol. 47, no. 11, pp. 5931–5940, Nov. 2020, doi: https://doi.org/10.1002/mp.14322.

[98] X. Liu, L. Song, S. Liu, and Y. Zhang, “A Review of Deep-Learning-Based Medical Image Segmentation Methods,” *Sustainability 2021, Vol. 13, Page 1224*, vol. 13, no. 3, p. 1224, Jan. 2021, doi: 10.3390/SU13031224.

[99] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, May 2015, Accessed: Feb. 11, 2022. [Online]. Available: https://arxiv.org/abs/1505.04597v1

[100] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, “Medical image segmentation based on U-Net: A review,” *Journal of Imaging Science and Technology*, vol. 64, no. 2, Mar. 2020, doi: 10.2352/J.IMAGINGSCI.TECHNOL.2020.64.2.020508.

Chapter 2

Development and validation of an automated radiomic CT signature for detecting COVID-19

Based on Guiot, J.; **Vaidyanathan, A.**; Deprez, L.; Zerka, F.; Danthine, D.; Frix, A.-N.; Thys, M.; Henket, M.; Canivet, G.; Mathieu, S.; Eftaxia, E.; Lambin, P.; Tsoutzidis, N.; Miraglio, B.; Walsh, S.; Moutschen, M.; Louis, R.; Meunier, P.; Vos, W.; Leijenaar, R.T.H.; Lovinfosse, P., *Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19*. Diagnostics 2021, 11, 41. https://doi.org/10.3390/diagnostics11010041

Our research presented in this chapter exploits the use of an AI-based auto-segmentation model in combination with radiomics and AI-based classification models to analyze features from lungs, for diagnosis of Covid-19. The coronavirus disease 2019 (COVID-19) outbreak reached a pandemic status in early 2020. Drastic measures of social distancing were enforced in society and healthcare systems were being pushed to and beyond their limits. To help in the fight against this threat to human health, a fully automated AI framework was developed to extract and analyze radiomics features from volumetric chest computed tomography (CT) exams. In this study, we hypothesize that a fully automated solution comprising an AI-based segmentation model to segment whole lungs from the CT and the radiomic analysis of features extracted from the segmented lung region can identify a diagnostic signature for COVID-19 infection, based on standard-of-care chest CT imaging. The detection model was developed on a dataset of 1381 patients (181 COVID-19 patients plus 1200 non-COVID control patients). A second, independent dataset of 197 RT-PCR confirmed COVID-19 patients and 500 control patients was used to assess the performance of the model. Diagnostic performance was assessed by the area under the receiver operating characteristic curve (AUC). The model had an AUC of 0.882 (95% CI: 0.851–0.913) in the independent test dataset (641 patients). The optimal decision threshold, considering the cost of false negatives twice as high as the cost of false positives, resulted in an accuracy of 85.18%, a sensitivity of 69.52%, a specificity of 91.63%, a negative predictive value (NPV) of 94.46% and a positive predictive value (PPV) of 59.44%. Benchmarked against RT-PCR confirmed cases of COVID-19, our AI framework was able to accurately differentiate COVID-19 from routine clinical conditions in a fully automated fashion. Thus, providing a rapid accurate diagnosis in patients suspected of COVID-19 infection, facilitating the timely implementation of isolation procedures and early intervention.

1 BACKGROUND

The rapid outbreak of coronavirus disease 2019 (COVID-19), originating from severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) infection, had been a public health emergency of international concern [1]. The outbreak of COVID-19 had a terrible impact on the economy and society all around the world. Globally there had been 71,554,018 confirmed cases and 1,613,671 deaths as of December 20, 2020 [2]. The presence of the disease was confirmed by reverse-transcription polymerase chain reaction (RT-PCR) [3]. There was, however, evidence that the sensitivity of RT-PCR may not be optimal for the objective of very early detection and early intervention of COVID-19 patients [4]. Due to the limited supply of RT-PCR kits, the lengthy turnaround times, and the emergence of false-negative cases, some experts propose to diagnose suspected cases using the widely available, time-saving, and non-invasive imaging approach of chest computed tomography (CT) rather than RT-PCR [5,6]. CT can capture imaging features from the lung, associated with COVID-19 [7], in the early stages of the disease [8]; CT could thus serve as an efficient and effective way to flag, diagnose, and possibly triage COVID-19 patients, in a more timely manner compared to traditional confirmation tests. Despite these advantages, there are several open questions on the use of CT for these purposes [9,10], due to increased radiation exposure of the population and the risk of cross-infection if disinfection is not properly implemented. Notwithstanding these concerns, the use of chest CT for COVID-19 diagnosis needs a proper toolset, to allow clinicians to exploit fully this technology. In the medical imaging domain, Artificial intelligence (AI) coupled with machine learning technology has accomplished impressive results due to the intrinsic properties of machine vision [11–14] and can be leveraged in this scenario. More so, the radiomics approach which was already proved to be extremely successful for cancer diagnosis and prognosis [15] might be also applied in this context. Radiomics is the high-throughput mining of quantitative image features from standard-of-care medical imaging that enables data to be extracted and applied within clinical decision support systems to improve diagnostic, prognostic, and/or predictive accuracy [16]. Conceptually, radiomics is a bridge between imaging and precision medicine [17].

In this study, we hypothesize that a fully automated solution comprising an AI-based segmentation model to segment whole lungs from the CT and the radiomic analysis of features extracted from the segmented lung region can identify a diagnostic signature for COVID-19 infection, based on standard-of-care chest CT imaging. As a result, we present a fully automated AI framework to detect COVID-19 using chest CT, referred to as COVIA ('coronavirus intelligence artificielle'), and validate its performance in an independent test cohort. This model has been built in a clinical real-life environment, the first Belgian wave of COVID-19 infection. This was mainly used for symptomatic patients with the European standard of care. Contrary to what is seen in other countries, we used CT scans from all patients reducing the bias found in some studies where clinicians reserved CT only for severe cases.

2 MATERIALS AND METHODS

2.1 ETHICS

The study has been approved by the local ethics committee of the CHU-Liège (EC number 116/2020). The institutional review board waived the requirement to obtain written informed consent for this retrospective case series since all analyses were performed on de-identified (i.e., pseudo-anonymized) data and there was no potential risk to patients.

2.2 SUBJECTS

Three cohorts of patients were included retrospectively in this study. Cohorts came from two sites (CHU Sart-Tilman and CHU Notre Dame des Bruyères) in Liège, Belgium. The first cohort (label: COVID) consists of all patients with COVID-19 infection confirmed by RT-PCR that underwent chest CT imaging before March 28th, 2020. The second cohort (label: Control) consisted of consecutive patients that underwent chest CT imaging between October 1st, 2019, and October 24th, 2019, which ensures that none of these patients were infected by COVID-19. The third cohort (label: Test) consisted of 697 consecutive patients that underwent chest CT imaging between August 12th, 2019, and April 6th, 2020. The Test cohort presents no overlap with COVID and Control cohorts and was acquired at a different time point. Within this cohort, 197 patients had RT-PCR confirmed COVID-19, whereas the remaining 500 patients tested negative for COVID-19. The first (COVID) and second (Control) cohorts were used for model development, and the third cohort (Test) was used as an independent test set. No other inclusion or exclusion criteria were considered while collecting the data. This resulted in sets of CT images from either COVID-19 infected patients or non-infected patients (normal and with a variety of diseases) representing real-life conditions.

2.3 RADIOMICS

The hypothesis is that quantitative analysis of medical image data *via* automatic or semi-automatic software can provide more and better information than that of a physician [18,19]. The schematic representation in Figure. 2.1 depicts the radiomics workflow applied in this study. The following sections will detail each step in the workflow.

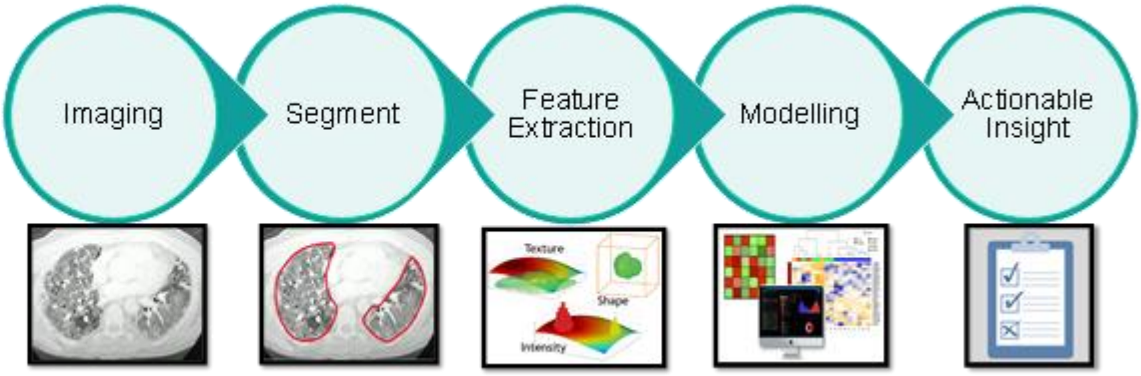


Figure 2.1 Schematic representation of the radiomics analysis steps: Imaging: Chest CT scans of healthy and COVID-19 infected patients were collected and divided between training and testing cohorts. Segment: The scan were automatically segmented to delineate the region of interest in the lung. Feature extraction: Hand-crafted radiomics features were extracted from the region of interest. The radiomics features were used to train the AI model and the performances were validated in the test set. Actionable insight: The model discrimination performances were assessed in terms of accuracy, sensitivity, specificity, NPV, and PPV.

2.4 IMAGING

All CT images used in the study were acquired on one of five multidetector CT scanners (Siemens Edge Plus (2), GE Revolution CT (1), and GE Brightspeed (2)) available at the sites. Since CT images were collected retrospectively, no standardized scan protocol was available over the complete dataset. To prevent excess variability in the imaging used for model generation, the following criteria for radiomic analysis were used:

- Lungs completely visible in the scan
- Slice increment less than 1.5 mm
- No missing slices
- For GE scans: STANDARD reconstruction kernel
- For Siemens scans: B30-range reconstruction intervals

2.5 LUNG SEGMENTATION

The lungs were segmented as a single structure using an AI model based on a 2D Feature Pyramid Network [20] adapted with ResNext blocks [21] in the encoder. The model was trained and validated on the following datasets,

1) Publicly available dataset with 888 CT scans and the corresponding reference annotations for lungs available from the LUNA16 challenge [22]
2) Publicly available data from the cancer imaging archive [23] containing CT scans of 422 confirmed non-small cell lung cancer cases, along with manual segmentations of the left and right lungs. The segmentations were performed by an experienced radiologist and these segmentations were used as a reference standard.

The network was trained with the 2D axial slices clipped at a window width of 1500 HU and a window level of –600 HU and with their corresponding reference labels. The network's weights were updated by using the Adam optimizer at an initial learning rate of $1e^{-5}$ [24]. The model was trained using customized Jaccard loss [25] as an objective function where the loss is calculated in a mini-batch of 8 images per iteration. The network was trained for 5 epochs and at the end of each epoch, the Jaccard loss was calculated on the model's predictions to ensure validation loss convergence.

The deep learning-based lung segmentation achieved a mean Dice similarity coefficient score of 0.92 across the publicly available datasets which indicate adequate precision (i.e. no significant over or under segmentation). The predicted segmentations by the AI model were used for the extraction of radiomics features. Figure 2.2 shows example segmentations for four patients from both the COVID and Control groups.

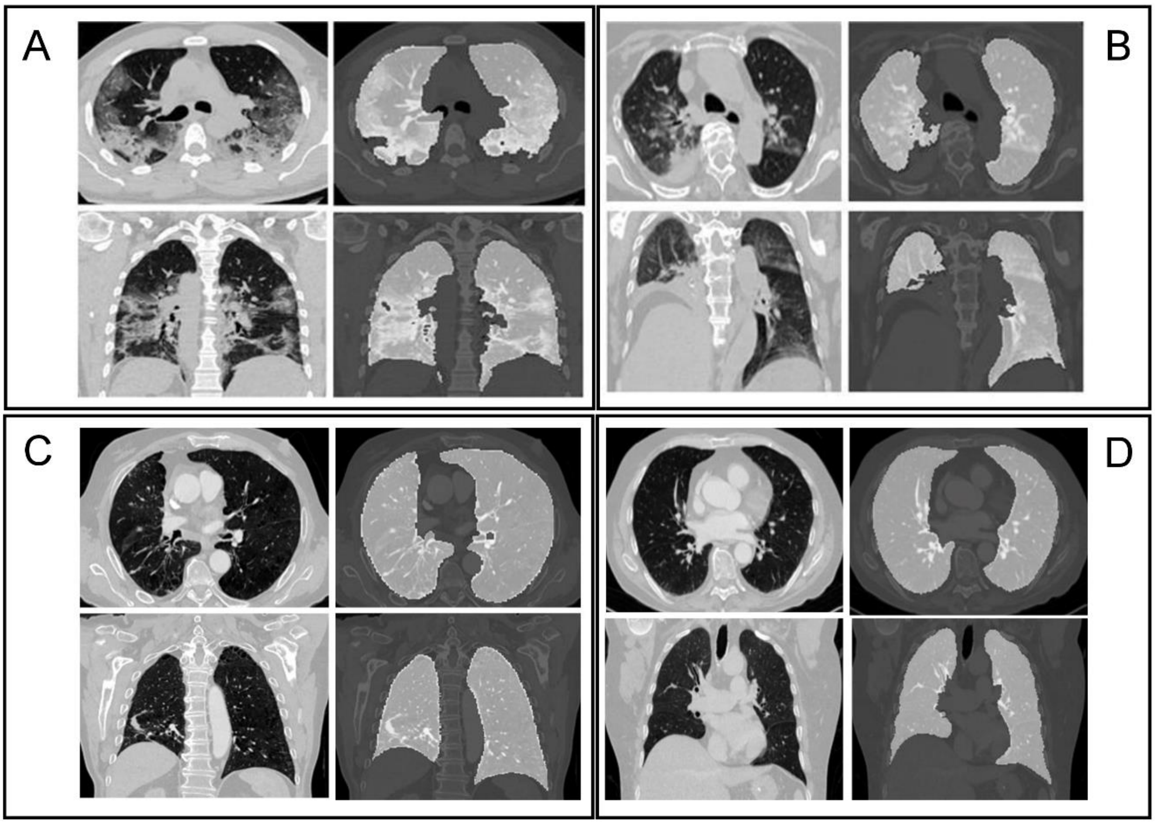


Figure 2.2 Axial and coronal slices with accompanying segmentation masks. A) Typical aspect of COVID-19 pneumonia characterized by bilateral multilobe ground-glass opacities of peripheral/subpleural distribution, with intralesional reticulations, presenting a “crazy paving” aspect. Also found are subpleural atelectasis and retraction bronchiectasis, typical of organizing pneumonia; B) Atypical aspect of COVID-19 pneumonia, with posterior right lower lobe condensation and retraction of the ipsilateral diaphragm. Central and peripherical ground-glass opacities in the right lower lobe, right upper lobe, and left upper lobe; C) Typical chronic obstructive pulmonary disease COPD chest CT characterized by severe centrilobular and para-septal emphysema, associated with cylindrical bronchiectasis and bronchial walls thickening. Right peripherical upper lobe tree in bud pattern seen in bronchiolitis. Middle lobe crescent-shaped atelectasis condensation; D) Normal chest CT.

2.6 FEATURE EXTRACTION

For each patient, 166 image features were extracted from the lung segmentation using RadiomiX (OncoRadiomics SA, Liège, Belgium) based upon quantitative image analysis technology. The extracted features comprised first order and intensity histogram statistics, texture (gray-level-co-occurrence, gray-level-run-length, gray-level-size-zone, gray-level-distance-zone, neighborhood gray-tone difference, and neighboring gray-level dependence matrix-based features), and shape. A bin width of 25 Hounsfield units was used for image intensity discretization. No further image pre-processing was performed. The mathematical descriptions of all features are reported in [17].

2.7 MODELING

For model development, multivariable logistic regression with Elastic Net regularization was performed in the training data set. Highly correlated features, features with near zero variance, and linear combinations between features were first eliminated from further analysis. For each highly correlated feature pair (Pearson correlation coefficient $\rho > 0.9$), the variable with the largest mean absolute correlation with all remaining features was removed. Model training was performed using 100 times repeated 10-fold cross-validation to select the optimal model hyperparameters, optimizing for AUC. All features were standardized before modeling. To further reduce the chance of overfitting on the training data, we selected the simplest candidate model (i.e., the model with the fewest non-zero coefficients) within one standard error of the best-performing model. Model performance was validated in the test data set. Here, the AUC was used to assess model performance in discriminating between COVID-19 positive and COVID-19 negative patients. Additionally, a hard classification was performed (i.e., classifying patients as either COVID-19 positive or negative) by applying different decision thresholds on the continuous scores (probabilities) predicted by the model on the test data set. Classification performance was then assessed by determining accuracy, sensitivity, specificity, NPV, and PPV for each decision threshold, assuming a disease prevalence of 15%. All statistical analysis was performed in R (version 3.6.2).

3 RESULTS

3.1 STUDY POPULATION

Table 2.1 lists the study population characteristics for the model development data (the COVID and Control cohorts), and the independent test dataset (the Test cohort), as well as the main CT findings as scored by radiologists. For the model development data, the COVID-19 positive and control patients have a similar mean age and male/female distribution. Of the COVID-19-infected patients, 69% needed O_2 at admission, resulting in 37% of patients ending up in the ICU. 17% of COVID-19 patients needed mechanical ventilation and 4% died. The comorbidity summary for the COVID-19 patients is presented in Table 2.2 For the independent test data set, the COVID-19 positive and control patients have a similar mean age and male/female distribution and 41% of the COVID-19 patients were admitted to the ICU.

Table 2.1 Summary of patient characteristics (age, gender, and CT findings scored by radiologist) per cohort.

	Training set (n=1381)		**Independent validation set (n=697)**	
	CONTROL (n=1200)	*COVID (n=181)*	*CONTROL (n=500)*	*COVID (n=197)*
Age (years)	63.8±14.4	64.4±15.8	64.2±14.0	69.1±13.3
Gender (% Male)	52	56	51	56
Normal (%)	33	4.41	25.2	25
Neoplasia (%)	8.73	0	0	0
CAP (%)	12.50	8.10	6.6	8.6
COPD (%)	26	19.33	33.4	11.7
Isolated pleurisy (%)	6.2	1.10	4.2	4
Pulmonary embolism (%)	0.77	1.10	0	0
Nodule (%)	19	6.62	17.2	6.6
Chronic inflammation (%)	8.48	5.52	13.6	3
Pneumothorax (%)	0.68	0	0.6	0
Isolated atelectasis (%)	3.68	3.31	5.4	1.0

Table 2.2 Baseline characteristics of the COVID-19 patients used for model training

Comorbidity	COVID training set (n=181)
Neoplasia (%)	23.7
Acute Respiratory Failure (%)	26.7
Cardiac disorder (%)	15.9
Hypertension (%)	6.8
Diabetes (%)	4.7
Chronic renal failure (%)	1.8
Obesity (%)	0

3.2 DATA CURATION

After an automated quality check on the inclusion criteria, CT images and lung segmentations for a total number of 1224 patients for model development and 641 patients for independent model testing were included for further processing. A flow chart describing the overall workflow from data collection to model training and testing is shown in Figure 2.3.

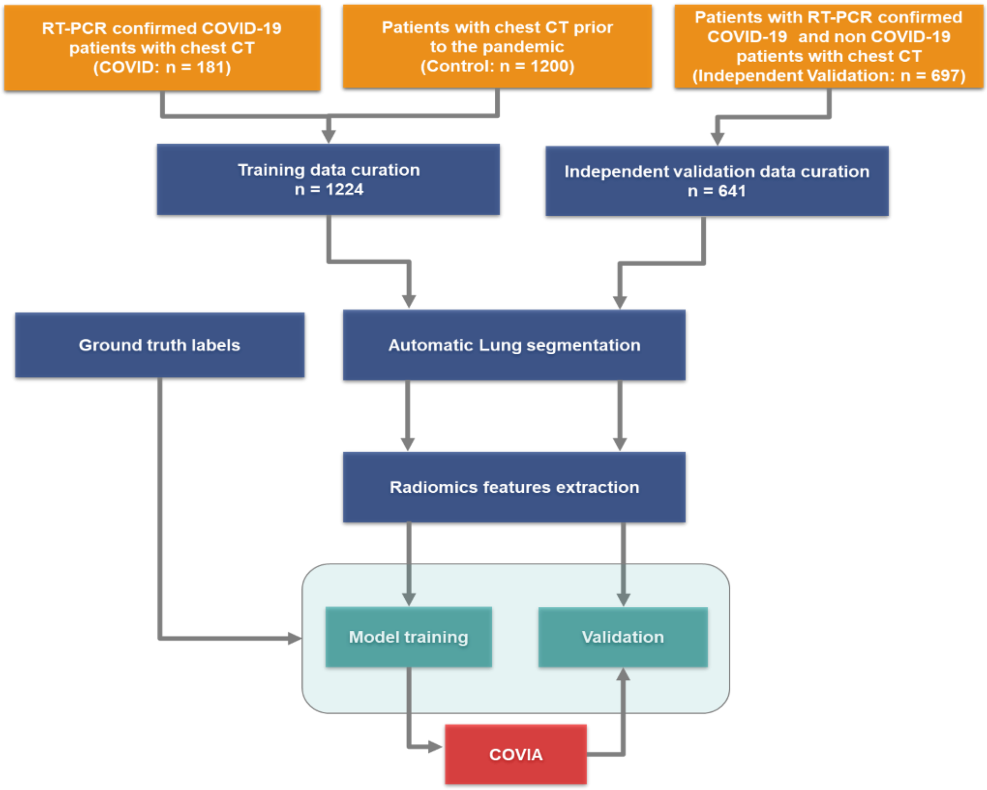

RT-PCR confirmed COVID-19 patients with chest CT (COVID: n = 181)
Patients with chest CT prior to the pandemic (Control: n = 1200)
Patients with RT-PCR confirmed COVID-19 and non COVID-19 patients with chest CT (Independent Validation: n = 697)
Training data curation n = 1224
Independent validation data curation n = 641
Ground truth labels
Automatic Lung segmentation
Radiomics features extraction
Model training
Validation
COVIA

Figure 2.3 Flow diagram: Training and validation data was collected, and the COVID and Control cohorts were combined. Lungs were segmented from both the training and validations datasets respectively, and radiomics features were extracted. The independent validation data was used to test the performance of COVIA with unseen patient CTs.

3.3 COVID-19 INFECTION STATUS PREDICTION

The final prediction model included 45 radiomics features with a non-zero regression coefficient. Included features and their importance, in terms of the absolute regression coefficient, are shown in Figure 2.4A while the ROC curve for the independent test data set is shown in the Figure. 2.4B. The corresponding AUC value for discriminating between COVID-19 positive and negative cases is 0.882 (95% CI: 0.851-0.913). Assuming a disease prevalence of 15%, the classification performance in the test dataset, in terms of accuracy, sensitivity, specificity, NPV, and PPV for different decision thresholds is shown in Figure 2.5. For example, a threshold of 0.11 corresponds to the optimal decision threshold in terms of the Youden Index, when considering the cost of false negatives twice as high as the cost of false positives. This particular decision threshold results in an accuracy of 85.18%, a sensitivity of 69.52, a specificity of 91.63%, an NPV of 94.46%, and a PPV of 59.44% for COVID-19 classification. Figure 2.6 depicts a chest CT of a typical COVID-19 positive patient (Figure. 2.6 A), and a normal chest CT (Figure. 2.6 B) alongside their corresponding heat-maps extracted from an end-to-end conventional black-box AI-based model trained to screen COVID. The heatmaps were obtained from a conventional CNN model based on VGG16 architecture trained to classify COVID from other CT images. A technique called Gradient-based localization [26] was used to obtain the heatmaps which explain the model's decision to classify the image in Figure 2.6 A as a COVID case.

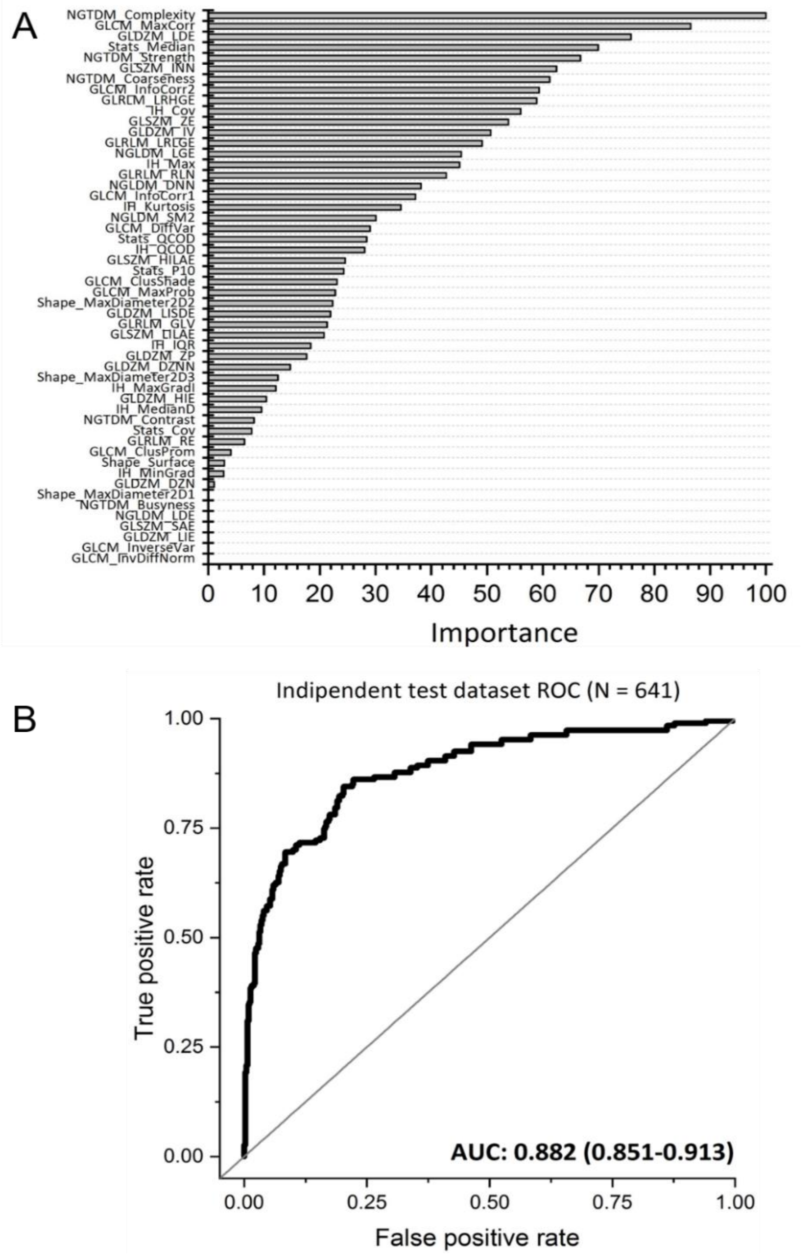


Figure 2.4 A) Features with a non-zero regression coefficient in the model and their importance, based on their absolute regression coefficient, and scaled between 0 and 100; B) ROC plot illustrating the performance (black curve) of the AI framework to discriminate between COVID-19 positive and negative cases in the independent test data set with an AUC of 0.882 (95% CI: 0.851-0.913).

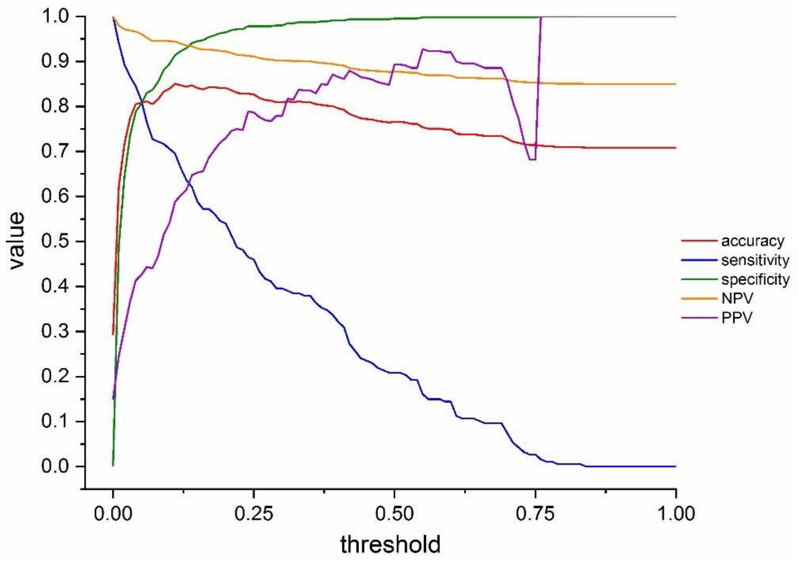


Figure 2.5 Classification performance plot. The classification performance in the test dataset, assuming a disease prevalence of 15%, in terms of accuracy (red line), sensitivity (blue line), specificity (green line), NPV (orange line), and PPV (purple line) for different decision thresholds.

Table 2.3 lists the values of the top 5 radiomics features and model scores (SCORE) of cases depicted in Figures 2.6 A and B. The top 5 features are a measure of texture complexity, quantifying non-uniformity and sudden changes in intensity values within the region of interest (NGTDM_Complexity; Neighborhood gray-tone difference matrix, Complexity); a texture measure of the correlation of the grey-level co-occurrence matrix (GLCM_MaxCorr; grey level co-occurrence matrix, maximal correlation coefficient); a texture measure emphasizing larger distances to the edge of the region of interest of connected voxels of similar intensity value (GLDZM_LDE; grey level distance zone matrix, Large distance emphasis); the median image intensity in the lungs (Stats_Median; First order statistics, Median); a measure of texture strength, quantifying how definable or visible the texture is within the image (NGTDM_Strength; Neighborhood gray-tone difference matrix, Strength). Figure 2.6 C-G report the box plots for the distribution of features among the COVID and non-COVID group.

Table 2.3 Top 5 radiomics features and model scores of cases depicted in Figure 2.6.

	Normal chest CT	**COVID-19 positive**
NGTDM_Complexity	7794.055	1147.344
GLCM_MaxCorr	0.8684842	0.9147317
GLDZM_LDE	143.07153	57.53219
Stats_Median	-839	-755
NGTDM_Strength	0.033166649	0.008062981
SCORE	0.01119137	0.765581

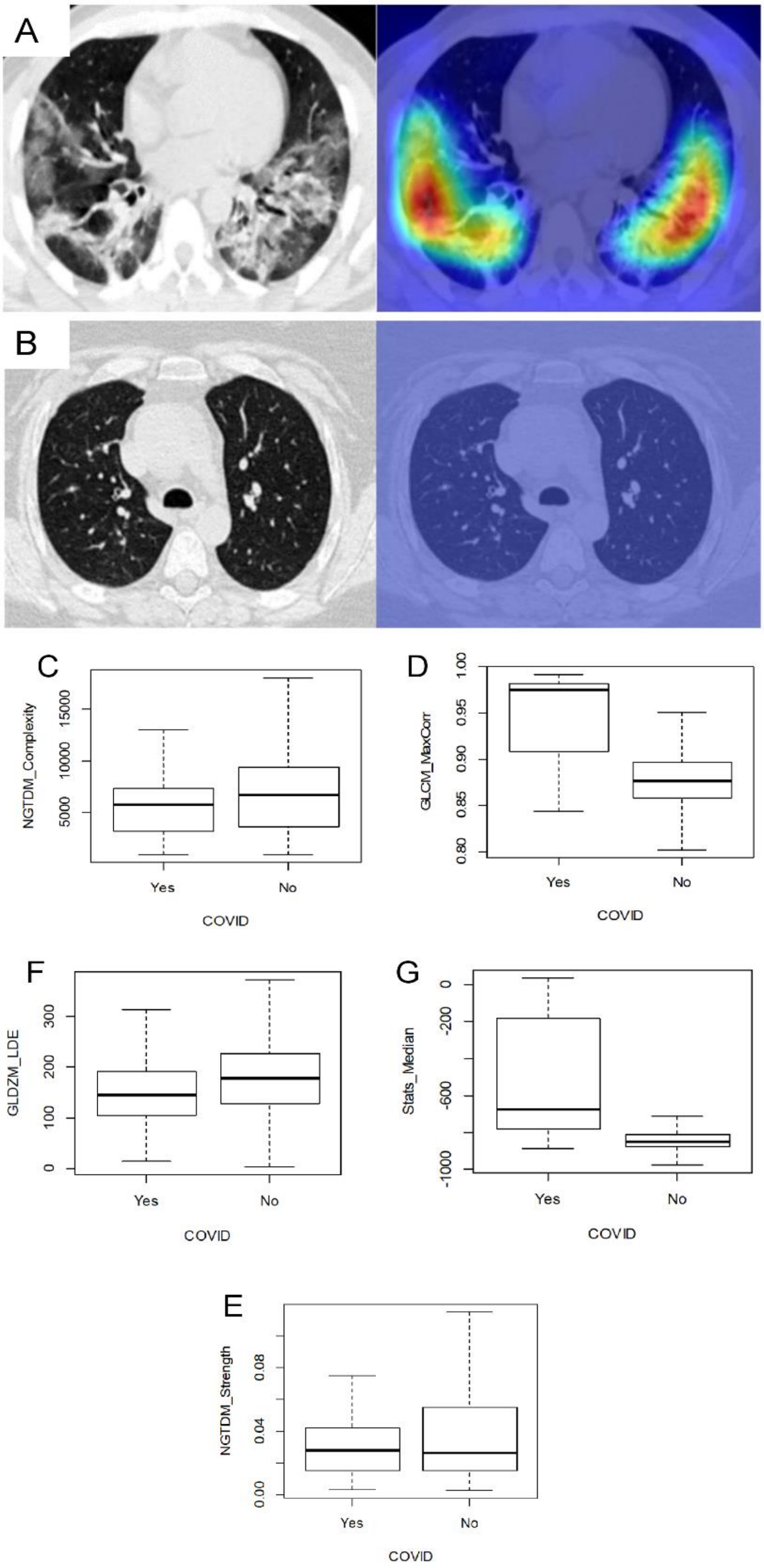


Figure 2.6 Chest CTs of a typical COVID-19 positive patient (A: original scan – left; heat-map - right) with evident reticulation, ground glass opacities, and condensations compared to a healthy patient CT scan (B: original scan – left; heat-map - right). Heat maps underline the more relevant areas for model prediction. Box plots comparing the distribution of the top 5 features among COVID and non-COVID cases (C – NGTDM_Complexity; D – GLCM_MaxCorr; E – NGTDM_Strenght; F – GLDZM_LDE; G – Stats_Median).

4 DISCUSSION

RT-PCR was considered the gold standard for COVID-19 identification However, there were reports of false negatives occurring which were eventually confirmed as true-positive by repeated swab tests [27]. False negatives could be a significant problem in high-throughput settings operating under severe pressure [28]. The correct operation of the test was crucial and there was ambiguity concerning the kinetics of SARS-CoV-2 viral shedding, thus the timing of the test could very well dictate the result. Furthermore, it was also unclear what kind of clinical sample was most appropriate as nasopharyngeal swabs might offer greater consistency than sputum samples [29]. When considering the limited supply of RT-PCR kits, the growing backlog, and the likely increasing pressure and turnaround times in laboratories along with the issues of false negatives, prompted the experts to suggest that diagnosing suspected cases using the widely available, time-saving, and non-invasive imaging approach of chest CT was justified. This approach proved useful sensitively and specifically in identifying COVID-19 patients [30]. We have shown that our model was able to achieve a high NPV (94.46 %), which provided further justification for using CT imaging-based diagnosis as a primary tool for COVID-19 patient management.

Whereas similar studies in COVID-19 focus mainly on the detection of various diseased regions (including ground-glass opacification, consolidation, bilateral involvement, and peripheral and diffuse distribution amongst others) in the lung [31,32] [33], our approach performs an easy segmentation of the lungs as one single structure, which is by far an easier task to automate with AI. Features for quantitative image analysis are extracted from this whole lung structure and subsequently used for prediction model application and COVID-19 infection status classification. In the end, this constituted a fully automated clinical decision support tool for the diagnosis of COVID-19, which was able to provide an objective, robust (i.e., no user variability), and easy-to-interpret classification (yes-no) of COVID-19 infection status. The complete workflow took between 40-60s, providing a rapid and accurate diagnosis in patients with suspicion of COVID-19 infection, and facilitating the timely implementation of isolation procedures and early intervention.

We developed a machine learning model that was able to discriminate between COVID-19 positive and negative patients, and which was trained and validated using a regularized logistic regression model. Elastic net logistic regression has been used, for its relatively straightforward interpretation of linear models and its demonstrated discriminative performance [32]. The continuous prediction scores of the model can be utilized for binary classification of patients (COVID-19 infected or not). Given this continuous output of the underlying model, it is possible to optimize the decision threshold used for hard classification based on more appropriate prevalence and costs of misclassification, which may vary, for instance, per geographic area. Although this study focuses solely on using image data for COVID-19 diagnosis, it is possible to imagine that, combining the model's continuous score with other clinical data, an even more accurate determination of the overall probability of diagnosis could be achieved.

We plan to test the capability of the AI algorithm in the diagnosing of COVID-19 against that of radiologists in a virtual clinical trial setting. This aspect is vital in the context of incidental findings, which is of increasing relevance [33]. An automated AI solution could help assist the accurate identification of potentially COVID-19-positive patients, alerting the radiologist who must prioritize the reading of this examination and the radiology department that a 'clean machine' now requires decontamination.

A general objection to AI methods is the lack of transparency and interpretability. This is not the case with our approach, as 'handcrafted' radiomics features are explicitly defined and linked to specified regions of interest within the images, driving the decision of the algorithm. Thus, clearly and intelligibly quantifying the imaging phenotype, has also been shown to provide a means of connecting to the underlying biology [34]. The interpretability of an AI-based classifier's decision is limited to highlighting image regions contributing to the decision, which allows only for qualitative interpretation (i.e., human/expert interpretation of these image regions). Our model proves to be more interpretable and explainable as the (top) features are associated with clearly pre-defined regions of interest and their values can be directly compared between different patients, as well as further interpreted based on their unambiguous mathematical definitions. For instance, the features listed in Table 2.3 clearly show the difference in values between normal and COVID patients' CTs. Hence, those features quantify a radiomics phenotype linked to the bilateral multilobe ground-glass opacities of peripheral/subpleural distribution, with intralesional reticulations seen on this typical COVID-19-positive chest CT.

Given the rapid development of serum-based tests for COVID-19, a critical contextualization is important. Serum analysis is dependent on logistics and takes a relatively long time to deliver results when compared with AI (near instantaneous). In the best case scenario serum takes hours and in the worst case several days [28]. Furthermore, serum analysis is practically limited to large centers with advanced biotechnology capabilities in developed countries (small centers have increased logistical challenges). In the case of an emergency procedure (e.g., surgery), the probable COVID-19 status of the patient must be immediately addressed to safeguard the hospital concerning transmission. Considering beyond the current pandemic phase that we are in, serum analysis offers little value in the way of incidental findings as clinicians will be less proactive in ordering a test to determine COVID-19 infection. Concerning RT-PCR detection [35], the positive rate of the 2019-nCoV nucleic acid test on the nasopharyngeal swab is 38% (180/472 times), and the positive rate of the 2019-nCoV nucleic acid test on sputum is 49% (148/304 times), the positive rate of blood 2019-nCoV nucleic acid test is 3% (4/132 times), and the positive rate of 2019-nCoV nucleic acid test of feces is 10% (24/244 times). The positive rate of 2019-nCoV nucleic acid detection in anal swabs is 10% (12/120 times). A meta-analysis [36] showed, the pooled sensitivity was 94% (95% CI: 91%, 96%) for chest CT and 89% (95% CI: 81%, 94%) for RT-PCR. The pooled specificity was 37% (95% CI: 26%, 50%) for chest CT. The prevalence of COVID-19 outside China ranged from 1% to 23%. The PPV ranged from 2% to 31%, and the NPV ranged from 95% to 100%. COVIA was tested against an assumed prevalence of 15% and the classification results indicate competitive performance.

In the last months, the literature about AI-assisted diagnosis and classification of COVID-19 infection has boomed like never seen before [37,38]. Many relevant papers have been published, reporting multicentric validation studies with remarkable performances [39–41], along with a new insight into the clinical aspect of CT scan COVID-19 characteristics [42]. In this fast-evolving field, where much innovation goes along sometimes with overly enthusiastic reports [43], our method has several advantages over other reported AI-based diagnostic tools: first of all the automatic segmentation of the whole lung does not require human input, speeding up the process and unburdening medical staff. More important, however, is the use of robust and validated radiomics features, compared to other parameters used in other approaches like consolidation and ground-glass opacity alone [44–46] which are not specific for the disease [9].

Compared to other radiomics signatures published in the last months [47,48], our signature was trained and tested on a wider dataset, acquired at different time points, to account for the small variability that might be present in scan acquisition at different dates. This is considered a more reliable strategy [49] as it closely mimics what happens in a real-world clinical scenario. The robust testing strategy of the model, coupled with the interpretability of the radiomics features, assures the reliability of the proposed model.

It is worth pointing out, however, that the study has still some limitations. Firstly, COVID-19 is caused by SARS-CoV-2 and may have similar imaging characteristics as pneumonia caused by other types of viruses. However, due to the lack of laboratory confirmation of the etiology for each of these cases, we were not able to select other viral pneumonias for comparison in this study. Although our Control group of non-COVID-19 patients contains several patients (see CAP in Table 2.1, 12.5%) with pneumonia (either viral, bacterial, or pneumonia from any other cause), it would be desirable to test the performance of our algorithm in distinguishing COVID-19 from other viral pneumonias that have RT-PCR confirmation methods for the viral agent.

Also, the population of patients with COVID-19 was selected after clinical evaluation of patients with respiratory symptoms such as dyspnea and desaturation. The degree of severity justified the fact that imaging analysis was left to clinical judgment and depending on local resources [50]. Therefore, COVIA was partially developed in a population of patients with disease at the moderate to severe end of the spectrum. Further analysis into the benefit, if any, of COVIA for patients with mild or no symptoms is required.

Future work is planned to collect additional chest CTs to externally validate the performance of our algorithm in an international multi-center prospective external validation to produce evidence level 1 [36] for the clinical utility of COVIA. The study protocol is in development and will be registered on clinicaltrials.gov.

Ultimately, this study was focused on diagnosis whereas prognosis on the future disease trajectory is an even more urgent unmet clinical need that would enable improved resource management (including management decisions regarding the allocation of resources). This is the next step for our collaborative research.

5 CONCLUSIONS

Benchmarked against RT-PCR confirmed cases of COVID-19, our AI framework constituting a combination of the Lung segmentation model and Radiomic feature analysis model can accurately detect COVID-19. Thus, it provides rapid accurate diagnosis in patients suspected of COVID-19 infection, facilitating the timely implementation of isolation procedures and early intervention. The proposed model, trained on a diverse and robust dataset, showed good performance (AUC of 0.882) with the added value of being explainable, linking the radiomics results with real clinical evidence, like lung abnormalities (ground glass opacities, consolidations, and others). This approach will be extended and improved, including the distinction between different types of pneumonia, streamlining the staging and therapy planning of patients. A further improvement could comprise the creation of a prognostic model along with the diagnostic one, to assess the severity of newly admitted patients and the probability of developing serious symptoms or admission to the ICU.

6 REFERENCES

[1] WHO. Landing page 2020.

[2] Johns Hopkinds University & Medicine. Coronavirus Resource Center 2020.

[3] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020;395:689–97. https://doi.org/10.1016/S0140-6736(20)30260-9.

[4] Yang S, Shi Y, Lu H, Xu J, Li F, Qian Z, et al. Clinical and CT features of early-stage patients with COVID-19: a retrospective analysis of imported cases in Shanghai, China. Eur Respir J 2020:2000407. https://doi.org/10.1183/13993003.00407-2020.

[5] Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology 2020;296:E32–40. https://doi.org/10.1148/radiol.2020200642.

[6] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology 2020:200432. https://doi.org/10.1148/radiol.2020200432.

[7] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

[8] Ruili Li, Guangxue Liu, Xiaochun Zhang HL. Letter to the Editor: Chest CT and RT-PCR: radiologists' experience in the diagnosis of COVID-19 in China. Eur Radiol 2020.

[9] Hope M. A role for CT in COVID-19? What data really tell us so far. Lancet 2020;395:1189–90. https://doi.org/10.1016/S0140-6736(20)30728-5.

[10] Huang Y, Cheng W, Zhao N, Qu H, Tian J. Correspondence CT screening for early diagnosis of SARS-CoV-2. Lancet Infect Dis 2020;51:30241. https://doi.org/10.1016/S1473-3099(20)30241-3.

[11] 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 2014;5:4006. https://doi.org/10.1038/ncomms5006.

[12] Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6:837–45. https://doi.org/10.1016/S2213-2600(18)30286-8.

[13] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89–94. https://doi.org/10.1038/s41586-019-1799-6.

[14] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25:954–61. https://doi.org/10.1038/s41591-019-0447-x.

[15] Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys 2018;45:3449–59. https://doi.org/10.1002/mp.12967.

[16] 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;48:441–6. https://doi.org/10.1016/j.ejca.2011.11.036.

[17] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–62. https://doi.org/10.1038/nrclinonc.2017.141.

[18] Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. PLoS One 2014;9:e102107.

[19] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep 2015;5:13087. https://doi.org/10.1038/srep13087.

[20] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017- Janua, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 936–44. https://doi.org/10.1109/CVPR.2017.106.

[21] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. 2017 IEEE Conf. Comput. Vis. Pattern Recognit., 2017, p. 5987–95. https://doi.org/10.1109/cvpr.2017.634.

[22] Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard C van den, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 2016;42:1–13. https://doi.org/10.1016/j.media.2017.06.015.

[23] Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. Sci Data 2017;4:170124. https://doi.org/10.1038/sdata.2017.124.

[24] Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015, p. 1–15.

[25] Bertels J, Eelbode T, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory & Practice. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2019;11765 LNCS:92–100. https://doi.org/10.1007/978-3-030-32245-8_11.

[26] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[27] Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing. Radiology 2020:200343. https://doi.org/10.1148/radiol.2020200343.

[28] Sheridan C. Coronavirus and the race to distribute reliable diagnostics. Nat Biotechnol 2020. https://doi.org/10.1038/d41587-020-00002-2.

[29] Revel MP, Parkar AP, Prosch H, Silva M, Sverzellati N, Gleeson F, et al. COVID-19 patients and the radiology department – advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI). Eur Radiol 2020. https://doi.org/10.1007/s00330-020-06865-y.

[30] Liu J, Yu H, Zhang S. The indispensable role of chest CT in the detection of coronavirus disease

2019 (COVID-19). Eur J Nucl Med Mol Imaging 2020:1. https://doi.org/10.1007/s00259-020-04795-x.

[31] Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. MedRxiv 2020:2020.02.23.20026930. https://doi.org/10.1101/2020.02.23.20026930.

[32] Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). MedRxiv 2020:2020.02.14.20023028. https://doi.org/10.1101/2020.02.14.20023028.

[33] Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. MedRxiv 2020:2020.02.25.20021568. https://doi.org/10.1101/2020.02.25.20021568.

[34] Leijenaar RTH, Bogowicz M, Jochems A, Hoebers FJP, Wesseling FWR, Huang SH, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: A multicenter study. Br J Radiol 2018;91:20170498. https://doi.org/10.1259/bjr.20170498.

[35] Wu J, Liu J, Li S, Peng Z, Xiao Z, Wang X, et al. Detection and analysis of nucleic acid in various biological samples of COVID-19 patients. Travel Med Infect Dis 2020:101673. https://doi.org/10.1016/j.tmaid.2020.101673.

[36] Oxford Centre for Evidence-based Medicine - Levels of Evidence (March 2009) - CEBM n.d.

[37] Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Rev Biomed Eng 2020:1. https://doi.org/10.1109/RBME.2020.2987975.

[38] Kundu S, Elhalawani H, Gichoya JW, Kahn CE. How Might AI and Chest Imaging Help Unravel COVID-19's Mysteries? Radiol Artif Intell 2020;2:e200053. https://doi.org/10.1148/ryai.2020200053.

[39] Mei X, Lee HC, Diao K yue, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26:1224–8. https://doi.org/10.1038/s41591-020-0931-3.

[40] Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 2020;11:1–7. https://doi.org/10.1038/s41467-020-17971-2.

[41] Ozsahin I, Sekeroglu B, Musa MS, Mustapha MT, Uzun Ozsahin D. Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence. Comput Math Methods Med 2020;2020:1–10. https://doi.org/10.1155/2020/9756518.

[42] Jalaber C, Lapotre T, Morcet-Delattre T, Ribet F, Jouneau S, Lederlin M. Chest CT in COVID-19 pneumonia: A review of current knowledge. Diagn Interv Imaging 2020;101:431–7. https://doi.org/10.1016/j.diii.2020.06.001.

[43] Hope MD, Raptis CA, Henry TS. Chest Computed Tomography for Detection of Coronavirus Disease 2019 (COVID-19): Don't Rush the Science. Ann Intern Med 2020;173:147–8. https://doi.org/10.7326/M20-1382.

[44] Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at

Chest CT. Radiology 2020;296:E156–65. https://doi.org/10.1148/radiol.2020201491.

[45] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 2020;296:E65–71. https://doi.org/10.1148/radiol.2020200905.

[46] Xie C, Ng M-Y, Ding J, Leung ST, Lo CSY, Wong HYF, et al. Discrimination of pulmonary ground-glass opacity changes in COVID&#x2010;19 and non-COVID-19 patients using CT radiomics analysis. Eur J Radiol Open 2020;7. https://doi.org/10.1016/j.ejro.2020.100271.

[47] Fang M, He B, Li L, Dong D, Yang X, Li C, et al. CT radiomics can help screen the Coronavirus disease 2019 (COVID-19): a preliminary study. Sci China Inf Sci 2020;63:172103. https://doi.org/10.1007/s11432-020-2849-3.

[48] Fu L, Li Y, Cheng A, Pang P, Shu Z. A Novel Machine Learning-derived Radiomic Signature of the Whole Lung Differentiates Stable From Progressive COVID-19 Infection: A Retrospective Cohort Study. J Thorac Imaging 2020.

[49] van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomogr (Ann Arbor, Mich) 2016;2:361–5. https://doi.org/10.18383/j.tom.2016.00208.

[50] Revel M-P, Parkar AP, Prosch H, Silva M, Sverzellati N, Gleeson F, et al. COVID-19 patients and the radiology department – advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI). Eur Radiol 2020:1–7. https://doi.org/10.1007/s00330-020-06865-y.

Chapter 3

Deep learning algorithm to classify COVID-19 and other pneumonias on chest CT based on automatic segmentation of lung abnormalities

Based on **Akshayaa Vaidyanathan**, Julien Guiot, Fadila Zerka, Flore Belmans, Ingrid Van Peufflik, Louis Deprez, Denis Danthine, Gregory Canivet, Philippe Lambin, Sean Walsh, Mariaelena Occchipinti, Paul Meunier, Wim Vos, Pierre Lovinfosse & Ralph T.H. Leijenaar, An externally validated fully automated deep learning algorithm to classify COVID-19 and other pneumonias on chest CT, ERJ open research vol. 8,2 00579-2021. 3 May. 2022, doi:10.1183/23120541.00579-2021

In this chapter, we propose an Artificial Intelligence framework based on a 3D Convolutional Neural Network (CNN) to classify CT scans of patients with COVID-19, Influenza/CAP, and no-infection, after automatic segmentation of the lungs and lung abnormalities. The AI classification model is based on inflated 3D Inception architecture and was trained and validated on retrospective data of CT images of 667 adult patients (No infection: 188, COVID-19: 230, Influenza/CAP: 249) and 210 adult patients (No infection: 70, COVID-19: 70, Influenza/CAP: 70), respectively. The model's performance was independently evaluated on an internal test set of 273 adult patients (No infection: 55, COVID-19: 94, Influenza/CAP: 124) and an external validation set coming from a different center (305 adult patients, COVID-19: 169, No infection: 76, Influenza/CAP: 60). The model showed excellent performance in the external validation set with an AUC of 0.90, 0.92 and 0.92 for COVID-19, Influenza/CAP and No infection respectively. The selection of the input slices based on automatic segmentation of the abnormalities in the lung reduces the analysis time (56 seconds per scan) and computational burden of the model. The TRIPOD score of the proposed model is 47% (15 out of 32 TRIPOD items). This AI solution provides rapid and accurate diagnosis in patients suspected of COVID-19 infection and influenza, facilitating the timely implementation of isolation procedures and early intervention.

1 BACKGROUND

Imaging with computed tomography (CT) plays a central role in the diagnosis of respiratory diseases [1,2]. After the outbreak of COVID-19 in 2020, more emphasis has been given to the different types of pneumonias and the distinctive features of COVID-19 from all others [3,4]. Viral pneumonias, either COVID-19 or others, can all present with reticulation, Ground Glass Opacities (GGO), and consolidations at chest CT scan, creating a challenge for radiologists in their routine differential diagnosis. Previous studies on the performance of radiologists in discriminating between COVID-19 and other pneumonias on chest CT scans have shown high variability in both sensitivity (73%-94%) and specificity (24%-100%), with on average high sensitivity and moderate specificity [5]. This variability of interpretation of CT findings of pneumonia still creates a routine challenge for clinicians in their differential diagnosis, which is key to properly treating the patients and preventing infection spread during pandemics and in the next future.

In this context, the development of innovative artificial intelligence (AI) imaging solutions to support radiologists in swift and precise differential diagnosis would be of invaluable help. Convolutional Neural Networks (CNN) have shown great potential in detection, segmentation, and classification tasks in radiological images [6]. A recent study demonstrated the application of CNN for the differentiation among Influenza, COVID-19, and no-infection, on chest CT scans with an overall accuracy of 86.7%. The proposed method incorporated training on image patches extracted from CT volumes where each image patch required manual labeling as “pneumonia” or “irrelevant information” [7]. Another study compared the performance of different AI models in classifying COVID-19 from other atypical and viral pneumonias, showing 99.5% accuracy in classifying COVID-19 [8]. However, these approaches involve all manual detection (i.e. drawing boxes around the lesions), labeling of the lesions in all the slices, and training the models on the patches of detected lesions and manual labels. The time required to perform these manual operations is usually not considered when addressing the real-world application of these models and represents probably one of the major hurdles to widespread clinical adoption.

A fully automatic tool running on chest CT images for the differential diagnosis of pneumonias can represent an important step forward for decreasing the variability of interpretation among clinicians and for a faster diagnostic process. This will unburden medical staff and in turn provide a better and faster diagnosis for patients, reducing the use of hospital resources. Better allocation of both material and human resources can be essential in a time of crisis as the COVID-19 pandemic demonstrated with dramatic clarity [9]. To attain this goal, we developed and externally validated a fully automated deep learning framework with a 3D CNN, able to classify chest CT scans of patients with COVID-19, Influenza/CAP, or no infection without manual intervention. Individual AI-based whole lung and lung abnormalities segmentation models were used to pre-process the CT images to train the 3D CNN model and are an integral part of the workflow to assure that only the patients presenting abnormalities in the lung volume are processed by the model, saving time and computational power.

2 MATERIAL AND METHODS

The study was approved by the local ethics committee of the CHU-Liège (EC number 116/2020). The institutional review board waived the requirement to obtain written informed consent for this retrospective case series since all analyses were performed on de-identified (i.e., anonymized) data and there was no potential risk to patients.

2.1 SUBJECTS

Three cohorts of patients were included retrospectively in this study for model training, validation, and testing. Cohorts came from two University Hospitals (CHU Sart-Tilman and CHU Notre Dame des Bruyères) in Liège, Belgium. The first cohort (label: COVID-19) consisted of all patients with COVID-19 infection confirmed by RT-PCR that underwent chest CT imaging before March 28th, 2020. The second cohort (label: Influenza/CAP) consisted of patients with influenza, parainfluenza, or community-acquired pneumonia (CAP) infection confirmed by RT-PCR or positive antigen testing that underwent chest CT imaging between March 2014 and March 2020. The third cohort (label: “No infection”) consisted of consecutive patients that underwent chest CT imaging during October 2019, with confirmed no infection in the lungs disregarding any other lung disease. The three cohorts were pooled together and randomly split between training, validation, and testing set (see Figure. 3.1). Additionally, the open source dataset COVID-CT-MD was used as an external validation set [10]. The final population consisted of 169 RT-PCR confirmed positive COVID-19 cases (from February 2020 to April 2020), 60 Community-Acquired Pneumonia (CAP) cases (from April 2018 to November 2019), and 76 No Infection cases (from January 2019 to May 2020): all the patients were treated at Babak Imaging Center in Tehran, Iran, and labeled by three experienced radiologists.

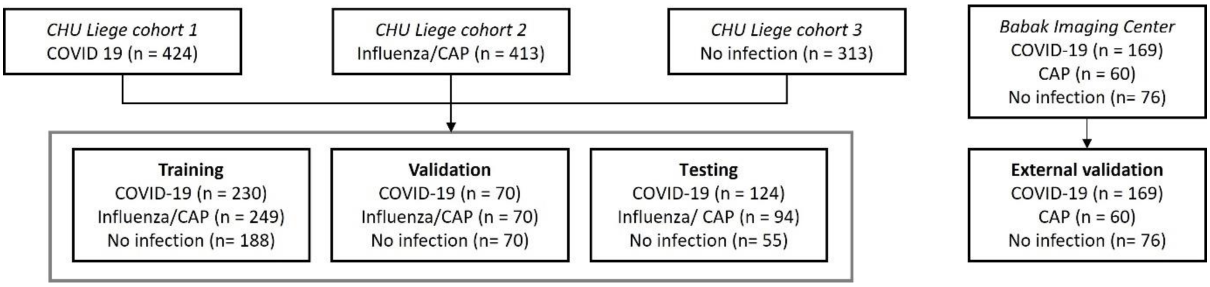


Figure 3.1 Flow chart of patient cohorts division

2.2 IMAGING SCANS

In this retrospective study, CT scans of the three cohorts of patients included were acquired from different scanners (Siemens and GE) with diverse reconstruction kernels (soft and sharp). In case of the presence of more than one series per case, all the available series were used in training the model (as the reconstruction kernel corresponding to the series was considered as a form of image augmentation). The slice thickness of the scans ranged between 0.5 mm and 2 mm while pixel spacing was between 1 and 2.5 mm. A complete summary of the imaging parameters of both the training and external validation set is reported in Table 3.1.

Table 3.1 Summary of imaging parameters for the training and external validation datasets

	Training set	**External validation set**
Manufacturer		
SIEMENES	60 %	100 %
GE MEDICAL SYSTEM	40 %	-
PHILIPS	< 1 %	-
Kernel		
B30f	-	7 %
B41s	-	45 %
D40s	-	48 %
B30s	2 %	-
Br32f	17 %	-
Br59f	12 %	-
Tr20f	13 %	-
LUNG	11 %	-
STANDARD	31 %	-
Others	14 %	-
Pixel spacing (mm)		
< 0.5	1 %	-
0.5 to 0.6	7 %	11 %
0.6 to 0.7	21 %	48 %
0.7 to 0.8	16 %	38 %
> 0.8	55 %	3 %
Slice thickness (mm)		
2	100 %	100 %

2.3 LUNG ABNORMALITIES SEGMENTATION

The segmentation model is based on 2D U-Net combined with Res Next as encoder and deep supervision and was trained on axial unenhanced chest CT scans of 199 COVID-19 patients coming from three different centers in three different countries [11]. The model's performance was evaluated on an external test set of 50 COVID-19 patients coming from several different centers in Moscow, Russia [12]. All datasets are open source, and freely available online. An automatic in-house lung segmentation model (see above 1 Lung segmentation) was used to crop the lung region from the CT volumes. Axial slices with no segmented lung regions were removed from the volumes. Different sets of 48 consecutive axial slices with an overlap of 10 slices between one set and the other (extracted from the whole volume) were used to train the model. Each set contains at least one slice with lung abnormalities. Each data point containing the consecutive axial slices was pre-processed in the following ways to obtain a three-channel input to the model:

- The first channel contains slices with intensities clipped at lung window level settings (W:1500 HU, L:-600 HU) with lungs and the abnormalities cropped.
- The second channel contains the slices with original intensities with lungs and abnormalities cropped.
- The third channel contains slices with intensities clipped at Mediastinal window level settings (W:350 HU, L:50 HU) with the region containing the lungs cropped. A rectangular crop was obtained with *x_min* = minimum x value for which lungs or lung abnormalities pixels are

present, *x_max* = maximum x value for which lungs or lung abnormalities pixels are present, and *y_min* = minimum x value for which lungs or lung abnormalities pixels are present, *y_max* = maximum y value for which lungs or lung abnormalities pixels are present.

The automatic deep learning segmentation algorithm achieved good performances (mean DSC 0.6 ± 0.1) on the external test set.

2.4 IMAGE PRE-PROCESSING

The prevalence of COVID-19 cases in the three datasets was adjusted to avoid class imbalance and bias in classification [13]. COVID-19 cases represented between 35 and 45% of the whole cohort for each dataset. A fully automated lung segmentation model (see Chapter 1) was used to filter out the slices not containing lungs from the CT scan series. The presence of abnormalities in each filtered slice was confirmed using the lung abnormalities segmentation model. If no abnormalities were present in the filtered slices, the scan was discarded from model processing. Different sets of 48 consecutive axial slices with an overlap of 10 slices between one set and the other (extracted from the whole volume with axial slices containing lungs) were obtained, while each set including at least one slice containing abnormalities in the lung was used to train the model. The workflow for the pre-processing protocol is depicted in Figure. 3.2. The entirety of datasets provided by clinicians were used in the model training and validations, without any prior scan quality selection

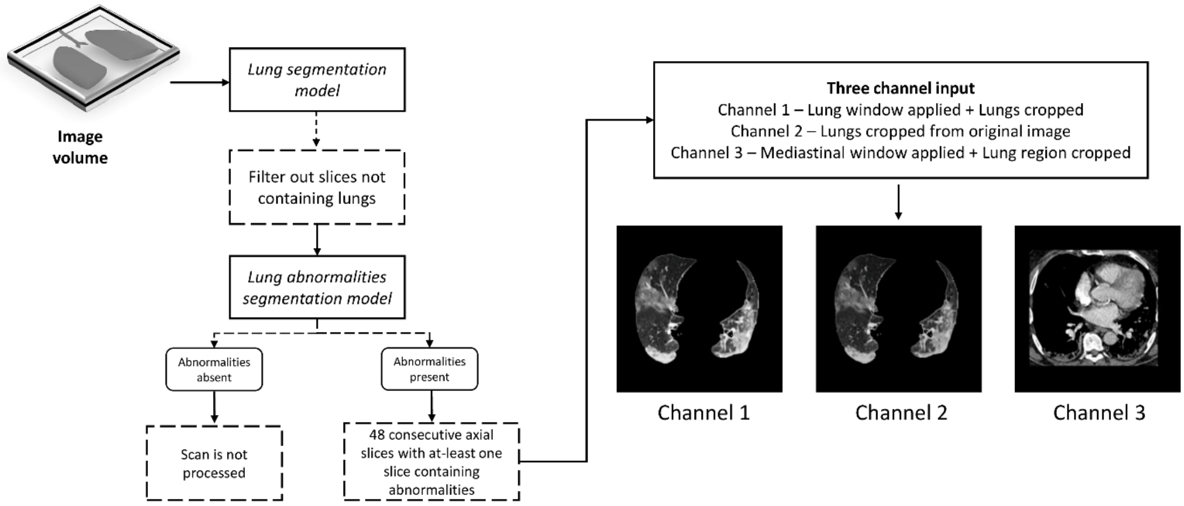


Image volume
Lung segmentation model
Filter out slices not containing lungs
Lung abnormalities segmentation model
Abnormalities absent
Abnormalities present
Scan is not processed
48 consecutive axial slices with at-least one slice containing abnormalities
Three channel input
Channel 1 – Lung window applied + Lungs cropped
Channel 2 – Lungs cropped from original image
Channel 3 – Mediastinal window applied + Lung region cropped
Channel 1
Channel 2
Channel 3

Figure 3.2 Scheme of the pre-processing workflow applied

Each data point containing the 48 consecutive axial slices was processed in three different ways to obtain a three-channel input for the model:

- The first channel (Channel 1) contained slices with intensities clipped at Lung window level settings (W:1500 HU, L:-600 HU) with lungs and the abnormalities cropped.
- The second channel (Channel 2) contained the slices with the original intensities of lungs along with the abnormalities cropped.
- The third channel (Channel 3) contained slices with intensities clipped at Mediastinal window level settings (W:350 HU, L:50 HU) within the region containing the cropped lungs, for which

the bounding rectangular crop within which lungs or lung abnormalities pixels are present was obtained. This operation was performed to better assess pleural effusion [14].

Finally, the slices were center cropped to a slice size of 448 by 448 pixels. An example of the resulting lung and abnormalities segmentation is reported for Influenza/CAP (Figure. 3.3) COVID-19 (Figure. 3.4) and No infection (Figure. 3.5) patients.

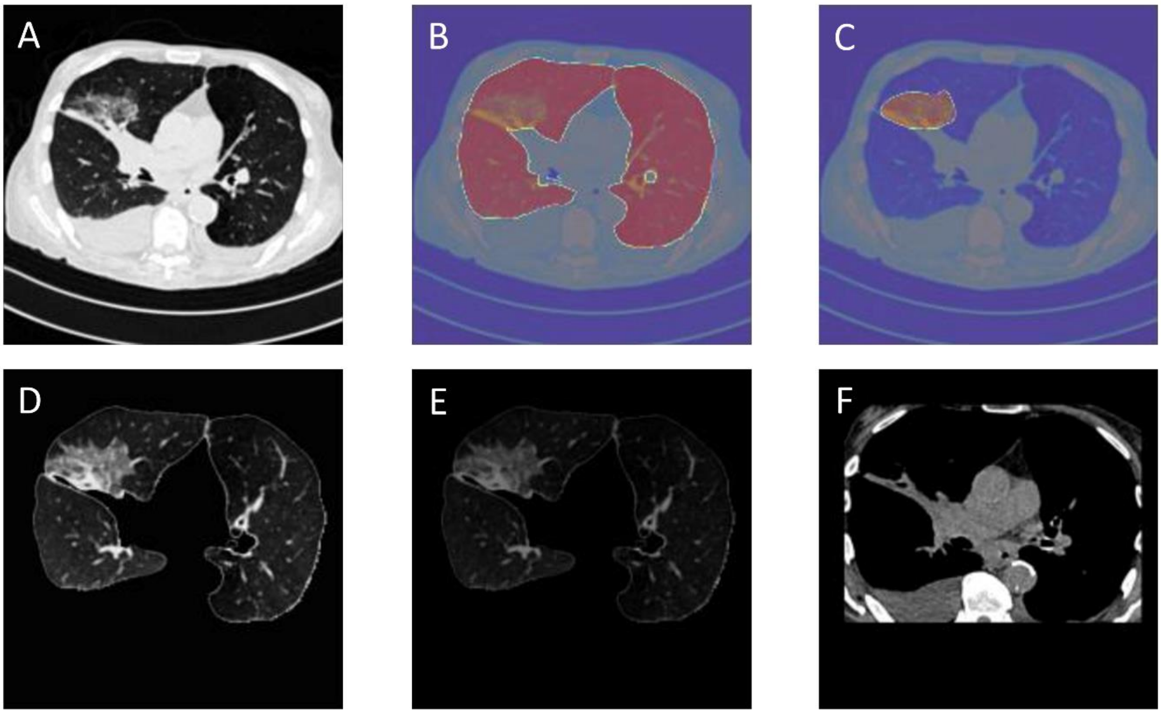


Figure 3.3 Lungs plus abnormalities segmentation on a slice from an Influenza/CAP patient. A) Original axial slice from a case with Influenza/CAP label; B) Lung segmentation obtained on the same slice; C) Ground Glass Opacities segmented by the lung abnormalities model. Three channel input obtained from the same slice, D) Channel 1; E) Channel 2; F) Channel 3.

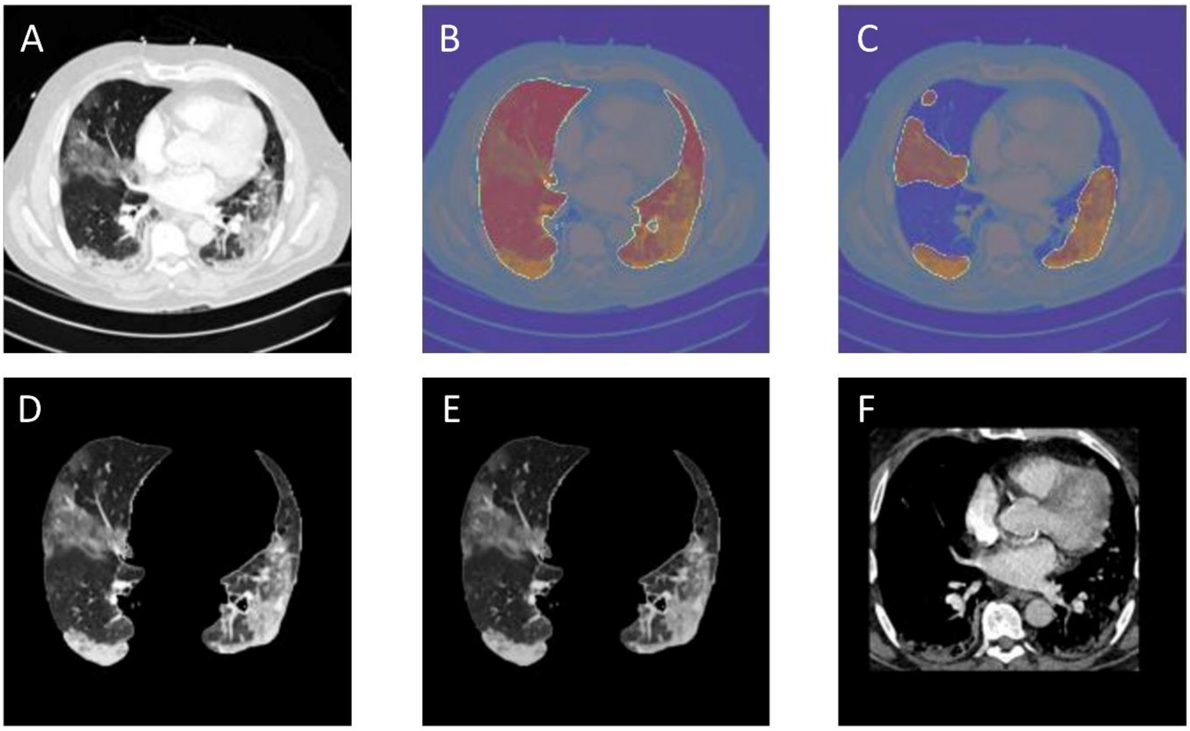


Figure 3.4 Lungs plus abnormalities segmentation on a slice from a COVID-19 patient. A) Original axial slice from a case with COVID-19 label; B) Lung segmentation obtained on the same slice; C) Ground Glass Opacities segmented by the lung abnormalities model. Three channel input obtained from the same slice, D) Channel 1; E) Channel 2; F) Channel 3.

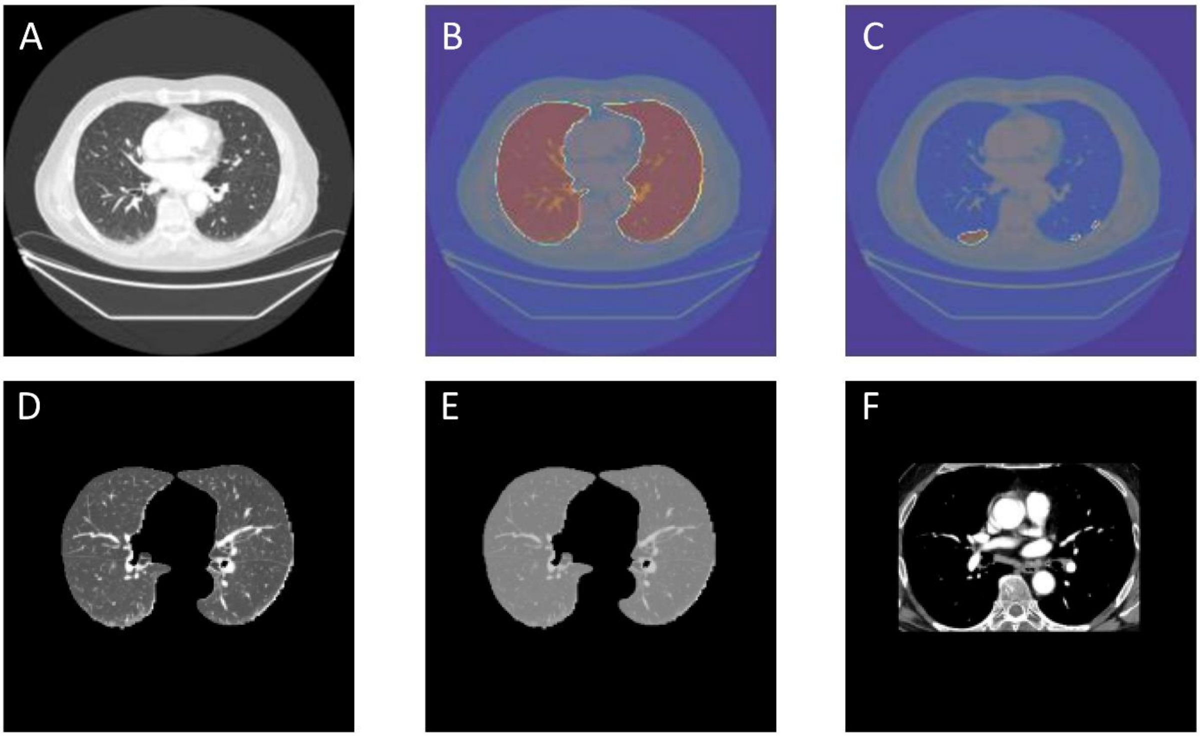


Figure 3.5 Lungs and abnormalities segmentation on a slice from a No Infection patient. A) Original axial slice from a case with No Infection label; B) Lung segmentation obtained on the same slice; C) Aspecific abnormalities

segmented by the lung abnormalities model; Three channel input obtained from the same slice, D) Channel 1; E) Channel 2; F) Channel 3

2.5 3D CNN ARCHITECTURE

An inflated 3D Inception model [15], pre-trained on the Kinetics dataset [16], was trained on 48 consecutive axial slices as 3D input. Inflated 3D inception, also known as 'Two-Stream Inflated 3D ConvNets', is based on the Inception v1 architecture [17] and consists of inflated filters and pooling kernels into 3D, leading to very deep, naturally spatiotemporal classifiers. The model is trained for five epochs and early stopping was performed after the 5th epoch as the validation loss started to increase while the training loss decreased, using the categorical cross-entropy loss as an objective function at a batch size of 2. A batch size of 2 was preferred to fit GPU memory of 11 GB. The model was trained on 10,500 data points (which are different sets of 48 consecutive axial slices obtained from the image volume with an overlap of 10 slices between one set and the other) and validated on 6000 data points. The network weights were updated by using an Adam optimizer at a constant learning rate of $1e^{-4}$ [18]. The model's architecture is depicted in Figure. 3.6.

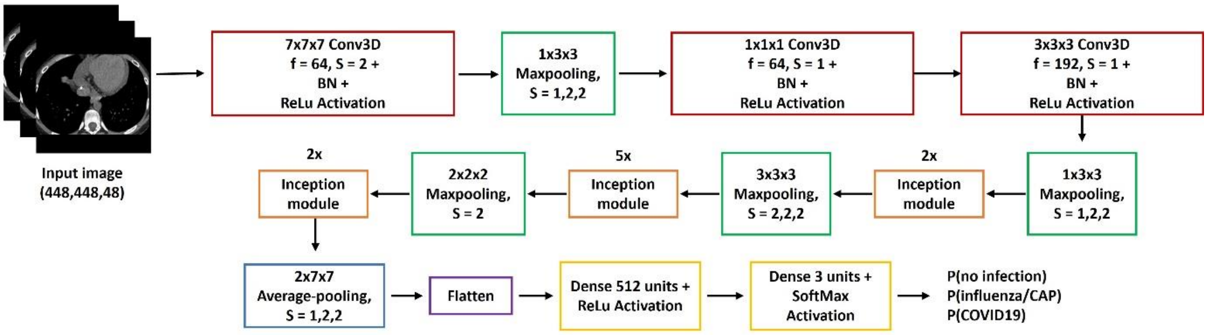


Figure 3.6 3D CNN model network. Inception module architecture is based on the implementation described in [15]. Convolution blocks (red); Maxpooling blocks (green); Inception modules (orange); Average pooling layer (blue); Flatten layer (purple); Fully connected layers (yellow).

2.6 MODEL'S PREDICTION

The model's predictions on the probability of each class were obtained on all 48 consecutive axial slices of the test datasets. The overall class and the overall class probability were computed: if more than 20 % of the predictions correspond to the class COVID-19, then the patient is assigned to that class. If the probabilities for the class Influenza/CAP are higher than 20% then the patient is assigned to the class Influenza/CAP. Otherwise, the scan is classified in the "No infection" class.

2.7 PERFORMANCE METRICS

Classification performances of the deep learning model in the internal testing set and external validation set are expressed in terms of Area under the Curve (AUC), Specificity, and Sensitivity. AUC, Sensitivity, and Specificity are calculated for each class by considering the respective class as positive and the rest of the classes as negative. For instance, the AUC of the class 'influenza' is calculated by considering the class influenza as positive and the class 'no infection' and 'COVID-19' as negatives. All data elaborations were performed in Python (version: 3.6.5) with Keras API. The computation time was calculated on average per scan on the external test set for an RTX 2080 ti 11GB GPU. The model was evaluated according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) [19] (see Appendix 3.1).

2.8 CLINICAL SUMMARY REPORT

The clinician is presented with an automatically generated report containing the results of the classification algorithm. The report presents basic patient data (Patients ID, Scan number, and Scan date) along with the diagnosis (No infection, Influenza/CAP, COVID-19) and the probability calculated by the model for each class. The reports also show the 48 consecutive slices with the corresponding lung and lung abnormalities segmentation masks used by the model to make the classification.

3 RESULTS

3.1 STUDY POPULATION

Table 3.2 lists the study population characteristics for the COVID-19, Influenza/CAP, and no-infection cohorts for the training, validation, internal testing, and external validation set. In the training set, for the COVID-19 patients, 69% needed O_2 therapy at admission with 37% of patients being admitted to the ICU. 17% of COVID-19 patients needed mechanical ventilation and 4% died.

Table 3.2 Study population characteristics

	Training and validation set	**Internal test set**	**External validation set**
Age (years)	63.8 ± 14.44	64.4 ± 15.8	50.67 ± 5.87
Gender (Female %)	48	44	40
Pixel Spacing (mm)	0.71 ± 0.10	0.70 ± 0.07	0.67 ± 0.07
Slice Thickness (mm)	1.19 ± 0.61	1.19 ± 0.59	2 ± 0

3.2 PERFORMANCE ON THE INTERNAL TEST SET

Model performance is reported in Figure. 3.7. The ROC curves for each class (COVID-19, Influenza/CAP, and no Infection) are depicted in Figure. 3.7A. The performance for COVID-19 classification in the internal test set has an AUC of 0.91 (Sensitivity: 87.90 %, Specificity: 88.01 %). Influenza/CAP and No infection classes present an AUC of 0.89 (Sensitivity: 82.97 %, Specificity: 88.79 %) and 0.98 (Sensitivity: 78.18 %, Specificity: 97.72 %) respectively. The confusion matrix (Figure. 3.7B) reports the classification performances (i.e., predicted vs real values) for each class.

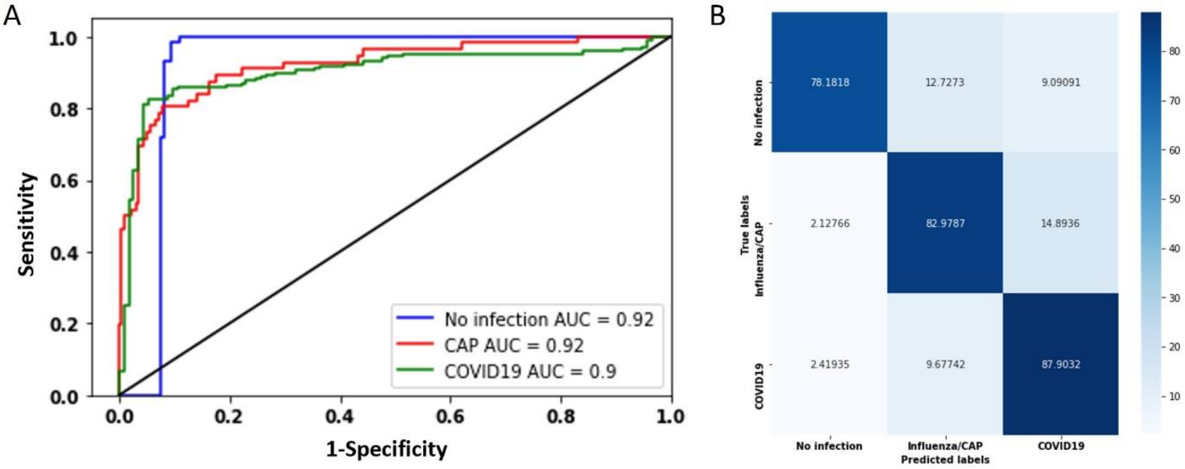


Figure 3.7 Confusion matrix (A) and ROC curve (B) for internal test set

Table 3.3. Performance metrics results

	AUC		**Sensitivity (%)**		**Specificity (%)**	
	Int.	***Ext.***	***Int.***	***Ext.***	***Int.***	***Ext.***
No infection	0.98	0.92	78.18	84.21	97.72	92.59
Influenza/CAP	0.89	0.92	82.97	78.57	88.79	89.44
COVID-19	0.91	0.90	87.90	83.43	88.01	91.07

3.3 PERFORMANCE ON THE EXTERNAL VALIDATION SET

The lung abnormalities segmentation model identified 19 cases with no abnormalities in the external validation set. These scans were not processed by the DL architecture: performance metrics report in Figure. 3.8A and B are based on the 57 cases from the no infection class which presented abnormalities in the lung. Classification for COVID-19 class had an AUC of 0.90 (Sensitivity: 83.43%, Specificity: 91.07%) while Influenza/CAP presents an AUC of 0.92 (Sensitivity: 78.57 %, Specificity: 89.44 %) and No infection with an AUC of 0.92 (Sensitivity: 84.21 %, Specificity: 92.59 %) (Figure. 3.8A). The confusion matrix on the external validation set is reported in Figure. 3.8B.

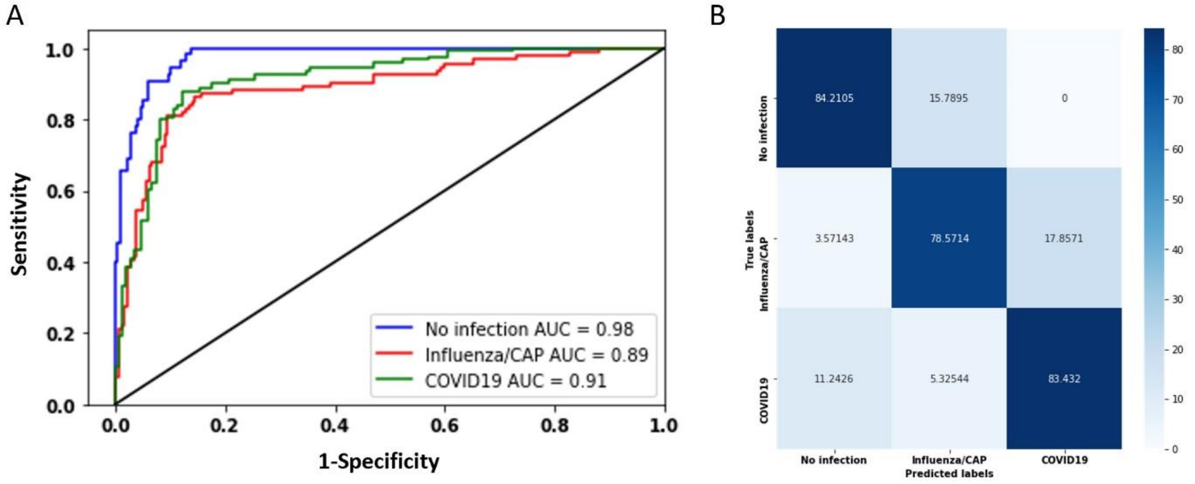
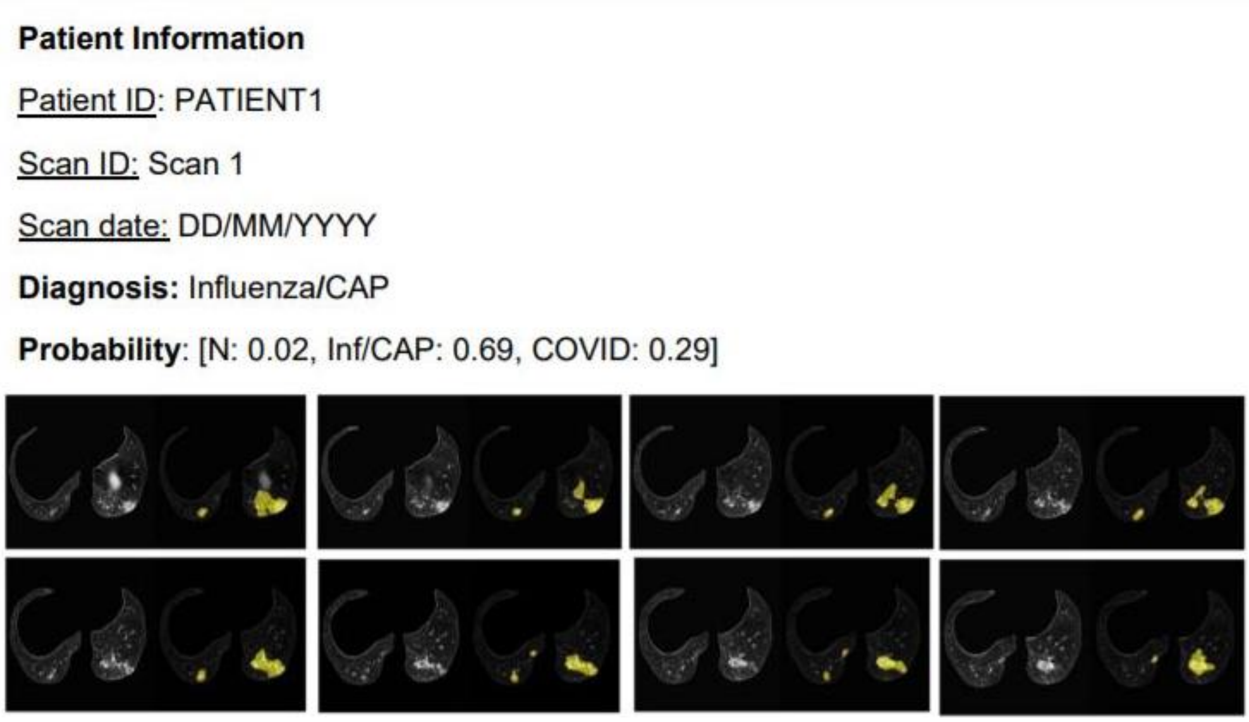


Figure 3.8 Confusion matrix (A) and ROC curve (B) for external test set

The performance in the external validation set is in good agreement with the internal testing set. A summary of the performance metrics for both the internal test set and external validation set is presented in Table 3.3. The TRIPOD score of the proposed model is 47% (15 out of 32 TRIPOD items). The output of the classification workflow is also reported in the Clinical summary report. A sample report for Influenza/CAP and COVID-19 patients is presented in Figure. 3.9 and 3.10.



Patient Information

Patient ID: PATIENT1

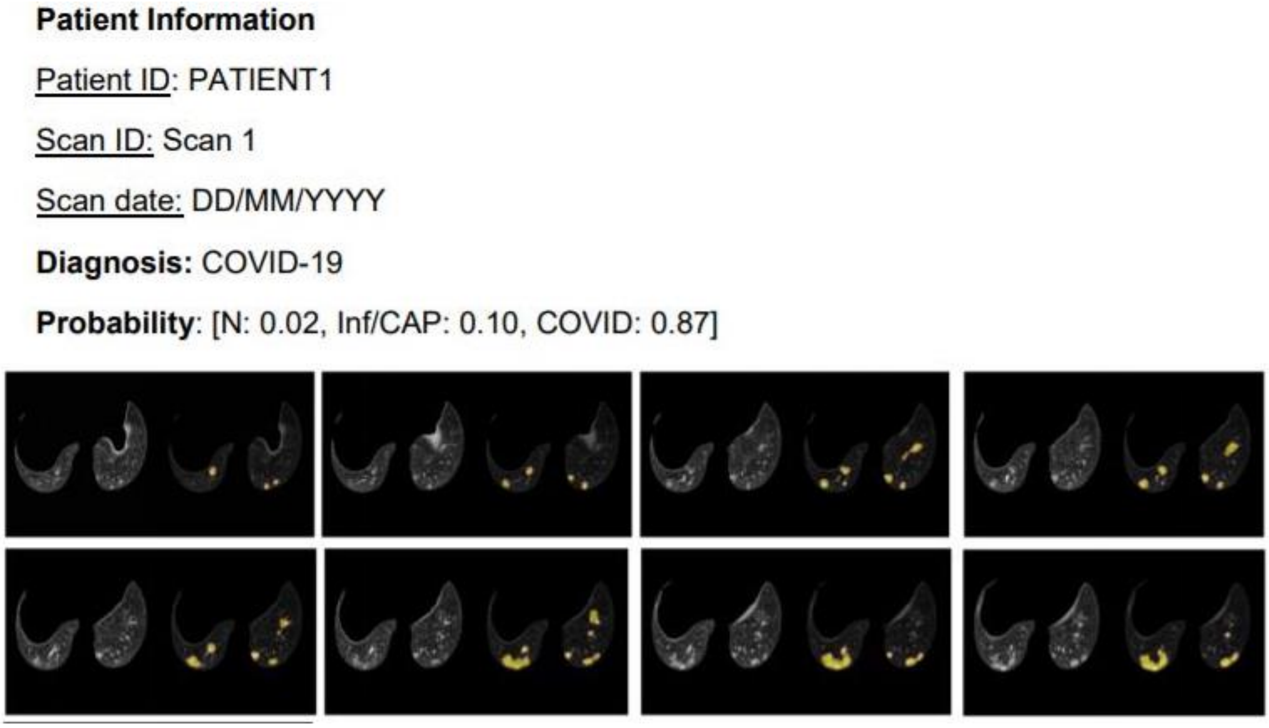
Scan ID: Scan 1

Scan date: DD/MM/YYYY

Diagnosis: Influenza/CAP

Probability: [N: 0.02, Inf/CAP: 0.69, COVID: 0.29]

Figure 3.9 Example of clinical summary report for Influenza/CAP patient.



Patient Information

Patient ID: PATIENT1

Scan ID: Scan 1

Scan date: DD/MM/YYYY

Diagnosis: COVID-19

Probability: [N: 0.02, Inf/CAP: 0.10, COVID: 0.87]

Figure 3.10 Example of the clinical summary report for a COVID-19 patient.

3.4 DISCUSSION

We developed and validated a deep learning AI model for the classification of no-infection, COVID-19, or Influenza/CAP cases based upon CT imaging. The model showed a performance of AUC of 0.90, 0.92, and 0.92 for COVID-19, Influenza/CAP, and No Infection respectively in the external validation. The proposed workflow automatically segments and detects both lung and lung abnormalities, reducing the time and computational burden of the classification task. Moreover, the network produces an automatic clinical summary report, that can be used by the clinician to verify the model decision.

The datasets used for this study come from different countries and different centers. The training cohort is from the University Hospital in Liege while the external validation set is from Babak Imaging Center in Tehran. The training dataset presents a certain homogeneity in imaging acquisition parameters, barring the use of different scanners at different centers. However, the validation data presents different characteristics as coming from a different country with a different standard of care and thus image acquisition protocols. This is an indication of the difference existing among the dataset and indirect proof of the generalizability of the performances of our model which attained good performances also in the external validation dataset.

Several deep learning COVID-19 classification networks have been published thus far, both 2D [20] and 3D [21], also based on automatic segmentation of the lungs [22,23]. Both Machine learning [24] Deep learning [25,26] or a combination of both [27] have been explored for this classification task. The models' performances are high to very high for all the published approaches (AUC between 0.8 and 0.95) and several authors compared the AI workflow with clinicians' performances [28,29], reporting comparable if not better performances from the AI models, and faster and more reproducible diagnosis. Our model has a performance of around AUC 0.9 for all the classes in line with those reported in the literature [23,30].

The possibility to integrate a fully automatic tool for the evaluation of pneumonias source in the clinical workflow can be instrumental to improve patient management and hospital resource allocation. Automatic identification of COVID-19, Influenza/CAP, and no infection patients can reduce the diagnostic errors, related to the human reader experience. The possibility of fast throughput of CT scan analysis will unburden medical staff and free resources to be allocated to more urgent needs. The dubious cases will have to be confirmed by the clinicians upon examination, but the time and effort required to do so will be drastically reduced. A careful evaluation of the real cost-benefit of these tools is sorely needed to promote their application in clinical practice.

However, these automatic tools still have important limitations of applicability in the clinical setting. Overfitting, lack of generalizability, and explainability are the most relevant ones for deep learning models [31,32]. In this study, several techniques were used to prevent overfitting. The model was trained on a multi-vendor (GE, Siemens) dataset with diverse acquisition protocols and differently reconstructed series of the same case. In this way, the model learnt how to generalize in varying image acquisition parameters, which is well reflected by the high sensitivity when evaluated on a held-out internal test set with diverse acquisition protocols and on the external validation set, coming from a different medical center. The ability of the model architecture to generalize to images with diverse imaging parameters is a desired property for real-world clinical applications. Another important aspect of deep learning applied to medical image analysis is explainability, with the "black box" perception hampering the widespread adoption of these methods by clinicians. The production of parsimonious models (i.e. clinicians comprehend and agree with how the model reached the result to support a clinical decision) is instrumental to build confidence and acceptance [33,34]. In the field of AI, there are two main explainability approaches: *post-hoc systems* which provide an explanation for a single specific decision and make it possible to obtain it on demand, and *ante-hoc systems* (also known as "glass box") in which the model is built to be intrinsically explainable, so it is possible to follow each step that the model takes to reach its classification decision [34–36]. Usually (gradient) class activation maps are used to visualize the region of the scan on which the model-based its classification decision [37]: thus, this explainability approach falls under the post-hoc systems category. In the present study, the use of pre-selected and segmented slices containing lung abnormalities can be seen as an ante-hoc explainability system, as the model is specifically looking at the abnormal areas of the lung, segmented by the lung abnormalities segmentation model. In this way, the end user can verify on which slices and on which areas of the slice (i.e., the abnormalities) the model based its classification decision. This can be easily confirmed by the clinicians by looking at the 48 consecutive slices along with lung and lung abnormalities segmentation masks, used by the model for the classification, and reported in the automatic clinical summary report.

Indeed, our model selected only those slices containing abnormalities in the lungs, while most deep learning models published in the literature [7,38] are still based on manual segmentation of the CT scans and use as input all the slices containing lungs or the whole 3D lung volume when automatic segmentation is implemented. Moreover, in previous studies the identification of the regions of the slice used by the model to make its classification decision are the output of the model, helping with interpretability. In our model the identification of the abnormalities in the lungs, linked to the different kinds of pneumonia, is done a-priori*,* removing irrelevant informatio'n (e.g., other pathological presentations in the lung). An additional advantage of our approach is the possibility to select up-front the scans for the model to process. If the selected slices do not present any abnormalities, the model will not process the image, saving time and computational power. This was verified in the external validation set. The "No-infection" patients cohort (n = 76) of the COVID-CT-MD dataset is composed also of healthy patients: our segmentation model correctly identified all the slices without abnormalities and the corresponding scans were not processed by the model (19 out of 76 cases). Furthermore, the pre-selection of slices to be evaluated by the model allows a reduction of the computational burden, also researched in this study by using Inception architecture. Indeed, the use of Inception architecture compared to other approaches based on ResNet or ResNext reduces the computational burden of the model, while maintaining equivalent performances [39]. This approach can allow shallow networks to achieve results comparable to their deeper and more complex counterparts with shorter training times, enabling good classification performances, even when using limited hardware [40]. The computation time (57 s per scan), which can be seen as an indication of the computational burden of the model, was faster than the alternatives reported in the literature. Moreover, compared to other studies that used Inception architecture for similar classification tasks (see Table 3.4), our network showed comparable performances [41,42] and was validated on an external testing set. This validation step is very important to verify the generalizability of the model to patients other than those used for model development (i.e., training and testing).

Table 3.4 Performances of other classification models to distinguish COVID-19 and other source of pneumonia, based on Inception modules

	AUC		**Sensitivity (%)**		**Specificity (%)**		**Sample size**		**Computation time (s)**
	Int.	***Ext.***	***Int.***	***Ext.***	***Int.***	***Ext.***	***Int.***	***Ext.***	
This work	**0.91**	**0.90**	**87**	**83**	**88**	**91**	**273**	**305**	**57**
Wang *et al.* [38]	0.93	0.81	88	83	87	67	455	290	n.r.
El Asnaoui *et al.* [42]	n.r.	-	92	-	96	-	n.r.	-	262
Gifani *et al.* [41]	0.85	-	77	-	n.r.	-	186	-	n.r.

The computation time is calculated as average time per scan on the external validation set.
Int = internal test set, Ext = external validation set, n.r. = not reported

3.5 LIMITATIONS

Considering the limitations of this study, a relevant point related to the external validation test set is the presence of only CAP cases for the Influenza/CAP class. This could lead to a misestimation of the model performance for this classification task. However, influenza cases were present in the internal validation and testing cohorts, and the performances of the model were tested there. An additional external validation dataset with a direct clinician assessment of the source of pneumonia would strengthen the generalizability and add credibility to our approach.

The further distinction between bacterial and viral (non-COVID) pneumonia would represent an additional step forward, allowing the clear identification of the best treatment for each patient. This can also result in better therapeutic management, regarding for example the administration of antibiotics. The misuse and abuse of antibiotics are a cause of great concern in the research and clinical communities. The insurgence of antimicrobial resistance (AMR) is regarded as one of the top 10 global public health threats for the near future [43]. The timely identification of patients with pneumonias that do not require antibiotics can inform better therapy decisions and procedures, contributing to easing the burden of healthcare-associated infections (HAS) from resistant strains of bacteria [44].

Looking at the dataset used for this study, the provenance of all scans from scanners from only two different vendors might limit somehow the generalizability of our approach, even though the image were acquired with two of the most diffuse scanner manufacturers on the market. Adding more data of different vendors, and different acquisition and reconstruction settings might improve the model performances. Ideally, these kinds of clinical decision-making support tools need to be continuously updated with new and heterogeneous data to attain accuracy, specificity, and sensitivity comparable to the latest implementation of diagnostic and therapeutic state-of-the-art, for example *via* Distributed learning [45,46].

To verify the real clinical utility of the proposed tool, a prospective clinical validation study should be carried out comparing performance and time to diagnosis of the AI tool to the current standard of care. Moreover, the clinical use of this tool might need to be updated and modified according to the development of the COVID-19 pandemic. We can expect that pneumonia from COVID-19 infection will become endemic and recurring in the future. Our approach could be adapted to spot the undiagnosed cases or to provide a second independent verification of the occurrence of the disease, also past the emergency status of this pandemic.

4 CONCLUSION

COVID-19-associated lung diseases can mimic other viral lung diseases such as (para-influenza or CAP which may result in misdiagnosis, and delayed and improper treatment. In this context, the development of new diagnostic tools based on AI could become critical for deployment in daily practice shortly. The proposed Inception architecture assured remarkable performances, equal to or higher than AUC 0.9 on the external validation set. Benchmarked against RT-PCR confirmed cases of COVID-19, our AI framework can accurately classify CT scans with COVID-19, Influenza/CAP, or no-infection. This approach could be exploited also for other types of pulmonary diseases, fine-tuning the abnormalities segmentation model to only recognize and select the slices which contain the abnormalities relevant to the investigated disease. To reach this goal a close collaboration between clinicians and data scientist is essential and will also promote the future application of these decision support tools in the clinic.

5 REFERENCES

[1] Marchiori E, Zanetti G, Hochhegger B, Rodrigues RS, Fontes CAP, Nobre LF, et al. High-resolution computed tomography findings from adult patients with Influenza A (H1N1) virus-associated pneumonia. Eur J Radiol 2010;74:93–8. https://doi.org/10.1016/j.ejrad.2009.11.005.

[2] Gao L, Zhang J. Pulmonary High-Resolution Computed Tomography (HRCT) Findings of Patients with Early-Stage Coronavirus Disease 2019 (COVID-19) in Hangzhou, China. Med Sci Monit 2020;26:e923885. https://doi.org/10.12659/MSM.923885.

[3] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395:507–13. https://doi.org/10.1016/S0140-6736(20)30211-7.

[4] Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020;382:1708–20. https://doi.org/10.1056/NEJMoa2002032.

[5] Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology 2020:200823. https://doi.org/10.1148/radiol.2020200823.

[6] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[7] Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia. Engineering 2020;6:1122–9. https://doi.org/https://doi.org/10.1016/j.eng.2020.04.010.

[8] Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med 2020;121:103795. https://doi.org/10.1016/j.compbiomed.2020.103795.

[9] Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. Fair Allocation of Scarce Medical Resources in the Time of Covid-19. N Engl J Med 2020;382:2049–55. https://doi.org/10.1056/NEJMsb2005114.

[10] Afshar P, Heidarian S, Enshaei N, Naderkhani F, Rafiee MJ, Oikonomou A, et al. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. Sci Data 2021;8:121. https://doi.org/10.1038/s41597-021-00900-3.

[11] An P, Xu S, Harmon S, Turkbey E, Sanford T, Amalou A, et al. CT Images in Covid-19 [Data set]. Cancer Imaging Arch 2020. https://doi.org/https://doi.org/10.7937/tcia.2020.gqry-nc81.

[12] Morozov SP, Andreychenko AE, Pavlov NA, Vladzymyrskyy A V, Ledikhova N V, Gombolevskiy VA, et al. MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset 2020.

[13] Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 2018;106:249–59. https://doi.org/https://doi.org/10.1016/j.neunet.2018.07.011.

[14] Ilsen B, Vandenbroucke F, Beigelman-Aubry C, Brussaard C, de Mey J. Comparative Interpretation of CT and Standard Radiography of the Pleura. J Belgian Soc Radiol 2016;100:1–10. https://doi.org/10.5334/jbr-btr.1229.

[15] Carreira J, Zisserman A, Com Z, Deepmind †. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. n.d.

[16] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The Kinetics Human Action Video Dataset. ArXiv 2017.

[17] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12- June, IEEE Computer Society; 2015, p. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

[18] Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015, p. 1–15.

[19] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.

[20] Ibrahim DM, Elshennawy NM, Sarhan AM. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. Comput Biol Med 2021;132:104348. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104348.

[21] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 2020;296:E65–71. https://doi.org/10.1148/radiol.2020200905.

[22] Ghaderzadeh M, Asadi F. Deep Learning in the Detection and Diagnosis of COVID-19 Using Radiology Modalities: A Systematic Review. J Healthc Eng 2021;2021:6677314. https://doi.org/10.1155/2021/6677314.

[23] Mohammad-Rahimi H, Nadimi M, Ghalyanchi-Langeroudi A, Taheri M, Ghafouri-Fard S. Application of Machine Learning in Diagnosis of COVID-19 Through X-Ray and CT Images: A Scoping Review . Front Cardiovasc Med 2021;8:185.

[24] Wu Z, Li L, Jin R, Liang L, Hu Z, Tao L, et al. Texture feature-based machine learning classifier could assist in the diagnosis of COVID-19. Eur J Radiol 2021;137. https://doi.org/10.1016/j.ejrad.2021.109602.

[25] Privor-dumm LA, Poland GA, Barratt J, Durrheim DN, Deloria M, Vasudevan P, et al. Automatic distinction between COVID-19 and common pneumonia using multi-scale convolutional neural network on chest CT scans. Chaos, Solitons and Fractals 2020;140.

[26] Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020;56. https://doi.org/10.1183/13993003.00775-2020.

[27] Wang H, Wang L, Lee EH, Zheng J, Zhang W, Halabi S, et al. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. Eur J Nucl Med Mol Imaging 2021;48:1478–86. https://doi.org/10.1007/s00259-020-05075-4.

[28] Liu H, Ren H, Wu Z, Xu H, Zhang S, Li J, et al. CT radiomics facilitates more accurate diagnosis of COVID-19 pneumonia: compared with CO-RADS. J Transl Med 2021;19:1–12. https://doi.org/10.1186/s12967-020-02692-3.

[29] Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell 2020;181:1423-1433.e11. https://doi.org/10.1016/j.cell.2020.04.045.

[30] Ozsahin I, Sekeroglu B, Musa MS, Mustapha MT, Uzun Ozsahin D. Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence. Comput Math Methods Med 2020;2020:1–10. https://doi.org/10.1155/2020/9756518.

[31] Ying X. An Overview of Overfitting and its Solutions 2019:22022. https://doi.org/10.1088/1742-6596/1168/2/022022.

[32] Caruana R, Lawrence S, Giles L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Adv. Neural Inf. Process. Syst., 2001.

[33] Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. J Imaging 2020;6. https://doi.org/10.3390/jimaging6060052.

[34] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9:e1312. https://doi.org/https://doi.org/10.1002/widm.1312.

[35] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? ArXiv 2017:1–28.

[36] Holzinger A. Explainable AI and Multi-Modal Causability in Medicine. I-Com 2021;19:171–9. https://doi.org/10.1515/icom-2020-0024.

[37] Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-18685-1.

[38] Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). Eur Radiol 2021. https://doi.org/10.1007/s00330-021-07715-1.

[39] Bianco S, Cadene R, Celona L, Napoletano P. Benchmark Analysis of Representative Deep Neural Network Architectures. IEEE Access 2018;6:64270–7. https://doi.org/10.1109/ACCESS.2018.2877890.

[40] Bressem KK, Adams LC, Erxleben C, Hamm B, Niehues SM, Vahldiek JL. Comparing different deep learning architectures for classification of chest radiographs. Sci Rep 2020;10:13590. https://doi.org/10.1038/s41598-020-70479-z.

[41] gifani P, Shalbaf A, Vafaeezadeh M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. Int J Comput Assist Radiol Surg 2021;16:115–23. https://doi.org/10.1007/s11548-020-02286-w.

[42] El Asnaoui K, Chawki Y. Using X-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn 2020;0:1–12. https://doi.org/10.1080/07391102.2020.1767212.

[43] Global Action Plan on Antimicrobial Resistance. Microbe Mag 2015;10:354–5. https://doi.org/10.1128/microbe.10.354.1.

[44] Haque M, McKimm J, Sartelli M, Dhingra S, Labricciosa FM, Islam S, et al. Strategies to Prevent Healthcare-Associated Infections: A Narrative Overview. Risk Manag Healthc Policy 2020;13:1765–80. https://doi.org/10.2147/RMHP.S269315.

[45] Zerka F, Urovi V, Vaidyanathan A, Barakat S, Leijenaar RTH, Walsh S, et al. Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM). IEEE Access 2020;8:183939–51. https://doi.org/10.1109/ACCESS.2020.3029445.

[46] Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Informatics 2020:184–200. https://doi.org/10.1200/cci.19.00047.

Chapter 4

Externally validated deep learning model for the diagnosis and detection of pulmonary embolism on chest CTPA images

Akshayaa Vaidyanathan, Flore Belmans, Fabio Bottari, François Blistein, Ingrid van Peufflik, Wim Vos, Mariaelena Occhipinti, Philippe Lambin, Julien Guiot, Sean Walsh, Externally validated deep learning model for the diagnosis and detection of pulmonary embolism on chest CTPA images

Our research presented in this chapter exploits the use of an AI-based classifier for the diagnosis and detection of pulmonary embolism in CTPA scans. A 2D classifier based on the ResNext50 architecture was trained and validated using the RSNA-STR Pulmonary Embolism CT (RSPECT) multicentric dataset composed of 7169 patients. From these retrospective data, 85,000 slices positive for PE and 123,428 negatives for PE were extracted for training. For internal validation, 9,922 slices were used for each class. The model was externally validated at the patient level using a dataset of 156 adult patients from 3 different public sources, having all emboli segmented by at least one experienced radiologist. To gain insight into the model predictions, activation maps were extracted using the Grad-CAM method. Comparing these maps with the ground truth (GT) segmentations, it was determined if the activated regions corresponded to regions of PE by computing the percentage of GT PE that is activated and the percentage of activated regions corresponding to GT PE. The PE classification model reached an area under the curve (AUC) of 0.86 [0.800-0.919], a sensitivity of 82.68 % [75.16 - 88.27] and a specificity of 79.31 % [61.61 - 90.15] on the external validation set. The activation maps of the slices rightly predicted positive by the PE classifier showed good visual correspondence with areas of PE. This was also quantitatively confirmed as 79.2% of PE regions in the GT were highlighted in the activation maps and the percentage of activated regions corresponding to GT PE is 80.3%. Our deep learning-based classifier can identify patients with pulmonary embolism with high accuracy and can localize the emboli by extracting the activation maps from the network. The activation maps help explain the features used by the deep learning model to make the diagnosis, increasing the likelihood of acceptance by clinicians for clinical routine.

1 BACKGROUND

Pulmonary embolism is one of the most fatal cardiovascular diseases, causing more than 60,000 deaths in the United States [1] and around 40,000 deaths in Europe [2] each year. A prompt and accurate diagnosis is key to avoiding fatal consequences for missed diagnosis on one hand and risks associated with anticoagulation effects for overdiagnoses on the other hand. Diagnosis relies on the assessment of pre-test clinical probabilities and then on imaging examinations, with computed tomography pulmonary angiography (CTPA) being the most widely available worldwide. Symptoms at presentation are aspecific and rapid diagnosis is critical to care decisions. Computed tomography pulmonary angiography (CTPA) has become the gold standard for the diagnosis of PE in recent years. However, the interpretation of CTPA is a time-consuming process and it presents a high inter and intra-reader variability [3], [4]. Timely interpretation is of essence in patients with concurrent oncological conditions or other respiratory pathologies. To expedite diagnosis and unburden medical staff, the use of computer-aided diagnosis (CAD) tools has gathered considerable momentum in recent years, especially in the pulmonary medicine field [5], [6]. Among the different approaches for automated detection of pulmonary embolism, Deep Learning (DL) networks have been employed since the very beginning, exploiting clinical covariates and/or ventilation-perfusion scans, with good results but rather poor robustness and generalizability [7]–[9]. With the development of Artificial Intelligence and in particular DL, the analysis of clinical images, mostly CTPA alone [10], [11] or in combination with clinical covariates [12], has attracted the interest of the scientific community. One of the main hurdles to the integration of these methods in the current clinical practice is the lack of interpretability and explainability of the inner workings of the DL models [13], [14]. Trust and acceptance of these innovative tools might benefit from a clearer view of the process with which the model reaches the diagnostic decision, linking the algorithm output to real clinical or imaging evidence. The goal of our study was the development and validation of a deep learning-based classifier for the detection of pulmonary embolisms (PE) in chest CTPA images. Particular care has been devoted to the explainability of the decision-making process of the model, using Grad-CAM activation maps to identify the regions in the scan contributing the most to the model output. This would allow improving the clinical adherence to diagnostic protocol and standard of care, with the ultimate goal of having an integrated Computer Aided Diagnostic (CAD) workflow with dynamic and adaptive prioritization of severe cases, improving patients' management.

2 MATERIALS AND METHODS

2.1 DATASET CHARACTERISTICS

The dataset for model training and validation consisted of retrospective data of 7,169 patients from the RSNA-STR Pulmonary Embolism CT (RSPECT) dataset [15]. The training set was composed of 6,316 patients while the validation set contained 853 patients. The external validation set is composed of 156 patients (127 PE positive, 19 PE negative) coming from Ferdowsi University of Mashhad’s dataset (FUMPE) (35 patients) [16], CAD-PE dataset (91 patients) [17], and Linköping dataset (30 patients) [18]. All the CTPA scans in the external validation dataset have been segmented and annotated by at least one experienced radiologist (5+ years of experience). A summary of the external patients' cohort division is reported in Figure 4.1. Patients' demographics are presented in Table 4.1.

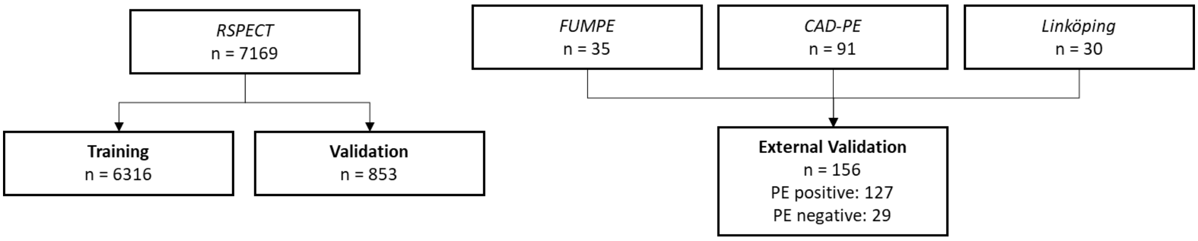


Figure 4.1 Flow chart of the patient cohorts division between training, validation, and external validation sets.

Table 4.1. Patients' characteristics

Dataset	Patients number	Sex	Age (mean +SD)
RSPECT	7169	N. R.	N. R.
FUMPE	35	M = 48 %, F = 52 %	52 ± 19
CAD-PE	91	M = 81 %, F = 19 %	65 ± 18
Linköping	30	N. R.	45-93

N.R.: not reported

2.2 IMAGING

The imaging characteristics of the datasets used in this study are as follows,

RSPECT: The CTPA scans were collected from different scanners with varying reconstruction kernels (STANDARD, B, FC08-H, etc.) with slice thickness between 0.5 and 0.8 mm and slice spacing between 1 and 2 mm in the axial plane. The annotation of this dataset was a collaborative effort between the RSNA and the Society of Thoracic Radiology (STR). A panel of 190 thoracic imaging experts was recruited for the data annotations, with 50 to 150 cases to annotate for each participant [15].

FUMPE: All CTPA images were acquired in one breath hold with slice-thickness between 1 mm and 2 mm and slice-interval ≤1.5 mm in the caudocranial direction. The images were acquired with NeuViz 16 multi-slice helical CT scanners (Philipps and Neusoft Medical Imaging). For each image, two expert radiologists provided the ground truth (GT) with the assistance of a semi-automated image processing software tool [16].

CAD-PE: All CTPA images were acquired in one breath in the caudocranial direction. Image pixel size ranges from 0.58 to 0.85 mm and reconstruction slice thickness between 0.75 and 1.5 mm. All studies were performed with SIEMENS Somaton Sensation 40 scanner. The ground truth (i.e. PE positive or negative) was provided by three experienced radiologists on slice level [19].

Linköping: The scans have been acquired on Philips Brilliance 64 CT or GE Lightspeed VCT with a slice thickness between 0.625 and 1 mm. The scans are contrast-enhanced and acquired in the pulmonary arterial phase; emboli were delineated by an experienced radiologist [18].

2.3 IMAGE PRE-PROCESSING

To create the training and validation set to train the DL model, 2D slices were extracted from the 7,169 patient scans in the RSPECT dataset. All slices in the dataset that were annotated as being positive for PE, were included. To achieve a balanced dataset, negative slices with lung area larger than or equal to 75% of the mean lung area were selected. This avoids overfitting of the model towards the presence of a part of the abdomen in the image. The lung areas were computed using an automatic in-house lung segmentation model (Section 2.5 of Chapter 2). The above approach resulted in a total of 208,428 slices coming from 6,316 patients for training (85,000 PE positive and 123,428 PE negative slices) and 19,844 slices coming from 853 patients (9,922 PE positive and 9,922 PE negative slices) for internal validation. It was ensured that there was no overlap between patients in the training and validation set. A summary of the composition of the training and validation set is reported in Figure. 4.2. Before being exposed to the model, the intensities of every image were clipped at mediastinal window level settings (W:350 HU, L:50 HU), which is the most optimal to detect PE. All slices were cropped to size (448,448).

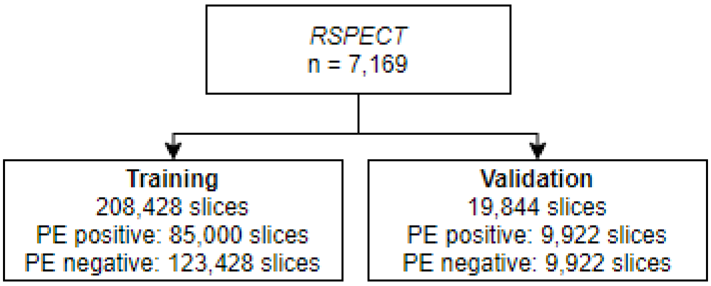

RSPECT
n = 7,169
Training
208,428 slices
PE positive: 85,000 slices
PE negative: 123,428 slices
Validation
19,844 slices
PE positive: 9,922 slices
PE negative: 9,922 slices

Figure 4.2 Flow chart of the composition of training and validation set.

2.4 MODEL ARCHITECTURE

A 2D classifier based on the ResNext50 architecture was used for slice classification (Figure. 4.3) [20], [21]. The model's input is divided into three channels which are filled with three consecutive slices during training. In this way, the model training takes into account also the following two slices for each slice. The classifier was trained with ImageNet pre-trained weights, minimizing categorical cross-entropy loss for a total of 4 epochs. During the first two epochs, the slices in the 3-channel input were cropped using a rectangular crop around the lung region. The lung regions were extracted using an automatic in-house lung segmentation model (Section 2.5 of chapter 2). The rectangular crop was obtained with x_min = minimum x value for which lungs pixels are present, x_max = maximum x value for which lungs pixels are present and y_min = minimum y value for which lungs pixels are present, y_max = maximum y value for which lungs are present. During the last two epochs, the model was trained without cropping. The last layer in the ResNext50 model was followed by a Global Average pooling layer which reduces the image spatial resolution, followed by a fully connected layer with 1024 units and ReLu activation, which is followed by a classification layer containing 2 units with Softmax activation. The network weights are updated by using the Adam optimizer at an initial learning rate of $1e^{-5}$. At each epoch during training, 85,000 PE negative slices were randomly selected from the 123,428 images to maintain a balance between PE positive and PE negative slices.

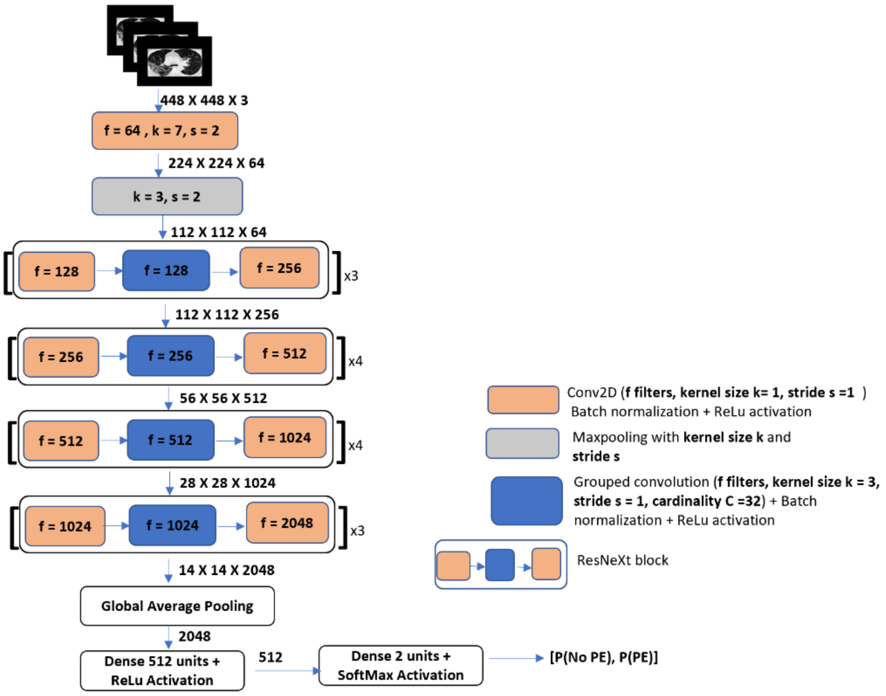

448 X 448 X 3
f = 64 , k = 7, s = 2
224 X 224 X 64
k = 3, s = 2
112 X 112 X 64
f = 128
f = 128
f = 256
x3
112 X 112 X 256
f = 256
f = 256
f = 512
x4
56 X 56 X 512
f = 512
f = 512
f = 1024
x4
28 X 28 X 1024
f = 1024
f = 1024
f = 2048
x3
14 X 14 X 2048
Global Average Pooling
2048
Dense 512 units + ReLu Activation
512
Dense 2 units + SoftMax Activation
[P(No PE), P(PE)]
Conv2D (f filters, kernel size k= 1, stride s =1) Batch normalization + ReLu activation
Maxpooling with kernel size k and stride s
Grouped convolution (f filters, kernel size k = 3, stride s = 1, cardinality C =32) + Batch normalization + ReLu activation
ResNeXt block

Figure 4.3. CNN architecture. [ ] x X denotes that the block is repeated X times. The grouped convolution block is equivalent to the implementation described by Xie et al [22]. Notations in blue text highlight the spatial resolution and the feature map count.

2.5 PERFORMANCE METRICS

The performance of the 2D DL classifier was first evaluated in a slice-based way on the internal validation set. Second, the performance was explored after aggregating the 2D classifier predictions to a patient-level result on both the internal validation and external validation set. To translate the classifier predictions on 2D slices to a patient-level probability of a positive PE diagnosis, the following workflow has been employed for each patient:

1. Select all slices containing lungs using an in-house lung segmentation model
2. Compute the average probability of the classifier predictions on all lung-containing slices

Performances are reported in terms of Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, the confusion matrix, sensitivity, and specificity. More details are reported in Appendix 4.1. The model was evaluated according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) [23] and Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [24] (see Appendix 4.2 and 4.3).

2.6 EXPLAINABILITY OF THE MODEL PREDICTION

To increase the transparency of the decision-making process of the DL classifier, the Gradient-weighted Class Activation Mapping (Grad-CAM) method [25] was employed to identify regions of interest in the model (activated regions). Using this approach, activation maps were extracted and visualized in the image as heat maps for true positive, false positive, and false negative results on the external validation set. Besides a qualitative evaluation, the clinical relevance of the activation maps was also quantitatively measured by comparing them to the GT manual segmentations present in these datasets. To quantify the correspondence of the activated regions to regions of PE, two additional metrics were computed: the percentage of GT PE that is activated and the percentage of activated regions corresponding to GT PE.

3 RESULTS

3.1 SLICE-LEVEL RESULTS

The performance of the 2D PE classifier on the slices in the internal validation set is presented using the ROC curve and confusion matrix in Figure. 4.4A and B. The model achieved an AUC of 0.95 [0.952-0.958] and after applying a 50% probability threshold, a sensitivity of 87.89% [87.23-88.51] and specificity of 89.13% [88.50-89.72]. The calibration plot (Figure. 4.4C) shows that the model is well calibrated with an intercept close to 0 (0.023) and a slope approaching 1 (0.95).

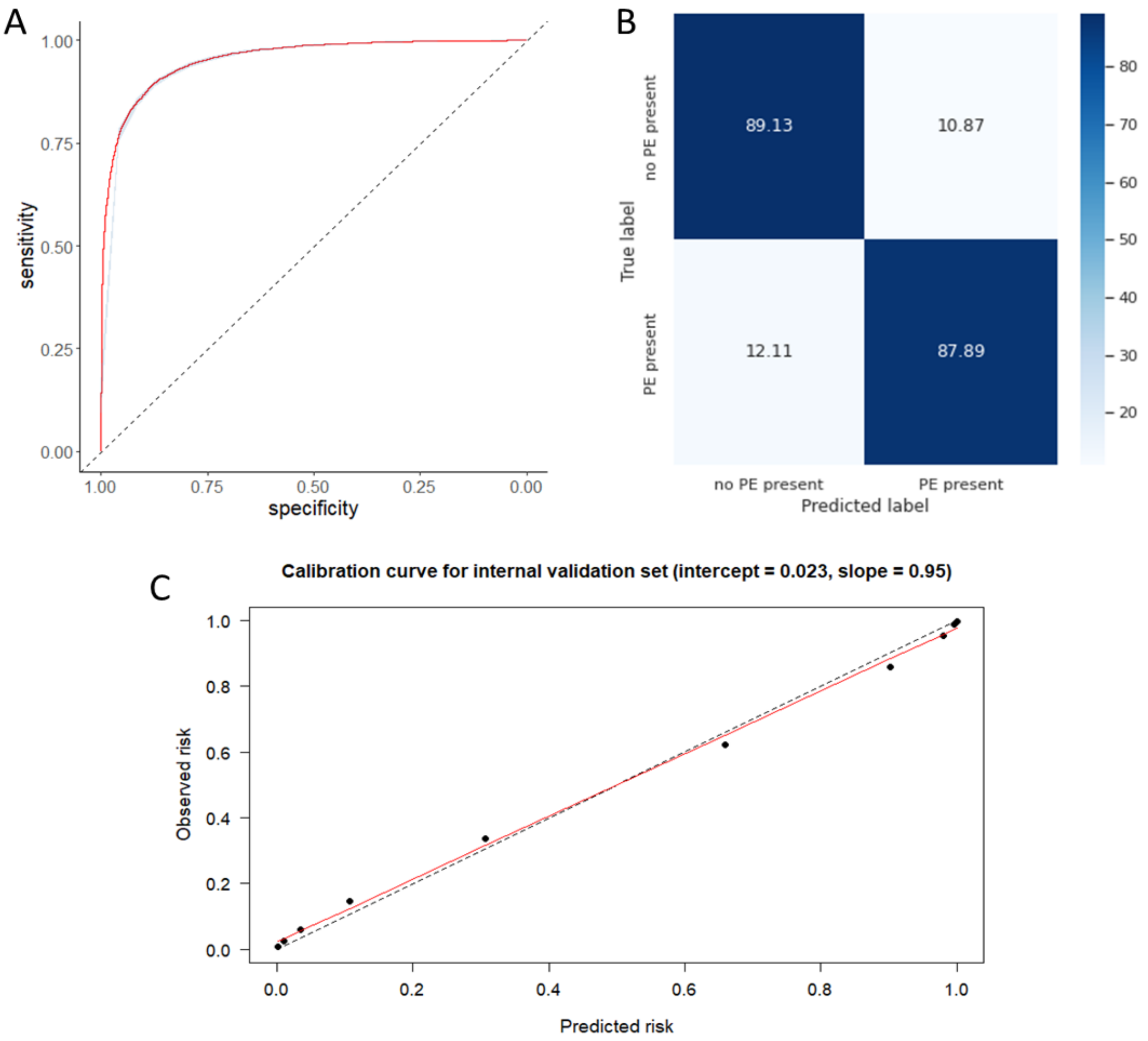


Figure. 4.4 ROC curve (AUC = 0.95) with bootstrap confidence intervals for the slice-level PE classifier (A); confusion matrix (B) on the internal validation set; (C) Calibration curve for the slice-level PE classifier on the internal validation set (intercept = 0.023, slope = 0.95).

3.2 PATIENT-LEVEL RESULTS

After applying the workflow for aggregating the 2D model predictions to a final probability per patient, the performances on the patient-level were assessed. The ROC curves for this approach were computed on the internal validation set (Figure. 4.5A) and external validation set (Figure. 4.5B). AUC values of respectively 0.87 [0.842 - 0.893] and 0.86 [0.800-0.919] were achieved. To define the decision threshold at which the classifier would operate to diagnose PE-positive patients, a threshold curve analysis was performed on the internal validation set (Figure. 4.6). The analysis involves the computation of the sensitivity and specificity at different decision thresholds to identify the optimal decision threshold as the intersection of both curves, maximizing both the sensitivity and specificity. The analysis revealed an optimal threshold of 22.58%. Using this probability threshold of 22.58%, the confusion matrices were computed for the internal validation (Figure. 4.5C) and external validation set (Figure 4.5D). On the internal validation set, a sensitivity of 78.74% [74.68 - 81.74] and specificity of 78.42% [74.68 - 81.74] were reached. On the external validation set, the achieved sensitivity was 82.68% [75.16 - 88.27] and specificity 79.31% [61.61 - 90.15]. The performance results are summarized in Table 4.2. As an example, the analysis time for a standard chest CTPA scan with an axial resolution of 1.25mm was around 1 min, of which half was dedicated to lung segmentation and a half to the classification itself, using an RTX 2080 ti 11GB GPU.

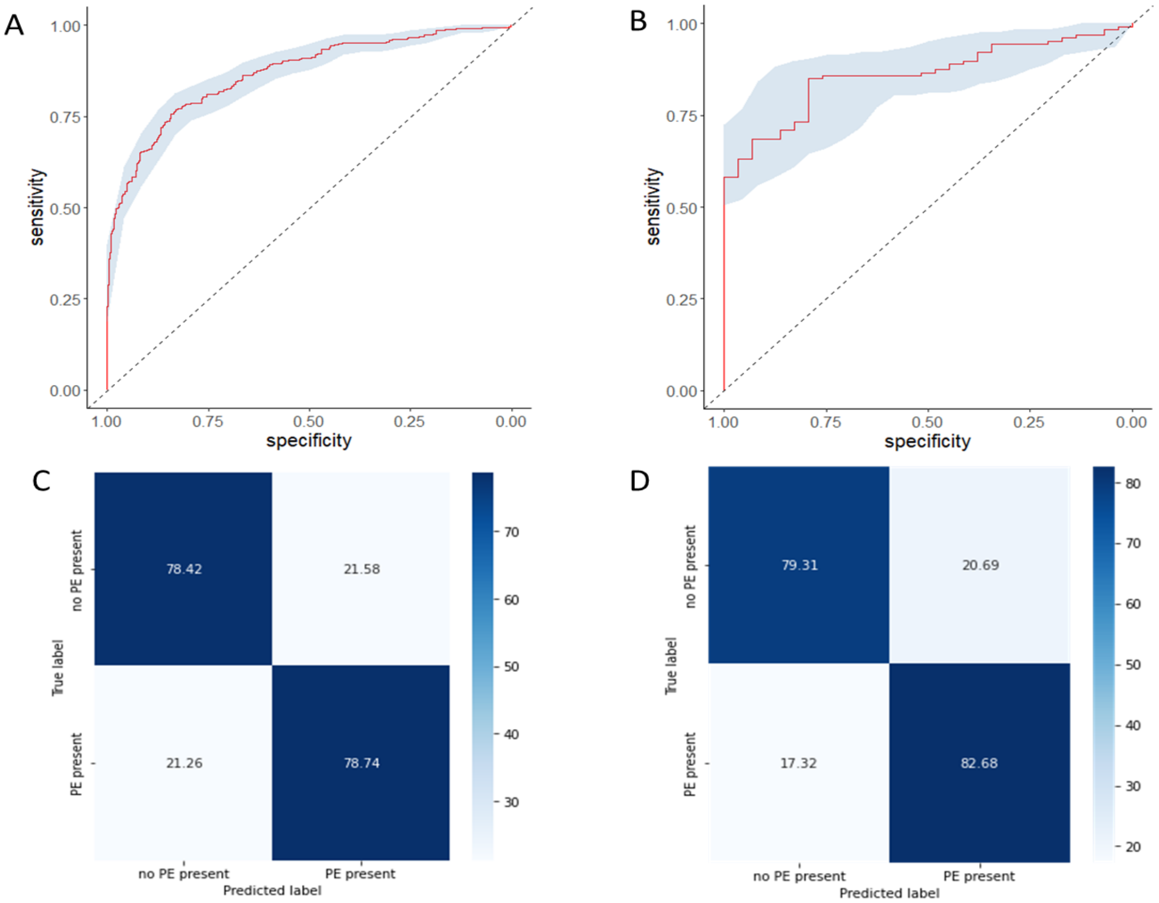


Figure. 4.5 Performances of the PE classifier on patient-level analysis: ROC curve (AUC = 0.87) with bootstrap confidence intervals on the internal validation set (A); ROC curve (AUC = 0.86) with bootstrap confidence intervals on the external validation set (B); confusion matrix computed with probability threshold of 22.58% on the internal validation set (C) and external validation set (D).

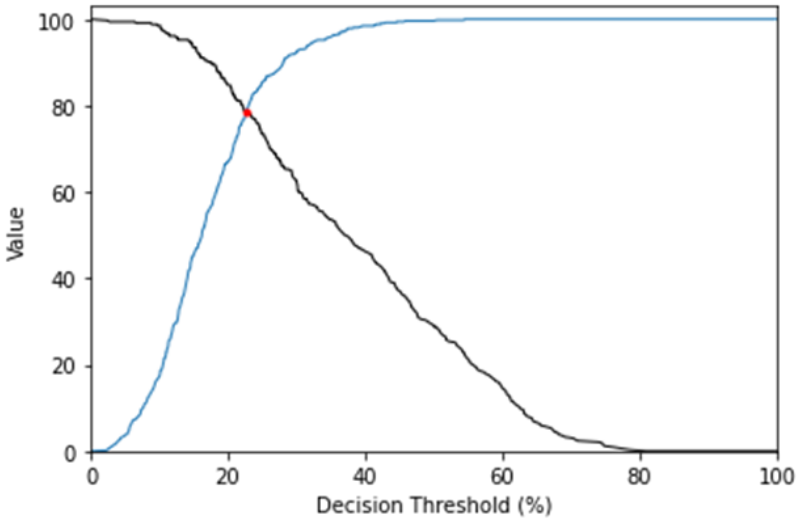


Figure. 4.6 Threshold curve analysis on the internal validation set; Sensitivity (black line), Specificity (blue line), Optimal threshold (red dot) 22.58 %

Table 4.2. Performance metrics for the patient-level PE classification model with 95% confidence intervals at a decision threshold of 22.58%.

	AUC	**Sensitivity (%)**	**Specificity (%)**
Internal validation set (n=853)	0.87 [0.842 - 0.893]	78.74 [74.68 - 81.74]	78.42 [74.68 - 81.74]
External validation set (n=156)	0.86 [0.800-0.919]	82.68 [75.16 - 88.27]	79.31 [61.61 - 90.15]

Calibration plots of the patient-level classification model have been produced for both the internal validation set and external validation set (Figures. 4.7A and B). The plots show apparent poor calibration of the method, with intercept slope pairs of respectively (–0.11, 1.94) and (0.42, 1.12). This is justified by the nature of our method which takes the average prediction over all slices. As PE is only present in a limited number of slices in the volume, the averaging induces a reduction of the final probability for positive cases.

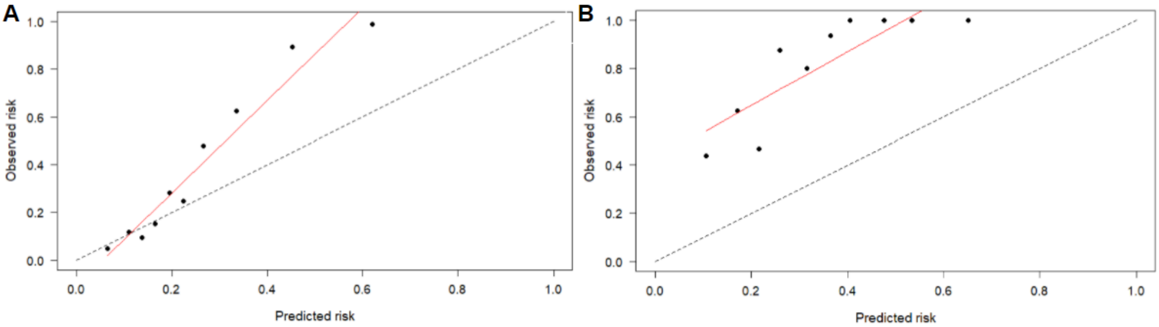


Figure 4.7. Calibration plots for the patient-based PE classifier on the internal validation (A) and external test (B) sets. The intercept and slope pairs are respectively (–0.11, 1.94) and (0.42, 1.12).

The TRIPOD score of the proposed model is 58 % (18 out of 31 TRIPOD items) while the CLAIM score is 60 % (25 out of 42 items).

In Figure 4.8, three examples of extracted Grad-CAM activation maps are reported for true positive, false negative, and false positive slices. The original slice is compared with the activation map and the radiologist's manual segmentation in parallel. The activation maps are visualized as heat maps superimposed on the image where the color code indicates the areas of the slice used by the model for the classification, graded in terms of importance (from red to blue). In the external validation set, the percentage of GT PE that is activated is 79.2%. Using the same data, the percentage of activated regions corresponding to GT PE is 80.3%.

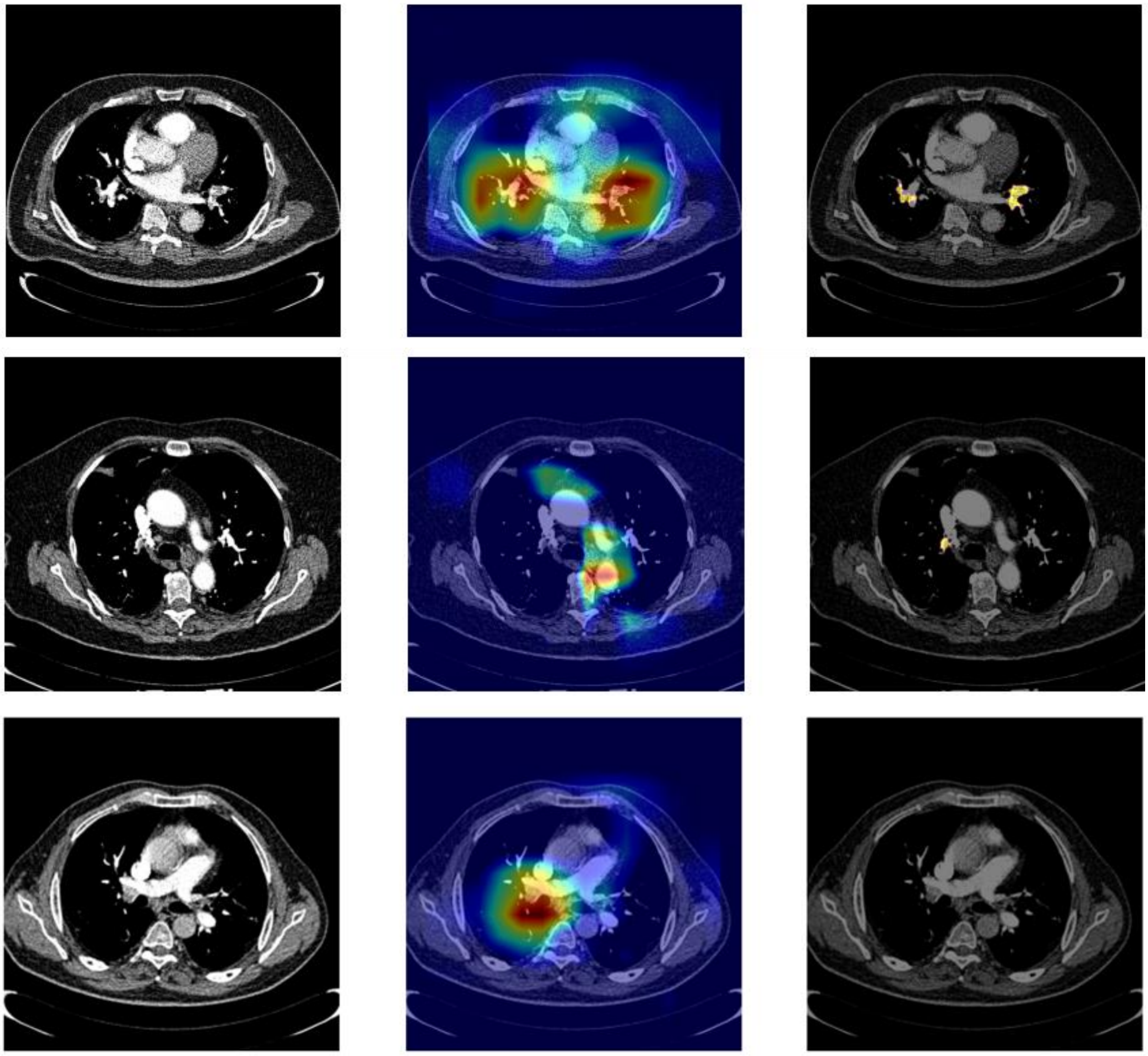


Figure 4.8 Examples of explainability by Grad-CAM activation maps for the PE classifier. Axial slice at the mediastinal window (left), Grad-CAM activation map (center), ground truth from radiologist manual segmentation (right). True positive case (top row), P = 0.98: the presence of filling defects within the segmental pulmonary arteries on axial CT image (on the left) corresponding to the red zones in the Grad-CAM activation maps (center) and the colored areas segmented by radiologists (on the right). False Negative case (middle row), P = 0.020, False positive case (bottom row), P = 0.90: although no emboli can be identified on the axial CT scan at the mediastinal window and none was segmented by radiologists (on the right), the classifier spotted an area in the right lower lobe.

4 DISCUSSION

We have developed and externally validated an AI model for the classification of pulmonary embolism in CTPA images. The model architecture was based on a 2D CNN and was trained on slice-level input, internally and externally validated at a patient-level input. Patient-level predictions were obtained by aggregating and thresholding the probabilities across all the slices per image per patient. Furthermore, the performance of the model was investigated for explainability using the Grad-CAM method. The activated regions are compared with manual segmentations of the embolisms for quantitative assessment of the explainability of the model's predictions.

Several studies have previously shown that the prompt diagnosis and subsequent treatment of PE patients can reduce morbidity and mortality [26], [27]. In the last decades, the use of CTPA as a first-line diagnostic tool for PE management has increased exponentially. However, patients might still incur more than 6 days of delay in diagnosis and 26% of patients are misdiagnosed during their first visit [28], [29]. In this scenario, methods to expedite and automatize the interpretation of CTPA scans might result in better triage of urgent cases of PE, improve time to diagnosis and treatment, and, at the same time, ease the pressure on medical staff and hospital resources.

Other research groups have tackled this unmet clinical need in the recent past. For example, Huang *et al.* reported a study that used 3D CNN for automated diagnosis of pulmonary embolism using CT images and utilized 24 consecutive axial slices for training and validation at the patient level. The method achieved a comparable AUC of 0.84 on detecting PE on internal validation and 0.85 on an external validation dataset. The method also explored the explainability of the model’s decisions based on Grad-CAMs showing visual representations of activations maps overlapped on images [30]. Another study that uses ResNeSt-50 CNN for PE classification at slice level performed at an AUC of 0.95 and used sequence modeling for PE classification at patient level with an AUC of 0.90 on an internal validation dataset [31]. However, performances on an external validation dataset and the explainability of the models were unexplored. A comparison between our approach and other recently published models for PE detection is reported in Table 4.3. Recently, Schmuelling and coworkers reported the technical implementation in the emergency department of a DL-based PE detector from a CTPA scan [32]. The performances of their method are remarkable, especially in terms of specificity (95%). However, this clinical implementation study did not show the expected results: there was no significant impact on the measured clinical performances (reading times, radiology report communication time, time to anticoagulation, and patient turnaround times). In other words, the introduction of the DL model in the clinical workflow did not impact significantly the clinical routine triaging of PE patients. This praiseworthy research is a cautionary tale for all future clinical implementation of said PE detectors in the clinics. From the author's own words, the issues evidenced by this study regarded “*the implementation of the DL-algorithm not accompanied by a transition between the implementation time periods, a training for doctors, a change in standard operating procedures (SOPS) or other measures that support a structured introduction of DL- algorithms and their handling*”.

Table 4.3. Comparison of different DL approaches for identification of PE on CTPA scans

	APPROACH	PERFORMANCES (INTERNAL/EXTERNAL)	REMARKS
This work	2D CNN (Resnext50), predictions are translated to a patient output	AUC = 0.87/ 0.86	Explainability using Grad-CAMs
Liu *et al.* [33]	U-net for segmentation of PE	AUC =0.92 / -	No external validation, only acute PE, only high-quality images considered
Huang *et al.* [34]	3D CNN with 24 consecutive slices as input, predictions are translated to a patient output	AUC =0.84/0.85	Explainability using Grad-CAMs
Huang *et al.* [12]	PENet with the integration of clinical data from the electronic medical record	AUC =0.95 / -	No external validation set from another institution
Weikert *et al.* [35]	3D CNN (ResNet-based)	F1 = 0.86 / -	No external validation
Pan [31]	2D CNN (ResNeSt-50) + sequence modeling for patient output	AUC = 0.90 / -	No external validation, no explainability, computationally too intensive for real implementation
Tajbakhsh *et al.* [36]	Vessel-oriented image representation (VOIR)	AUC = 0.90 / -	No external validation

Compared to other methods present in the literature, our approach presents several advantages. The upfront selection of the slices containing lungs for both positive and negative cases assures that the model is trained only on the organ of interest, eliminating possible confounding structures also present in the scan (e.g. the abdominal region). Moreover, the provenance of the patient cohort (multinational and multicentric) assures the robustness of the model, which is trained on relatively heterogeneous scans coming from different centers and acquired with different imaging parameters. In addition, the datasets contributing to the external validation set are all publicly available implying the easy benchmarking of future or current methods. Regarding the explainability of the model, we performed an additional validation step on the use of the Grad-CAM methodology, comparing the activation maps with the real manual delineation of the emboli by a human reader, achieving a high level of correspondence (79.2%). This is another indirect proof of the proficiency of the model in mimicking the approach used by radiologists in identifying the PE areas in the CTPA scans. Also, the threshold curve analysis on the patients in the internal validation set assures that the choice of the threshold to distinguish between PE positive and negative patients optimizes both sensitivity and specificity of the model and limits over or underdiagnosis. We also proved that this threshold is robust on new, unseen datasets as we reached comparable performance on the external validation set.

To get a final probability for each patient, the model probabilities on all lung slices are averaged. As PE events are rather rare and the emboli can be small, presenting only on a limited number of slices in the total volume, the final probability will be reduced and a threshold lower than 50% should be considered to diagnose a patient as PE positive. More specifically, the threshold curve analysis suggested a predicted risk of 22.58% to classify a patient as PE positive. For positive cases, the predicted risk is thus not aligned with the observed risk as the observed risk is higher. The final method at the patient-level is thus not well calibrated, caused by the nature of the approach, the final probability should not be interpreted as a scaled uncertainty indication. Referring to the slice-level results, it’s shown that the actual DL classifier, predicting the probability per slice, is well calibrated.

This approach presents several limitations that need to be properly addressed. Considering that the published overall rate of a positive diagnosis of PE on pulmonary CTPA usually ranges between 12% and 22% [37], [38] the patient cohort used for external validation might not represent accurately the reality as it presents a higher percentage of positive cases (127 out of 155). Also, the model performance in patients with concurrent pulmonary diseases or non-thrombotic emboli was not explored as well as the differentiation between chronic and acute PE cases. Finally, the ability to localize the emboli using the activation map extraction method is limited by the size of the activated areas. The limited precision of the activation maps cannot rule out that the model is spotting suspicious artifacts in the neighborhood of the embolus.

While patient-level results are obtained by averaging the slice-level predictions, a more refined approach could have been explored by directly training a 3D CNN model on the slice-level activation maps extracted from the 2D CNN to predict patient-level probabilities for the presence of PE. However, this approach will need an additional set of training data to train the 3D model, to prevent the bias from using the same dataset used to train the 2D model.

Further improvement of the presented methodology could include other PE severity assessment methods such as the Geneva score [39] which, combined with the DL model results, might yield more trustworthy predictions, with higher sensitivity, promoting at the same time the confidence of the clinicians towards the results.

5 CONCLUSIONS

Our deep learning-based classifier can identify patients with pulmonary embolism with high accuracy and can localize the emboli by extracting the activation maps from the network. The activation maps help explain the features used by the deep learning model to make the diagnosis, increasing the likelihood of acceptance by clinicians for clinical routine. Further prospective validation is required before the algorithm can be used in the clinic.

6 REFERENCES

[1] M. K. A., M. Rebecca, C. M. J., D. K. R., S. D. R., and K. S. S., “Time Trends in Pulmonary Embolism Mortality Rates in the United States, 1999 to 2018,” *Journal of the American Heart Association*, vol. 9, no. 17, p. e016784, Sep. 2020, doi: 10.1161/JAHA.120.016784.

[2] S. Barco *et al.*, “Trends in mortality related to pulmonary embolism in the European Region, 2000-15: analysis of vital registration data from the WHO Mortality Database,” *The Lancet Respiratory Medicine*, vol. 8, no. 3, pp. 277–287, Mar. 2020, doi: 10.1016/S2213-2600(19)30354-6.

[3] C. Zhou *et al.*, "Variabilities in Reference Standard by Radiologists and Performance Assessment in Detection of Pulmonary Embolism in CT Pulmonary Angiography," *Journal of Digital Imaging*, vol. 32, no. 6, pp. 1089–1096, 2019, doi: 10.1007/s10278-019-00228-w.

[4] L. Salehi, P. Phalpher, M. Ossip, C. Meaney, R. Valani, and M. Mercuri, "Variability in practice patterns among emergency physicians in the evaluation of patients with a suspected diagnosis of pulmonary embolism.," *Emergency Radiology*, vol. 27, no. 2, pp. 127–134, 2020, doi: 10.1007/s10140-019-01740-w.

[5] H.-P. Chan, L. Hadjiiski, C. Zhou, and B. Sahiner, "Computer-Aided Diagnosis of Lung Cancer and Pulmonary Embolism in Computed Tomography - A Review," *Academic Radiology*, vol. 15, no. 5, pp. 535–555, May 2008, doi: 10.1016/j.acra.2008.01.014.

[6] M. Remy-Jardin *et al.*, "Machine Learning and Deep Neural Network Applications in the Thorax: Pulmonary Embolism, Chronic Thromboembolic Pulmonary Hypertension, Aorta, and Chronic Obstructive Pulmonary Disease," *Journal of Thoracic Imaging*, vol. 35, no. May, pp. S40–S48, 2020, doi: 10.1097/RTI.0000000000000492.

[7] D. Tourassi, R. Edward, and E. Floyd, "Thoracic Radiology Acute Neural Pulmonary Network Embolism : Approach Artificial for Diagnosis '," *Thoracic Radiology*, vol. 189, pp. 555–558, 1993.

[8] H. Holst *et al.*, "Automated interpretation of ventilation-perfusion lung scintigrams for the diagnosis of pulmonary embolism using artificial neural networks," *European Journal of Nuclear Medicine*, vol. 27, no. 4, pp. 400–406, 2000, doi: 10.1007/s002590050522.

[9] G. Serpen, D. K. Tekkedil, and M. Orra, "A knowledge-based artificial neural network classifier for pulmonary embolism diagnosis," *Computers in Biology and Medicine*, vol. 38, no. 2, pp. 204–220, 2008, doi: https://doi.org/10.1016/j.compbiomed.2007.10.001.

[10] F. Zhang *et al.*, "A deep-learning-based prognostic nomogram integrating microscopic digital pathology and macroscopic magnetic resonance images in nasopharyngeal carcinoma: a multi-cohort study.," *Therapeutic advances in medical oncology*, vol. 12, p. 1758835920971416, 2020, doi: 10.1177/1758835920971416.

[11] X. Yang *et al.*, "A Two-Stage Convolutional Neural Network for Pulmonary Embolism Detection From CTPA Images," *IEEE Access*, vol. 7, pp. 84849–84857, 2019, doi: 10.1109/ACCESS.2019.2925210.

[12] S. C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020, doi: 10.1038/s41598-020-78888-w.

[13] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging* , vol. 6, no. 6. 2020. doi: 10.3390/jimaging6060052.

[14] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery* , vol. 9, no. 4, p. e1312, Jul. 2019, doi: https://doi.org/10.1002/widm.1312.

[15] E. Colak *et al.*, “The RSNA Pulmonary Embolism CT Dataset,” *Radiology: Artificial Intelligence*, vol. 3, no. 2, p. e200254, Jan. 2021, doi: 10.1148/ryai.2021200254.

[16] M. Masoudi, H.-R. Pourreza, M. Saadatmand-Tarzjan, N. Eftekhari, F. S. Zargar, and M. P. Rad, “A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism,” *Scientific Data*, vol. 5, no. 1, p. 180180, 2018, doi: 10.1038/sdata.2018.180.

[17] G. G. Serrano, “CAD-PE,” *IEEE Dataport*, 2019, doi: https://dx.doi.org/10.21227/9bw7-6823.

[18] T. Sjöblom, N. Sladoje, and T. F. Ali Teymur Kahraman, Dimitris Toumpanakis, “Computed Tomography Pulmonary Angiography (CTPA) Data.” 2019. doi: 10.23698/aida/ctpa.

[19] D. Jimenez-Carretero *et al.*, “Computer Aided Detection for Pulmonary Embolism Challenge (CAD-PE),” *arXiv*, no. August, 2019.

[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 5987–5995. doi: 10.1109/CVPR.2017.634.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-January, pp. 5987–5995. doi: 10.1109/CVPR.2017.634.

[23] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement,” *BMC Medicine*, vol. 13, no. 1, p. 1, 2015, doi: 10.1186/s12916-014-0241-z.

[24] J. Mongan, L. Moy, and C. E. Kahn, “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers,” *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e200029, Mar. 2020, doi: 10.1148/ryai.2020200029.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[26] A. K. Tarbox and M. Swaroop, “Pulmonary embolism.,” *International journal of critical illness and injury science*, vol. 3, no. 1, pp. 69–72, Jan. 2013, doi: 10.4103/2229-5151.109427.

[27] N. J. Giordano, P. S. Jansson, M. N. Young, K. A. Hagan, and C. Kabrhel, “Epidemiology, Pathophysiology, Stratification, and Natural History of Pulmonary Embolism.,” *Techniques in vascular and interventional radiology*, vol. 20, no. 3, pp. 135–140, Sep. 2017, doi: 10.1053/j.tvir.2017.07.002.

[28] J. L. Alonso-Martínez, F. J. A. Sánchez, and M. A. U. Echezarreta, “Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism.,” *European journal of internal medicine*, vol. 21, no. 4, pp. 278–282, Aug. 2010, doi: 10.1016/j.ejim.2010.04.005.

[29] J. M. T. Hendriksen *et al.*, “Clinical characteristics associated with diagnostic delay of pulmonary embolism in primary care: a retrospective observational study.,” *BMJ open*, vol. 7, no. 3, p. e012789, Mar. 2017, doi: 10.1136/bmjopen-2016-012789.

[30] S. C. Huang *et al.*, “PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging,” *npj Digital Medicine*, vol. 3, no. 1, pp. 1–9, 2020, doi: 10.1038/s41746-020-0266-y.

[31] I. Pan, “Deep Learning for Pulmonary Embolism Detection: Tackling the RSNA 2020 AI Challenge,” *Radiology. Artificial intelligence*, vol. 3, no. 5, pp. e210068–e210068, Jun. 2021, doi: 10.1148/ryai.2021210068.

[32] L. Schmuelling *et al.*, “Deep learning-based automated detection of pulmonary embolism on CT pulmonary angiograms: No significant effects on report communication times and patient turnaround in the emergency department nine months after technical implementation,” *European Journal of Radiology*, vol. 141, Aug. 2021, doi: 10.1016/j.ejrad.2021.109816.

[33] W. Liu *et al.*, “Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning,” *European Radiology*, vol. 30, no. 6, pp. 3567–3575, 2020, doi: 10.1007/s00330-020-06699-8.

[34] S.-C. Huang *et al.*, “PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging,” *npj Digital Medicine*, vol. 3, no. 1, p. 61, 2020, doi: 10.1038/s41746-020-0266-y.

[35] T. Weikert *et al.*, “Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm.,” *European radiology*, vol. 30, no. 12, pp. 6545–6553, Dec. 2020, doi: 10.1007/s00330-020-06998-0.

[36] N. Tajbakhsh, J. Y. Shin, M. B. Gotway, and J. Liang, “Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation,” *Medical Image Analysis*, vol. 58, p. 101541, 2019, doi: https://doi.org/10.1016/j.media.2019.101541.

[37] A. A. Donato, S. Khoche, J. Santora, and B. Wagner, “Clinical outcomes in patients with isolated subsegmental pulmonary emboli diagnosed by multidetector CT pulmonary angiography,” *Thrombosis Research*, vol. 126, no. 4, pp. e266–e270, 2010, doi: https://doi.org/10.1016/j.thromres.2010.07.001.

[38] G. C. Hui, A. Legasto, and C. Wittram, “The Prevalence of Symptomatic and Coincidental Pulmonary Embolism on Computed Tomography,” *Journal of Computer Assisted Tomography*, vol. 32, no. 5, 2008.

[39] G. Le Gal, M. Righini, R. Pierre-Marie, O. Sanchez, D. Aujesky, and H. Bounameaux, “Prediction of Pulmonary Embolism in the Emergency Department,” *Annals of Internal Medicine*, vol. 144, no. 3, pp. 165–171, 2006.

Chapter 5

Prospective validation of a prognostic radiomics signature for patients with non-metastatic NSCLC treated with standard of care non-surgical therapy

Based on Ralph T.H. Leijenaar, **Akshayaa Vaidyanathan**, Dirk De Ruysscher, Andre Dekker, Lizza E.L. Hendriks, Petros Kalendralis, Flore Belmans, Martin Gueuning, Fabio Bottari, Mariaelena Occhipinti, Julien Guiot, Pierre Lovinfosse, Wim Vos, Philippe Lambin, & Sean Walsh, Prospective validation of a prognostic radiomics signature for patients with non-metastatic NSCLC treated with standard of care non-surgical therapy, under review

In this chapter, we provide clinical evidence of level 1b by prospectively validating the original prognostic radiomics signature for chemoradiotherapy in patients with non-metastatic non-small cell lung cancer (NSCLC) and evaluate the potential of this radiomics signature to complement current prognostic factors. A total of 228 patients with inoperable stage I-III NSCLC treated with chemoradiotherapy were randomly sampled from the observational SDC-lung clinical trial (NCT01855191). The gross tumor volumes were used as input to the radiomics signature. Segmentations were performed both manually and automatically using a deep learning model. The primary outcome was overall survival. The signature was used to classify patients as survivors or non-survivors (high/low prognostic score based on the coefficients proposed by Aerts et al., 2014). Predefined statistical tests were performed to prospectively validate the performance of the published signature without any recalibration. The prognostic value of the signature was compared with TNM staging and the gross tumor volume. Discrimination in the model was assessed by Harrell's concordance index (c-index = 0·66 (95% CI: 0·60-0·71). Kaplan-Meier survival curves between patients classified by the radiomics signature as survivors/non-survivors were significantly different (log-rank test p-value = $3{\cdot}670e^{-6}$). The calibration slope (β) on the linear predictor of the signature in a Cox proportional hazards model was 1·404 (H0: β = 1, p = 0·146), indicating a valid relative risk model. The prognostic performance of both Signature-0 and volume features are sensibly superior to the overall stage (6th ed TNM) (p < 0.01). To the best of our knowledge, this study demonstrates clinical evidence level 1b for a prognostic radiomics signature for NSCLC patients. This has implications for the wider field as it demonstrates that other signatures could also be prospectively validated. This signature could be practically used as a clinical decision support tool to evaluate the likelihood of survival after chemoradiotherapy. Potential applications of this signature include use as a stratification tool in future trials or for better therapy planning.

1 BACKGROUND

Lung cancer is the most common cancer worldwide (excluding non-melanoma skin cancer), with over 2 million newly diagnosed cases annually and almost 1.9 million deaths.[1,2] Approximately 80-85% of patients with lung cancer are identified as non-small cell lung cancer (NSCLC).[3] Therapeutic approaches (surgery, radiotherapy, chemotherapy, targeted agents, immunotherapy, or a combination of these modalities) depend upon staging and risk assessment.[4] Clinical decisions regarding treatment regimens are supported by guidelines [5–7] based on the best available evidence. However, it is clear that artificial intelligence (AI) will play an important role here in the future.[8–10] In oncology, computational imaging utilizing AI has produced remarkable results in recent times.[11–13] This rapidly maturing field is known as 'radiomics' which stands for quantitative image analysis with two subfamilies: handcrafted radiomics and deep learning based radiomics sometimes called deep radiomics. Radiomics enables actionable insights to be obtained and applied within clinical-decision support systems to improve diagnostic, prognostic, and predictive accuracy, to further push forward personalized medicine [14]. Radiomic signatures (i.e., quantitative image biomarkers linked to a biological or clinical endpoint) have shown their importance for numerous tumor types.[15–17] Signature-0 is the original radiomic signature,[18] published in 2014. Signature-0 identifies a general prognostic phenotype of patients with NSCLC captured from routine computed tomography (CT) scans (i.e., favorable or unfavorable to reach 2-year overall survival). Signature-0 has been already externally validated [19,20] and its success can be attributed to a spatially and temporally robust methodology used in model development (e.g., multiple segmentations, test/re-test scans, and appropriate machine learning techniques) [21–23].

However, Timmeren et al. used the signature to test the interchangeability of CT and cone-beam CT for the extraction of radiomics features[24] while de Jong et al. tested the Signature-0 in stage IV NSCLC patients. Thus far no evidence level 1b [25] of the original prognostic Signature-0 has been reported.

For these reasons, here we present a prospective validation of Signature-0 in the observational clinical study SDC-lung (NCT01855191) [26]. We hypothesized that the prognostic value of Signature-0 would validate in a prospective cohort using a manual and an automated segmentation method. Furthermore, as TNM staging [27] is a well-established prognostic factor, we compared this with Signature-0 and with tumor volume alone, which has been demonstrated to be a promising prognostic factor for survival prediction [28–30]. The use of Signature-0 can be further extended by stratifying the patients in several different prognostic groups, to better assess the likelihood of survival after chemoradiotherapy and inform clinical decision-making on the best therapy planning options.

2 MATERIAL AND METHODS

2.1 CLINICAL TRIAL REGISTRATION

An observational standardized data collection of lung cancer patients was registered on clinicaltrials.gov (SDC-lung: NCT01855191) to improve the performance of prediction models. The study was approved by the institutional review board and informed consent was obtained from all patients before treatment. A subgroup of 228 patients with stage I-III NSCLC was extracted from the whole SDC-lung cohort, to validate the performance of Signature-0 for chemoradiotherapy survival prediction [31].

2.2 PATIENT POPULATION

From June 2013 until June 2018, patients eligible for stereotactic radiotherapy (Stage I), sequential or concurrent chemoradiotherapy (Stage II-III) were entered in this prospective study conducted at MAASTRO clinic (Maastricht, The Netherlands), where the radiotherapy was delivered, [32] chemotherapy was administered in the referring hospitals. Included were patients with Stage I-III (6th TNM edition [33]), histological or cytological confirmed NSCLC, no prior thoracic radiation, and a work-up according to national guidelines, including a staging whole-body FDG-PET-CT scan and an MRI or contrast-enhanced CT scan of the brain. A WHO Performance Status (WHO PS) of 0 to 2 was required. All patients had to have moderate to good lung function (FEV1 ≥30% and DLCO ≥30% of predicted value). The presence of supraclavicular lymph nodes, pleural fluid that was negative for malignancy on cytological examination, and cardiac comorbidities including arrhythmia or a decreased ejection fraction were no exclusion criteria. Patients with other invasive cancers within the last five years were also allowed provided they were in clinical complete remission at the time of enrolment.

2.3 TREATMENT

Chemotherapy was given in the referring hospital. It consisted of 1 cycle of cisplatin or carboplatin–gemcitabine (cisplatin 75 mg/m2, carboplatin AUC 5, gemcitabine 1250 mg/m2), followed by concurrent cisplatin–vinorelbine (cisplatin 40–50 mg/m2, vinorelbine 15–20 mg/m2) or concurrent cisplatin–etoposide every 3 weeks for 3 cycles (cisplatin 75–80 mg/m2 day 1 or carboplatin AUC 5 depending on the cardiovascular history or limited renal function, etoposide 100 mg/m2 day 1-3) with radiotherapy. The regimen depended on the referring hospital. Dose-reduction was applied according to guidelines and in case of decreasing renal function, cisplatin was substituted by carboplatin. Radiation treatment planning was performed during the first cycle of chemotherapy and radiotherapy was intended to start on the first day of the second cycle of chemotherapy, according to Dutch guidelines.

2.4 PROSPECTIVE VALIDATION

To validate the performance of Signature-0 for survival after chemoradiotherapy, we conducted two different analyses, using a randomly selected patient cohort from SDC-lung dataset. The first case was the validation of the original Signature-0 with CT scans manually segmented, of the tumour and the lymph nodes if involved, by the treating radiation oncologist and verified by a second one, as per clinical protocol. The second case was the validation of the original Signature-0 on the same datasets, but automatically segmented by a deep learning algorithm. A flow chart describing the overall workflow from data collection to model validation is shown in Figure 5.1.

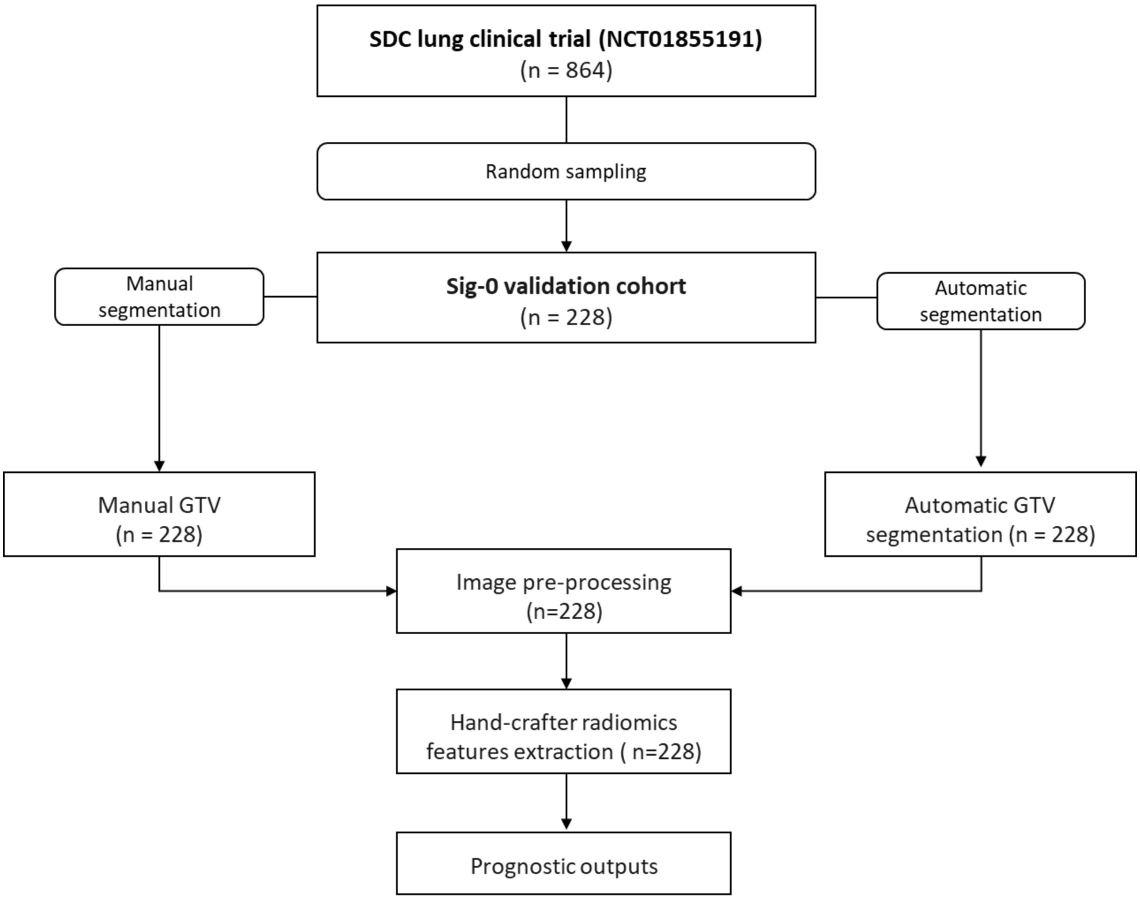


Figure 5.1: Scheme of the workflow used in this study

A validation cohort composed of 228 cases was collected by randomly sampling patients from the SDC-lung clinical trial cohort (n = 864). The only selection criteria used were the presence of a chest CT scan, acquired following the image acquisition protocol, and completeness of the TNM staging. Before radiomics feature extraction, the images were pre-processed to optimize feature extraction [34]. The pre-processed images with the segmentation masks were used for hand-crafted radiomics feature extraction. Finally, radiomics features were used in a multivariate Cox proportional hazards regression model to compute the prognostic score linked to the probability of 2-year survival. To compare the prospective validation with the original external validation performed on Signature-0 [18] the prognostic performances of the signature for 2-year survival were compared with the one obtained with TNM staging. Additionally, a comparison of the volume feature alone was performed, to verify the conclusion reached in the original paper, which reported good performances for volume alone on all the investigated datasets. The prognostic power of TNM, tumor volume, and Signature-0 were explored by Kaplan– Meier survival analysis, on both the manual and automatic segmentations. Additional Kaplan– Meier survival analyses were performed, identifying four different prognostic groups based on different thresholds to expand the clinical usability of Signature-0. The radiomics signature validation was evaluated following the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [35] and the Radiomics Quality Score (RQS) [36].

2.5 IMAGE ACQUISITION

All CT images used in the study were acquired on multidetector CT scanners available at the sites. Since CT images were collected prospectively, a standardized scan protocol was available over the complete dataset. To prevent excessive variability in the imaging used for model validation the following criteria for radiomic analysis were used: tumor completely visible in the scan, slice increment less than 1·5 mm, and no missing slices. The full overview of the different parameters used for image acquisition and reconstruction in the validation dataset is reported in Table 5.1.

Table 5.1 overview of the imaging parameters

	Distribution (%)
Manufacturer	
SIEMENS	92.5
Varian Medical System	7.5
Kernel	
B30f	47
B19f	15
B19s	4.8
B41f	2.6
Others	30.6
Pixel spacing (mm)	
0.976	98.7
1.3	1.3
Slice thickness (mm)	
3	90
2	10

2.6 MANUAL SEGMENTATION

The manual segmentations of the gross tumor volume (GTV), performed for therapy planning purposes at the different medical centers as per standard of care, were used for the extraction of radiomics features and the application of Signature-0.

2.7 AUTOMATIC DEEP LEARNING SEGMENTATION

The GTV was segmented using RadiomiX (Radiomics SA, Liège, Belgium) based on convolutional neural networks by combining 3D and 2D architectures. Publicly available data from the Cancer Imaging Archive [37] was used to train and validate the model. The specific dataset (Lung1) [38] contains CT scans of 422 confirmed non-small cell lung cancer cases, along with manual segmentations of the primary lesion and involved lymph nodes. The segmentations were performed by an experienced radiologist and these segmentations were used as a reference standard. The data was randomly partitioned into a training set (n = 337) and a validation set (n = 85). The model was then externally tested in the SDC-Lung dataset (n = 220). Details on the 2D and 3D architectures are reported below:

2.7.1 2D Architecture

The model is based on Feature Pyramid Network [39] with ResNext blocks [40] in the encoder part of the network and was trained on 2D axial slices as input. Both positive slices (containing tumor) and negative slices (not containing tumor) were used to train the model.

2.7.2 3D Architecture

A 3D U-Net with residual connections in the encoder and decoder part of the network was trained on 3D volumes containing 16 consecutive axial slices with at least one slice containing a portion of the primary tumor.

The predicted segmentations of each architecture (i.e., the segmentation output from both the 3D and the 2D segmentation models) were ensembled and the intersection constitutes the final total GTV segmentation which is used for the extraction of radiomics features. The deep learning-based GTV segmentation achieved a mean Dice similarity coefficient score of 0.82 on the external testing set which indicates adequate precision (i.e. no significant over or under segmentation).

2.8 STATISTICAL ANALYSIS

To assess the performance of Signature-0, pre-specified statistical tests were performed. A log-rank test indicates a significant split. Discrimination by the signature was assessed by calculating Harrell's concordance metric. Cox regression was performed on the signature to determine the calibration slope and a likelihood ratio test indicates the relative risk. Additionally, the coefficients of the individual variables of the signature were jointly tested to indicate that the performance in the prospective validation cohort could be improved by adjusting the original coefficients of the features. The linear predictors of the prospective validation dataset were determined. Linear predictors are defined as $\Sigma ix i\beta i$, which is the sum of the model's variables x multiplied by the regression coefficients β. To determine the calibration slope, Cox regression was performed, and the unit value of the slope was tested through a log-rank test. Afterward, a joint log-rank test on all the predictors plus the offset was performed and tested for non-significance, which would indicate a good fit for our model [41]. To evaluate the clinical utility of Signature-0 predictions, compared to volume and TNM staging, decision curve analysis [42,43] was performed by quantifying the net benefits for a range of threshold probabilities in the whole validation dataset. All statistical analyses were performed in R (3.2).

3 RESULTS

3.1 DEEP LEARNING AUTOMATIC SEGMENTATION MODEL

The deep learning segmentation achieved a mean DICE similarity coefficient score of 0·82 for lung tumor volume across the SDC-lung dataset, which indicates adequate precision and accordance with expert delineation (i.e., no significant over or under-segmentation). Figure 5.2 shows an example of segmentations for two patients, comparing manual and automatic segmentation.

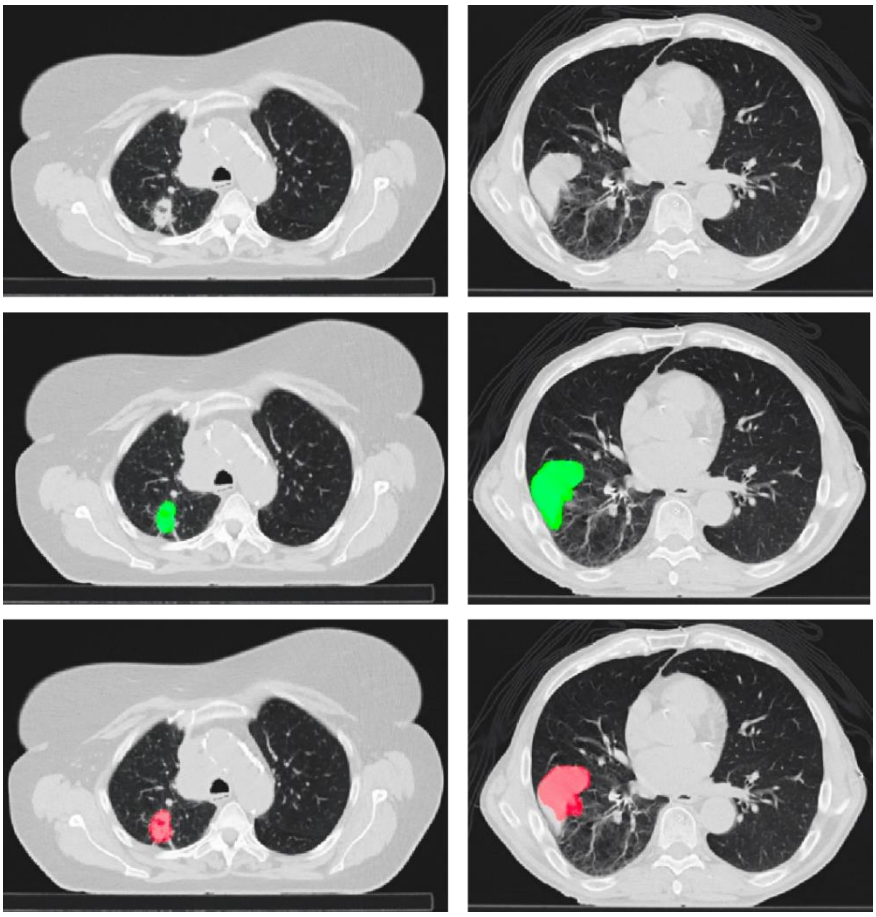
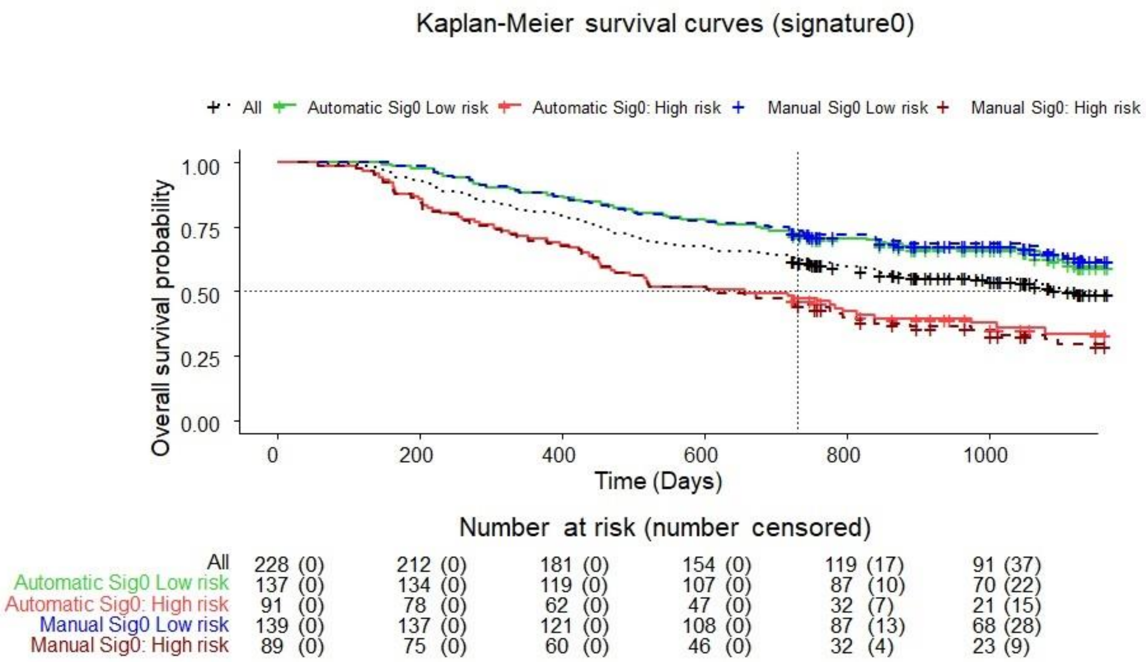


Figure 5.2: Lung Gross tumor volume (GTV) segmentation. Example on CT scan slice from two different patients: top row) original CT scan; center row) manual segmentation of the GTV (in green); bottom row) automatic segmentation of the GTV (in red). DICE score for the automatic segmentation of 0.78 (left) and 0.79 (right).

3.2 PROGNOSTIC PERFORMANCES

3.2.1 Signature-0 prospective validation

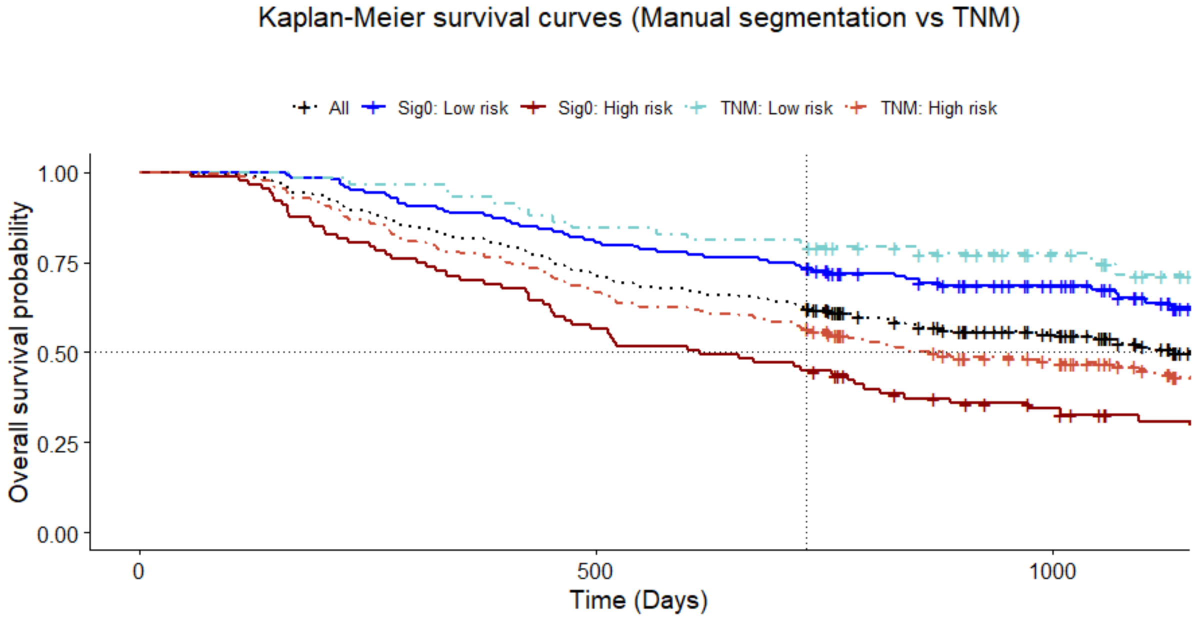
The KM curves analysis for the pure prospective validation of Signature-0 visualized a clear split between groups classified as survivors/non-survivors (high/low prognostic score based on a median prediction threshold of the original signature [18]) (Figure 5.3). A log-rank test indicates a significant split with a *p*-value of $3.670e^{-6}$ for the manual segmentation compared to $9.005e^{-5}$ for the automatic segmentation. Discrimination by the signature was assessed by calculating Harrell's concordance metric. The performance of the two segmentation methods was comparable with a c-index of 0.66 (95% CI: 0.60-0.71) for the manual segmentation and of 0.63 (95% CI: 0.54-0.69) for the automatic one. Cox regression was performed on the signature to determine the calibration slope and a likelihood ratio test indicates a valid relative risk model as the slope is close to 1 and not significantly different from 1 (Table 5.1).



	0	200	400	600	800	1000
All	228 (0)	212 (0)	181 (0)	154 (0)	119 (17)	91 (37)
Automatic Sig0 Low risk	137 (0)	134 (0)	119 (0)	107 (0)	87 (10)	70 (22)
Automatic Sig0: High risk	91 (0)	78 (0)	62 (0)	47 (0)	32 (7)	21 (15)
Manual Sig0 Low risk	139 (0)	137 (0)	121 (0)	108 (0)	87 (13)	68 (28)
Manual Sig0: High risk	89 (0)	75 (0)	60 (0)	46 (0)	32 (4)	23 (9)

Figure 5.3 Kaplan-Meier survival curves stratified into low- and high-risk groups for Signature-0 with manual segmentation (blue and brown lines) and automatic segmentation (green and red lines) on the SDC-lung validation dataset (n = 228)

Following the rationale of the original research, the prognostic performances of the signature on the SDC-lung validation dataset were compared to the overall tumor stage results, obtained with TNM staging (6th edition). The c-index of the overall stage is comparable with Signature-0 (0.65, 95% CI: 0.56-0.74): however, Kaplan-Meier curves stratification is sensibly different (Figure. 5.4).



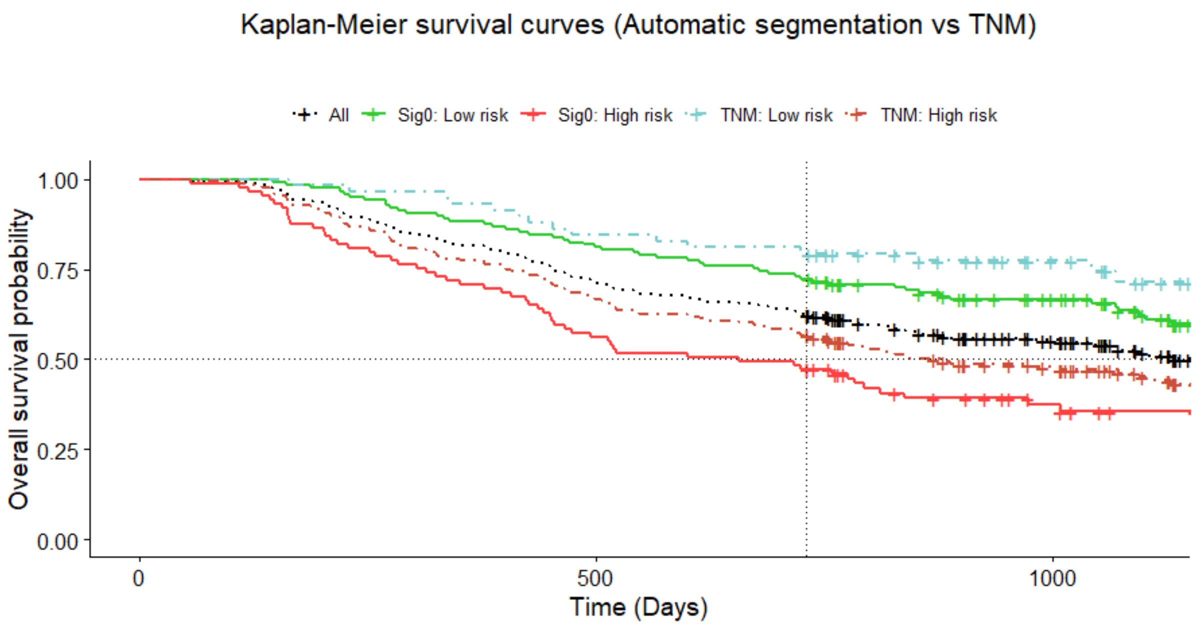
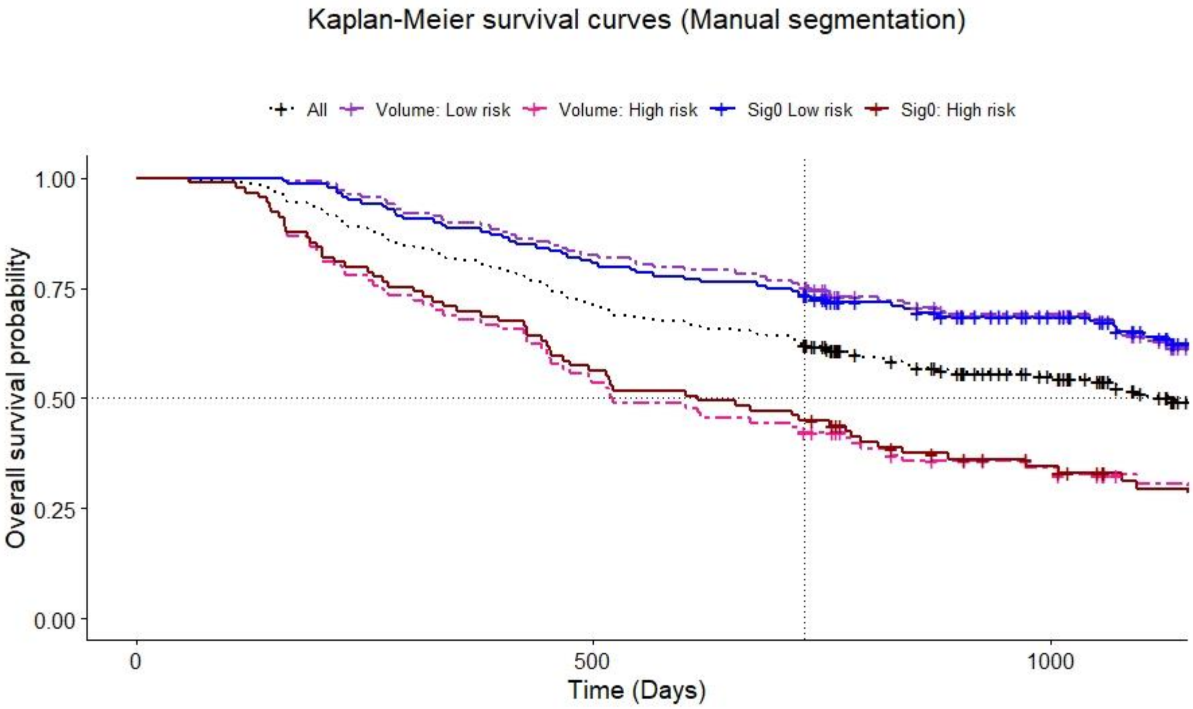


Figure 5.4 Kaplan-Meier survival curves stratified into low- and high-risk groups for Signature-0 vs Overall TNM stage (manual segmentation) (top) and Signature-0 vs overall TNM stage (automatic segmentation) (bottom) on the SDC-lung dataset (n = 228)

Previous reports showed the prognostic power of tumor volume as an independent variable [44,45]: for this reason, we compared the patients' stratification on Kaplan-Meier curves for the volume feature alone vs Signature-0. The comparison was done for both manual and automatic segmentation. The c-index for stratification of volume feature on manual segmentation is 0.67 (95% CI: 0.62-0.73) and 0.65 (95% CI: 0.59-0.71) for automatic segmentation (Figure. 5.5). The results obtained with volume feature and Signature-0, for both manual and automatic segmentation, are not statistically different (confidence interval of C-index falls in the same range – see Table 5.2).

Table 5.2 Prognostic performances on the SDC-lung dataset

	Signature-0 ManualSeg		**Signature-0 AutomaticSeg**		**Volume ManualSeg**		**Volume AutomaticSeg**		**Overall stage**	
C-index	0.66 (0.61 – 0.71)		0.63 (0.57 – 0.69)		0.67 (0.62 – 0.73)		0.65 (0.59 – 0.71)		0.65 (0·56 – 0·74)	
P-value (long-rank)	$3.7e^{-6}$		$9.005e^{-5}$		$1.23\ e^{-6}$		$1.67\ e^{-4}$		$3.65\ e^{-3}$	
Calibration	*Slope*	*p-value*	*Slope*	*p-value*	*Slope*	*p-value*	*Slope*	*p-value*	*Slope*	*p-value*
	1.40	0.15	1.04	0.87	1.07	0.65	0.91	0.61	3.65	0.04
Joined test of model coefficients	0.003		0.003		0.649		0.612		0.055	



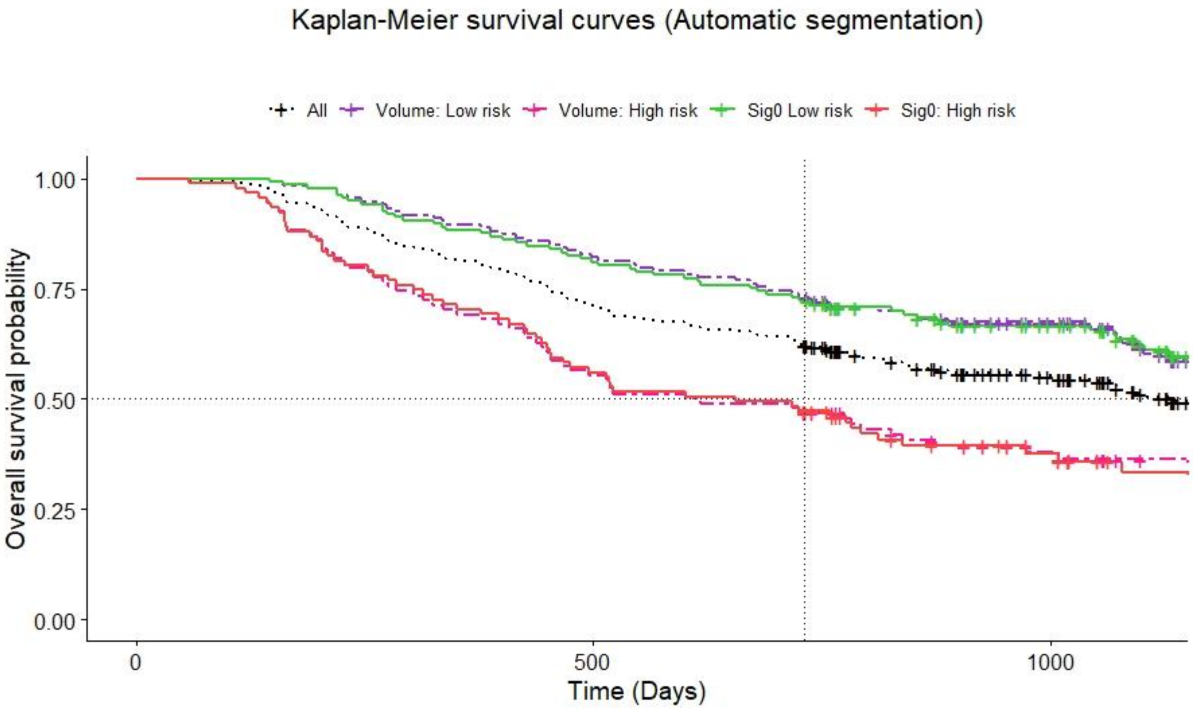
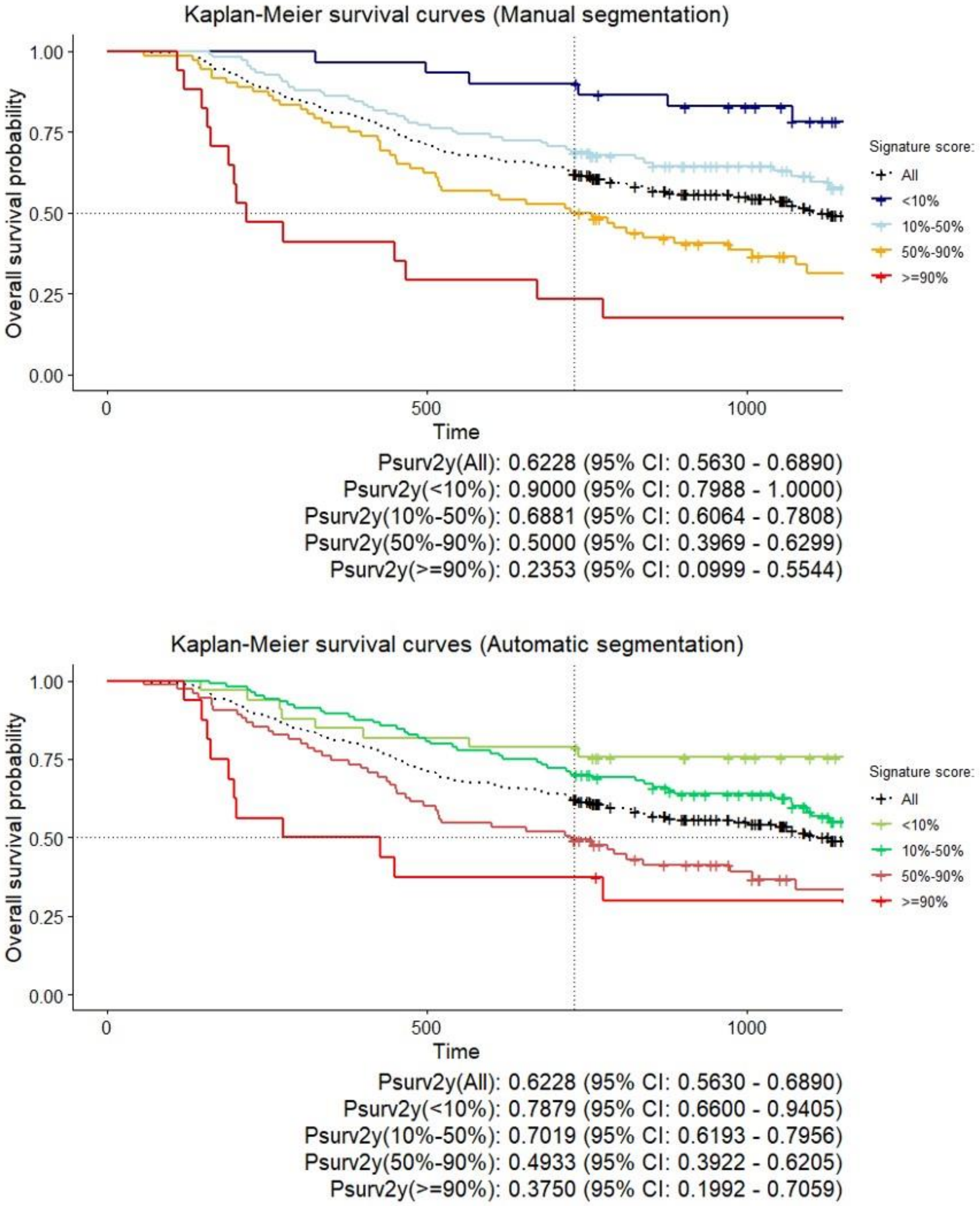


Figure 5.5 Kaplan-Meier survival curves stratified into low- and high-risk groups for Signature-0 vs volume feature (manual segmentation) (top) and Signature-0 vs volume feature (automatic segmentation) (bottom) on SDC-lung dataset (n = 228)

The prognostic performance of both Signature-0 and volume features are sensibly superior to the overall stage (6th ed TNM) ($p < 0.01$). To compare the possible clinical utility of the proposed signature with volume feature and TNM staging, we performed a decision curve analysis (Appendix 5.1 Figure 1), comparing once again manual and automatic segmentation, with no sensible differences found. Also, the time-dependent AUC and calibration curves analysis (Appendix 5.1 Figure 2 and 3) confirmed the superior performance of the radiomics approach over the prognostic power of TNM staging alone.

The radiomics quality score (RQS) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) score are both 58%.

To improve and expand the applicability of the proposed signature for patient staging, four prognostic categories have been defined to provide a more accurate prognosis.



Psurv2y(All): 0.6228 (95% CI: 0.5630 - 0.6890)
Psurv2y(<10%): 0.9000 (95% CI: 0.7988 - 1.0000)
Psurv2y(10%-50%): 0.6881 (95% CI: 0.6064 - 0.7808)
Psurv2y(50%-90%): 0.5000 (95% CI: 0.3969 - 0.6299)
Psurv2y(>=90%): 0.2353 (95% CI: 0.0999 - 0.5544)

Psurv2y(All): 0.6228 (95% CI: 0.5630 - 0.6890)
Psurv2y(<10%): 0.7879 (95% CI: 0.6600 - 0.9405)
Psurv2y(10%-50%): 0.7019 (95% CI: 0.6193 - 0.7956)
Psurv2y(50%-90%): 0.4933 (95% CI: 0.3922 - 0.6205)
Psurv2y(>=90%): 0.3750 (95% CI: 0.1992 - 0.7059)

Figure 5.6. KM curves with four prognostic groups

Three different threshold values for the predictions are applied to stratify patients into four groups with different overall survival. The thresholds are defined as the 10th percentile (T10), the 50th percentile (T50), and the 90th percentile (T90) of prognostic signature scores in the original model development cohort [18]. A predicted signature score ≤ T10 indicates a favorable prognosis. A predicted score > T10 and ≤T50 indicates a likely favorable prognosis. A predicted score > T50 and ≤ T90 indicates a likely unfavorable prognosis. A predicted score > T90 indicates an unfavorable prognosis. Kaplan-Meier survival analysis has been performed for all patients (as reference) and for the four prognostic categories defined by the signature and Kaplan-Meier survival curves are shown in Figure 5.6. Considering, for example, the validation cohort was manually segmented, the overall survival difference across all four prognostic groups was statistically significantly different ($p < 0.0001$). Considering all patients, the probability of surviving 2 years after the start of treatment is 62.2% (i.e., when there is no stratification of patients based on the radiomics signature). Patients predicted by the signature to have a favorable outcome have a probability of 90.0% of surviving 2 years after treatment. For patients predicted by the signature to have a likely favorable, likely unfavorable, or unfavorable outcome, the probability of surviving 2 years after the start of treatment is 68.8%, 50.0%, and 23.5%, respectively.

4 DISCUSSION

The results of this study showed that the original radiomics signature, Signature-0, developed in 2012, outperforms the contemporaneous standard of care (TNM 6th edition) producing superior stratification between survivors and non-survivors. This holds true across time displayed in the Appendix 5.1 Figure 1. Furthermore, the added value of Signature-0 above TNM and, to a lesser extent, volume is addressed by decision curve analysis displayed in the Appendix 5.1 Figure 1 (i.e., what is the proportion of patients that are classified better concerning OS with the use of Signature-0 affecting treatment choice?). In brief, the clinical net benefit of Signature-0 in comparison to TNM and volume is calculated across a range of threshold probabilities, defined as the minimum probability of survival at which reconsideration of treatment choice would be warranted. This study demonstrates a real but limited clinical net benefit (<10%).

From the results reported is evident that volume is a dominant component in the signature even though it is not directly included. However, this does not diminish the validity and significance of our results. The prospective validation of Signature-0 is a step forward on the path to reproducible and translational science [46].

In the meantime, the AI field as well as the clinical understanding of NSCLC prognosis and the clinical tools available have advanced considerably. The original Signature-0 has now improved opportunities, thanks to the novel avenues of research in AI and radiomics. An improved signature should then be benchmarked against the current standard of care, as in the latest edition of the TNM staging (8th), which still presents all the limitations and drawbacks of a subjective and experienced dependent assessment tool, which were true in 2012 for the 6th edition as well.

The prospective validation of a prognostic radiomics signature for survival prediction of NSCLC is the culmination of 8 years of scientific/clinical investigation. Thus, we also present this analysis in light of the proposed regulatory framework of AI-based software as a medical device (dynamic AI) [47]. The FDA is considering a total product lifecycle-based regulatory framework for these technologies that would allow for modifications to be made from real-world learning and adaptation, while still ensuring that the safety and effectiveness of the software as a medical device are maintained. Prognostic factors are very useful to get information about disease evolution and to construct homogeneous groups of patients. They can be used to guide the therapy and identify subgroups of patients where more aggressive therapy is needed.[48] The most reliable and used prognostic factor for NSCLC is cancer staging according to TNM classification [27]. The automatization of prognostic factors evaluation and quantification has been explored in the recent past, applying several computer-based methods for the automatic or semi-automatic staging of patients based on TNM criteria[49] or integrating histopathological, molecular, and clinical data via AI approaches.[50,51] These previous studies however only leveraged in different ways the same prognostic factors while here is presented a novel and additional tool, which can successfully supplement the current clinical practice, regarding therapy decision making. This approach has been recently explored also for oropharyngeal cancer patients, with a prospectively validated radiomics signature for survival prediction in locally advanced Squamous Cell Carcinoma of the Head & Neck [52]. Specifically, in the lung cancer space, several methods to improve the prognostic stratification has been proposed, among these the Lung Cancer Prognostic score (LCPI) was built using stage, histology, mutation status, performance status, weight loss, smoking history, respiratory comorbidity, sex, and age [53]. The discrimination power of this approach is good (c-index equal to or higher than 0.7) and was externally validated with multicentric data. However, the collection of the complete set of clinical covariates needed could be cumbersome and the lack of one or more of this information might hamper the final prognostic results. Relying on a single source of information (i.e., the chest CT scan) could represent a viable alternative when the clinical and laboratory data are not readily available or not complete.

4.1 LIMITATIONS

This study utilizes 228 patients randomly sampled from the 864 patients recruited to the SDC-lung trial, this is due to funding constraints concerning the time and expertise required of clinical and technical professionals to acquire, segment, curate, quality control, transfer, store, process, analyze, and report on the patient data. While not ideal, it is extremely unlikely that this sample does not accurately represent the characteristics of the complete SDC-lung dataset.

This study reports the performance of Signature-0 contextualized by the standard-of-care and compared with the TNM staging system, both have changed substantially since the discovery of Signature-0 in 2014. [54,55] Immunotherapy is now ubiquitous and the TNM staging has progressed from the 6th to the 8th edition. Nevertheless, the ability to accurately and precisely identify patients that are prognostically favorable to survive beyond 2 years of concurrent chemoradiotherapy is of benefit.[32] The current standard of care is adjuvant immunotherapy after this treatment, which can be burdensome for patients, is expensive, and can cause toxicity. Identifying patients that most benefit from immunotherapy with a cost-effective radiomics methodology would be ideal to facilitate the personalization of consolidation therapy [56] (the alternative of circulating tumor DNA is promising but is intrinsically more expensive).

This study uses 2-year overall survival as a reasonable surrogate for long-term outcomes in patients with non-metastatic NSCLC treated with non-surgical therapy (i.e., chemoradiotherapy), with uncensored data for 2 years and censored data for up to 3 years. Due to the marked improvement of systemic treatments (e.g., immunotherapy) many patients are surviving longer with disease and the utility of 2-year overall survival as a reasonable surrogate for long-term outcomes is unclear.

This study uses the radiotherapy planning CT acquired after the first cycle of chemotherapy, which likely impacts the features of the tumor relative to the baseline CT (acquired before the first cycle of chemotherapy) and introduces the following issues/biases concerning staging/treatment: Stage I patients did not receive chemotherapy and so the radiotherapy planning CT is equivalent to the baseline CT (i.e., unaltered GTV); Stage II-III did receive chemotherapy, if sequential only responders were irradiated (i.e., greatly altered GTV), if concurrent CT at planning is taken after one cycle of chemotherapy (i.e., modestly altered GTV).

This study demonstrates the possibility of using AI to segment the GTV, which includes all areas of gross disease, including clinically involved lymph nodes (more and a higher status are associated with worse survival). However, the AI does not distinguish between these components and thus it is not possible to report here the contribution of each segmented component concerning the performance of the signature.

5 CONCLUSIONS

To the best of our knowledge, this study demonstrates the first clinical evidence level 1b for a prognostic radiomics signature in lung cancer (individual inception cohort study with > 80% follow-up; clinical decision rule to split on the radiomics Signature-0 score as low/high likelihood for survival; validated in a single population testing the quality of a specific radiomics signature based on prior evidence). This has implications for the wider field as it demonstrates that other signatures could also be prospectively validated. This potentially fully automatic signature could be used as a clinical decision support tool at the multi-disciplinary-tumor board to evaluate the likelihood of survival of a patient with non-metastatic NSCLC patients treated by chemoradiotherapy. Potential applications of this signature include use as a stratification tool in future trials or for better therapy planning. The signature requires updating and evaluation in the context of immuno-oncology.

6 REFERENCES

[1] Max Roser, Hannah Ritchie. Cancer - Our World in Data. Cancer 2019:1.

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 2021;71:209–49. https://doi.org/https://doi.org/10.3322/caac.21660.

[3] Cancer Research UK. Types of lung cancer | Cancer Research UK. Cancer Research UK 2020. https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types.

[4] Rami-Porta R, Call S, Dooms C, Obiols C, Sánchez M, Travis WD, et al. Lung cancer staging: A concise update. European Respiratory Journal 2018;51. https://doi.org/10.1183/13993003.00190-2018.

[5] ESMO CLINICAL PRACTICE GUIDELINES: LUNG AND CHEST TUMOURS n.d. https://www.esmo.org/guidelines/lung-and-chest-tumours.

[6] Thoracic Cancer | ASCO n.d. https://www.asco.org/research-guidelines/quality-guidelines/guidelines/thoracic-cancer.

[7] Ettinger DS, Wood DE, Aggarwal C, Aisner DL, Akerley W, Bauman JR, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 1.2020. Journal of the National Comprehensive Cancer Network : JNCCN 2019;17:1464–72. https://doi.org/10.6004/jnccn.2019.0059.

[8] Walsh S, de Jong EEC, van Timmeren JE, Ibrahim A, Compter I, Peerlings J, et al. Decision Support Systems in Oncology. JCO Clinical Cancer Informatics 2019;3:1–9. https://doi.org/10.1200/cci.18.00001.

[9] Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nature Reviews Clinical Oncology 2020;17:771–81. https://doi.org/10.1038/s41571-020-0417-8.

[10] Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. Nature Reviews Clinical Oncology 2013;10:27–40. https://doi.org/10.1038/nrclinonc.2012.196.

[11] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nature Reviews Cancer 2018;18:500–10. https://doi.org/10.1038/s41568-018-0016-5.

[12] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 2019;25:954–61. https://doi.org/10.1038/s41591-019-0447-x.

[13] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89–94. https://doi.org/10.1038/s41586-019-1799-6.

[14] Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix A-N, et al. A review in radiomics: Making personalized medicine a reality via routine imaging. Medicinal Research Reviews 2021;n/a. https://doi.org/https://doi.org/10.1002/med.21846.

[15] O’Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nature Reviews Clinical Oncology 2017;14:169–86. https://doi.org/10.1038/nrclinonc.2016.162.

[16] Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiotherapy and Oncology 2018;127:349–60. https://doi.org/10.1016/j.radonc.2018.03.033.

[17] 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. European Journal of Cancer 2012;48:441–6. https://doi.org/10.1016/j.ejca.2011.11.036.

[18] 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. Nature Communications 2014;5:4006. https://doi.org/10.1038/ncomms5006.

[19] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and Oncology 2017;123:363–9. https://doi.org/10.1016/j.radonc.2017.04.016.

[20] de Jong EEC, van Elmpt W, Rizzo S, Colarieti A, Spitaleri G, Leijenaar RTH, et al. Applicability of a prognostic CT-based radiomic signature model trained on stage I-III non-small cell lung cancer in stage IV non-small cell lung cancer. Lung Cancer 2018;124:6–11. https://doi.org/10.1016/j.lungcan.2018.07.023.

[21] JE van T, RTH L, W van E, J W, Z Z, A D, et al. Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016;2:361–5. https://doi.org/10.18383/j.tom.2016.00208.

[22] Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncologica 2018;57:1070–4. https://doi.org/10.1080/0284186X.2018.1445283.

[23] Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Medical Physics 2018;45:3449–59. https://doi.org/10.1002/mp.12967.

[24] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and Oncology 2017;123:363–9. https://doi.org/10.1016/j.radonc.2017.04.016.

[25] OCEBM Levels of Evidence Working Group. Oxford Centre for Evidence-Based Medicine: Levels of Evidence (2009) n.d.

[26] SDC for Lung Cancer Patients Treated With Curative Primary or Postoperative Radiotherapy or Chemoradiation (SDC lung) n.d.

[27] Chansky K, Sculier J-P, Crowley JJ, Giroux D, Van Meerbeeck J, Goldstraw P. The International Association for the Study of Lung Cancer Staging Project: Prognostic Factors and Pathologic TNM Stage in Surgically Managed Non-small Cell Lung Cancer. Journal of Thoracic Oncology 2009;4:792–801. https://doi.org/10.1097/JTO.0b013e3181a7716e.

[28] Koo TR, Moon SH, Lim YJ, Kim JY, Kim Y, Kim TH, et al. The effect of tumor volume and its change on survival in stage III non-small cell lung cancer treated with definitive concurrent chemoradiotherapy. Radiation Oncology (London, England) 2014;9:283. https://doi.org/10.1186/s13014-014-0283-6.

[29] Stinchcombe TE, Morris DE, Moore DT, Bechtel JH, Halle JS, Mears A, et al. Post-chemotherapy gross tumor volume is predictive of survival in patients with stage III non-small cell lung cancer treated with combined modality therapy. Lung Cancer 2006;52:67–74. https://doi.org/10.1016/j.lungcan.2005.11.008.

[30] Dehing-Oberije C, De Ruysscher D, van der Weide H, Hochstenbag M, Bootsma G, Geraedts W, et al. Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than TNM stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. International Journal of Radiation Oncology, Biology, Physics 2008;70:1039–44. https://doi.org/10.1016/j.ijrobp.2007.07.2323.

[31] SDC for Lung Cancer Patients Treated With Curative Primary or Postoperative Radiotherapy or Chemoradiation (SDC lung) n.d.

[32] De Ruysscher D, van Baardwijk A, Wanders R, Hendriks LE, Reymen B, van Empt W, et al. Individualized accelerated isotoxic concurrent chemo-radiotherapy for stage III non-small cell lung cancer: 5-Year results of a prospective study. Radiotherapy and Oncology 2019;135:141–6. https://doi.org/10.1016/j.radonc.2019.03.009.

[33] Frederick L. Greene MD, David L. Page MD, Irvin D. Fleming MD, April G. Fritz, C.T.R. RHIT, Charles M. Balch MD, Daniei G. Haller MD, et al., editors. AJCC Cancer Staging Manual, 6th edition. Springer; 2002.

[34] Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncologica 2017;56:1544–53. https://doi.org/10.1080/0284186X.2017.1351624.

[35] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Medicine 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.

[36] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 2017;14:749–62. https://doi.org/10.1038/nrclinonc.2017.141.

[37] Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. Scientific Data 2017;4:170124. https://doi.org/10.1038/sdata.2017.124.

[38] NSCLC-Radiomics - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki n.d.

[39] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017- Janua, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 936–44. https://doi.org/10.1109/CVPR.2017.106.

[40] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017- Janua, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 5987–95. https://doi.org/10.1109/CVPR.2017.634.

[41] Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA 2017;318:1377–84. https://doi.org/10.1001/jama.2017.12126.

[42] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making : An International Journal of the Society for Medical Decision Making 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.

[43] Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making 2008;8:53. https://doi.org/10.1186/1472-6947-8-53.

[44] Alexander BM, Othus M, Caglar HB, Allen AM. Tumor Volume Is a Prognostic Factor in Non&#x2013;Small-Cell Lung Cancer Treated With Chemoradiotherapy. International Journal of Radiation Oncology, Biology, Physics 2011;79:1381–7. https://doi.org/10.1016/j.ijrobp.2009.12.060.

[45] van Laar M, van Amsterdam WAC, van Lindert ASR, de Jong PA, Verhoeff JJC. Prognostic factors for overall survival of stage III non-small cell lung cancer patients on computed tomography: A systematic review and meta-analysis. Radiotherapy and Oncology 2020;151:152–75. https://doi.org/10.1016/j.radonc.2020.07.030.

[46] Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Scientific Data 2019;6:218. https://doi.org/10.1038/s41597-019-0241-0.

[47] FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. US Food & Drug Administration 2019:1–20.

[48] Ludwig JA, Weinstein JN. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. Nature Reviews Cancer 2005;5:845–56. https://doi.org/10.1038/nrc1739.

[49] Sieswerda MS, Bermejo I, Geleijnse G, Aarts MJ, Lemmens VEPP, De Ruysscher D, et al. Predicting Lung Cancer Survival Using Probabilistic Reclassification of TNM Editions With a Bayesian Network. JCO Clinical Cancer Informatics 2020:436–43. https://doi.org/10.1200/cci.19.00136.

[50] Lai YH, Chen WN, Hsu TC, Lin C, Tsao Y, Wu S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific Reports 2020;10:1–11. https://doi.org/10.1038/s41598-020-61588-w.

[51] Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communications 2016;7:1–10. https://doi.org/10.1038/ncomms12474.

[52] Keek SA, Wesseling FWR, Woodruff HC, van Timmeren JE, Nauta IH, Hoffmann TK, et al. A Prospectively Validated Prognostic Model for Patients with Locally Advanced Squamous Cell Carcinoma of the Head and Neck Based on Radiomics of Computed Tomography Images. Cancers 2021;13. https://doi.org/10.3390/cancers13133271.

[53] Alexander M, Wolfe R, Ball D, Conron M, Stirling RG, Solomon B, et al. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. British Journal of Cancer 2017;117:744–51. https://doi.org/10.1038/bjc.2017.232.

[54] Faivre-Finn C, Vicente D, Kurata T, Planchard D, Paz-Ares L, Vansteenkiste JF, et al. Four-Year Survival With Durvalumab After Chemoradiotherapy in Stage III NSCLC-an Update From the PACIFIC Trial. Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer 2021;16:860–7. https://doi.org/10.1016/j.jtho.2020.12.015.

[55] Bradley JD, Hu C, Komaki RR, Masters GA, Blumenschein GR, Schild SE, et al. Long-Term Results of NRG Oncology RTOG 0617: Standard- Versus High-Dose Chemoradiotherapy With or Without Cetuximab for Unresectable Stage III Non-Small-Cell Lung Cancer. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology 2020;38:706–14. https://doi.org/10.1200/JCO.19.01162.

[56] Moding EJ, Liu Y, Nabet BY, Chabon JJ, Chaudhuri AA, Hui AB, et al. Circulating Tumor DNA Dynamics Predict Benefit from Consolidation Immunotherapy in Locally Advanced Non-Small Cell Lung Cancer. Nature Cancer 2020;1:176–83. https://doi.org/10.1038/s43018-019-0011-0.

Chapter 6

Deep learning based identification of bone scintigraphies containing metastatic bone disease foci

*Based on Abdalla Ibrahim, **Akshayaa Vaidyanathan**, Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Refaee, Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Henry C. Woodruff, Helle D. Zacho, Sean Walsh, Wim Vos, Mariaelena Occhipinti, François-Xavier Hanin, Philippe Lambin, Felix M. Mottaghy, Roland Hustin., Deep learning based identification of bone scintigraphies containing metastatic bone disease foci*

In this chapter, we have investigated a DL algorithm that can classify areas of increased uptake on bone scintigraphy scans, with automated reporting of the body region containing the lesion(s). We collected 2365 BS from three European medical centers. The model was trained and validated on 1203 and 164 BS scans respectively. Furthermore, we evaluated its performance on an external testing set composed of 998 BS scans. We further aimed to enhance the explainability of our developed algorithm, using activation maps. We compared the performance of our algorithm to that of 6 nuclear medicine physicians. The developed DL-based algorithm can detect MBD on BSs, with high specificity and sensitivity (0.80 and 0.82 respectively on the external test set), in a shorter time compared to the nuclear medicine physicians (2.5 minutes for AI and 30 minutes for nuclear medicine physicians to classify 134 BSs), that could be applied to any BS regardless of the patient's gender and history of cancer. Further prospective validation is required before the algorithm can be used in the clinic.

.

1 BACKGROUND

Metastatic bone disease (MBD) is the most common form of metastatic lesions [1,2]. The incidence of bone metastasis varies depending on the cancer type [3], yet around 80% of MBD arise from breast and prostate cancers [4]. MBD, as the name implies, is due to the propensity of these tumors to metastasize to bones, and it results in eventually difficulty treating painful lesions. Henceforth, early diagnosis is necessary for individualized management that could significantly improve a patient's quality of life [5].

MBD is usually detected using radionuclide bone scintigraphy (or bone scans, BS). BS are nuclear medicine images, which are used frequently to evaluate the distribution of active bone formation, related to benign or malignant processes, in addition to physiological processes. BS scans are indicated in a spectrum of clinical scenarios including exploring unexplained symptoms, diagnosing a specific bone disease or trauma, and the metabolic assessment of patients before and during the treatment[6,7]. BS combining whole-body planar images and tomographic acquisition (SPECT – single photon emission computed tomography) on selected body parts are highly sensitive, as they detect metabolic changes earlier than conventional radiologic images, with lower sensitivity to lytic lesions. However, depending on the pattern it may lack the specificity to identify the underlying causes. Therefore, a SPECT/CT that correlates the findings of bone scintigraphy anatomically is often useful and leads to a more specific diagnosis of the changes noted [8], although MRI scans may also be additionally requested to clarify the diagnosis. Hence, a tool to improve the specificity of decisions based on BS, and reduce the need for further imaging is a relevant unmet clinical need.

Deep learning (DL) is a branch of machine learning (ML) and refers to data-driven modeling techniques, which apply the principles of simplified neuron interactions [9]. The application of imaging analysis techniques using artificial neurons in medical imaging started to draw attention decades ago [10], but it only became major research focus recently due to the advancement in computational capacities and imaging techniques [11,12]. 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. 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 [13], and the 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.

Several studies investigated the potential of DL-based algorithms for analyzing bone scintigraphy scans [14–16]. The majority of these studies applied DL algorithms on BS scans of diagnosed (specific) cancer patients, which could limit the learning ability of the DL algorithm to differentiate MBD from other bone diseases. To the best of our knowledge, no study combined both male and female patients, with no-cancer patients included.

In this study, we hypothesize that DL-based algorithms can learn the pattern of metastatic bone disease on bone scintigraphy scans, and differentiate it from other non-metastatic bone diseases. We investigate the potential of a DL-based algorithm to detect MBD on BS not limited to those of cancer patients using weakly-supervised detection based on activation maps obtained using the gradient weighted class activation mapping (Grad-CAM) method [17,18]. By doing so, we aim to develop a generalizable tool that can classify scans containing metastases and detect MBD on BS, regardless of the gender and malignancy status of the patient. Moreover, by extracting activation maps with the Grad-CAM method [19] and superimposing these maps to the original BD scans, we explored the explainability of the deep learning model's predictions. This is very important to promote the application of these methods in the clinic and avoid the common misconception that sees DL models as "black boxes" without any real connection to clinical and imaging characteristics. As a complementary step, we explored the development of an automated label generator for the location of the detected metastatic foci.

2 MATERIALS AND METHODS

2.1 IMAGING DATA

The imaging data were retrospectively collected from different European centers: Aachen RWTH University Clinic (Aachen, Germany), Aalborg University Hospital (Aalborg, Denmark), and Namur University Hospital (Namur, Belgium). The electronic medical records of these hospitals were searched for patients who underwent BS between 2010 and 2018. Patients for whom a definitive classification of the foci was available, mostly through further investigations, were further included. All images were acquired with anteroposterior (AP) and posteroanterior (PA) whole-body views. The imaging analysis was approved by the Aachen RWTH institutional review board (No. EK 260/19), and informed consent was obtained from all included patients. According to Danish National Legislation, the Danish Patient Safety Authority can waive informed consent for retrospective studies (approval 31-1521-110). All methods were carried out following the relevant guidelines and regulations [20]. The study protocol was published on clinicaltrials.gov (NCT: NCT05110430)

2.2 STUDY POPULATION

The imaging data were retrospectively collected from different European centers: Aachen RWTH University Clinic (Aachen, Germany), Aalborg University Hospital (Aalborg, Denmark), and Namur University Hospital (Namur, Belgium). The electronic medical records of these hospitals were searched for patients who underwent BS between 2010 and 2018. Patients for whom a definitive classification of the foci was available, mostly through further investigations, were further included. All images were acquired with anteroposterior (AP) and posteroanterior (PA) whole-body views. The imaging analysis was approved by the Aachen RWTH institutional review board (No. EK 260/19), and informed consent was obtained from all included patients. According to Danish National Legislation, the Danish Patient Safety Authority can waive informed consent for retrospective studies (approval 31-1521-110). All methods were carried out following the relevant guidelines and regulations [20]. The study protocol was published on clinicaltrials.gov (NCT: NCT05110430)

2.3 IMAGE PRE-PROCESSING

Every data point containing acquisition at two views (AP and PA) was resized to size (length = 256, height = 512) and the intensities were normalized to the range [0-1] using the minimum and maximum intensity of each image. For all the data points, image acquisitions at both views are appended beside each other as shown in Figure 6.1.

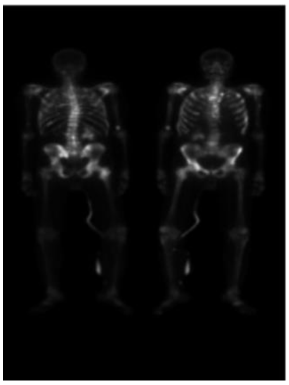


Figure 6.1. Example of pre-processed BS scans used as input for model training

2.4 MODEL ARCHITECTURE, TRAINING, AND TESTING

The training and validation datasets are composed of 1203 and 164 images respectively, coming from Centre A (Aachen) and B (Aalborg). The external test cohort is composed of 998 images collected at center C (Namur). A full overview of the patients' cohort division between the different datasets is reported in Table 6.1.

Table 6.1. Division of the patient's cohort between training, validation, and external test

	Training (n = 1203)	**Validation (n = 164)**	**External test (n = 998)**
Centre A (Achen)	235 with metastasis 668 normal	58 with metastasis 58 normal	-
Centre B (Albourg)	94 with metastasis 206 normal	24 with metastasis 24 normal	-
Centre C (Namur)	-	-	411 with metastasis 587 normal

The model was trained on 329 images containing metastasis from Centre B (94) and A (235). At each epoch, the 874 images without any metastasis were shuffled and 329 images were randomly selected to train the model with balanced labels. VGG16 architecture with ImageNet pre-trained weights [21] was trained with categorical cross-entropy loss for 6 epochs with 200 steps per epoch. The model was trained with 3-channel input. The pre-processed input was duplicated in all the channels. During the training, the images were augmented [22] by flipping along the vertical axis so that the views at AP and PA were randomly represented on the left or right in the images.

The last Max Pooling layer in the VGG16 model was followed by a Global Average pooling layer, followed by a fully connected layer with 512 units and ReLu activation, which is followed by a classification layer containing 2 units with Softmax activation [23] as shown in Figure 6.2. The network weights are updated by using the Adam optimizer at an initial learning rate of 1e^{-4} [24].

The trained model's performance was evaluated on an external test dataset (n = 998).

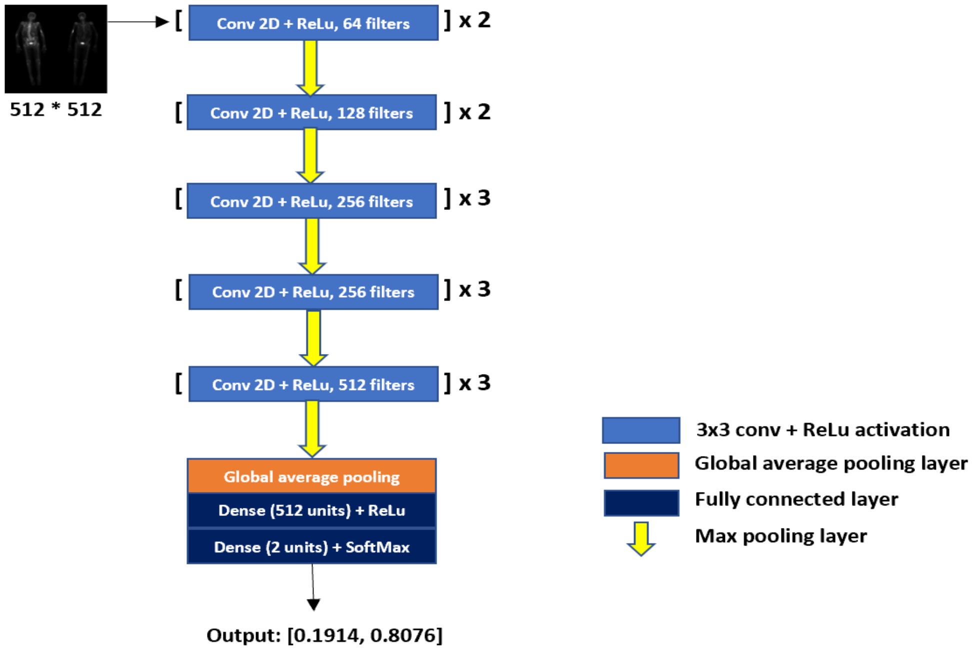

512 * 512
[Conv 2D + ReLu, 64 filters] x 2
[Conv 2D + ReLu, 128 filters] x 2
[Conv 2D + ReLu, 256 filters] x 3
[Conv 2D + ReLu, 256 filters] x 3
[Conv 2D + ReLu, 512 filters] x 3
Global average pooling
Dense (512 units) + ReLu
Dense (2 units) + SoftMax
Output: [0.1914, 0.8076]
3x3 conv + ReLu activation
Global average pooling layer
Fully connected layer
Max pooling layer

Figure 6.2. The architecture used in the study. Pre-processed BS scans resized to 512 * 512 dimensions were provided as input to the network. The network outputs a probability score for the presence and absence of metastasis on BS images. X = block repetitions, Conv = Convolution kernel, ReLU = rectified linear unit, 3x3 =the size of the 2D CNN kernels.

2.5 AUTOMATIC LABELER FOR THE LOCATION OF METASTASIS IN BONE SCINTIGRAPHY SCANS

A dataset of BS was provided by the University of Aachen and contained the scans of 20 patients, each containing both AP and PA views. All scans had annotations for six anatomical regions (head, thorax, pelvis, shoulders, upper limbs, and lower limbs), as shown in Figure 6.3. The total of 40 scans was split into a training (32) and validation (8) set.

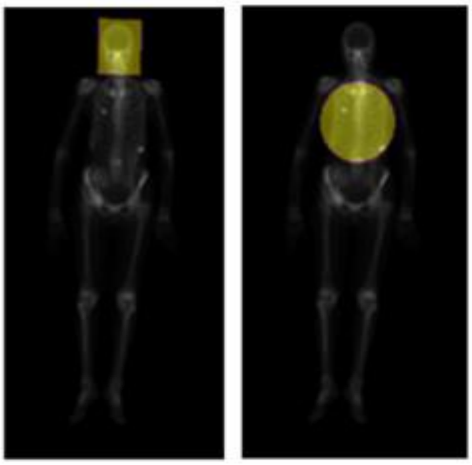


Figure 6.3. Examples of annotations for two anatomical regions: head and thorax.

A ResNext50 architecture [25] with ImageNet pre-trained weights [26] was trained with categorical cross-entropy loss. A 3-channel input was used where the first channel contained the scan while the two others contained a segmented region. The segmented region was artificially created from the region annotations that came with the dataset. An example of a scan with a segmented region is shown in Figure 6.4. The last convolutional layer in the ResNext50 model was followed by a Global Average pooling layer which reduces the image spatial resolution, followed by a fully connected layer with 512 units and ReLu activation, which is followed by a classification layer containing 6 units with Softmax activation. The network weights are updated by using the Adam optimizer at an initial learning rate of 1e^{-5}. Due to the limited number of scans, the fact that metastasis can occur in a lot of different locations, and the fact that metastatic regions are much smaller than the region annotations, extensive augmentation was applied during training. Three different augmentations were applied during training:

1. Variation in the highlighted region of the scan (head, thorax, pelvis, shoulders, upper limbs, and lower limbs)
2. Variation in the number of pixels highlighted
3. Variation in the shape of the highlighted region
4. Left/right flip of the scan

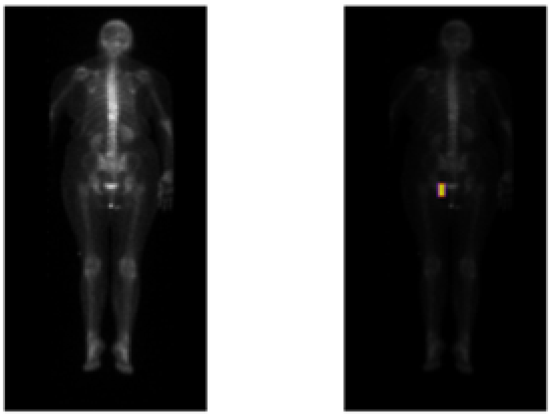


Figure 6.4. Example of a scan with a segmented region in the pelvic area.

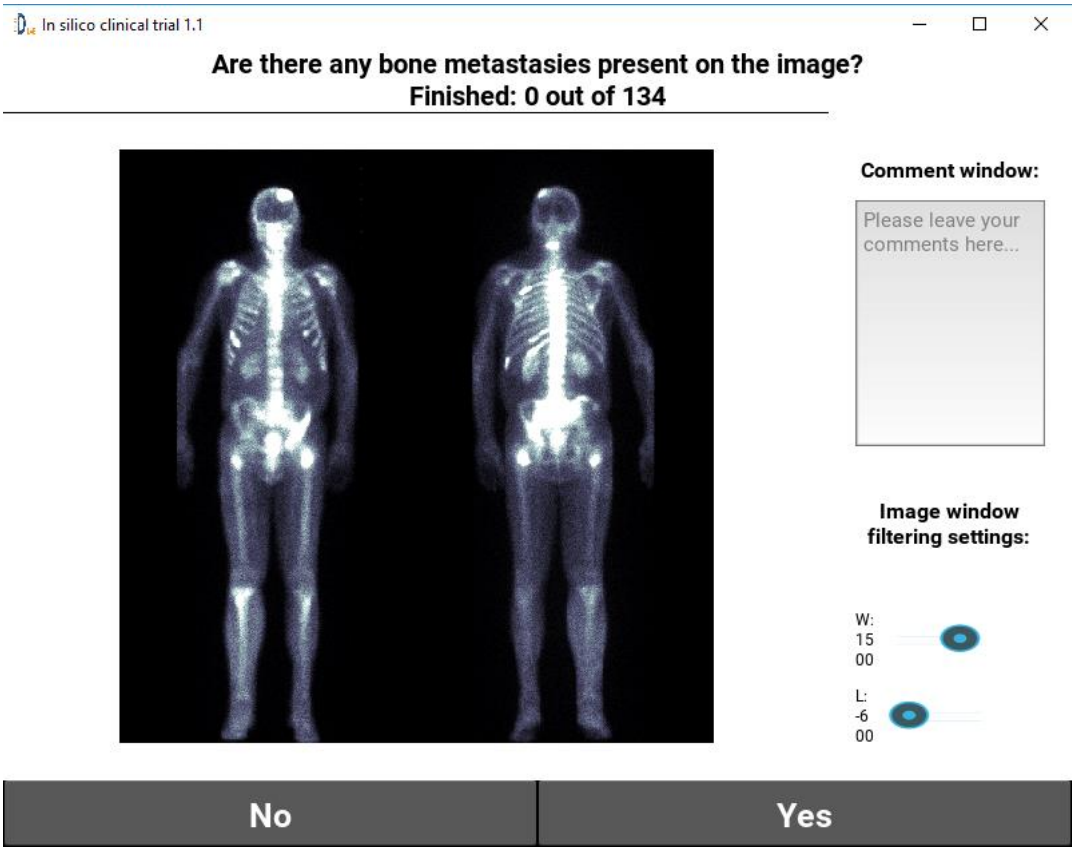
2.6 QUANTITATIVE METRICS

The quantitative model performance in this study was assessed using ROC AUC, sensitivity and specificity of the classifier, and confusion matrix (true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR)). The model was evaluated according to the Checklist for AI in Medical Imaging (CLAIM) [27]and Standards for Reporting Diagnostic accuracy studies (STARD) [28].

2.7 IN SILICO CLINICAL TRIAL

To better gauge the proposed DL model performance, we developed an application allowing the creation of a reference performance point by collecting nuclear medicine physicians' feedback based on the visual assessment of BS scans. We have enrolled 6 nuclear medicine physicians to measure their performance on the evaluation dataset of 134 BS images. This dataset was sampled from the Centre C images with an equal number of negative and positive cases. To collect participants' feedback, the application was displaying BS image, comment window and window filtering settings

(Figure. 6.5). At the end of the feedback assessment excel file was generated. For better visual comparison we have evaluated DL-based AUC on the same dataset that has been used for visual assessment (134 BS images). We used bootstrapping with 100 iterations to generate DL-based AUC distribution.


In silico clinical trial 1.1
Comment window:
Please leave your comments here...
Image window filtering settings:
No
Yes

Are there any bone metastasies present on the image?
Finished: 0 out of 134

W: 1500
L: -600

Figure 6.5. Screenshot of the application feedback window.

3 RESULTS

3.1 MODEL PERFORMANCE

The classification performances of the DL model were evaluated on the external test set coming from Centre C, in terms of Area under the Curve (AUC). The AUC gives the diagnostic ability of a binary classifier to discriminate between true and false values, in this case, metastatic and non-metastatic bone disease. Figure. 6.6 (left) represents the ROC curve of the DL classification model, while Figure. 6.6 (right) is the confusion matrix, which reports the percentages of correct and incorrect classification for each class (metastatic and non-metastatic).

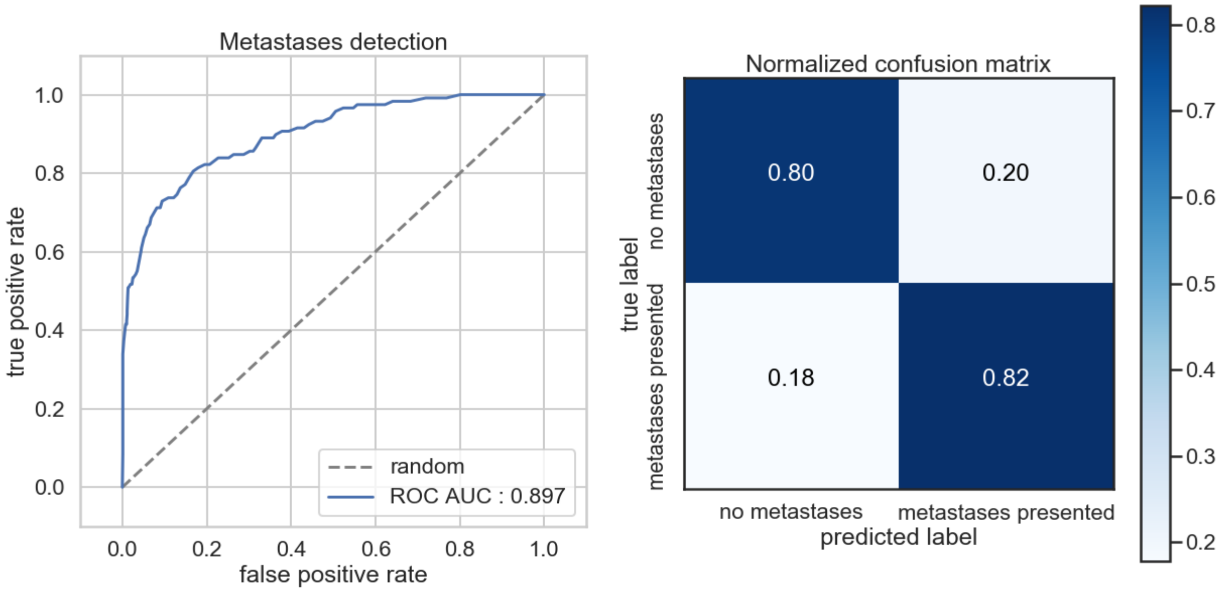


Figure 6.6. ROC curve for the classification DL model (left) and Confusion matrix (right)

The model achieved an AUC of 0.897, TPR of 82.2%, TNR of 80.45 %, FPR of 19.55%, and FNR of 17.79 % on the external test set. The model achieved a CLAIM score of 64 % (27 out of 42 items) and a STARD of 50 % (15 out of 30 items).

3.2 EXPLAINABILITY OF TRAINED MODEL BASED ON ACTIVATION MAPS

During the testing phase of the trained model, for the scans that were predicted positive (i.e. metastatic disease), activation maps were extracted using the Grad-CAM method. The method uses the gradients extracted corresponding to the class with the highest predicted probability, flowing through the last convolutional layer, to produce the activation map. The map was then resized to the size of the input image and superimposed on the original BS scan, allowing visual inspection of activated zones on the image as shown in Figures 6.7 and 6.8. The activated regions are compared with radiologists' segmentation of metastatic spots for qualitative assessment of the explainability of the model's predictions.

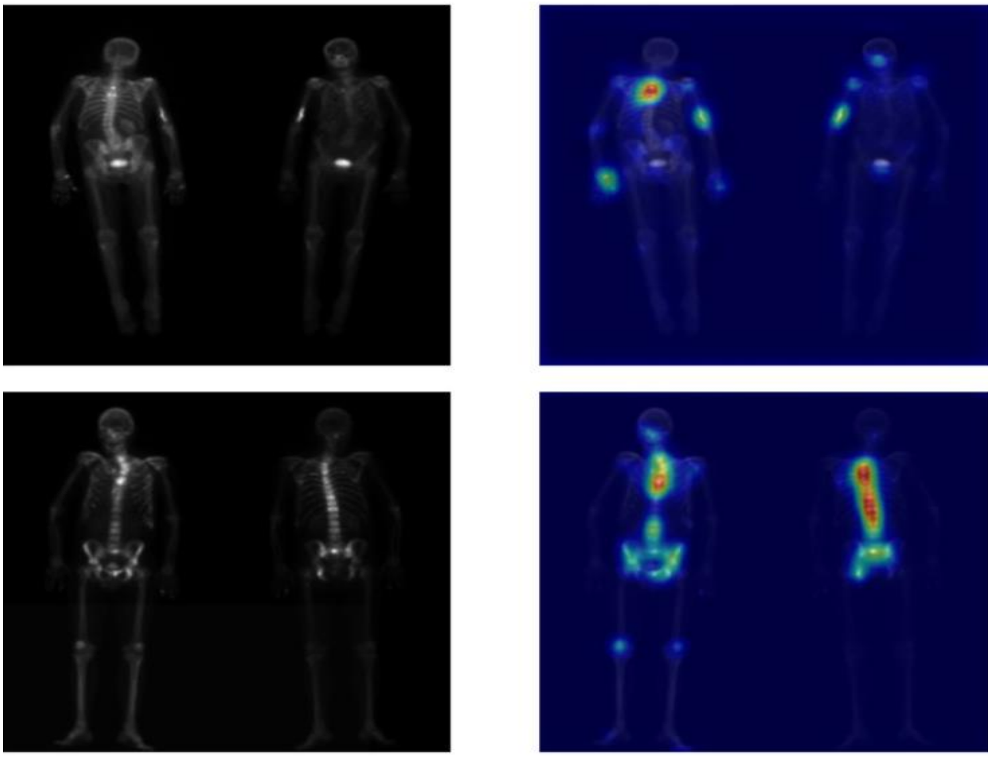


Figure 6.7. BS images which are correctly classified along with their corresponding activation maps extracted using the GRAD-CAM method. Left) original BD scan, Right) Grad-CAM activation maps were obtained from the DL model. Scan correctly classified with a probability of 0.78 (top) and 0.99 (bottom)

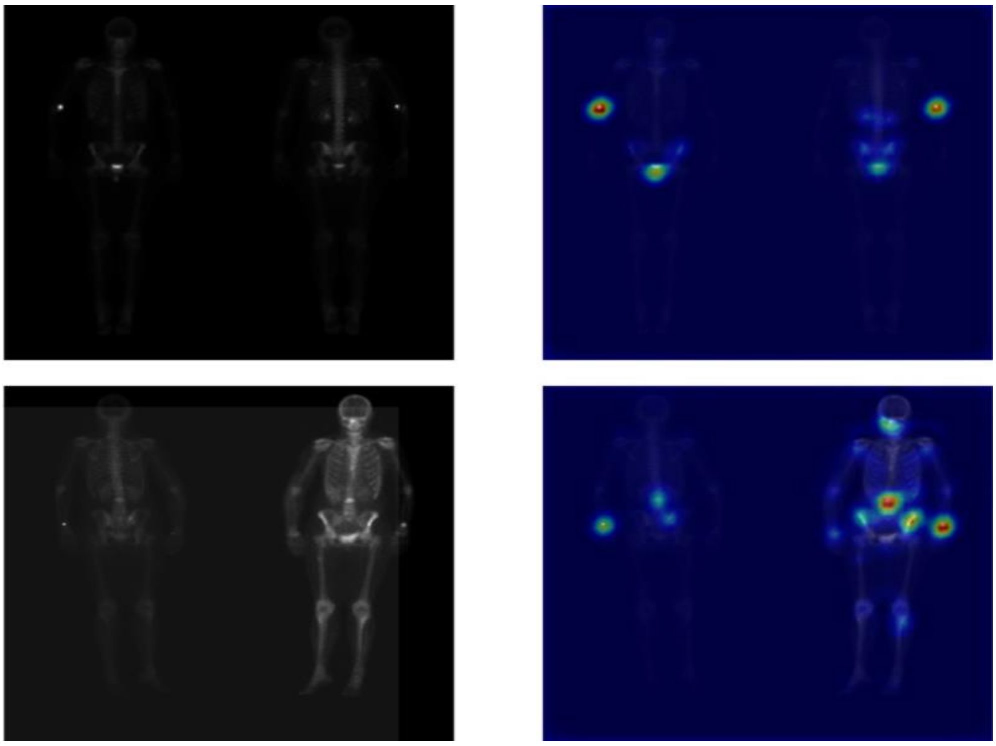


Figure. 6.8. BS images that are wrongly classified along with their corresponding activation maps extracted using the GRAD-CAM method, Left) original BD scan, Right) Grad-CAM activation maps were obtained from the DL model. Scan incorrectly classified with a probability of 0.79 (top) and 0.63 (bottom)

3.3 AUTOMATIC LABELER FOR THE LOCATION OF METASTASIS IN BONE SCINTIGRAPHY SCANS

We developed an automatic labeler for the location of metastasis in BS after the metastatic regions have been extracted. This objective is of great interest as it would allow automated completion of the clinical report with the location of metastasis. The approach proposed here automatically predicts the anatomic locations of metastasis in BS, given the scan and metastatic region as input. For this purpose, a model was built to distinguish between 6 different anatomic regions: head, thorax, pelvis, shoulders, upper limbs, and lower limbs. At the end of the training, a categorical accuracy of 0.92 was reached on the validation set. However, segmented spots for the scans in the validation set were also artificially created. The trained model was therefore tested on an external dataset (n = 462) of BS scans with indications of metastatic regions extracted from the activation maps of the MBD classifier. The resulting labels were qualitatively evaluated. A few examples are shown in Figures 6.9 and 6.10.

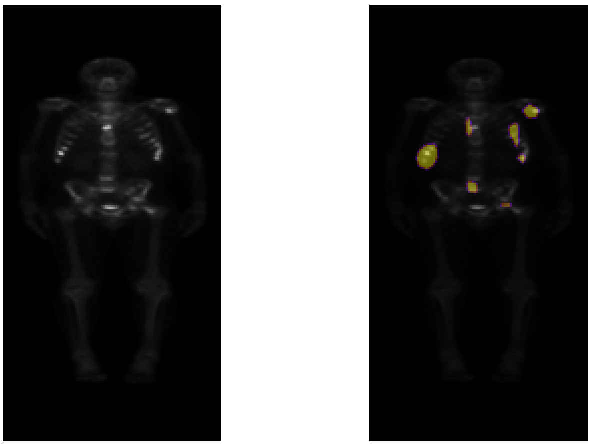


Figure 6.9. Example of a test scan with corresponding metastasis segmentation.
The labels predicted by the location labeler model are ‘upper limbs’, ‘pelvis’, and ‘thorax’.

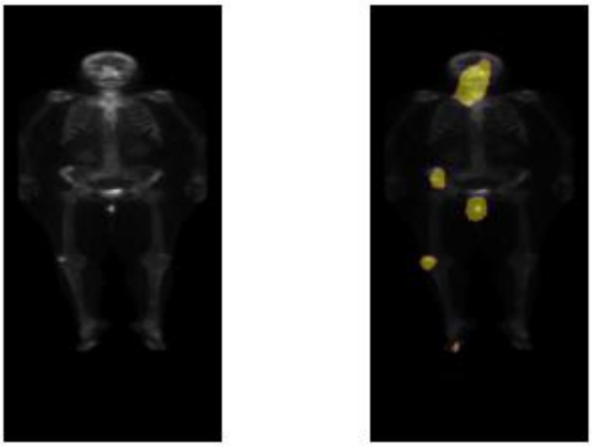


Figure 6.10. Example of a test scan with corresponding metastasis segmentation.
The labels predicted by the location labeler model are ‘lower limbs’, ‘pelvis’, ‘skull’, and ‘thorax’.

3.4 IN SILICO CLINICAL TRIAL

The performance of nuclear medicine physicians based on the BS images was evaluated using AUC, where the median performance of the nuclear medicine physician was 0.895 (IQR = 0.087) and the median performance of DL based method was 0.95 (IQR = 0.024) (Figure. 6.11).

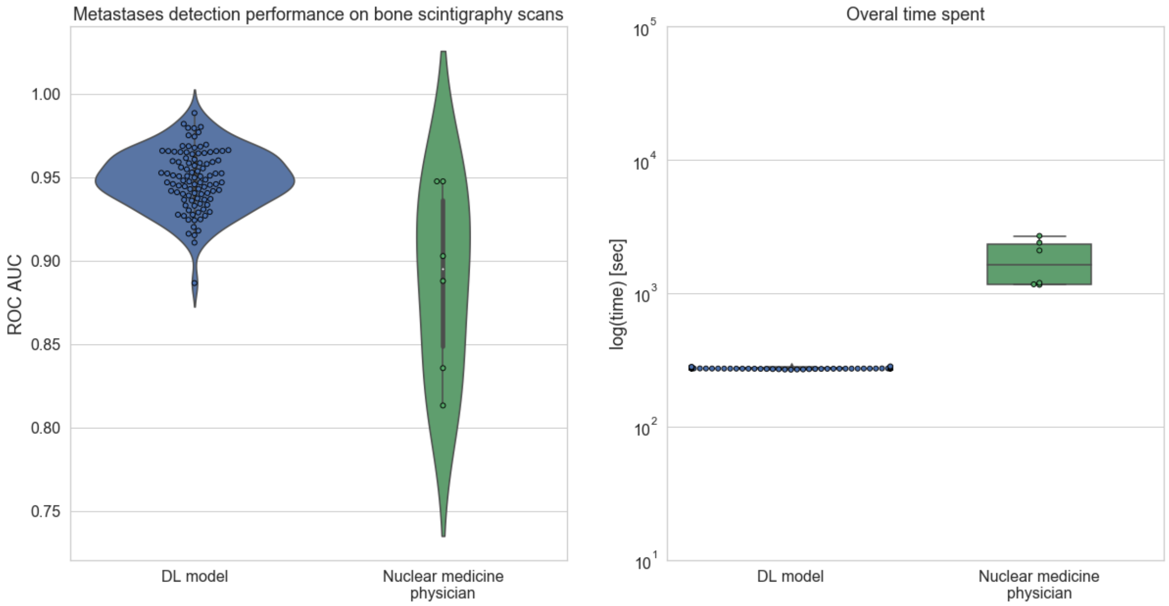


Figure. 6.11. Violin plots showing the distributions of AUC scores for DL-based and manual (across physicians) metastases detection on BS (left); boxplots of the log of the time needed by DL algorithm and nuclear medicine physicians (right).

On average, nuclear medicine physicians spent 30 mins to classifying all the 134 scans. Given that the physicians had no access to clinical information about the patients, it takes on average 15 seconds to review one scan. In comparison, our developed algorithm takes 2 and a half minutes to classify all the 134 scans, which is around 2 seconds per patient scan.

4 DISCUSSION

In this study, we investigated the potential of DL-based algorithms to detect MBD on BSs collected from different centers without limiting the study population to cancer patients. Our results show that DL-based algorithms have a great potential to be applied as clinical decision aid tools, which could minimize the time needed by a nuclear physician to assess BSs and increase the diagnostic specificity of BSs. The application of the state-of-the-art classification techniques has yielded a performance similar to nuclear physicians with no background about the patient's history, which was further endorsed by the results of the in silico clinical trial.

Besides classification and the extraction of activation maps, the first exploratory steps were taken towards the development of a model to automatically label the location of metastasis which can be extended further to automatic report generation in a clinical setting. This latter objective is of great interest as it would allow us to automatically complete the clinical report file with the location of metastasis. For this purpose, a classification model, based on ResNet50 architecture, was built to distinguish between 6 different anatomic regions: head, thorax, pelvis, shoulders, upper limbs, and lower limbs. The ground truths to train the classifier consisted of images with indicated regions at the aforementioned locations. To create a robust model from the available labels, augmentation techniques were applied during training. These include variation in the highlighted region of one scan, variation in the number of pixels highlighted, and variation in the shape of the highlighted region [29]. This preliminary work resulted in a DL model able to classify activated metastatic regions into 6 anatomical categories with a performance of AUC 0.92. These preliminary results showed the potential of a DL-based classifier to automatically label the location of metastasis in bone scintigraphy scans which can be used to finalize clinical reports. However, further validation of this model is needed in the future.

Some studies previously investigated the potential of DL algorithms to classify lesions on BSs. A study investigated the potential of a DL algorithm trained on 139 patients to detect MBD on BSs of prostate cancer patients [16]. The authors reported that the nuclear medicine physicians participating in the study achieved higher sensitivity and specificity compared to the DL algorithm, though the differences were not statistically significant, and highlighted the possibility of involving DL in this clinical aspect. Another study also investigated the ability of DL algorithms to detect MBD in BS of prostate cancer patients [15]. The authors trained the algorithm on 778 BS that could accurately (accuracy of 91.61% ± 2.46%) detect MBD for prostate cancer patients on BS. However, the authors did not report on the comparison with the performance of nuclear medicine physicians. Another study investigated the performance of two DL architectures for classifying BS in prostate cancer patients [30]. The study included a large number of scans, and the authors reported that the best model achieved an overall accuracy of 0.9. Anand et al. reported on the performance of EXINI bone software, a classification tool for classifying BS of prostate cancer patients based on bone scan index, on simulated and patient scans [31]. The authors reported that the software was more consistent in classifying BS compared to visual assessment. Uniquely, we trained our model on patients with and without a history of cancer. The use of our developed algorithm resulted in better classification results on the external test set compared to the median nuclear medicine physician performance, in a significantly shorter time. These results highlight the potential of such algorithms to become reliable clinical decision support tools that minimize the time a clinician needs to review bone scintigraphy scans. Furthermore, our automatic labeling function and the Grad-CAM maps allow the nuclear physicians to rapidly check the spots based on which the classification was made.

While our study included a relatively large number of scans for training and externally testing the algorithm, several limitations of this study should be noted. Although the explainability of the model's predictions was explored with qualitative assessment, this study lacks quantitative assessment of the activations due to the limited number of manual segmentations of metastasis (c.a. 25) on the external test dataset. Also, as shown in figure 6.7, the activated zones correspond to the injected spot in the hand, which shows the model's overfitting [32] on features that are not relevant to the metastatic spot to classify the presence or absence of metastasis in images. Secondly, prospective validation is required to properly assess the impact of using the algorithm on the current standard of care. Lastly, the physicians 'performances in the in-silico trial are only indicative, as they dealt with planar images only, without SPECT and CT, and without any clinical input. This merely approximates the actual routine clinical setting, but it provides a fair indication of the potential added value of DL in this setting.

5 CONCLUSIONS

We developed a DL-based algorithm that can detect MBD on BSs, with high specificity and sensitivity, that could be applied to any BS regardless of the patient's gender and history of cancer. Further prospective validation is required before the algorithm can be used in the clinic.

6 REFERENCES

[1] Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research 2006;12:6243s–9s. https://doi.org/10.1158/1078-0432.CCR-06-0931.

[2] Migliorini F, Maffulli N, Trivellas A, Eschweiler J, Tingart M, Driessen A. Bone metastases: a comprehensive review of the literature. Molecular Biology Reports 2020;47:6337–45. https://doi.org/10.1007/s11033-020-05684-0.

[3] Huang J-F, Shen J, Li X, Rengan R, Silvestris N, Wang M, et al. Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study. Annals of Translational Medicine 2020;8:482. https://doi.org/10.21037/atm.2020.03.55.

[4] Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. Cancer Treatment Reviews 2001;27:165–76. https://doi.org/10.1053/ctrv.2000.0210.

[5] Macedo F, Ladeira K, Pinho F, Saraiva N, Bonito N, Pinto L, et al. Bone Metastases: An Overview. Oncology Reviews 2017;11:321. https://doi.org/10.4081/oncol.2017.321.

[6] Ryan PJ, Fogelman I. Bone scintigraphy in metabolic bone disease. Seminars in Nuclear Medicine 1997;27:291–305. https://doi.org/10.1016/s0001-2998(97)80030-x.

[7] Ziessman HA, O'Malley JP, Thrall JHBT-NM (Fourth E, editors. Chapter 7 - Skeletal Scintigraphy, Philadelphia: W.B. Saunders; 2014, p. 98–130. https://doi.org/https://doi.org/10.1016/B978-0-323-08299-0.00007-9.

[8] Van den Wyngaert T, Strobel K, Kampen WU, Kuwert T, van der Bruggen W, Mohan HK, et al. The EANM practice guidelines for bone scintigraphy. European Journal of Nuclear Medicine and Molecular Imaging 2016;43:1723–38. https://doi.org/10.1007/s00259-016-3415-4.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[10] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 1943;5:115–33. https://doi.org/10.1007/BF02478259.

[11] Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing 2014;3:e2. https://doi.org/DOI: 10.1017/atsip.2013.9.

[12] Aslam Y, N S. A Review of Deep Learning Approaches for Image Analysis. 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 2019, p. 709–14. https://doi.org/10.1109/ICSSIT46314.2019.8987922.

[13] Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. Schedae Informaticae 2016;25:49–59. https://doi.org/10.4467/20838476SI.16.004.6185.

[14] Cheng D-C, Hsieh T-C, Yen K-Y, Kao C-H. Lesion-Based Bone Metastasis Detection in Chest Bone Scintigraphy Images of Prostate Cancer Patients Using Pre-Train, Negative Mining, and Deep Learning. Diagnostics 2021;11. https://doi.org/10.3390/diagnostics11030518.

[15] Papandrianos N, Papageorgiou E, Anagnostis A, Papageorgiou K. Efficient Bone Metastasis Diagnosis in Bone Scintigraphy Using a Fast Convolutional Neural Network Architecture. Diagnostics (Basel, Switzerland) 2020;10. https://doi.org/10.3390/diagnostics10080532.

[16] Aoki Y, Nakayama M, Nomura K, Tomita Y, Nakajima K, Yamashina M, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. Annals of Nuclear Medicine 2020;34:926–31. https://doi.org/10.1007/s12149-020-01524-0.

[17] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[18] Dubost F, Adams H, Yilmaz P, Bortsova G, Tulder G van, Ikram MA, et al. Weakly supervised object detection with 2D and 3D regression neural networks. Medical Image Analysis 2020;65:101767. https://doi.org/https://doi.org/10.1016/j.media.2020.101767.

[19] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision 2016;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[20] World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310:2191–4. https://doi.org/10.1001/jama.2013.281053.

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings 2015:1–14.

[22] Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data 2019;6:60. https://doi.org/10.1186/s40537-019-0197-0.

[23] Calin O. Activation Functions BT - Deep Learning Architectures: A Mathematical Approach. In: Calin O, editor., Cham: Springer International Publishing; 2020, p. 21–39. https://doi.org/10.1007/978-3-030-36721-3_2.

[24] Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[25] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017- Janua, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 5987–95. https://doi.org/10.1109/CVPR.2017.634.

[26] Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

[27] Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artificial Intelligence 2020;2:e200029. https://doi.org/10.1148/ryai.2020200029.

[28] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799. https://doi.org/10.1136/bmjopen-2016-012799.

[29] Abdollahi B, Tomita N, Hassanpour S. Data Augmentation in Training Deep Learning Models for Medical Image Analysis BT - Deep Learners and Deep Learner Descriptors for Medical Applications. In: Nanni L, Brahnam S, Brattin R, Ghidoni S, Jain LC, editors., Cham: Springer International Publishing; 2020, p. 167–80. https://doi.org/10.1007/978-3-030-42750-4_6.

[30] Han S, Oh JS, Lee JJ. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. European Journal of Nuclear Medicine and Molecular Imaging 2021. https://doi.org/10.1007/s00259-021-05481-2.

[31] Anand A, Morris MJ, Kaboteh R, Båth L, Sadik M, Gjertsson P, et al. Analytic Validation of the Automated Bone Scan Index as an Imaging Biomarker to Standardize Quantitative Changes in Bone Scans of Patients with Metastatic Prostate Cancer. Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine 2016;57:41–5. https://doi.org/10.2967/jnumed.115.160085.

[32] M.R.Narasinga Rao D, Venkatesh Prasad V, Sai Teja P, Zindavali M, Phanindra Reddy O. A Survey on Prevention of Overfitting in Convolution Neural Networks Using Machine Learning Techniques. International Journal of Engineering & Technology 2018;7:177. https://doi.org/10.14419/ijet.v7i2.32.15399.

Chapter 7

A non-invasive, automated diagnosis of Menière’s disease using radiomics and machine learning on conventional magnetic resonance imaging: A multicentric, case-controlled feasibility study

Based on van der Lubbe, M.F.J.A., **Vaidyanathan, A.**, de Wit, M. et al. A non-invasive, automated diagnosis of Menière’s disease using radiomics and machine learning on conventional magnetic resonance imaging: A multicentric, case-controlled feasibility study. Radiol med 127, 72–82 (2022). https://doi.org/10.1007/s11547-021-01425-w

In this chapter, we investigated the feasibility of Radiomics, as a computer-aided diagnostic tool for Menière’s disease. This study included 119 patients with unilateral or bilateral Menière’s disease and 141 controls from four centers in the Netherlands and Belgium. Multiple radiomic features were extracted from conventional MRI scans and used to train a machine learning-based, multi-layer perceptron classification model to distinguish Menière’s disease from the control group. The primary outcomes were accuracy, sensitivity and specificity, positive predictive value, and negative predictive value of the classification model. The classification accuracy of the machine learning model on the validation set was 82%, with a sensitivity of 83% and a specificity of 82%. The positive and negative predictive values were 71% and 90%, respectively. The multi-layer perceptron classification model yielded a high-diagnostic performance in identifying patients with Menière’s disease based on radiomic features extracted from conventional T2-weighted MRI scans. This solution could serve as a fast and noninvasive decision support system, next to clinical evaluation in the diagnosis of Menière’s disease.

1 BACKGROUND

Menière’s disease (MD) is a multifactorial condition of the inner ear characterized by recurrent episodes of vertigo and fluctuating aural symptoms like hearing loss, aural fullness, and tinnitus. The exact etiology of the disease is unknown. However, MD is strongly associated with the classical histological finding known as endolymphatic hydrops (EH) which is a distention of the endolymphatic compartment of the labyrinth [1]. The consistent finding of EH in temporal bones of patients with MD [2,3] led to defining EH as the pathological basis of MD. However, it also marked the beginning of a diagnostic challenge, as EH could only be identified post-mortem. As a consequence, MD remained a clinical diagnosis, and different symptom-based classification methods emerged over time. The most widely accepted diagnostic criteria are those proposed in 1995 by the American Academy of Otolaryngology-Head and Neck Surgery (AAO-HNS) and the revised criteria by the classification committee of the Bárány Society in 2015 [4].

The clinical diagnosis of MD is complicated due to the diverse clinical presentation of the disease. MD appears to be a continuum from a single initial symptom, to the full-blown symptom spectrum. The time delay between the first presenting symptoms can reach up to five years [5], which hampers the clinical diagnosis. Moreover, patients’ complaints are subjective and difficult to objectify in testing. For example, due to the prominent and incapacitating aspect of vertigo spells, hearing loss is not always noticed and can also recover before audiometric measurements are performed [6], which may lead to an underestimation of MD cases.

A variety of additional audio-vestibular tests are available to support the clinical diagnosis, including electrocochleography and vestibular evoked myogenic potentials. Still, it remains challenging to differentiate between MD and other causes of vertigo due to substantial symptom overlap and the lack of specific biomarkers [7]. Therefore, new imaging techniques are under investigation as an MD diagnostic, which includes cone beam computed tomography (CBCT) [8] and magnetic resonance imaging (MRI) enhanced by contrast agents [9]. Until recently, MRI was only to exclude other diseases with a similar symptom presentation, such as vestibular schwannoma. However, the developments in gadolinium-based contrast-enhanced MRI enabled the in vivo visualization of EH as a biomarker in humans. Nowadays, this technique is broadly investigated, and its clinical utility as a new diagnostic test is debated in the literature [10].

Still, there is no consensus concerning the best method of how to grade EH. In addition, the administration of gadolinium is an invasive and time-consuming procedure. Imaging is performed four hours after intravenous and 24 hours after intratympanic administration [9]. Furthermore, intravenous administration of gadolinium is contraindicated in patients with contrast allergies or renal failure [11]. Although no ototoxicity has been reported [12], other adverse effects such as gadolinium depositions in the brain have been observed [11,13], hence, a non-invasive imaging technique to diagnose MD would be preferable.

Increasing evidence indicates that with new image analyzing techniques, diagnostic, prognostic, and predictive information can be extracted from conventional image modalities [14–16] The process of converting standard medical images into mineable high-dimensional data by extracting quantitative image features and linking them to clinical outcomes is referred to as Radiomics [15,17]. To analyze such large amounts of quantitative image features, machine learning (ML) methods are often used to find patterns in the data. ML is a subdiscipline in the field of artificial intelligence that focuses on the ability of algorithms to “learn” from data (e.g., by adapting their structure) rather than through explicit programming [18]. Particularly in oncology, [16,19] Radiomics is an emerging field, but it is also applicable to other medical disciplines.

Recently, a proof-of-concept study demonstrated the possible value of Radiomics within the diagnosis of MD by detecting differences in image features between patients with MD and controls in conventional MRI scans [20]. To further explore the application of Radiomics, the objective of this study was to develop a computer-aided diagnostic tool for MD by using a Radiomics approach combined with ML. Its performance and feasibility as a new diagnostic tool for MD were evaluated.

2 MATERIAL AND METHODS

2.1 ETHICAL CONSIDERATIONS

This study was performed following the guidelines outlined by Dutch and Belgium legislation. Subjects were enrolled and fully anonymized by the local investigators (Maastricht University Medical Center +, University Hospital Antwerp, VieCuri Hospital Venlo, and Apeldoorn Dizziness Center) and were therefore not asked for their consent. According to the Medical Research Involving Human Subjects Act (WMO), ethical approval was not required due to the retrospective nature and anonymization of the data.

2.2 STUDY DESIGN AND INCLUSION

A retrospective, diagnostic case-control study was performed on patients with unilateral and bilateral MD. Medical records in the following centers in the Netherlands and Belgium were searched for eligible subjects:

1. Maastricht University Medical Center + (MUMC+), The Netherlands
2. Antwerp University Hospital, Belgium
3. Apeldoorn Dizziness Center, The Netherlands
4. VieCuri Hospital Venlo, The Netherlands

Subjects were enrolled as “Patients” when MD was clinically diagnosed by an ENT-specialist as “Definite” MD according to the criteria of the American Academy of Otolaryngology-Head and Neck Surgery (AAO-HNS) [21] and/or Barany society (2015) [4], and in case a conventional MRI scan of the cerebellopontine angle was already available from the clinical setting. Both unilateral and bilateral cases of definitive MD were included. Subjects were enrolled as “Controls” when diagnosed by an ENT specialist with idiopathic asymmetric sensorineural hearing loss and when a conventional MRI scan of the cerebellopontine angle was available. The labyrinth least affected by hearing loss was considered to be the best representative of a ‘normal’ labyrinth and was included in the study.

These patients were chosen as controls since this was a retrospective study and no MRI scans from ‘healthy’ people without any hearing loss were available. Controls were excluded in case of a documented history of vertigo and/or balance disorders. Subjects (Patients and Controls) were excluded in case of motion artifacts and/or an unsharp delineation of the inner ear on the MRI scan.

2.3 STATISTICAL ANALYSIS

A chi-square test of independence was performed for between-group comparisons of gender distribution and independent samples t-test for age distributions. Statistical analyses were carried out using Statistical Package for the Social Sciences (SPSS) software version 25.0 (IBM Corp, Armonk, NY).

2.4 RADIOMICS WORKFLOW

The radiomic workflow applied in this study consisted of four steps, as illustrated in figure 7.1.

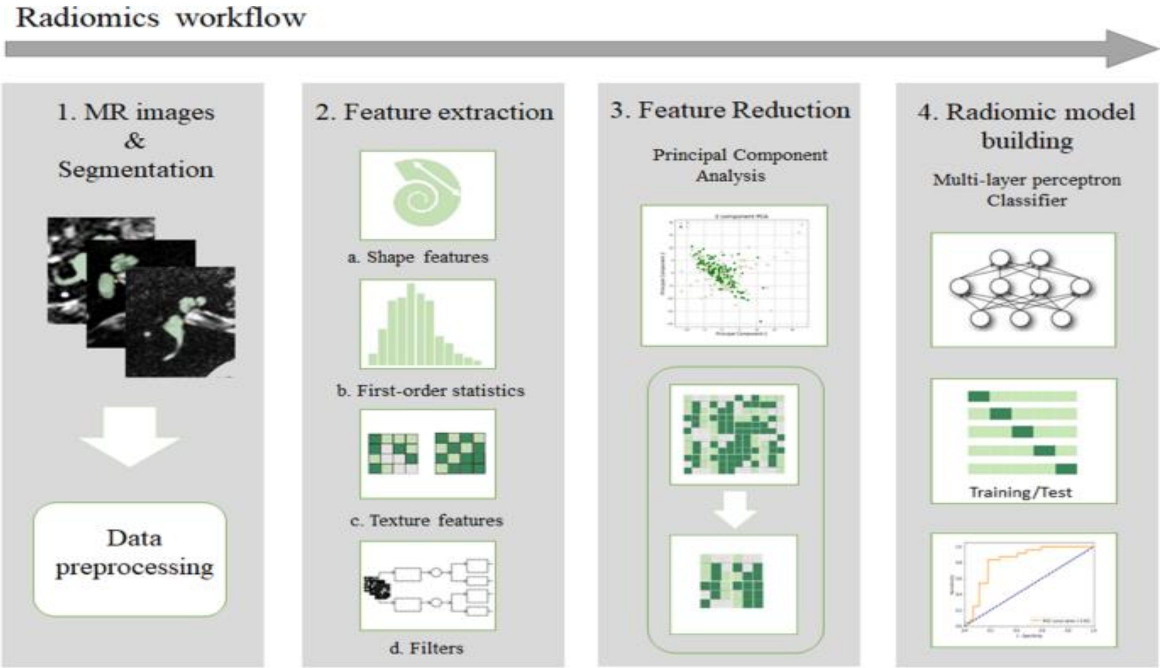

Radiomics workflow
1. MR images & Segmentation
Data preprocessing
2. Feature extraction
a. Shape features
b. First-order statistics
c. Texture features
d. Filters
3. Feature Reduction
Principal Component Analysis
4. Radiomic model building
Multi-layer perceptron Classifier
Training/Test

Figure 7.1. The workflow of Radiomics in this study is graphically presented in four steps. (1) T2-weighted MR images were collected from four different centers in the Netherlands and Belgium and manually segmented. The MR volumes and their corresponding segmentation masks were preprocessed into isotropic voxels. (2) Four types of features (a. Shape features, b. First-order statistic features, c. Texture features, and d. Features extracted after applying different filters) were extracted from the segmented masks. (3) Feature reduction was done by principal component analysis. (4) A multi-layer perceptron classifier was used for radiomic analysis

2.4.1 MR imaging and segmentation

Image acquisition and data anonymization were performed by the local investigators of the four centers. T2-weighted MR images were acquired with center-specific protocols on 1,5T and 3T scanners. 3D Slicer 4.8.1, an open-source software package for visualization and image analysis was used to segment the labyrinth from all MRI scans. Two authors (EB, MW) manually segmented all labyrinths using an inbuild region-growing algorithm (Grow from seeds) from 3D Slicer [22]. The first author (ML) cleaned the initial dataset and re-segmented the labyrinths in case of missing labels.

The following preprocessing steps were performed [23]. First, to normalize the voxel sizes across the volumes, the MR volumes and their corresponding segmentation masks were resampled to isotropic voxels of length 0.5 mm using cubic spline interpolation. Secondly, voxel intensities were transformed using Z-score normalization to minimize the influence of contrast or brightness variation among the images. Thirdly, the transformed voxel intensities were discretized using a fixed bin width of 0.5.

2.4.2 Feature extraction

In total, 812 radiomic features were extracted from the segmented masks using RadiomX, an In-House developed software toolbox in MATLAB 2014a (Mathworks, Natick, USA). First-order features were obtained from the intensity histograms using first-order statistics, including intensity mean, median, maximum, minimum, range, energy, entropy, kurtosis, and skewness. Shape features were obtained from the 3D shape of delineated volumes. Texture features were obtained from the spatial distribution of fractal dimensions and voxel intensities using 6 texture matrices, including grey-level co-occurrence (GLCM), gray-level distance-zone (GLDZM), grey-level run-length (GLRM), grey-level size-zone (GLSZM), neighboring grey-level dependence (NGLDM) and neighborhood grey-tone difference matrix (NGTDM). Furthermore, 3D wavelet, Laplacian, and Gaussian filters were applied to the original images to extract additional first-order, shape, and texture features. Mathematical descriptions of all features were previously published and presented as supplemental material with the permission of the corresponding authors [14,16,24].

2.4.3 Feature reduction

To reduce the dimensionality of the extracted features, a Principal Component Analysis (PCA) was performed. PCA is an unsupervised, linear dimensionality reduction technique in which small numbers of uncorrelated variables are extracted as “Principal Components” to explain most of the variation in the data in lower dimensions [25,26]. As a result, essential information holding most of the variation in the data was preserved and non-essential parts with fewer variations were removed. Ten Principal Components were extracted from the analysis and used to train the model. The inverse PCA was applied to identify the mean contribution of each feature's overall principal components to predict the most important features. A mean contribution of > 0.7 was chosen to identify 15 features that had the largest contribution to the PCA.

2.4.4 Machine learning classifier

The dataset (“Patients” and “Controls”) was divided into a training and validation set of 74% and 26%, respectively. The training set contained images from the centers located in Maastricht, Antwerp, and, Apeldoorn. The validation set contained images from Venlo, complemented with randomly selected scans from the other centers which were excluded from training. Next to this, a 10-fold cross-validation was performed.

A Multi-Layer Perceptron classifier with 500 units in the hidden layer was trained with Adam optimizer at a learning rate of 0.001. Input to the model was the extracted Principal Components. The output layer consisted of a single neuron for each prediction class (Patients = 1 and Control = 0), which used the Softmax function to output a value between 0 and 1. The output represented the probability of the predicted classes. The regularization method “early stopping” was adopted during training to avoid overfitting the model [27].

2.5 OUTCOME MEASUREMENTS

The primary outcomes of this study were accuracy, sensitivity and specificity, positive predictive value, and negative predictive value, of the classification model to distinguish MD from the control group. The precision (i.e., confidence interval) of each parameter was determined.

3 RESULTS

3.1 STUDY POPULATION

This retrospective study included 119 patients with MD (59 men, 60 women, aged 16-84; mean age 58 ± 14.0) and 141 controls with asymmetric sensorineural hearing loss (69 men, 31 women, aged 6-88; mean age 59 ± 14.3, in 41 controls gender was unknown) over four centers. There were 73 labyrinths included from MUMC+ (65.8% MD, 34.2% Control), 56 from Antwerp University Hospital (57.1% MD, 42.9% Control), 107 from Apeldoorn Dizziness Center (30.8% MD, 69.2% Control), and 24 from VieCuri Hospital Venlo (25% MD, 75% control) There was no significant difference in age distribution between the patient and the control group and between the training and test cohort. The proportion of known males versus females did not differ between the test and training cohorts. However, significantly more males were included in the control group (chi-square test, p = 0,004). A significant difference in scan data between the training and test cohort was found (Independent sample t-test: p = 0.019) with MRI scans of the training cohort being performed on earlier dates. No significant differences in scan date between all patients with MD and controls were found. Details of the training and test cohort are presented in Table 7.1.

Table 7.1. Details of the study cohort

Group	**n**	**Center**	**Menière's (n)**	**Controls (n)**	**Age (years)**	**Gender** (M/F)	**Date MRI**
Training cohort (74%)	192	A	25	20	60 ± 8	93/67*	2004-2017*
		B	31	56			
		C	40	20			
		Total	96	96			
Test cohort (26%)	68	A	7	4	61 ± 9	34/25	2004-2017*
		B	2	18			
		C	8	5			
		D	6	18			
		Total	23	45			

Demographic details of the study cohorts. N = number of ears, Age is median age with median absolute deviation, * Significant difference between cohorts.

3.2 PRINCIPAL COMPONENT ANALYSIS

By applying the inverse PCA, the mean contribution of each feature overall principal components is illustrated in figure 7.2. As a result, the features with the most substantial influence on the principal components could be identified.

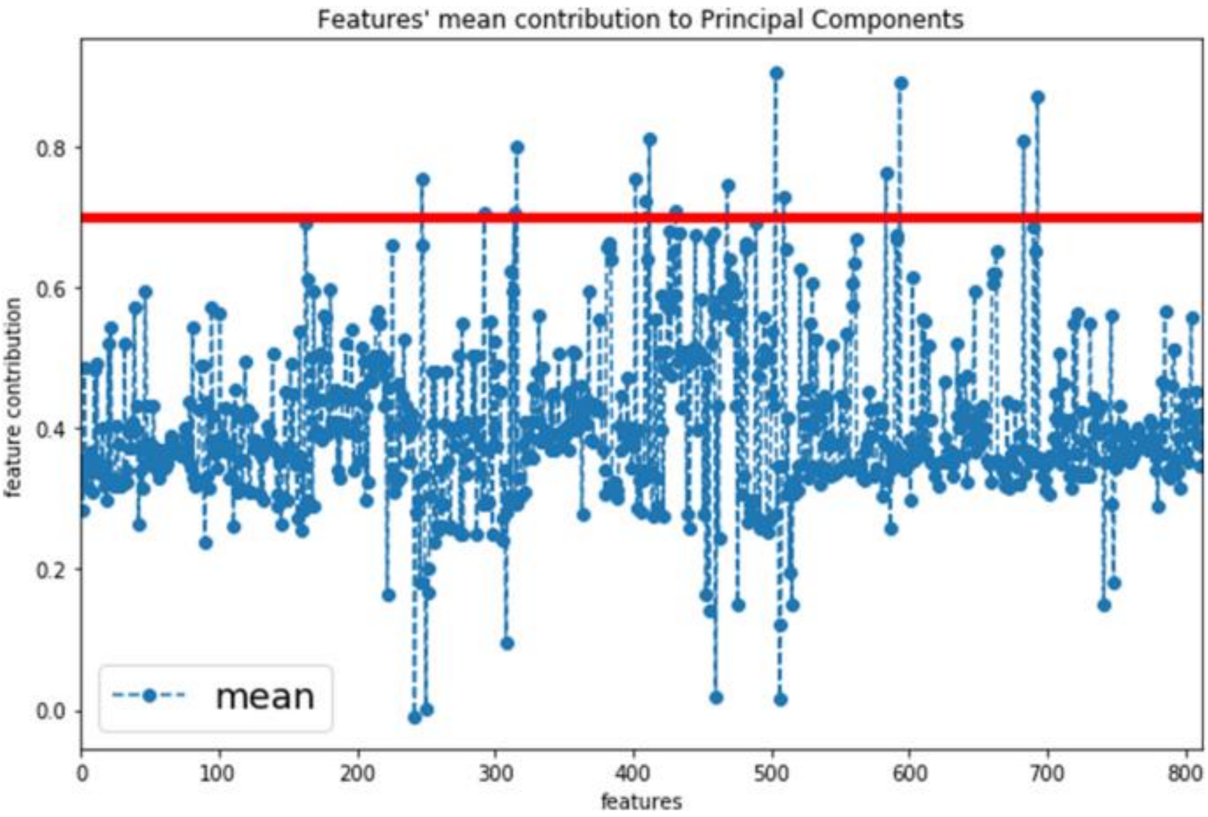


Figure 7.2. The mean contribution overall principal components aggregated for each feature. The red line indicates the cut-off value (< 0.7) for the most important features that contributed to the PCA

3.3 MACHINE LEARNING CLASSIFIER

The ML model´s performance in classifying patients with MD and controls is demonstrated in Table 7.2. The classification accuracy of the validation set was 82% with a sensitivity of 83%, specificity of 82%, and AUC of 0.83. The positive and negative predictive values were 71% and 90%, respectively. The ROC curve and the confusion matrix are shown in Figures 7.3 and 7.4. The results of the 10-folds cross-validation are also presented in Table 7.2. The mean classification accuracy across the 10-folds was 80%, with a mean sensitivity and specificity of 78% and 77%, respectively. The mean AUC was 84% and the mean positive predictive value was 77% and the negative predictive value was 78%.

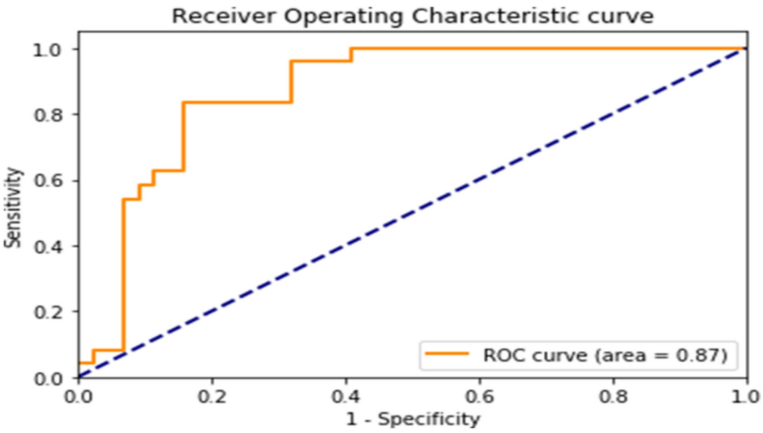


Figure 7.3. The Receiver Operator Characteristic curve of the test cohort of the multi-layer perceptron classifier

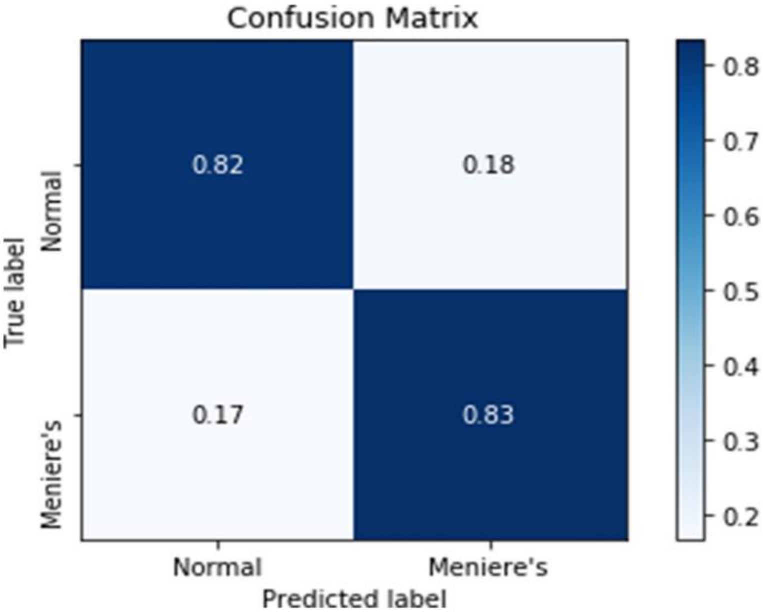


Figure 7.4. *The confusion matrix of the test cohort of the multi-layer perceptron classifier. The true labels are the diagnostic labels after subject inclusion. The predicted labels are the labels predicted by the classifier*

Table 7.2. Classification performance

	Training cohort	**Test cohort**	**10-fold cross-validation**
Patients vs. Controls	96 vs. 96	24 vs. 44	
Accuracy (%)	72.9	82.3	80.0
AUC (95% CI)	80.6 (80.5-81.2)	86.9 (86.6-88.8)	83.6 (77.9-89.3)
Sensitivity (95% CI)	80.2 (80.0-81.1)	83.4 (82.6 -86.9)	78.3 (71.4-85.3)
Specificity (95% CI)	65.6 (65.3-66.3)	81.8 (81.4-83.7)	77.5 (70.5-84.5)
Positive predictive value (95% CI)	70.0 (69.7-70.6)	71.4 (70.4-74.1)	77.6 (69.9-85.4)
Negative predictive value (95% CI)	76.8 (67.5-77.8)	90.0 (89.7 -92.3)	78.4 (70.6-86.3)

Performance of the multi-layer perceptron classification metric to distinguish MD from healthy controls showing the area under the curve of the Receiver Operating Curve, sensitivity, specificity, positive predictive value, and negative predictive value. Abbreviations: CI = confidence interval, AUC = area under the curve

4 DISCUSSION

The purpose of this study was to explore the feasibility of Radiomics as a new diagnostic tool in MD to assist the diagnostic trajectory and to provide a better understanding of the underlying process of the disease. This study demonstrated that radiomic features extracted from conventional MRIs can be used to discriminate MD patients from 'normal' controls. In this small, multicentric dataset, a machine learning-based multi-layer perceptron network yielded a precise, high-diagnostic performance in identifying patients with MD with an accuracy of 82%. This implies that a computer-aided diagnosis of Menière's disease might be possible. In the future, Radiomics could be implemented as valuable decision support next to clinical evaluation.

Within neuro-otology, Radiomics is a very new concept. This study pioneered the development of a computer-aided diagnostic tool for MD by using a radiomic approach. The results of this study are in line with the earlier published proof-of-concept [20], where it was demonstrated that significant differences in radiomic image features existed between MD patients and a control group.

One of the main benefits of Radiomics is that no contrast agents or expert radiologists are required. Other studies have also investigated the feasibility of the radiological diagnosis of MD without the use of contrast agents, by evaluating the morphoanatomy of the membranous labyrinth [8], or by manually measuring the length and width of the saccule and/or the utricle. One publication showed changes in the membranous labyrinth between patients with MD and healthy subjects on 3D CBCT and suggested the usefulness of 3D CBCT imaging for the objective diagnosis of MD [8]. However, the diagnostic value has not been clinically evaluated yet. Three other publications have documented and evaluated measurements of the vestibule on T2-weighted MRI of patients with definite MD and healthy controls [28–30]. One of these papers reported the maximum saccular height in healthy volunteers to be 1.6 mm [28]. Another reported a high specificity (95%) but a low sensitivity (63%) for a cut-off value of 1.51 mm for saccular height [29]. Measurements of the absolute utricle area and the utricle-to-vestibule area ratio were also identified as predictors of MD and yielded a sensitivity of 44% and 75% and a specificity of 81% and 53%, respectively. These results suggest that enlargement of the endolymphatic space, and thus EH, can be detected using non-contrast T2-weighted MRI by human readers. However, the main disadvantage of these techniques is that only vestibular hydrops is evaluated while the Radiomics method in this study, assessed the entire labyrinth. Moreover, human performances on non-contrast MR imaging seem to exhibit lower diagnostic performance compared to Radiomics. To prove this assumption, prospective studies that will perform Radiomics and vestibular measurements on the same dataset with non-contrast T2-weighted MRIs, are needed.

Another benefit of Radiomics is that it is less prone to interference from human-induced factors compared to contrast-enhanced MR imaging, which requires specific expertise. Therefore, Radiomics is not just reserved for specialized tertiary centers. It can be used as a standardized decision support system that might reduce interobserver variability within and between centers and allows more widely accessible diagnostic care for patients with MD.

4.1 LIMITATIONS

This study has several limitations. First, it is important to recognize that no gold standard test is available to compare the Radiomics method with. This retrospective study included patients clinically diagnosed with definite MD according to the AAO-HNS criteria [4]. These patients, however, do not represent the full clinical spectrum. After all, patients who do not fulfill these criteria due to the fluctuating aspect of hearing loss (not captured by audiometry) or atypical symptom presentation might be an interesting group to explore with Radiomics. Especially since they might be very difficult to diagnose with the current clinical diagnostic tools available.

Secondly, the duration of the disease was not considered in this study cohort. Disease duration might alter the morphology of the labyrinth. For example, the severity of endolymphatic hydrops in patients with MD seems to increase with the duration of the disease [34,35]. Perhaps, the disease duration could also alter the composition of the endolymphatic fluid as well. The patient cohort in this study probably contained patients with different disease stages. Early disease stages might be challenging to recognize since important image features were not yet significantly present. Adding clinical information about disease duration would probably have improved the model's performance regarding classification. Although changes in the vestibular aqueduct have been proven to be valuable in the diagnosis of MD [36], the vestibular aqueduct was not considered in this study due to difficulties in manually segmenting this structure. Extracting Radiomic features from the vestibular aqueduct might have improved the model's performance as well.

Thirdly, the study dealt with a relatively small dataset consisting of MRI scans from four independent centers. In the absence of sufficient data points, it was inevitable to divide the data into just a training and validation set, where ideally a third 'test' data set is required. Therefore, in this study, an independent dataset (Venlo) was included for validation to better detect overfitting. Overfitting happens when the model learns details and noise to fit the training data with high accuracy but fails to a new set of data [27,37]. The addition of an external validation set helps to apply early stopping when the model starts to overfit on the training dataset (i.e., when the training loss decreases, and validation loss starts to increase). Due to the small size of the training dataset, overfitting could not be avoided. In this study, the risk of overfitting was contained by diversifying the training data. This was done by acquiring data from four different centers where each center had different scan parameters and by manually segmenting the labyrinth by three different observers. Therefore, the model should be more generalizable for differences in scan parameters and inter-reader segmentations. The common regularization method, "early stopping" was also adopted during training to avoid overfitting the model [27].

Lastly, the heterogeneities in voxel spacing and slice thickness between the images were handled by isotropic resampling. This could have induced noise due to interpolations. Further study with a larger training dataset and/or using convolutional neural networks for the direct extraction of deep features from the raw MRI [38] might improve the diagnostic accuracy and the generalizability of the model.

4.2 CLINICAL IMPLICATIONS AND FUTURE PERSPECTIVES

Radiomics is a new imaging analysis technique that enables a noninvasive, fast, and accurate diagnosis of MD. After validating the current results in a prospective study, it could easily be implemented in clinical routine since almost every patient who is suspected of MD receives a conventional MRI. The output of the multi-layer perceptron classifier provides a value between 0 and 1, which represents the probability of the predicted classes. This will allow clinicians to interpret the probability of having an MD based on the features extracted from MRIs together with the clinical profile of the patients. The potential role of Radiomics, for now, is mainly to aid the clinical diagnosis of MD as a clinical decision support system. However, there lie more perspectives in the future for Radiomics.

In the current study, only patients with MD were included. However, Radiomics might apply to other labyrinthine disorders as well. One study indicated that cochlea CT image features can be useful biomarkers for predicting sensorineural hearing loss in patients treated with chemoradiotherapy for head and neck cancer [39]. It would be valuable to study the relationship between radiomic features and hearing loss in different causes of sensorineural hearing loss. Performing Radiomics in more patients with different disorders will eventually allow a comparison between the general vestibular population and a healthy population. Performing a cluster analysis might reveal a reclassification of vestibular disorders based on similarities in Radiomic signatures. Compared to symptoms, Radiomic signatures might better classify vestibular disorders. Finally, to provide a fully automated computer-aided diagnosis for MD, automated segmentation of the labyrinth on MRI would be the next goal. Several studies already demonstrated the potential of auto-segmentation in medical images [40]. These methods will probably also apply to the labyrinth.

5 CONCLUSION

The automated extraction of Radiomic features from conventional MRI scans proved to be valuable to discriminate between patients with Menière's disease and 'normal' controls. In the current study, the machine learning-based multi-layer perceptron network yielded a precise, high-diagnostic performance in identifying patients with Menière's disease with an accuracy of 82%. In the future, Radiomics could be implemented as a fast, noninvasive, and accurate decision support system, next to clinical evaluation, in the diagnostic trajectory of Menière's disease.

6 REFERENCES

[1] Cairns H. Observations on the pathology of Meniere's syndrome. The Journal of Laryngology & Otology 1980. https://doi.org/10.1017/S002221510008960X.

[2] Merchant SN, Adams JC, Nadol JB. Pathophysiology of Ménière's syndrome: Are symptoms caused by endolymphatic hydrops? Otology and Neurotology 2005. https://doi.org/10.1097/00129492-200501000-00013.

[3] Foster CA, Breeze RE. Endolymphatic hydrops in Ménière's disease: Cause, consequence, or epiphenomenon? Otology and Neurotology 2013. https://doi.org/10.1097/MAO.0b013e31829e83df.

[4] Lopez-Escamez JA, Carey J, Chung WH, Goebel JA, Magnusson M, Mandalà M, et al. Diagnostic criteria for Menière's disease. Journal of Vestibular Research: Equilibrium and Orientation 2015. https://doi.org/10.3233/VES-150549.

[5] Pyykkö I, Nakashima T, Yoshida T, Zou J, Naganawa S. Ménière's disease: A reappraisal supported by a variable latency of symptoms and the MRI visualisation of endolymphatic hydrops. BMJ Open 2013. https://doi.org/10.1136/bmjopen-2012-001555.

[6] Vassiliou A, Vlastarakos P V, Maragoudakis P, Candiloros D, Nikolopoulos TP. Meniere's disease: Still a mystery disease with difficult differential diagnosis. Annals of Indian Academy of Neurology 2011. https://doi.org/10.4103/0972-2327.78043.

[7] Lopez-Escamez JA, Dlugaiczyk J, Jacobs J, Lempert T, Teggi R, von Brevern M, et al. Accompanying symptoms overlap during attacks in Menière's disease and vestibular migraine. Frontiers in Neurology 2014. https://doi.org/10.3389/fneur.2014.00265.

[8] Yamane H, Iguchi H, Konishi K, Sakamaoto H, Wada T, Fujioka T, et al. Three-dimensional cone beam computed tomography imaging of the membranous labyrinth in patients with Meniere's disease. Acta Oto-Laryngologica 2014. https://doi.org/10.3109/00016489.2014.913315.

[9] Naganawa S, Nakashima T. Visualization of endolymphatic hydrops with MR imaging in patients with Ménière's disease and related pathologies: Current status of its methods and clinical significance. Japanese Journal of Radiology 2014. https://doi.org/10.1007/s11604-014-0290-4.

[10] Conte G, Lo Russo FM, Calloni SF, Sina C, Barozzi S, Di Berardino F, et al. MR imaging of endolymphatic hydrops in Ménière's disease: Not all that glitters is gold. Acta Otorhinolaryngologica Italica 2018. https://doi.org/10.14639/0392-100X-1986.

[11] Rose TA, Choi JW. Intravenous Imaging Contrast Media Complications: The Basics That Every Clinician Needs to Know. American Journal of Medicine 2015. https://doi.org/10.1016/j.amjmed.2015.02.018.

[12] Louza J, Krause E, Gürkov R. Hearing function after intratympanic application of gadolinium-based contrast agent: A long-term evaluation. Laryngoscope 2015. https://doi.org/10.1002/lary.25259.

[13] Gulani V, Calamante F, Shellock FG, Kanal E, Reeder SB. Gadolinium deposition in the brain: summary of evidence and recommendations. The Lancet Neurology 2017. https://doi.org/10.1016/S1474-4422(17)30158-8.

[14] 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. European Journal of Cancer 2012. https://doi.org/10.1016/j.ejca.2011.11.036.

[15] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016. https://doi.org/10.1148/radiol.2015151169.

[16] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 2014. https://doi.org/10.1038/ncomms5006.

[17] Cho YS, Choi SH, Park KH, Park HJ, Kim JW, Moon J, et al. Prevalence of otolaryngologic diseases in South Korea: Data from the Korea national health and nutrition examination survey 2008. Clinical and Experimental Otorhinolaryngology 2010. https://doi.org/10.3342/ceo.2010.3.4.183.

[18] Sajda P. MACHINE LEARNING FOR DETECTION AND DIAGNOSIS OF DISEASE. Annual Review of Biomedical Engineering 2006. https://doi.org/10.1146/annurev.bioeng.8.061505.095802.

[19] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 2017. https://doi.org/10.1038/nrclinonc.2017.141.

[20] van den Burg EL, van Hoof M, Postma AA, Janssen AML, Stokroos RJ, Kingma H, et al. An exploratory study to detect ménière's disease in conventional MRI scans using radiomics. Frontiers in Neurology 2016. https://doi.org/10.3389/fneur.2016.00190.

[21] Committee on Hearing and Equilibrium guidelines for the diagnosis and evaluation of therapy in Meniere's disease*. Otolaryngology - Head and Neck Surgery 1995. https://doi.org/10.1016/S0194-5998(95)70102-8.

[22] Egger J, Kapur T, Fedorov A, Pieper S, Miller J V., Veeraraghavan H, et al. GBM volumetry using the 3D slicer medical image computing platform. Scientific Reports 2013. https://doi.org/10.1038/srep01364.

[23] Carré A, 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. Scientific Reports 2020. https://doi.org/10.1038/s41598-020-69298-z.

[24] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Lambin P. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncologica 2017. https://doi.org/10.1080/0284186X.2017.1350285.

[25] Song F, Guo Z, Mei D. Feature selection using principal component analysis. Proceedings - 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, ICSEM 2010, 2010. https://doi.org/10.1109/ICSEM.2010.14.

[26] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics 2015. https://doi.org/10.1155/2015/198363.

[27] Caruana R, Lawrence S, Giles L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in Neural Information Processing Systems, 2001.

[28] Venkatasamy A, Veillon F, Fleury A, Eliezer M, Abu Eid M, Romain B, et al. Imaging of the saccule for the diagnosis of endolymphatic hydrops in Meniere disease, using a three-dimensional T2-weighted steady state free precession sequence: accurate, fast, and without contrast material intravenous injection. European Radiology Experimental 2017. https://doi.org/10.1186/s41747-017-0020-7.

[29] Simon F, Guichard JP, Kania R, Franc J, Herman P, Hautefort C. Saccular measurements in routine MRI can predict hydrops in Menière’s disease. European Archives of Oto-Rhino-Laryngology 2017. https://doi.org/10.1007/s00405-017-4756-8.

[30] Keller JH, Hirsch BE, Marovich RS, Branstetter BF. Detection of endolymphatic hydrops using traditional MR imaging sequences. American Journal of Otolaryngology - Head and Neck Medicine and Surgery 2017. https://doi.org/10.1016/j.amjoto.2017.01.038.

[31] Morin O, Vallières M, Jochems A, Woodruff HC, Valdes G, Braunstein SE, et al. A Deep Look Into the Future of Quantitative Imaging in Oncology: A Statement of Working Principles and Proposal for Change. International Journal of Radiation Oncology Biology Physics 2018. https://doi.org/10.1016/j.ijrobp.2018.08.032.

[32] J.E. van T, R.T.H. L, W. van E, B. R, Lambin P. AO - van Timmeren JE; O http://orcid. org/0000-0002-8166-6853. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncologica 2017. https://doi.org/http://dx.doi.org/10.1080/0284186X.2017.1350285.

[33] Wang S, Yang M, Du S, Yang J, Liu B, Gorriz JM, et al. Wavelet entropy and directed acyclic graph support vector machine for detection of patients with unilateral hearing loss in MRI scanning. Frontiers in Computational Neuroscience 2016. https://doi.org/10.3389/fncom.2016.00106.

[34] Fukushima M, Kitahara T, Oya R, Akahani S, Inohara H, Naganawa S, et al. Longitudinal up-regulation of endolymphatic hydrops in patients with Meniere’s disease during medical treatment. Laryngoscope Investigative Otolaryngology 2017. https://doi.org/10.1002/lio2.115.

[35] Fiorino F, Pizzini FB, Beltramello A, Barbieri F. Progression of endolymphatic hydrops in ménière’s disease as evaluated by magnetic resonance imaging. Otology and Neurotology 2011. https://doi.org/10.1097/MAO.0b013e31822a1ce2.

[36] Attyé A, Barma M, Schmerber S, Dumas G, Eliezer M, Krainik A. The vestibular aqueduct sign: Magnetic resonance imaging can detect abnormalities in both ears of patients with unilateral Meniere's disease. Journal of Neuroradiology 2018. https://doi.org/10.1016/j.neurad.2018.10.003.

[37] Geras KJ. Exploiting diversity for efficient machine learning. 2018.

[38] Yun J, Park JE, Lee H, Ham S, Kim N, Kim HS. Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma. Scientific Reports 2019. https://doi.org/10.1038/s41598-019-42276-w.

[39] Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. Physica Medica 2018. https://doi.org/10.1016/j.ejmp.2017.10.008.

[40] Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging 2019. https://doi.org/10.1007/s10278-019-00227-x.

Chapter 8

Deep learning for the fully automated segmentation of the inner ear on MRI

Based on Vaidyanathan, A., van der Lubbe, M.F.J.A., Leijenaar, R.T.H. et al. Deep learning for the fully automated segmentation of the inner ear on MRI. Sci Rep 11, 2885 (2021). https://doi.org/10.1038/s41598-021-82289-y

In this chapter, we propose a deep-learning approach for the fully automated segmentation of the inner ear in MRI. A 3D U-net was trained on 944 MRI scans with manually segmented inner ears as a reference standard. The model was validated on an independent, multicentric dataset consisting of 177 MRI scans from three different centers. The model was also evaluated on a clinical validation set containing eight MRI scans with severe changes in the morphology of the labyrinth. The 3D U-net model showed precise Dice Similarity Coefficient scores (mean DSC- 0.8790) with a high True Positive Rate (91.5%) and low False Discovery Rate and False Negative Rates (14.8% and 8.49% respectively) across images from three different centers. The model proved to perform well with a DSC of 0.8768 on the clinical validation dataset. The proposed auto-segmentation model is equivalent to human readers and is a reliable, consistent, and efficient method for inner ear segmentation, which can be used in a variety of clinical applications such as surgical planning and quantitative image analysis.

1 BACKGROUND

The inner ear, also known as the labyrinth, is a complex structure located in the temporal bone. It roughly consists of the cochlea, the vestibule, and the semi-circular canals. Understanding changes and variations within these structures can help diagnose and predict several conditions [1], such as inflammatory and neoplastic processes. Technological developments in imaging techniques have allowed (neuro)radiologists to evaluate the human labyrinth, with recent advances increasing the level of detail [1]. Moreover, applications of artificial intelligence and the quantitative assessment of medical images for the non-invasive exploration of anatomical structures and the classification of diseases have remarkably increased in recent years [2].

The process of the automated extraction and analysis of large amounts of quantitative information from medical images is known as *radiomics* [3], [4]. A recent study investigated the value of *radiomics* for the diagnosis of Meniere's disease (MD), an inner ear disorder characterized by episodic vertigo spells, hearing loss, and tinnitus [5]. Other labyrinthine disorders such as sensorineural hearing loss might benefit from quantitative image analysis as well [6].

Image segmentation is a critical step to work toward fully automated diagnostic tools for inner ear disorders. Manual segmentation requires experienced readers, is time-consuming, and is prone to intra-and inter-observer variability [7]–[9]. Over the past years, several automatic and semi-automatic inner ear segmentation methods were proposed for both MRI and CT imaging [10]–[14], including region-growing, thresholding and edge detection [15], model-based [10], [14], atlas-based [12], [13] and machine-learning techniques [11]. The inner ear's small and complex structure makes segmentation challenging, especially in MR imaging due to non-homogenous image intensities [10], [11].

Recent work proposed a statistic shape model (SSM) for inner ear segmentation in MR images 10,14. However, the proposed methodology presents a high computational burden, both in terms of time and cost. Another recently published segmentation model showed very good agreement between an atlas-based segmentation and the manual gold standard [12], yet requires manual intervention. Additionally, the segmentation performance of atlas-based methods decreases for complex structures with variable shapes and sizes [16].

Recent studies have demonstrated the successful application of deep learning techniques for detection, segmentation, and classification tasks in the medical field [17]. Among deep learning techniques, the U-Net architecture is a specific type of convolutional neural network (CNN) consisting of multilayer neural networks. These networks have been implemented successfully, especially for auto-segmentation in medical images[18], [19]. Although U-Net-based deep learning approaches do exist for the segmentation of the inner ear [18], [20], they lack the incorporation of anatomical variations, pathological situations, or missing anatomical structures, which are part of daily clinical practice. Hence, there is currently no fully automated, generic segmentation method for the inner ear to meet the growing demand for developments in 3D visualization and quantitative image assessments.

Therefore, this study's objective was to develop a deep-learning approach for the automatic segmentation of the inner ear in clinical MR images, focusing on the robustness of the method in varying clinical situations, and to evaluate its performance and generalizability with manual segmentation as reference.

2 MATERIAL AND METHODS

2.1 ETHICAL CONSIDERATIONS

This study was performed following the guidelines outlined by Dutch and Belgian legislation. MRI scans were collected and fully anonymized by the local investigators of four centers. The ethics committee of University Hospital Antwerp approved the study (Approval number - 17/09/093) and written informed consent was obtained from the participants. The other centers waived the ethics approval due to the retrospective nature and full anonymization of the data according to the Medical Research Involving Human Subjects Act (WMO).

2.2 AUTOMATIC SEGMENTATION WORKFLOW

The workflow applied in this study consisted of four steps and is illustrated in Figure 8.1. Each step of the workflow is detailed in the following paragraphs.

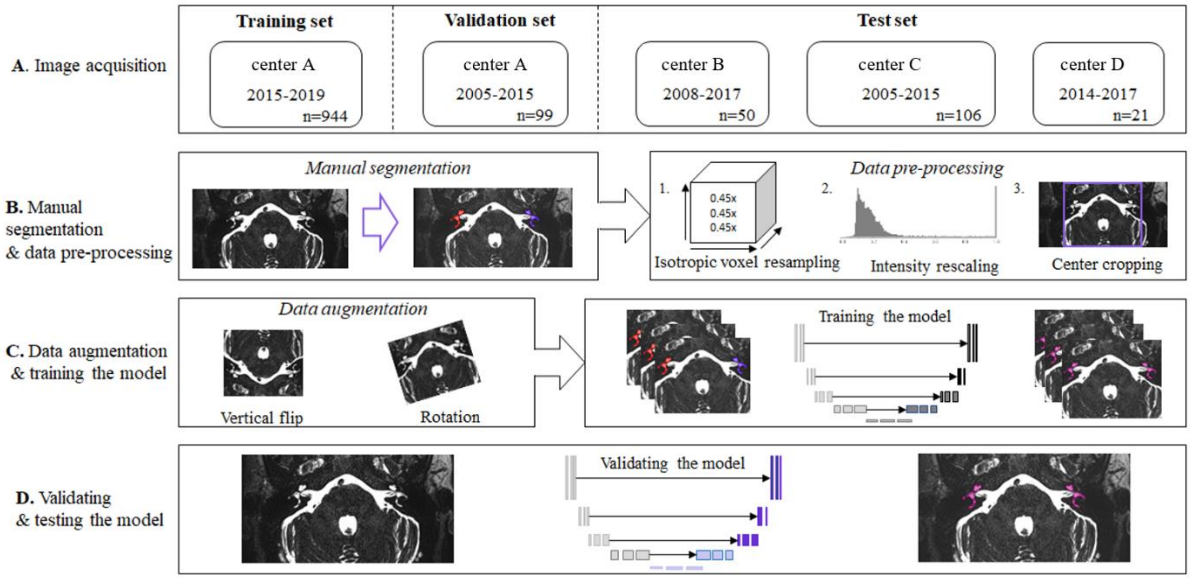


Figure 8.1. The workflow of auto segmentation of the inner ear in this study is graphically presented in four steps. **A**. The image acquisition from four different centers is divided into training, validation, and an independent test set. **B**. Manual segmentation of the labyrinth and pre-processing steps consisting of isotropic voxel resampling, intensity rescaling, and center cropping. **C**. Extending the data set (data augmentation) by flipping and rotating the input images and training the model. **D**. Validation and testing the model on an independent test cohort.

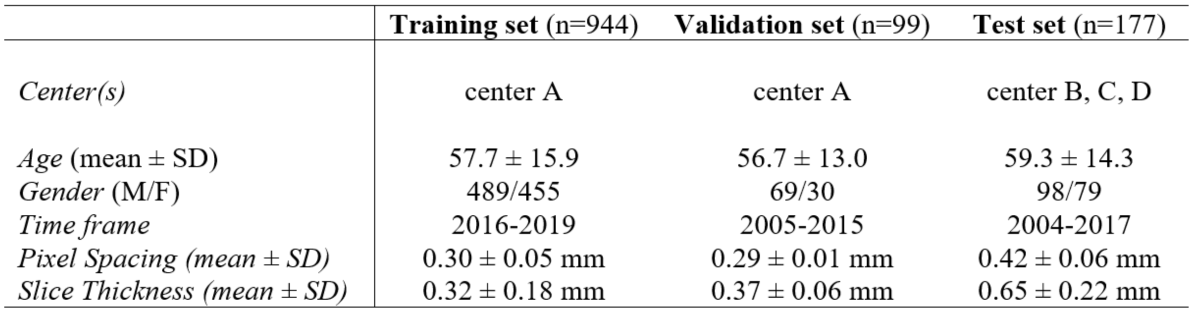
2.3 TRAINING DATASET

A total of 1203 images of patients who underwent an MRI scan of the cerebellopontine angle for diverse neuro-otological indications in the period of December 2015 to April 2019 in Maastricht Medical University center (center A) were collected and fully anonymized. All high-resolution T2-weighted images were acquired in 1.5 and 3 Tesla (T) MRI scanners, from different vendors with a variety of high-resolution T2-weighted sequences (3D cochlea, DRIVE, SPC_TRA_ISO), with local optimized protocols. MRI scans of the cerebellopontine angle were included if they allowed labyrinth visualization with at least a portion of the labyrinth recognizable and suitable for manual segmentation. MRI scans, which did not allow a clear manual segmentation, were excluded from this study. In total, 259 MRI images were excluded due to unsuitable sequences (DWI, T1, SURVEY MST), poor quality, or skewed MR images. The final training dataset included MRI scans of 944 cases (Table 8.1).

2.4 VALIDATION AND TEST DATASET

The validation dataset included MRI scans of 99 cases collected from Maastricht University Medical Center + (center A) in the period from 2005 to 2015 (Table 8.1). MRI scans collected from 3 different centers, University Hospital Antwerp (center B), Viecure Hospital Venlo (center C), and Apeldoorn dizziness center (center D) from 2005 to 2017 (Table 8.1) were used as an independent Test dataset. Both validation and test datasets consisted of T2-weighted MR images of the cerebellopontine angle of patients with uni- or bilateral definitive Meniere's disease and idiopathic asymmetric sensorineural hearing loss.

Table 1. Data set characteristics



	Training set (n=944)	**Validation set** (n=99)	**Test set** (n=177)
Center(s)	center A	center A	center B, C, D
Age (mean ± SD)	57.7 ± 15.9	56.7 ± 13.0	59.3 ± 14.3
Gender (M/F)	489/455	69/30	98/79
Time frame	2016-2019	2005-2015	2004-2017
Pixel Spacing (mean ± SD)	0.30 ± 0.05 mm	0.29 ± 0.01 mm	0.42 ± 0.06 mm
Slice Thickness (mean ± SD)	0.32 ± 0.18 mm	0.37 ± 0.06 mm	0.65 ± 0.22 mm

N = number, SD = standard deviation, M = male, F= female

2.5 MANUAL SEGMENTATION

A team of six readers was trained by the second author (MvdL), an experienced clinician and researcher in inner ear imaging, to manually segment the labyrinth on both sides in 3D Slicer 4.8.1[19]. Manual segmentation was facilitated by intensity-based thresholding and region-growing algorithms. The original MRI scans and the manually segmented masks were visualized by 3D maximum intensity projections as shown in Figure 8.2. This provided an overview of the manually segmented results which allowed for a thorough quality assessment. All segmentations were curated by the experienced reader (MvdL), where any missing or incorrectly segmented masks were re-segmented by the experienced reader (MvdL). The final segmentation results served as the ground truth for training CNN. The independent validation and test datasets were segmented and curated by the experienced reader (MvdL). The resulting manual segmentations on the test dataset were used as the ground truth (reference standard).

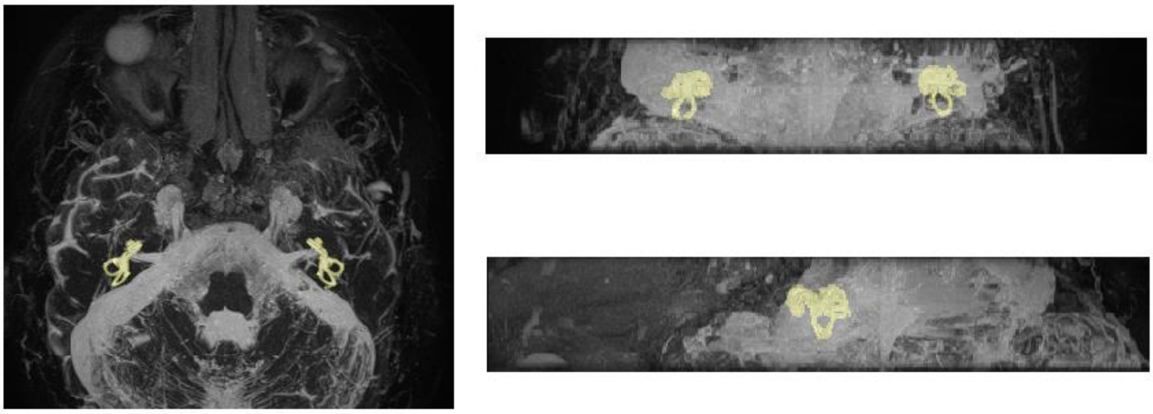


Figure 8.2. Maximum intensity projection of a sample MR in the axial, sagittal, and coronal plane showing a manual segmentation of the labyrinth in yellow. Left: axial plane, right top: Coronal plane, right bottom: sagittal plane

2.6 PRE-PROCESSING

To generate homogeneous MRI volumes as input for the model, the following pre-processing steps were performed. Firstly, all volumes were resampled by B-spline interpolation to an isotropic voxel size of 0.45 mm. Secondly, the intensities of the MRI volumes were normalized to the range [0-1] using the minimum and maximum intensity of each volume. Lastly, since the model's architecture required inputs of the same dimensions, a center crop of 256 x 256 x 64 pixels was obtained from the pre-processed volumes. This crop size was large enough to contain contextual information about the inner ear. Images smaller than 256 x 256 pixels in the transversal plane and 64 pixels in the slice direction were padded with zeros.

2.7 MODEL ARCHITECTURE

The model's architecture is based on a classical 3D U-net [21], as illustrated in Figures 8.3a and 8.3b. It comprises an encoder, a decoder block, and skip connections. The encoder network is a contracting path with convolution layers, which extracts high-level features, decreasing the spatial resolution at each layer. The decoder network is an expanding path, which increases the spatial resolution by up-sampling and uses the feature information to segment the pixels corresponding to the Region of interest. Skip connections, between encoder and decoder, allow retrieval of fine details, which might be lost during spatial down-sampling.

The model's architecture was adapted with attention gates, as the relevant features of the inner ear showed large shape variability and were very small compared to the surrounding structures [22].

The attention gates highlight the regions that correspond to the inner ear and suppress the regions that correspond to the background. The highlighted features are propagated by the skip connections from the deep stages of contracting paths to the expanding paths. More specifically, attention Gates are used to propagate the important spatial information corresponding to the inner ear from the encoding to the decoding part of the model. As shown in Figure 8.3b, the input feature maps from the encoder part of the network are scaled by the attention coefficients generated by the Attention Gates, thereby outputting the features relevant to the inner ear. The scaled features are then concatenated with the up-sampled output feature maps at each level in the decoder part of the network.

Since different components of the inner ear are more easily accessible at different scales, we additionally input the same volume at 3 different scales along the encoder path, which has been previously described as an input image pyramid by Oktay *et al.* [22].

Other network parameter changes included an increase in the number of convolutional filters from 16 to 128 in the encoder network. Each Maxpooling layer reduced the image spatial resolution by a factor of two. Along the decoder path, transposed convolutions were used for up-sampling which increased the image size by a factor of two at each layer. All the convolutional blocks included 3D convolutions [23], ReLu activation [24] and Instance Normalization [25].

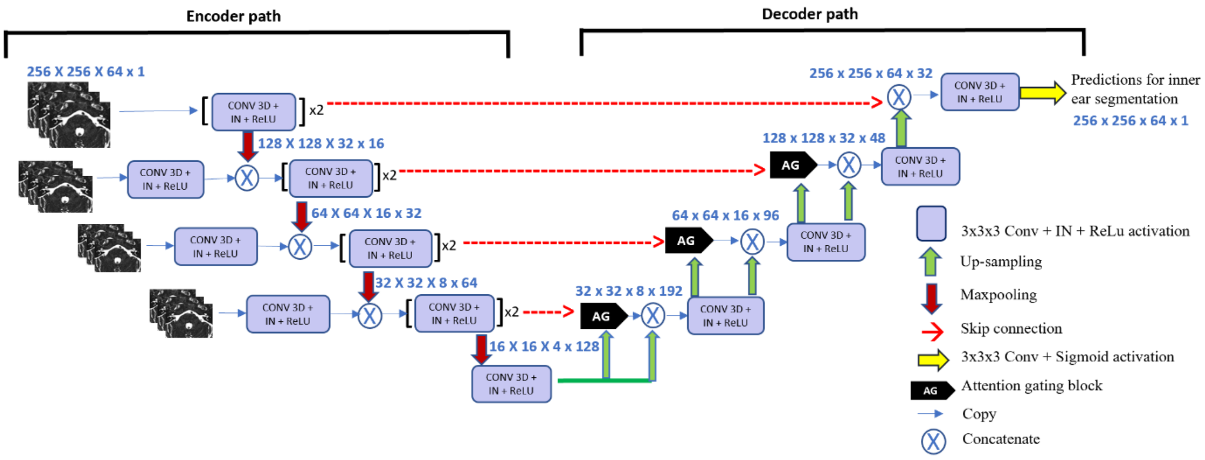


Figure 8.3a. The proposed 3D U-Net-based architecture used in the study. MRI volumes, at multiple scales, were provided as input to the encoder network. The decoder network outputs a score to classify each voxel as an inner ear or not. Notations in blue text (a x a x a x b) highlight the spatial resolution (a x a x a) and the feature map count (b). X = block repetitions, IN = Instance Normalization, Conv = Convolution kernel, ReLU = rectified linear unit, 3x3x3 =the size of the 3D CNN kernels.

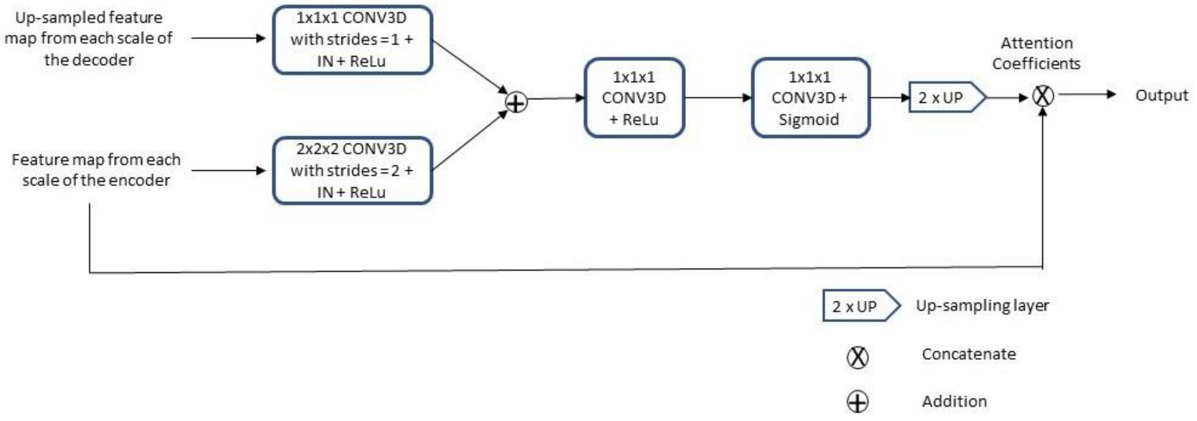

Up-sampled feature map from each scale of the decoder
Feature map from each scale of the encoder
1x1x1 CONV3D with strides = 1 + IN + ReLu
2x2x2 CONV3D with strides = 2 + IN + ReLu
1x1x1 CONV3D + ReLu
1x1x1 CONV3D + Sigmoid
2 x UP
Attention Coefficients
Output
2 x UP
Up-sampling layer
Concatenate
Addition

Figure 8.3b. Components of Attention Gating Block. The block receives as inputs, the up-sampled output feature map at each scale in the decoder and the feature map from each scale in the encoder. Attention coefficients generated, scale the input feature maps from the encoder.

2.8 TRAINING, VALIDATION, AND TESTING

The model was trained with the pre-processed volumes and their corresponding ground truth labels of the training dataset. Randomly selected input volumes were augmented by vertical flipping or rotation during training. The network weights were initialized by using the He-normal initialization method [26] and updated by using the Adam optimizer [27] at an initial learning rate of $1e^{-4}$.

Since the number of positive voxels (i.e. part of the inner ear) and the negative voxels were highly imbalanced, Tversky loss [28] was used as an objective loss function while training the model, which penalized false negatives more than false positives at a false positive penalty score (β) of 0.3 and a false negative penalty score (α) of 0.7. This approach emphasizes learning features corresponding to the positive voxels. The loss was calculated in a mini-batch of two images per iteration and at the end of each epoch, Tversky loss was calculated on the model's predictions on the validation dataset to ensure validation loss convergence (i.e., decrease in validation loss). The final model's performance was evaluated on the multicentric, independent test dataset.

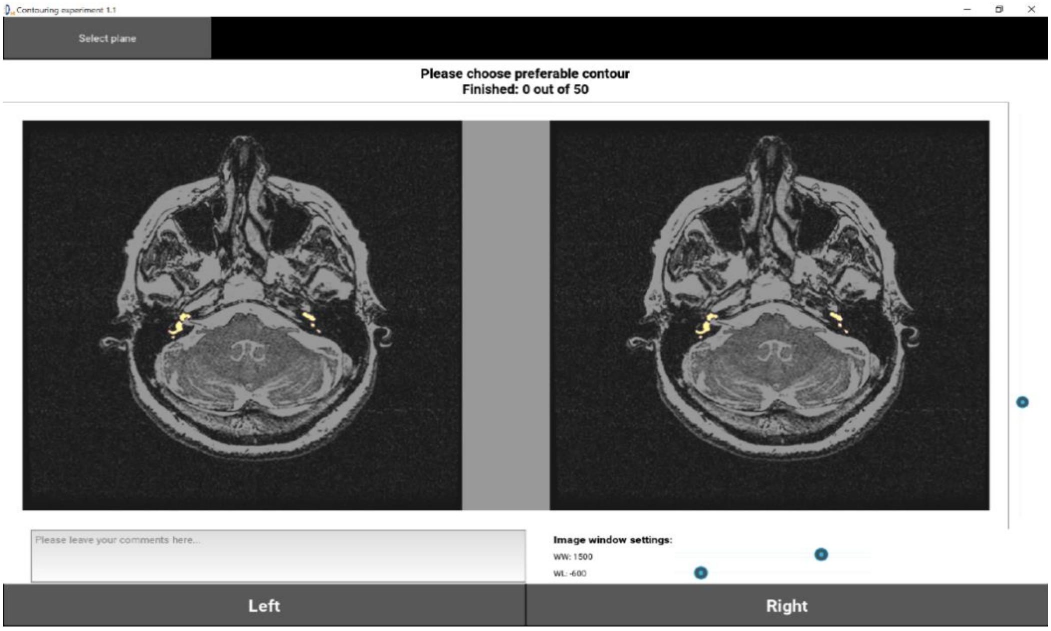
2.9 OUTCOME MEASUREMENTS

The main outcomes of this study were the Dice similarity coefficient (DSC), true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and false discovery rate (FDR)

As a secondary outcome, a subjective evaluation of clinical validation was performed by the second author (MvdL) in consensus with an experienced neuroradiologist (A.A.Postma). Towards clinical implementation, it is critical that a deep learning model can segment the inner ear under all conditions, including those that might alter the shape of the inner ear (e.g., by pathology). Therefore, eight MR images, with their corresponding masks, were selected by the second author (MvdL) in which the signal intensities of the inner ear were altered either by pathology or post-therapeutic changes. These scans were left out of the training dataset and were used for clinical validation of the performance of the model.

2.10 QUALITATIVE ASSESSMENT – IN SILICO CLINICAL STUDY

An in silico clinical study was performed to make a qualitative comparison between manual and model-generated segmentations for 50 MRI volumes randomly selected from the test cohort. In-house developed software was used to display pairs of segmentations (automated vs manual), at randomized screen positions (left or right) blinded to the participants, overlaid on MRI images, as shown in Figure 8.4. The software allowed for scrolling through all image slices and adjustment of window level settings. We enrolled 7 participants (3 computer scientists working in the field of medical imaging and 4 radiologists with an average experience of 2.5 years). For each image, the participants were asked to select their preferred segmentation. For each participant, the qualitative preference score was defined as the percentage of cases with preferred automated segmentation.



Please choose preferable contour
Finished: 0 out of 50

Please leave your comments here...

Image window settings:
WW: 1500
WL: -600

Figure 8.4. Example automated and manual segmentation overlaid on MRI volume as displayed by the software.

3 RESULTS

The final training dataset included MRI scans of 944 cases (489 men, 455 women, aged 41-74; mean age 57 ± 15). The final validation dataset included MRI scans of 99 cases from center A (69 men, 30 women, aged 43-69; mean age 56 ± 13) and 177 cases from centers B, C, and D (79 men, 59 women, aged 45-74; mean age 59 ± 14).

3.1 SEGMENTATION PERFORMANCE

The segmentation accuracy was evaluated against the ground truth by assessing the DSC. DSC measures the overlap between the reference and the model's output. The overall average metrics of segmentation accuracy, DSC, TPR, FNR, FDR, and FPR are summarized in Table 8.2. Figure 8.5 shows a comparison between ground truth volume and predicted true positive volume on the validation and test dataset. Figure 8.6 shows the distribution of DSCs on the validation and test dataset. The correlation between the true positive volume and ground truth volume was also investigated. In Figures 8.7a and 5.7b, agreements for ground truth volume and predicted volume are graphically displayed by Bland–Altman plots. Figures 8.8a and 8.8b show an example of a well-predicted and poorly predicted segmentation.

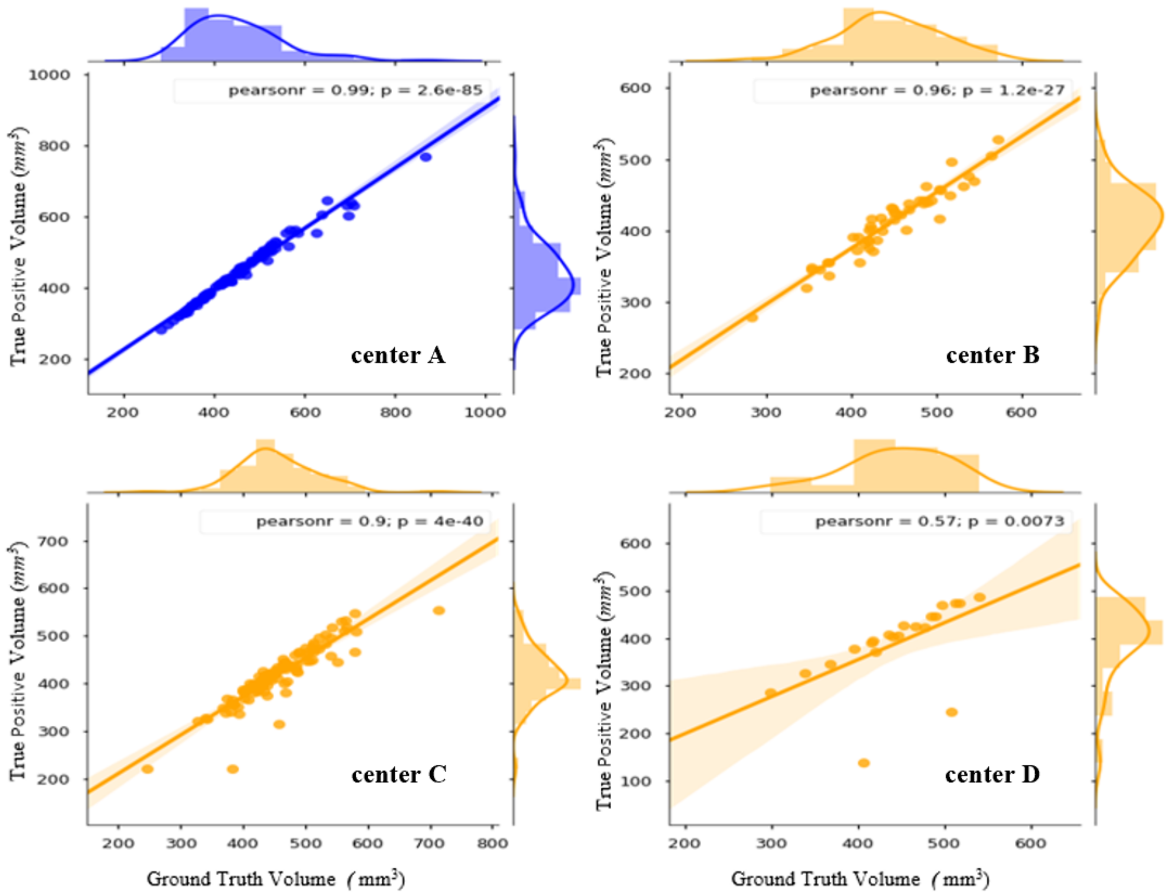


Figure 8.5. The quantitative analysis shows linear correlations between the Ground Truth Volume and the Predicted True Positive Volume for the validation (plot in blue) and the test sets (plots in orange). The plot of center D shows 2 clear outliers which do not fit the trendline. This suggests under-segmentation of the inner ear in 2 cases belonging to the test cohort from Center D.

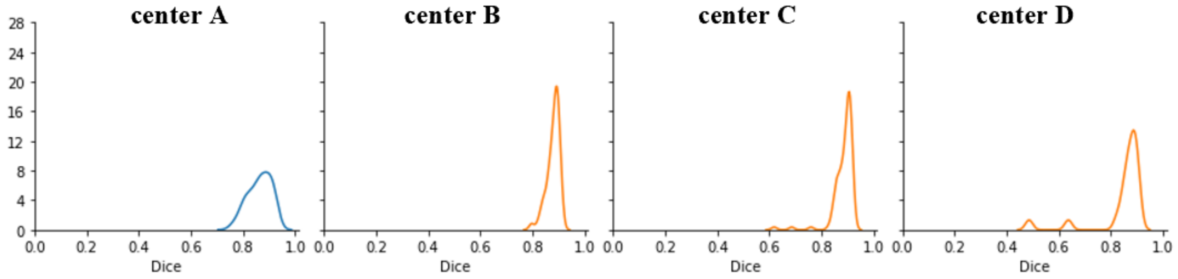


Figure 8.6. Distribution of DSC on the validation (blue curve) and the test dataset (orange curve). The distribution corresponding to Center C and D show outliers (DSC < 0.7) which means less overlap between Ground Truth and predicted segmentation. The distribution also shows that the majority of the predictions have DSC between 0.8 to 1.0.

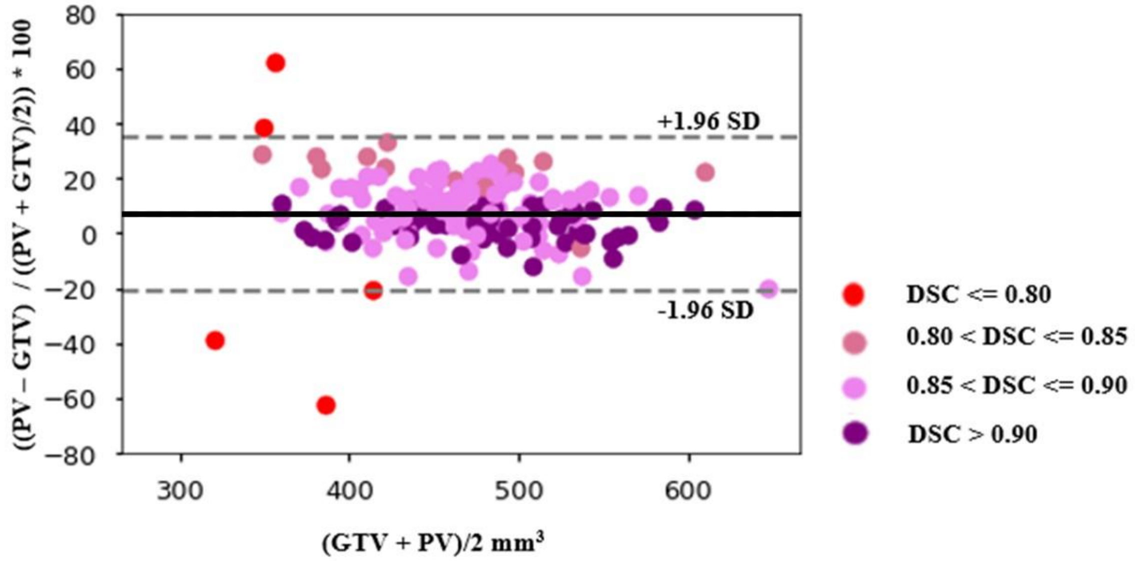


Figure 8.7a. Bland–Altman plot for inner ear volume of the entire test cohort showing percentage difference between Predicted Volume (PV) and Ground Truth Volume (GTV) as a function of the average of Ground Truth and Predicted Volume. The solid line shows the mean difference and the dotted line shows the limits of agreement. PV = Predicted Volume of the inner ear, GTV = Ground truth Volume of the inner ear. The plot shows five clear outliers (Red dots) with three cases that were under-segmented by 20%, 40%, and 60% and two cases that were over-segmented by 40% and 60% respectively. The plot also shows the relationship between the DSC metrics and the level of under/over-segmentation percentage. The outliers correspond to the DSC<=0.80.

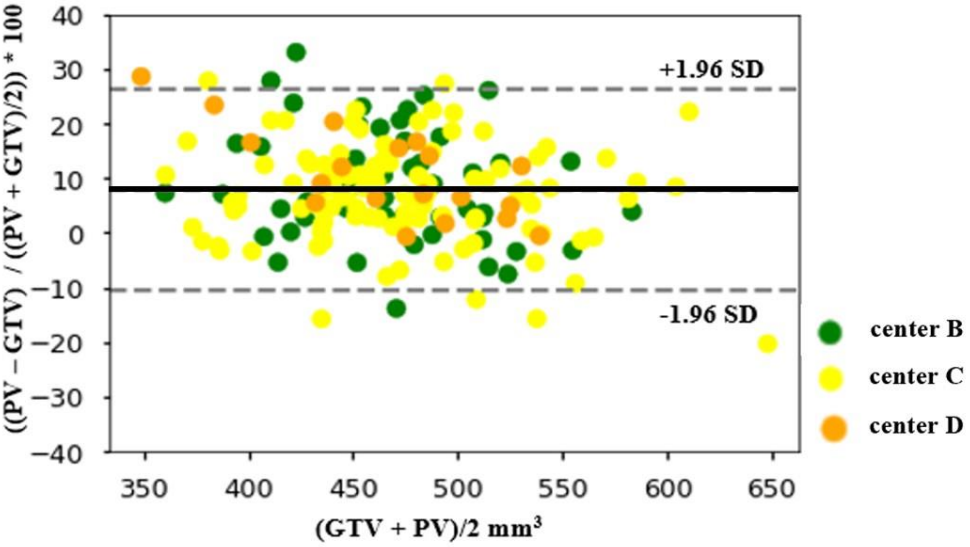


Figure 8.7b. Bland–Altman plot for inner ear volume of the entire test cohort showing percentage difference between Predicted Volume (PV) and Ground Truth Volume (GTV) as a function of the average of Ground Truth and Predicted Volume after excluding the outliers shown in Figure 8.5 (DSC <= 0.80). The solid line shows the mean difference and the dotted line shows the limits of agreement. PV = Predicted Volume of the inner ear, GTV = Ground truth Volume of the inner ear. The plot shows that the model, on average tends to over-segment by 9%.

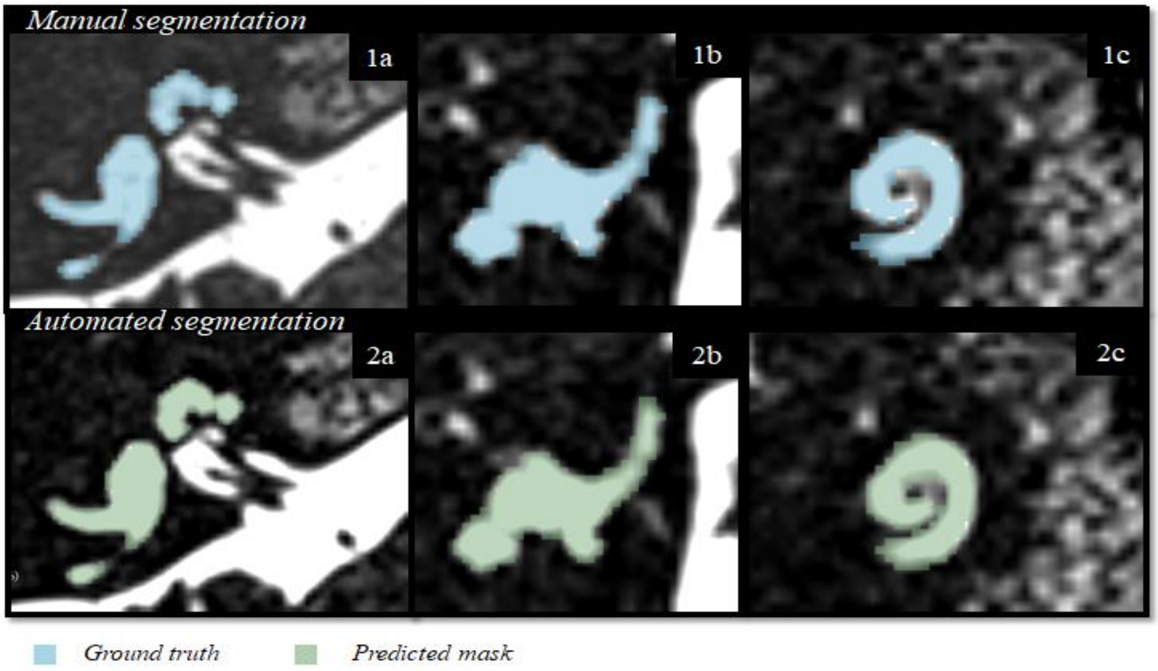


Figure 8.8a. Example of a well-predicted segmentation. The first row denotes the ground truth segmentation. The second row contains the model's segmentation. 1a. Ground truth, axial plane. 1b Ground truth, sagittal plane. 2c. Ground truth, coronal plane. 2a. Predicted mask, axial plane. 2b Predicted mask, sagittal plane. 2c. Predicted mask, coronal plane. DSC: 0.92, Ground Truth Volume: 465.37 mm^3, True Positive Volume: 445.32mm^3, True Positive Rate: 95.69%, False Negative Rate: 4.3%. False Discovery Rate: 11.7%

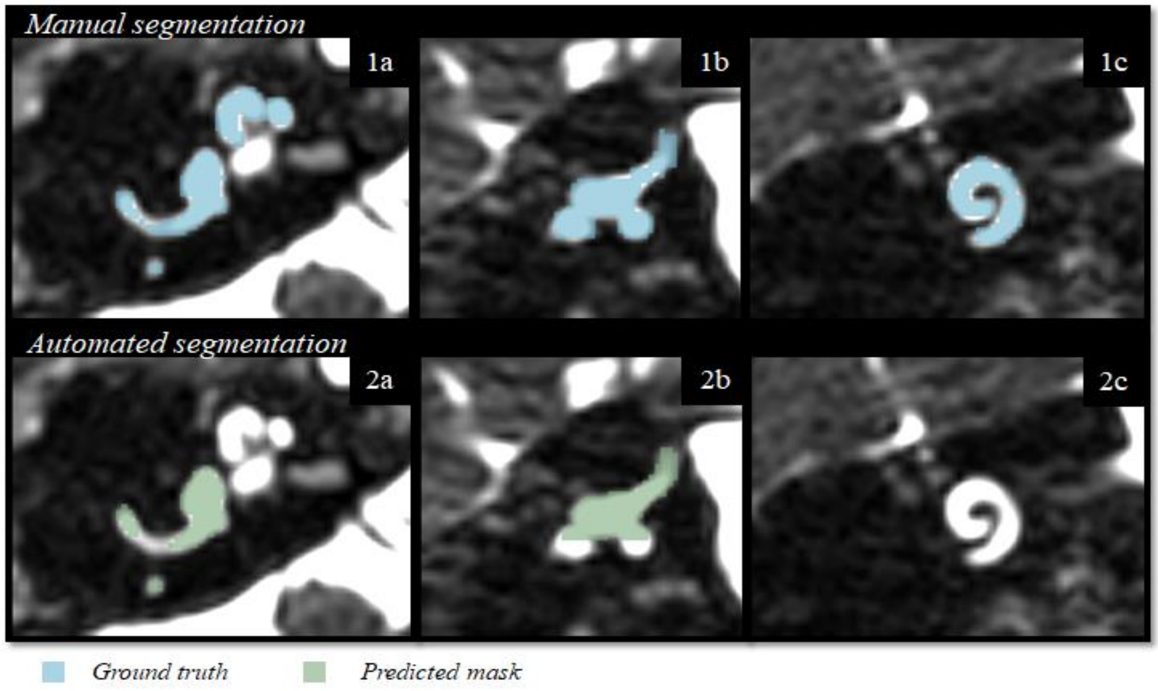


Figure 8.8b. Example of poor segmentation. The first row denotes the ground truth segmentation. The second row contains the model's segmentation. 1a. Ground truth, axial plane. 1b Ground truth, sagittal plane. 2c. Ground truth, coronal plane. 2a. Predicted mask, axial plane. 2b Predicted mask, sagittal plane. 2c. Predicted mask, coronal plane. DSC: 0.48, Ground Truth Volume: 406.05 mm^3, True Positive Volume: 137.96mm^3, True Positive Rate: 33.97%, False Negative Rate: 66.02%. False Discovery Rate: 1.5%

Table 8.2. Performance of the proposed 3D U-Net for the automatic segmentation of the inner ear

	Validation cohort	**Test cohort**
Manual vs. Fully automated	*99 vs. 99*	*177 vs. 177*
DSC	0.86 (CI = 0.85-087)	0.87 (CI = 0.87-0.88)
True Positive Volume (mm^3)	441 (CI = 424-459)	412 (CI = 403-421)
False Positive Volume (mm^3)	123 (CI = 113-134)	72 (CI = 67-76)
False Negative Volume (mm^3)	12 (CI = 8-16)	39 (CI = 34-44)
True Positive Rate (%)	97.7 (CI = 97.2-98.3)	91.50 (CI = 90-92.5)
False Discovery Rate (%)	21.8 (CI = 21.3-22.2)	14.8 (CI = 14.2-15.4)
False Negative Rate (%)	2.2 (CI = 1.6-2.7)	8.5 (CI = 7.4-9.6)

True Positive Volume: the volume correctly segmented as the inner ear, False Negative Volume: the volume incorrectly not segmented as the inner ear (under segmentation). False Positive Volume: the volume incorrectly segmented outside the inner ear (over-segmentation) True Positive Rate: the percentage of voxels correctly segmented as the inner ear, False Discovery Rate: the percentage of voxels incorrectly segmented outside the inner ear (over-segmentation), False Negative Rate: the percentage of voxels incorrectly not segmented as the inner ear (under segmentation), CI: 95% Confidence Interval.

3.2 PERFORMANCE ON THE CLINICAL VALIDATION DATASET

On the held-out clinical validation dataset, the model achieved an average DSC of 0.876, TPR of 87.86%, FDR of 15.2%, and FNR of 12.13%. The automated segmentations on this dataset are included in Appendix 8.1. It includes labyrinths in which parts of the semi-circular canals, the vestibule, or the cochlea were missing or not properly displayed. As an example, an MRI scan with vestibular schwannoma (a tumorous process growing from the vestibular nerve) was included in Figures 8.9a and 8.9b.

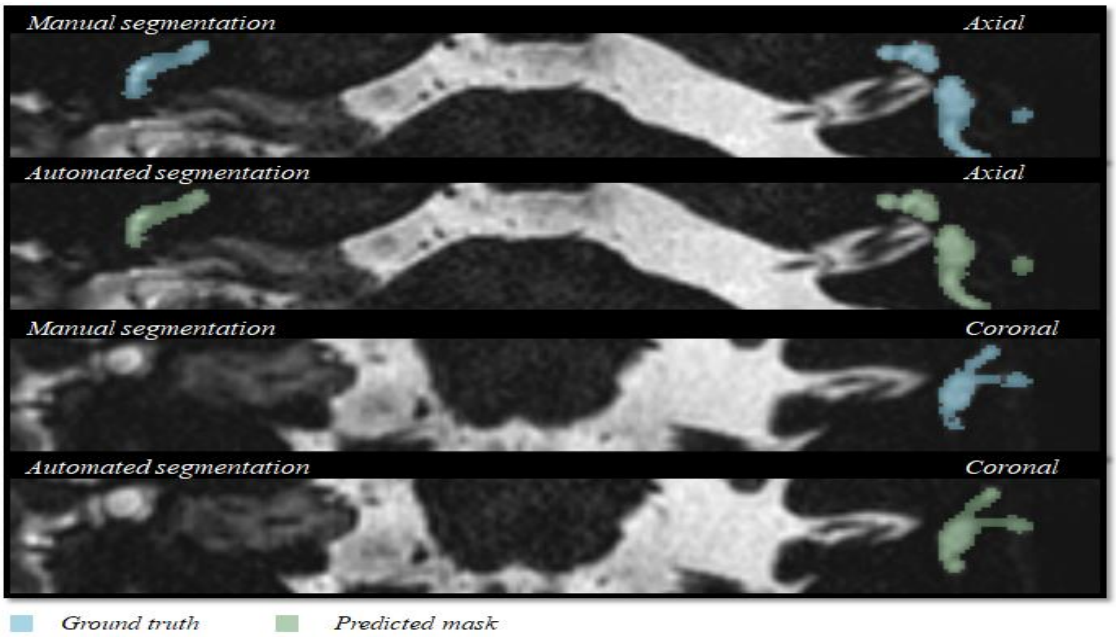


Figure 8.9a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows the presence of a vestibular schwannoma after a translabyrinthine resection on the right side. Therefore, the right semi-circular canals and vestibule are not segmented. DSC: 0.8973, Ground Truth Volume: 316.11 mm^3,

True Positive Volume: 294.69mm^3, True Positive Rate: 93.22%, False Negative Rate: 6.77%. False Discovery Rate: 7.3%

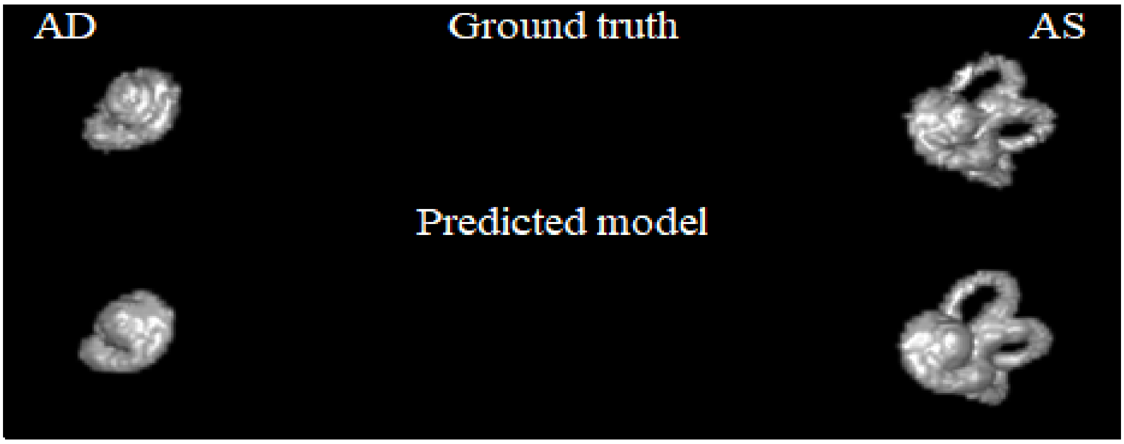


Figure 8.9b. The 3D volume rendering of the ground truth and the predicted mask. The semi-circular canals and the vestibule of the right inner ear were not displayed on MRI. The model has correctly not segmented the semi-circular canals and the vestibule. AD= auriculum dextra, AS=auriculum sinistra

3.3 QUALITATIVE ASSESSMENT - IN SILICO CLINICAL STUDY

On average, the participants preferred automated segmentation in 67% of the cases. A paired one-sided t-test for the hypothesis that this average score is greater than 50% was significant ($p= 9.82474e^{-17}$), indicating that expert users preferred the segmentations generated by the proposed model over the manual segmentations.

4 DISCUSSION

In this work, the first proof-of-concept of an artificial intelligence-based model for the fully automatic segmentation of the inner ear on MRI was demonstrated and validated.

The proposed model showed high performance, with a mean DSC of 0.87 between the manual and the automated segmentation validated across images from three different centers. The mean TPR of 91.5% implies accurate segmentation of the inner ear without significant over or under-segmentation as indicated by the low FDR and FNR metrics (14.8% and 8.49% respectively).

The *in silico* based qualitative analysis showed that on average, the expert users (radiologists and computer scientists) are more likely to prefer model-generated segmentations over manual segmentations. The Bland-Altman plot (Figure 8.6) shows 5 outliers. The fact that the slice thicknesses of those scans were high (mean slice thickness – 1.1 mm) compared to the mean slice thickness of the training cohort (0.32 mm) could explain the miss-segmentations for these cases. Also, the three scans from Center D contained either moving artifacts or had tight margins around the labyrinth, which might explain the lower limits of agreements. All the MRI scans of Center C had noticeably more hyperintense areas at the apex of the pas petrosa compared to the other centers. Although this might have 'challenged' the model, it does not explain why only two out of 21 scans had lower DSCs.

A prior study, that used deep learning to facilitate the auto-segmentation of the inner ear, compared the performance of a 3D Fully Connected Network (FCN) to a 2D-FCN [20]. The study reported an overall DSC of 0.66 and 0.58 when using 3D-FCN and 2D-FCN, respectively. Another recent study reported a high DSC of 0.95 using an SSMs-based level set [10]. However, their model was evaluated on a small dataset (10 cases out of 23 cases were held out for testing) and no independent validation was performed. Directly comparing the present approach with the already published methods in terms of DSC is not possible due to differences in datasets. Nevertheless, it is worth noting that our presented method achieves a state-of-the-art performance, which can be ascribed to the robust deep learning approach combined with a wide and varied dataset, both for training and validation, an aspect often neglected in similar studies.

There are several important strengths of this study. First of all, the model was trained on a diverse set of MR images of the cerebellopontine region. Although all MR images of the training dataset were collected in one center, they were acquired over a wide time span (2015-2019) and include different acquisition and reconstruction protocols [29]. Next to this, the training dataset was manually segmented by five independent readers. Therefore, the model learned to eliminate noise in the manually segmented labels caused by inter-reader variability. These methodological aspects resulted in a well-generalizable model, which is reflected in the high-validation performance. Past studies have shown high inter-reader and intra-reader variability in medical image segmentation tasks [30], [31]. Our method's consistency (i.e., no segmentation variability) alleviates this issue. Additionally, the interaction time was approximately 10 minutes per case for manual segmentation by an experienced reader compared to only 6.5 seconds for automatic segmentation.

One of the most important strengths of this study is the evaluation of the MR images containing deviant morphological shapes and decreased signal intensities of the labyrinth caused by cerebellopontine pathology. On this held-out clinical validation dataset, the model proved to generalize well with an average DSC and TPR of 0.8768 and 87.86% respectively. So far, previous auto-segmentation studies have trained their models on normal ears or small datasets 10–14. To the best of our knowledge, our study is the first to assess generalizability concerning pathologies.

4.1 LIMITATIONS

Several limitations of this study should be noted. First of all, the most important limitation is the lack of a gold standard for manual segmentations from highly experienced neuroradiologists. Due to the extent of the segmentation process, manual segmentation of approximately 1500 labyrinths by one or more senior radiologists was not feasible. Therefore, in this study, the authors chose to work with independent readers who were trained and supervised by an experienced clinical researcher in inner ear imaging (MvdL) to generate the first proof of concept. This could have induced noise in the manual segmentations. Also, the intra- and inter-observer variability of the segmentation team was not evaluated. Although manual segmentation was performed under the strict supervision of the second author and a curating process was performed to detect incorrectly segmented masks, the quality of the manual segmentation could not be fully guaranteed. Since the manually segmented masks were considered the reference standard for the evaluation of the model, lower DSC scores might have indicated better automated segmentation compared to manual segmentation.

Nevertheless, efforts have been made to contain this limitation by training a deep learning architecture with a large number of parameters and applying Early Stopping to prevent overfitting on the noise in the manual segmentation. Previous studies have proved that overparameterized networks are more robust against noisy labels when Early Stopping is applied [32].

Given the very small area occupied by the inner ear in the whole MRI volume, the performance of our model might be further improved by applying bounding box detection [33] or shape identification [34] before automated segmentation, especially for abnormal cases.

Secondly, poor generalizability is the most common problem of deep learning models [29]. In this study, attempts were made to prevent overfitting by training the model on a large dataset from one center and testing its generalizability by holding out 3 independent validation cohorts. Although the overall DSC scores were markedly high, the model performed poorly and failed to generalize in five cases out of 177 (3 cases from center C and 2 cases from center D had DSC < 0.70). This situation could have been mitigated by training the model on all of the centers. This would have made the training dataset more diverse (e.g., in terms of image acquisition and reconstruction) and the model's performance could have been evaluated by cross-validation techniques (i.e., holding out 20-30% of the data from each center for a single validation test data). However, this would degrade the credibility of the generalizability of the model due to concerns regarding overfitting.

Lastly, the model was trained and evaluated on datasets that included only the Dutch and Belgian populations. The generalizability of the model on MRI images from an international cohort is currently unexplored.

4.2 CLINICAL IMPLICATIONS AND FUTURE PERSPECTIVES

The future clinical advantages of automated 3D image segmentation of the inner ear are versatile. Image segmentation can be used for 3D visualization, allowing a better understanding of the spatial relations and morphological changes within the inner ear, assisting radiologists in the diagnostic process, and providing tools for surgical planning [35] or learning purposes [36]. Previous studies have proven the usability of auto-segmentation for pre-operative planning of cochlear implant surgery using CT imaging [37] and for the diagnosis of adolescent idiopathic scoliosis using MRI imaging [11]. Our model proved to be efficient in MRI imaging. However, the proposed methodology can be easily leveraged for similar auto-segmentation applications on different imaging modalities.

Nowadays, quantitative analysis of the inner ear is gaining more importance. Techniques like radiomics [6], volumetric assessment of fluid compartments in the labyrinth [12], [38], and the analysis of the morphoanatomy for the vestibular system [11] are used to aid the diagnosis of vestibular diseases. *Radiomics* refers to the process of the automated extraction and analysis of large amounts of quantitative features from medical images. These features are sometimes not perceptual for the human eye and might contain information that reflects underlying tissue heterogeneity and pathophysiology [4], [39]. Quantitative image features involve descriptors of shape, size, volume, intensity distributions, and texture heterogeneity patterns [39].

A histological feature strongly associated with Meniere's disease is endolymphatic hydrops (EH), distension of the endolymphatic compartment in the inner ear [40]. In conventional MRI, the endolymphatic compartment cannot be distinguished from the perilymphatic compartment, and thus, EH is not depicted [41]. The differences found in radiomic features between MD and controls could hypothetically be explained by the different compositions of the fluids in the labyrinth, causing a different distribution of signal intensities [5]. Possibly, EH is captured in the quantitative image features due to damage to or morphological changes to the endolymphatic space. Since Meniere's disease is still a clinical diagnosis challenge [42], discovering distinctive image features might benefit the diagnostic trajectory of MD. Another study showed that cochlea CT image features can be useful biomarkers for predicting sensorineural hearing loss in a patient with head and neck cancers who received chemoradiation therapy [6]. Different machine learning methods were used for feature selection, classification, and prediction. The advantage of using machine learning in combination with radiomics is that the analysis of the labyrinth could be done autonomously in the future [5]. However, for both studies, setting a Region Of Interest (ROI) by manual segmentation was necessary. The fully automated segmentation of the inner ear contributes to efficient research on quantitative image analysis of the inner ear.

Next to analyses of conventional MRI and CT imaging, the volumetric assessment of fluid compartments in the labyrinth is also promising for vestibular research [38]. Contrast-enhanced MR imaging allows the in vivo confirmation and quantification of endolymphatic hydrops [12], [43].

Several studies investigated the value of the 3D volumetric assessment of the endolymphatic space (ELS) to better monitor EH in vivo, for example in therapeutic trials in Meniere's disease, and to better compare the ELS in patients with different otological diseases [38], [44], [45]. However, the 3D reconstruction was rendered semi-automatic. Due to this time-consuming process, the applications for volumetric assessment are yet more scientifically than clinically relevant. A recent study proposed atlas-based segmentation for the volume-based quantification of the fluid spaces of the inner ear [12]. Which created fast, standardized (auto)segmentation. Further research is necessary to explore the option of the proposed U-net model can be leveraged for contrast-enhanced imaging as well, as to facilitate the volumetric assessment of the ELS in clinics.

Auto-segmentation in its current form is a step towards fully automated diagnostic tools for inner ear disorders.

5 CONCLUSION

In this study, a working first proof-of-concept is demonstrated regarding the fully automatic segmentation of the inner ear using deep learning. Overall, the proposed auto-segmentation model is equivalent to manual segmentation and is a reliable, consistent, and efficient method for inner ear segmentation which can be used in a variety of clinical applications, such as 3D visualization, surgical planning, and quantitative image analysis. Auto-segmentation of the inner ear in its current form might open doors toward automated diagnostic tools for inner ear disorders.

6 REFERENCES

[1] I. Pyykkö, J. Zou, R. Gürkov, S. Naganawa, and T. Nakashima, "Imaging of temporal bone," *Advances in Oto-Rhino-Laryngology*, 2019, doi: 10.1159/000490268.

[2] M. Sollini, L. Antunovic, A. Chiti, and M. Kirienko, "Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics," *European Journal of Nuclear Medicine and Molecular Imaging*, 2019, doi: 10.1007/s00259-019-04372-x.

[3] V. Kumar *et al.*, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012, doi: https://doi.org/10.1016/j.mri.2012.06.010.

[4] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data.," *Radiology*, 2016, doi: 10.1148/radiol.2015151169.

[5] E. L. van den Burg *et al.*, "An exploratory study to detect ménière's disease in conventional MRI scans using radiomics," *Frontiers in Neurology*, vol. 7, no. NOV, 2016, doi: 10.3389/fneur.2016.00190.

[6] H. Abdollahi, S. Mostafaei, S. Cheraghi, I. Shiri, S. Rabi Mahdavi, and A. Kazemnejad, “Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study,” *Physica Medica*, vol. 45, pp. 198–204, 2018, doi: 10.1016/j.ejmp.2017.10.008.

[7] N. Nogovitsyn *et al.*, “Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants,” *NeuroImage*, vol. 197, pp. 589–597, 2019, doi: https://doi.org/10.1016/j.neuroimage.2019.05.017.

[8] K. Men *et al.*, “Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning.,” *Physica Medica*, vol. 50, pp. 13–19, Jun. 2018, doi: 10.1016/j.ejmp.2018.05.006.

[9] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, “Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions,” *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, 2017, doi: 10.1007/s10278-017-9983-4.

[10] S. Zhu, W. Gao, Y. Zhang, J. Zheng, Z. Liu, and G. Yuan, “3D automatic MRI level set segmentation of inner ear based on statistical shape models prior,” 2018. doi: 10.1109/CISP-BMEI.2017.8301973.

[11] L. Shi *et al.*, “Automatic MRI segmentation and morphoanatomy analysis of the vestibular system in adolescent idiopathic scoliosis,” *NeuroImage*, 2011, doi: 10.1016/j.neuroimage.2010.04.002.

[12] V. Kirsch, F. Nejatbakhshesfahani, S. A. Ahmadi, M. Dieterich, and B. Ertl-Wagner, “A probabilistic atlas of the human inner ear’s bony labyrinth enables reliable atlas-based segmentation of the total fluid space,” *Journal of Neurology*, 2019, doi: 10.1007/s00415-019-09488-6.

[13] K. A. Powell, T. Liang, B. Hittle, D. Stredney, T. Kerwin, and G. J. Wiet, “Atlas-Based Segmentation of Temporal Bone Anatomy,” *International Journal of Computer Assisted Radiology and Surgery*, 2017, doi: 10.1007/s11548-017-1658-6.

[14] F. A. Reda, J. H. Noble, R. F. Labadie, and B. M. Dawant, “An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-implantation CT,” 2014. doi: 10.1117/12.2043260.

[15] C. Todd, M. Kirillov, M. Tarabichi, F. Naghdy, and G. Naghdy, “An analysis of medical image processing methods for segmentation of the inner ear,” 2009.

[16] N. Sharma *et al.*, “Automated medical image segmentation techniques,” *Journal of Medical Physics*, 2010, doi: 10.4103/0971-6203.58777.

[17] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[18] J. Fauser *et al.*, “Toward an automatic preoperative pipeline for image-guided temporal bone surgery,” *International Journal of Computer Assisted Radiology and Surgery*, 2019, doi: 10.1007/s11548-019-01937-x.

[19] A. Fedorov *et al.*, “3D Slicer as an image computing platform for the Quantitative Imaging Network,” *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[20] Z. Gong, X. Li, L. Zhou, and H. Zhang, “A 3D Fully Convolutional Network Based Semantic Segmentation for Ear Computed Tomography Images,” 2019. doi: 10.1109/CISP-BMEI.2018.8633242.

[21] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-net: Learning dense volumetric segmentation from sparse annotation,” 2016. doi: 10.1007/978-3-319-46723-8_49.

[22] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” 2018.

[23] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition.”

[24] A. M. Fred Agarap, “Deep Learning using Rectified Linear Units (ReLU).”

[25] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance Normalization: The Missing Ingredient for Fast Stylization,” 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” 2015. doi: 10.1109/ICCV.2015.123.

[27] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2015.

[28] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention u-net for lesion segmentation,” 2019. doi: 10.1109/ISBI.2019.8759329.

[29] R. Caruana, S. Lawrence, and L. Giles, “Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping,” 2001.

[30] B. Norman, V. Pedoia, and S. Majumdar, “Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry,” *Radiology*, vol. 288, no. 1, pp. 177–185, 2018, doi: 10.1148/radiol.2018172322.

[31] K. Yasaka, H. Akai, O. Abe, and S. Kiryu, “Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study,” *Radiology*, vol. 286, no. 3, pp. 887–896, 2018, doi: 10.1148/radiol.2017170706.

[32] M. Li, M. Soltanolkotabi, and S. Oymak, “Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks,” *arXiV*, 2019.

[33] X. Zhuang *et al.*, “Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge,” *Medical Image Analysis*, vol. 58, p. 101537, 2019, doi: https://doi.org/10.1016/j.media.2019.101537.

[34] G. Yang *et al.*, “Simultaneous left atrium anatomy and scar segmentations via deep learning in multiview information with attention,” *Future Generation Computer Systems*, vol. 107, pp. 215–228, 2020, doi: https://doi.org/10.1016/j.future.2020.02.005.

[35] F. Heutink *et al.*, “Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images,” *Computer Methods and Programs in Biomedicine*, vol. 191, p. 105387, 2020, doi: https://doi.org/10.1016/j.cmpb.2020.105387.

[36] A. Ferreira, F. Gentil, and J. M. R. S. Tavares, “Segmentation algorithms for ear image data towards biomechanical studies,” *Computer Methods in Biomechanics and Biomedical Engineering*, 2014, doi: 10.1080/10255842.2012.723700.

[37] J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, “Automatic segmentation of intracochlear anatomy in conventional CT,” *IEEE Transactions on Biomedical Engineering*, 2011, doi: 10.1109/TBME.2011.2160262.

[38] R. Gürkov *et al.*, “MR volumetric assessment of endolymphatic hydrops,” *European Radiology*, 2015, doi: 10.1007/s00330-014-3414-4.

[39] P. Lambin *et al.*, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, doi: 10.1016/j.ejca.2011.11.036.

[40] S. N. Merchant, J. C. Adams, and J. B. Nadol, “Pathophysiology of Ménière’s syndrome: Are symptoms caused by endolymphatic hydrops?,” *Otology and Neurotology*. 2005. doi: 10.1097/00129492-200501000-00013.

[41] R. K. Lingam, S. E. J. Connor, J. W. Casselman, and T. Beale, “MRI in otology: applications in cholesteatoma and Ménière’s disease,” *Clinical Radiology*. 2018. doi: 10.1016/j.crad.2017.09.002.

[42] J. A. Lopez-Escamez *et al.*, “Diagnostic criteria for Menière’s disease,” *Journal of Vestibular Research: Equilibrium and Orientation*, 2015, doi: 10.3233/VES-150549.

[43] S. Naganawa and T. Nakashima, “Visualization of endolymphatic hydrops with MR imaging in patients with Ménière’s disease and related pathologies: Current status of its methods and clinical significance,” *Japanese Journal of Radiology*. 2014. doi: 10.1007/s11604-014-0290-4.

[44] G. Homann *et al.*, “Semi-quantitative vs. volumetric determination of endolymphatic space in Menière’s disease using endolymphatic hydrops 3T-HR-MRI after intravenous gadolinium injection,” *PLoS ONE*, 2015, doi: 10.1371/journal.pone.0120357.

[45] H. Inui, T. Sakamoto, T. Ito, and T. Kitahara, “Magnetic resonance-based volumetric measurement of the endolymphatic space in patients with Meniere’s disease and other endolymphatic hydrops-related diseases,” *Auris Nasus Larynx*, 2019, doi: 10.1016/j.anl.2018.11.008.

Chapter 9

Discussion

1 GENERAL DISCUSSION

The research work presented in this thesis validates the overall hypothesis that **(semi) automated Radiomics and AI-based methodologies can produce generalizable performance, overall equivalent to that of an expert human charged with the same tasks and is exemplified in detection, diagnosis, and treatment response prediction use cases.** In chapters 2 and 3, we validated the hypothesis, benchmarked against RT-PCR confirmed cases of COVID-19, and proved superior performance compared to radiologists while considering the false negative RT-PCR rates [1], [2] and high variability among the radiologists in the differential diagnosis of COVID-19 from other pneumonias[3]. In chapter 4, we validated the hypothesis and proved equivalent performance compared to radiologists' performance on detection of embolism on CTPA [4], [5] [6]. In chapter 6, we validated the hypothesis and proved the superior performance of the model for metastasis detection on bone-scintigraphy data against human readers by performing an in-silico trial on the task. In chapter 5, we validated the hypothesis and proved superiority in the prognostic performance of the radiomics signature compared with the TNM staging (6th edition). Chapter 8 was a feasibility study proving the hypothesis for Meniere's disease diagnosis compared to the other clinically prevalent diagnostic methods [7].

Although the current scenario on AI models proved to be equivalent to expert human readers, the new hypothesis is that in the longer term, a combination of AI and an expert reader will surpass the individual performances of an AI and an expert reader. The same hypothesis has already been tested on several use cases in recent articles on breast cancer screening [8] [9] [10].

1.1 METHODOLOGY

Methodologically, chapters 2 and 5 present a specific type of radiomics pipeline which involves combining an AI-based automatic segmentation model and a statistical model to analyze features extracted from the segmented region. Furthermore, chapters 3, 4, and 6 present a type of radiomics pipeline involving an AI-based automatic segmentation model in combination with a Deep learning (DL)-based classifier for predicting clinical outcomes. Such pipelines allow for more flexibility over reusing the AI-based segmentation models for analyses involving different objectives (clinical outcomes). Lastly, chapter 7 describes the radiomics pipeline involving manual intervention for segmentation of the inner ear on MRI before analysis of features extracted from the segmented region. A substantial manual effort was needed in such an analysis and that was the motivation for the research work in chapter 8, particularly focussing on a methodology for AI-based automatic segmentation of the inner ear on MRI imaging. Radiomic pipelines involving manual intervention in segmenting region of interest also suffers from inconsistencies in the quality of segmentations. Research in chapter 8 proves that hypothesis through a qualitative in-silico study where the expert radiologists preferred segmentations generated by a trained AI compared to manual segmentations.

1.1.1 Auto-segmentation models

The research works presented in this thesis also highlight the utility of several auto-segmentation models. For instance, chapter 2 uses a lung segmentation model to extract features from the lungs. Chapter 3 uses a combination of the whole lungs and lung abnormalities model to select the axial slices containing lung abnormalities and in chapter 4, the lung segmentation model was used to crop the region around the lungs, to facilitate the model, learning features from relevant regions on lungs. In chapter 5, the lung tumor segmentation model was used to automatically segment the region for extracting radiomic features. Overall, this thesis also proves the importance of AI-based automatic segmentation models for fully automatizing the radiomic pipeline.

1.1.2 Generalizability

Attempts were made to validate the methods presented in this thesis specifically to prove generalizability. Generalizability is an essential aspect of the usability of a model in a real-world clinical setting where data variability and variety are inevitable. For instance, in chapter 8, the trained model was validated on a held-out test set containing images from 4 different centers where the distribution of acquisition protocols (pixel spacing and slice thickness) differed from that of the training dataset. Added to that, the model was also validated on a clinical validation cohort containing images with diverse comorbidities. In chapter 6, the classifier model was evaluated on a held-out dataset coming from a different center than that of the training dataset. Similarly, in chapter 4, the trained model was evaluated on datasets from 3 different centers with diverse and varying imaging parameters from that of the training dataset.

1.1.3 Explainability

Throughout my research, attempts were made to explain decisions made by the trained AI-based models. For instance, in chapter 6, a model trained to classify bone scintigraphy images with metastasis was explored for the detection of metastatic spots using activation maps extracted from the CNN-based model. However, we lacked ground truth to validate the detections derived from the activation maps. In chapter 4, the model which was trained to classify CT axial slices containing pulmonary embolism was explored for the detection of regions of pulmonary embolism using the activation maps extracted from the CNN-based model and we have also validated the performance on a limited test set. Similar approaches couldn’t be employed for other studies presented due to the technical complexity of the model architecture. For instance, in chapter 3, a 3D CNN, based on inception model architecture [11] was used for a differential diagnosis application. There were two challenges when applying 3D class activation mapping (3D-CAM) and 3D gradient-weighted class activation mapping (3D-Grad-CAM) as weighted visualizations of the activation maps in the convolutional layers. Firstly, they were limited by the low resolution of the convolutional layers. Secondly, the upsampled heatmaps didn’t provide enough detail to accurately identify important regions. Future studies on applying state-of-the-art methods on explainability of 3DD CNNs [12], [13] could be beneficial to circumvent the aforementioned challenges.

Other key challenges that were faced, the solutions that were explored as an attempt to solve the challenges, and which are still in store for the future are presented in the sections below.

2 CHALLENGES

2.1 IMAGING PARAMETERS AND ACQUISITION PROTOCOLS

2.1.1 Acquisition

Across different institutes and even within an institution, imaging systems are often from different manufacturers. Especially MRI imaging data suffers from significant inter- and intra-site variability, which hinders multi-site data analysis. A recent study on rectal MRI data investigated the sources of variation in multicenter MRI data and their effect on radiomics feature reproducibility [14]. The study showed that features derived from T2W-MRI and in particular ADC differ significantly between centers when performing multicenter data analysis. Another study showed that only less than half of the radiomic features extracted were reproducible when extracted from 5 different scanners[15]. This also influenced the analysis performed in chapter 7 because the MRI scans were acquired from 4 different centers with varying pixel spaces and slice thicknesses and other scanning parameters. The influence of the reconstruction kernel could have also affected the generalizability of the model presented in chapter 3, where the external validation dataset contained images with a distribution of the reconstruction kernels (B30f, B41s, and D40s) different from that of the training dataset. Ideally, images are all obtained using controlled acquisition settings, but the availability of this type of data is often limited. A prospective study could overcome this issue, whereas a phantom study could help to investigate the influence of image acquisition parameters on radiomic feature values and possibly allow for post-harmonization. When applying an end-to-end DL-based pipeline for radiomic analysis, image augmentation while training can also help in improving the generalizability of the model to test images with varying characteristics as that of the training data which is discussed later in this chapter.

2.1.2 Radiation dose and Reconstruction kernel

Some studies have investigated the influence of reconstruction settings and radiation dose on feature stability [16] [17], [18] and have shown that, of all technical parameters, reconstruction kernel and radiation dose had the largest impact on the reproducibility of radiomic features. The radiomic features in the shape category (including the maximum axial diameter and volume) were insensitive to changes in radiation dose and reconstruction CT settings, compared with radiomic features in the texture and, to a lesser extent, the intensity categories (including mean attenuation). This could have impacted the analysis of chapter 2 which includes data from two different manufacturers GE and SEIMENS with STANDARD and B30-range reconstruction kernels. There are many ways to circumvent this problem of feature variability, one option could be by performing prospective studies where the imaging parameters are standardized. In retrospective studies like the one in chapter 7, post-reconstruction feature harmonization techniques can be applied to eliminate the batch effect. For instance, in a recent study [19], the effect of the use of Reconstruction Kernel Normalization (RKN) and ComBat harmonization [20] on the reproducibility of radiomics features across scans acquired with different reconstruction kernels and have shown that the use of RKN resulted in a significant increment in the number of reproducible features.

2.1.3 Phases of contrast enhancement

Optimal contrast enhancement is important for a successful diagnostic CT scan. The purpose of contrast-enhanced CT (CECT) is to find pathology by enhancing the contrast between a lesion and the normal surrounding structures. Sometimes a lesion will be hypovascular compared to the normal tissue and in some cases, a lesion will be hypervascular to the surrounding tissue in a certain phase of enhancement. When deploying a model like the one described in chapter 4 in a real-world clinical setting, manual intervention is usually required to ensure that the input scan is at a particular contrast phase that is expected by the trained model, which takes up substantial manual effort and could be erroneous. Automated AI-based methods to identify the contrast phase on the scans would help in circumventing the challenge. For instance, in a recent study [21], DL–based Detection of Intravenous Contrast Enhancement on CT scans was investigated on Head and neck, and chest CT scans and have proved their model to be highly accurate on an external validation dataset. Such methods can be adopted in combination with the radiomic analysis pipeline when deployed in real-world clinical settings.

2.2 INPUT DATA QUALITY

Quality check of the input data plays a major role in the effective utilization of trained AI models for a particular use case. For instance, the models presented in chapters 3 & 4 were trained in chest CT scans and were not trained on CT acquisitions of other regions as negative data points, which might cause erroneous predictions from the model, given an abdomen CT scan for instance as input. The dataset available in real-world clinical settings is not usually sorted and might also contain erroneous metadata. For the models developed to be efficiently used in a real clinical setting, it would be beneficial to combine these models with an automated AI-based Quality check pipeline that can filter the images which are not compliant for a particular analysis. For instance, a recent study investigates the use of AI-based solutions for region identification on CT and MRI [22].

2.3 LARGE-SCALE AI MODELS - KNOWLEDGE DISTILLATION

The AI models presented throughout my thesis were trained on NVIDIA GeForce RTX 2080 Ti, 11GB GPU, and, the DL models consists of more than a million parameters. For instance, the lung segmentation model used in the development of the COVID model presented in chapter 2 consists of around 23 million trainable parameters. However, whilst training large models helps improve state-of-the-art performance, deploying such large trained models in a real clinical setting might get expensive when there is a consistent need for computational resources. Knowledge distillation [23] helps overcome these challenges by capturing and “distilling” the knowledge in a complex machine learning model or an ensemble of models into a smaller single model that is much easier to deploy without significant loss in performance [24]. In a recent study, a lightweight CNN model was developed by applying the knowledge distillation technique for cervical cell classification [25]. Such lightweight models which have relevant information distilled from a large-scale model trained on a huge dataset can be used for deployment to be efficiently used in real-world clinical settings with limited computational resources.

3 FUTURE PERSPECTIVES

3.1 DISTRIBUTED LEARNING

Medical data is greatly sensitive and highly protected by law and ethics; making access to such data harder and time-consuming. Therefore, more research on distributed learning methods in the future can circumvent these limitations while satisfying the concerns regarding sharing of clinical data by hospitals. The concept of distributed (federated, privacy-preserving) machine learning is not new in healthcare applications [26], [27] but has recently shown its potential for radiomics [28], [29]. For example, Shi *et al.* performed a decentralized multi-center study to develop a radiomic signature for lung cancer in one institution and validated the performance in an independent institution, without the need for data exchange [30]. In another recent case study, Bogowicz *et al.* developed and validated a radiomic signature for head and neck cancer, training the model remotely from 6 independent cohorts, showing that the performances of the distributed model were as good as the one obtained with the traditional radiomic approach [31]. Several state-of-the-art methods employing blockchain technology have proven the efficiency of distributed learning in the bio-medical field [32]–[34]. However, most of the existing research works lack validation of their framework in a real clinical setting with diversity in data distribution. As an attempt to prove the potential of training and validating AI-based segmentation models in a distributed fashion, during my research, a basic Unit based model [35] was trained in distributed settings [36] which proved no difference in performance between a model trained in centralized and distributed strategy. More research is still needed to assess the robustness of such a methodology when in presence of multiple centers with variability in data distribution where a quality check of the data would be substantially needed.

3.2 SYNTHETIC DATA GENERATION

Some of the studies (for instance chapter 7) presented in this thesis suffered from limited data for external validation of the proposed model. Another area of Artificial intelligence that has the potential to solve the challenge of limited data is AI Synthetic data generation. Generative Adversarial Networks (GANs) are gaining increasing attention as a means of synthesizing data. For example, Research work presented in [37] trained and validated a GAN to synthesize new T1-weighted brain MRI with comparable quality to real images, and [38] succeeded in generating high-resolution skin lesion images which experts could not reliably tell apart from real images. In [39] authors have shown that GAN-generated images of lung cancer nodules are nearly indistinguishable from real images, even by trained radiologists. However, more research needs to be done on validating the effect of artifacts created by GAN-generated synthetic images. Furthermore, GANs have been used for inter-site data harmonization as a preprocessing step to normalize the features extracted from the medical images [40]. However, more research to prove the stability and reproducibility of such features extracted from GAN-harmonized images on a larger cohort is to be done in the future.

3.3 VISION TRANSFORMERS AS A REPLACEMENT FOR CNNS

Recently, vision transformers (ViT) [41] have appeared as a competitive alternative to CNNs, yielding similar levels of performance while possessing several interesting properties that could prove beneficial for medical imaging tasks. A recent study [42] has investigated the performance of CNNs and ViTs on three different medical image tasks and has proved that the ViTs reached the same level of performance as CNNs in small medical datasets, provided the transfer learning is applied and on larger datasets, ViTs significantly outperformed CNNs. Hence Transformers could be a better and an efficient alternative to the CNN-based models built for medical imaging use-cases.

3.4 SEMI-SUPERVISED LEARNING

One major challenge in medical imaging analysis is the lack of label and annotation which usually requires medical knowledge and training. To ease the manual labeling burden, significant efforts have been devoted to annotation-efficient DL methods for medical image segmentation tasks by enlarging the training data through label generation [43], data augmentation [44], leveraging external related labeled datasets [45], and leveraging unlabelled data with semi-supervised learning. Among these approaches, semi-supervised segmentation is a more practical method by encouraging segmentation models to utilize unlabelled data which is much easier to acquire in conjunction with a limited amount of labeled data for training, which has a high impact on real-world clinical applications. However, without expert-examined annotations, it is still an open and challenging question on how to efficiently exploit useful information from these unlabelled data. Existing semi-supervised medical image segmentation approaches have achieved comparable results with fully supervised methods [46] [47]. However, these methods still need a small amount of well-annotated labeled data to guide the learning of unlabelled data. Acquiring such fully annotated training data can still be costly, especially for the tasks of medical image segmentation. To further alleviate the annotation cost, some more future research on integrating semi-supervised learning with other annotation efficient approaches like leveraging image-level, box-level, and pixel-level annotations [48] or scribble supervisions [49] could be beneficial.

3.5 CLINICAL XAI

Methodologically, the studies presented to prove explainability is limited to Gradient weighted class activation mapping (Grad-CAM)-based techniques. While Grad-CAM-based methods can be leveraged to visualize the relevant region of interest on medical images which contributed to the model's decision, further quantification of known semantic features from the region of interest can provide clinically relevant and explainable features which can further contribute to surrogate function explaining the model's decision. A schematic of this idea is shown in Figure 9.1, assuming the scenario: DL model used for diagnosis of respiratory abnormalities.

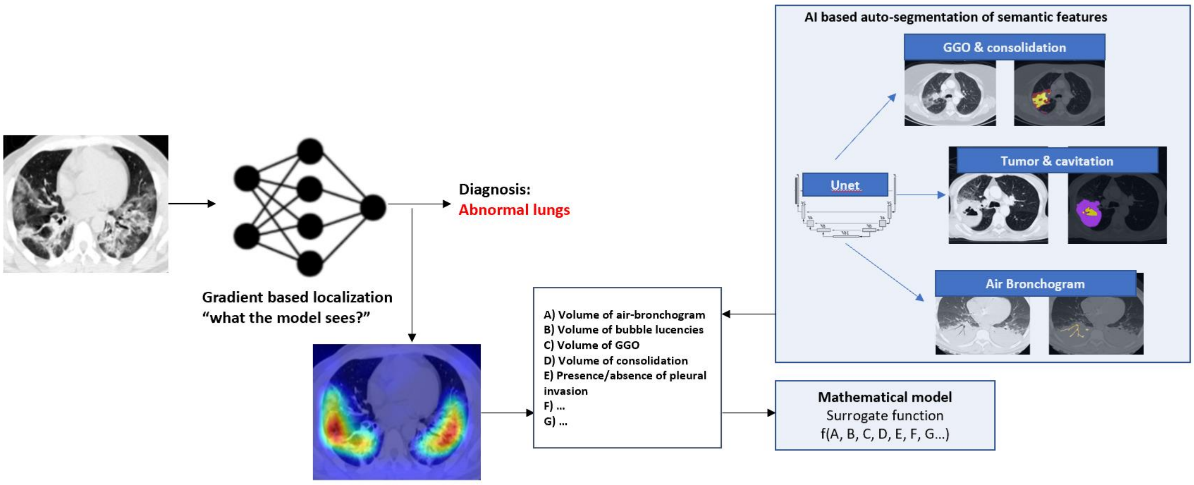

Diagnosis:
Abnormal lungs
Gradient based localization
"what the model sees?"
AI based auto-segmentation of semantic features
GGO & consolidation
Unet
Tumor & cavitation
Air Bronchogram
A) Volume of air-bronchogram
B) Volume of bubble lucencies
C) Volume of GGO
D) Volume of consolidation
E) Presence/absence of pleural invasion
F) …
G) …
Mathematical model
Surrogate function
f(A, B, C, D, E, F, G…)

Figure 9.1. Schematic on the prototype showing explanation of activation maps using a surrogate function of parameters corresponding to quantification of semantic features

3.6 VIRTUAL BIOPSY

The research presented in this thesis is limited to Radiomic data. The effect of combining radiomics with other available omics data in predicting clinical outcomes is unexplored. Specifically, in oncological studies, different parts of the tumor have distinct molecular characteristics, but also different lesions (metastases) from a tumor disease, which may have a role in terms of therapeutic efficacy, and such differences might change over time. As it is not possible to take samples of every part of each tumor at multiple time points, the optimal characterization of tumors is not achieved using biopsy[50]. However, radiomics might be used to "sample" different parts of the tumor at different time points (i.e. different scans) and, along with genomic data, used as a virtual biopsy tool [51], [52]. The combination of radiomics and genomics is called radiogenomics and more research in this field can be a way of augmenting the power of both approaches, for personalized medicine and treatment follow-up [53]–[55].

3.7 DELTA RADIOMICS

The vast majority of radiomics methods published including the works presented in this thesis focus on imaging data acquired at a single time point, mostly imaging tumors before the start of treatment. Delta-radiomics introduces a time component with the extraction of quantitative features from image sets acquired throughout treatment [56]–[58], which provides information on the evolution of feature values. Future research on Delta-radiomics while leveraging the methods introduced related to radiomics on a single imaging timepoint, can provide more relevant and significant biomarkers for disease diagnosis, prognosis, prediction, monitoring, image-based intervention, or assessment of therapeutic response [59], [60].

3.8 OPEN SCIENCE AND DATA SHARING

There is a pressing need to embrace knowledge and data-sharing technology [61], which transcends institutional and national boundaries [62]. This is especially true for radiomics whose potency is directly linked to the amount and quality of data available. A large dataset with deep clinical and molecular information and homogeneous imaging sources will result in more robust and reliable radiomics models. To unlock the full potential of radiomics for clinical decision-making, the research and clinical communities must strive for truly open science – sharing datasets, algorithms, and best practices and finding new ways to improve collaborations. One initiative to accomplish these goals is CancerLinQ [63], the ASCO data centralization approach. Other initiatives are worldCAT and its European counterpart euroCAT [64] which consist of a novel data-federated approach that successfully links radiotherapy institutes in the Netherlands, Germany, Belgium, Italy, Denmark, Australia, China, India, South Africa, Ireland, UK, USA and Canada [27], [65]. Other important links include The Cancer Imaging Archive (TCIA) [66], The Quantitative Imaging Network (QIN) [67], the Quantitative Imaging Biomarkers Alliance (QIBA) [68], the MEDomics consortium [69], and Quantitative Imaging in Cancer: Connecting Cellular Processes with Therapy (QuIC-ConCePT) [70]. The next step in this open science initiative for radiomics should be the creation of a database to store and cross-reference radiomics features and relevant clinical data (radiomics ontology [71], [72]). Extracted radiomic features must be stored in searchable databases to realize the unprecedented potential for RLHC that routine standard-of-care imaging represents. Hence, RLHC networks can dynamically capture multimodal data and share knowledge across departmental and institutional boundaries [73], to accumulate sufficient datasets of significant statistical power for model development and validation. Also, the accessibility of radiomics, in general, must be improved and some initiatives in this regard are already in place, especially from a software perspective. Several open source or freeware software is already available [74]–[77] and code sharing is becoming more and more accepted in the scientific community.

4 CONCLUSION

Continuous efforts are ongoing in improving the AI-based methodology for creating more transparency on the model’s decisions with clinically relevant explainability and reproducibility. Nevertheless, additional efforts are required to answer all the questions related to the robustness and usefulness of such research tools given a dynamic environment where standardization of data is still not possible in all the circumstances while also deriving clinically meaningful outcomes. There is also much work to do, especially to link fundamental research to current clinical practice. Physicians and healthcare personnel should be involved from the start of the process, along with relevant authorities. On the other hand, more effort should be devoted to the technological transfer, taking the published research and performing the necessary steps to bring it from a (validated) proof-of-concept to the clinic. This also emphasizes the need for comprehensive and universal indicators (such as the RQS) of the quality of a model. The normative framework is currently evolving along with innovations in the field of AI-driven healthcare. For example, FDA is gathering feedback and propositions to draft a novel regulatory framework for AI-based medical devices [78]. Paradigms need to be re-invented to allow these breakthroughs to reach the clinic very shortly, always putting patients’ welfare first. Personalized, patient-centric medicine is almost a reality and radiomics is playing a major role in it and will represent one of the key factors for the future of healthcare.

5 REFERENCES

[1] I. Arevalo-Rodriguez *et al.*, “False-negative results of initial RT-PCR assays for COVID-19: A systematic review,” *PLoS One*, vol. 15, no. 12 December, Dec. 2020, doi: 10.1371/JOURNAL.PONE.0242958.

[2] V. Pecoraro, A. Negro, T. Pirotti, and T. Trenti, “Estimate false-negative RT-PCR rates for SARS-CoV-2. A systematic review and meta-analysis,” *Eur J Clin Invest*, vol. 52, no. 2, Feb. 2022, doi: 10.1111/ECI.13706.

[3] H. X. Bai *et al.*, “Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT,” *Radiology*, vol. 296, no. 2, pp. E46–E54, Aug. 2020, doi: 10.1148/RADIOL.2020200823.

[4] J. Eng *et al.*, “Accuracy of CT in the Diagnosis of Pulmonary Embolism: A Systematic Literature Review,” *http://dx.doi.org/10.2214/ajr.183.6.01831819*, vol. 183, no. 6, pp. 1819–1827, Nov. 2012, doi: 10.2214/AJR.183.6.01831819.

[5] S. J. Kligerman, J. W. Mitchell, J. W. Sechrist, A. K. Meeks, J. R. Galvin, and C. S. White, “Radiologist performance in the detection of pulmonary embolism: Features that favor correct interpretation and risk factors for errors,” *J Thorac Imaging*, vol. 33, no. 6, pp. 350–357, Nov. 2018, doi: 10.1097/RTI.0000000000000361.

[6] M. Das *et al.*, “Computer-aided detection of pulmonary embolism: Influence on radiologists’ detection performance with respect to vessel segments,” *Eur Radiol*, vol. 18, no. 7, pp. 1350–1355, Jul. 2008, doi: 10.1007/S00330-008-0889-X.

[7] L. E. Ordonez-Ordonez *et al.*, “Diagnostic test validation: Cochlear hydrops analysis masking procedure in Ménière’s disease,” *Otology and Neurotology*, vol. 30, no. 6, pp. 820–825, Sep. 2009, doi: 10.1097/MAO.0B013E3181B11EB2.

[8] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, and L. Umutlu, “Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis,” *Lancet Digit Health*, vol. 4, no. 7, pp. e507–e519, Jul. 2022, doi: 10.1016/S2589-7500(22)00070-X.

[9] T. Schaffter *et al.*, “Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms,” *JAMA Netw Open*, vol. 3, no. 3, pp. e200265–e200265, Mar. 2020, doi: 10.1001/JAMANETWORKOPEN.2020.0265.

[10] Y. Wan *et al.*, “Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography,” *Front Oncol*, vol. 12, p. 1739, Apr. 2022, doi: 10.3389/FONC.2022.880150/XML/NLM.

[11] C. Szegedy *et al.*, “Going Deeper with Convolutions”.

[12] M. Ennab and H. Mcheick, “Designing an Interpretability-Based Model to Explain the Artificial Intelligence Algorithms in Healthcare,” 2022, doi: 10.3390/diagnostics12071557.

[13] F. Cruciani *et al.*, “Explainable 3D-CNN for Multiple Sclerosis Patients Stratification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12663 LNCS, pp. 103–114, 2021, doi: 10.1007/978-3-030-68796-0_8.

[14] N. W. Schurink *et al.*, “Sources of variation in multicenter rectal MRI data and their effect on radiomics feature reproducibility,” *Eur Radiol*, vol. 32, no. 3, pp. 1506–1516, Mar. 2022, doi: 10.1007/S00330-021-08251-8/FIGURES/4.

[15] R. Berenguer *et al.*, “Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters,” *Radiology*, vol. 288, no. 2, pp. 407–415, Aug. 2018, doi: 10.1148/RADIOL.2018172361.

[16] M. Meyer *et al.*, “Reproducibility of CT radiomic features within the same patient: Influence of radiation dose and CT reconstruction settings,” *Radiology*, vol. 293, no. 3, pp. 583–591, 2019, doi: 10.1148/RADIOL.2019190928.

[17] L. Rinaldi *et al.*, “Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters,” *Eur Radiol Exp*, vol. 6, no. 1, Dec. 2022, doi: 10.1186/S41747-021-00258-6.

[18] H. Kim *et al.*, “Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability,” *PLoS One*, vol. 11, no. 10, Oct. 2016, doi: 10.1371/JOURNAL.PONE.0164924.

[19] T. Refaee *et al.*, “CT Reconstruction Kernels and the Effect of Pre- and Post-Processing on the Reproducibility of Handcrafted Radiomic Features,” *J Pers Med*, vol. 12, no. 4, Apr. 2022, doi: 10.3390/JPM12040553.

[20] C. Müller *et al.*, “Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data,” *PLoS One*, vol. 11, no. 6, p. e0156594, Jun. 2016, doi: 10.1371/JOURNAL.PONE.0156594.

[21] Z. Ye *et al.*, “Deep Learning–based Detection of Intravenous Contrast Enhancement on CT Scans,” *https://doi.org/10.1148/ryai.210285*, vol. 4, no. 3, May 2022, doi: 10.1148/RYAI.210285.

[22] P. Raffy[1] *et al.*, “Deep Learning Body Region Classification of MRI and CT examinations,” Apr. 2021, doi: 10.48550/arxiv.2104.13826.

[23] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” *Int J Comput Vis*, vol. 129, no. 6, pp. 1789–1819, Jun. 2020, doi: 10.1007/s11263-021-01453-z.

[24] A. Alkhulaifi, F. Alsahli, and I. Ahmad, “Knowledge distillation in deep learning and its applications,” *PeerJ Comput Sci*, vol. 7, pp. 1–24, Apr. 2021, doi: 10.7717/PEERJ-CS.474.

[25] W. Chen, L. Gao, X. Li, and W. Shen, “Lightweight convolutional neural network with knowledge distillation for cervical cells classification,” *Biomed Signal Process Control*, vol. 71, Jan. 2022, doi: 10.1016/j.bspc.2021.103177.

[26] F. Zerka *et al.*, “Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care,” *JCO Clin Cancer Inform*, no. 4, pp. 184–200, 2020, doi: 10.1200/cci.19.00047.

[27] T. M. Deist *et al.*, “Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT,” *Clin Transl Radiat Oncol*, vol. 4, pp. 24–31, 2017, doi: https://doi.org/10.1016/j.ctro.2016.12.004.

[28] F. Zerka *et al.*, “Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM),” *IEEE Access*, vol. 8, pp. 183939–183951, 2020, doi: 10.1109/ACCESS.2020.3029445.

[29] A. Jochems *et al.*, “Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries,” *Int J Radiat Oncol Biol Phys*, vol. 99, no. 2, pp. 344–352, 2017, doi: 10.1016/j.ijrobp.2017.04.021.

[30] Z. Shi *et al.*, “Distributed radiomics as a signature validation study using the Personal Health Train infrastructure,” *Sci Data*, vol. 6, no. 1, p. 218, 2019, doi: 10.1038/s41597-019-0241-0.

[31] M. Bogowicz *et al.*, “Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer,” *Sci Rep*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-61297-4.

[32] T. Hai, J. Zhou, S. R. Srividhya, S. K. Jain, P. Young, and S. Agrawal, “BVFLEMR: an integrated federated learning and blockchain technology for cloud-based medical records recommendation system,” *Journal of Cloud Computing*, vol. 11, no. 1, p. 22, Dec. 2022, doi: 10.1186/S13677-022-00294-6.

[33] R. Durga and E. Poovammal, “FLED-Block: Federated Learning Ensembled Deep Learning Blockchain Model for COVID-19 Prediction,” *Front Public Health*, vol. 10, Jun. 2022, doi: 10.3389/FPUBH.2022.892499/FULL.

[34] A. Rahman, S. Hossain, G. Muhammad, D. Kundu, and G. Muhammad, “Federated learning-based AI approaches in smart healthcare : concepts , taxonomies , challenges and open issues Data management,” *Cluster Comput*, vol. 4, 2022, doi: 10.1007/S10586-022-03658-4.

[35] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[36] F. Zerka *et al.*, “PO-1744: Privacy preserving distributed liver tumor segmentation,” *Radiotherapy and Oncology*, vol. 152, pp. S968–S969, Nov. 2020, doi: 10.1016/s0167-8140(21)01762-x.

[37] A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, B. A. Landman, and C. Bermudez, “Learning Implicit Brain MRI Manifolds with Deep Learning,” *Proc SPIE Int Soc Opt Eng*, vol. 10574, p. 56, Mar. 2018, doi: 10.1117/12.2293515.

[38] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, “Augmenting data with GANs to segment melanoma skin lesions,” *Multimed Tools Appl*, vol. 79, no. 21–22, pp. 15575–15592, Jun. 2020, doi: 10.1007/S11042-019-7717-Y.

[39] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, “How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis,” *Proceedings - International Symposium on Biomedical Imaging*, vol. 2018-April, pp. 240–244, Oct. 2017, doi: 10.48550/arxiv.1710.09762.

[40] J. Zhong *et al.*, “Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: Application to neonatal white matter development,” *Biomed Eng Online*, vol. 19, no. 1, pp. 1–18, Jan. 2020, doi: 10.1186/S12938-020-0748-9/TABLES/2.

[41] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing Properties of Vision Transformers," *Adv Neural Inf Process Syst*, vol. 34, pp. 23296–23308, Dec. 2021, Accessed: Sep. 09, 2022. [Online]. Available: https://git.io/Js15X.

[42] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it Time to Replace CNNs with Transformers for Medical Images?," Aug. 2021, doi: 10.48550/arxiv.2108.09038.

[43] C. Fang *et al.*, "Label-free coronavirus disease 2019 lesion segmentation based on synthetic healthy lung image subtraction," *Med Phys*, vol. 49, no. 7, pp. 4632–4641, Jul. 2022, doi: 10.1002/MP.15661.

[44] L. Zhang *et al.*, "Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation HHS Public Access," *IEEE Trans Med Imaging*, vol. 39, no. 7, pp. 2531–2540, 2020, doi: 10.1109/TMI.2020.2973595.

[45] Y. Zhang, Q. Liao, L. Yuan, H. Zhu, J. Xing, and J. Zhang, "Exploiting Shared Knowledge From Non-COVID Lesions for Annotation-Efficient COVID-19 CT Lung Infection Segmentation and also with the Beijing Advanced Innovation Centre for Big Data-Based Precision Medicine," *IEEE J Biomed Health Inform*, vol. 25, no. 11, 2021, doi: 10.1109/JBHI.2021.3106341.

[46] Y. Xia *et al.*, "3D Semi-Supervised Learning with Uncertainty-Aware Multi-View Co-Training," Nov. 2018, Accessed: Sep. 11, 2022. [Online]. Available: http://arxiv.org/abs/1811.12506

[47] X. Li, L. Yu, H. Chen, C. W. Fu, L. Xing, and P. A. Heng, "Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation," *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 2, pp. 523–534, Feb. 2021, doi: 10.1109/TNNLS.2020.2995319.

[48] S. Reiß, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every Annotation Counts: Multi-label Deep Supervision for Medical Image Segmentation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9527–9537, Jun. 2021, doi: 10.1109/CVPR46437.2021.00941.

[49] K. Zhang and X. Zhuang, "CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision," Mar. 2022, Accessed: Sep. 11, 2022. [Online]. Available: http://arxiv.org/abs/2203.01475

[50] G. L. Banna *et al.*, "The Promise of Digital Biopsy for the Prediction of Tumor Molecular Features and Clinical Outcomes Associated With Immunotherapy," *Frontiers in medicine*, vol. 6. Oncology Department, United Lincolnshire Hospital Trust, Lincoln, United Kingdom., p. 172, 2019. doi: 10.3389/fmed.2019.00172.

[51] P. Martin-Gonzalez *et al.*, "Integrative radiogenomics for virtual biopsy and treatment monitoring in ovarian cancer," *Insights Imaging*, vol. 11, no. 1, 2020, doi: 10.1186/s13244-020-00895-2.

[52] B. Shofty *et al.*, "Virtual biopsy using MRI radiomics for prediction of BRAF status in melanoma brain metastasis," *Sci Rep*, vol. 10, no. 1, pp. 1–7, 2020, doi: 10.1038/s41598-020-63821-y.

[53] E. J. Limkin and R. Sun, "Radiomics to predict response to immunotherapy: an imminent reality?," *Future Oncol*, vol. 16, no. 23, pp. 1673–1676, 2020, doi: 10.2217/fon-2020-0015.

[54] L. Tselikas *et al.*, "Role of image-guided biopsy and radiomics in the age of precision medicine," *Chin Clin Oncol*, vol. 8, no. 6, pp. 6–13, 2019, doi: 10.21037/cco.2019.12.02.

[55] M. Ismail *et al.*, “Spatial-And-Context aware (SpACe) ‘virtual biopsy’ radiogenomic maps to target tumor mutational status on structural MRI,” pp. 1–10, 2020.

[56] X. Fave *et al.*, “Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer.,” *Sci Rep*, vol. 7, no. 1, p. 588, Apr. 2017, doi: 10.1038/s41598-017-00665-z.

[57] Y. Ma, W. Ma, X. Xu, and F. Cao, “How Does the Delta-Radiomics Better Differentiate Pre-Invasive GGNs From Invasive GGNs?,” *Front Oncol*, vol. 10, no. July, pp. 1–7, 2020, doi: 10.3389/fonc.2020.01017.

[58] H. Chen *et al.*, “MRI Radiomics for Prediction of Tumor Response and Downstaging in Rectal Cancer Patients after Preoperative Chemoradiation,” *Adv Radiat Oncol*, pp. 1–10, 2020, doi: 10.1016/j.adro.2020.04.016.

[59] H. Nasief *et al.*, “A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer,” *NPJ Precis Oncol*, vol. 3, no. 1, pp. 1–10, 2019, doi: 10.1038/s41698-019-0096-z.

[60] P. Lin *et al.*, “A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma,” *Cancer Imaging*, vol. 20, no. 1, pp. 1–12, 2020, doi: 10.1186/s40644-019-0283-8.

[61] J. O. Deasy *et al.*, “Improving normal tissue complication probability models: the need to adopt a ‘data-pooling’ culture.,” *Int J Radiat Oncol Biol Phys*, vol. 76, no. 3 Suppl, pp. S151-4, Mar. 2010, doi: 10.1016/j.ijrobp.2009.06.094.

[62] T. Skripcak *et al.*, “Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets.,” *Radiother Oncol*, vol. 113, no. 3, pp. 303–309, Dec. 2014, doi: 10.1016/j.radonc.2014.10.001.

[63] R. L. Schilsky, D. L. Michels, A. H. Kearbey, P. P. Yu, and C. A. Hudis, “Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ.,” *J Clin Oncol*, vol. 32, no. 22, pp. 2373–2379, Aug. 2014, doi: 10.1200/JCO.2014.56.2124.

[64] “http://www.eurocat.info/.”

[65] P. Lambin *et al.*, “‘Rapid Learning health care in oncology’ – An approach towards decision support systems enabling customised radiotherapy’,” *Radiotherapy and Oncology*, vol. 109, no. 1, pp. 159–164, 2013, doi: https://doi.org/10.1016/j.radonc.2013.07.007.

[66] The Cancer Imaging Archive, “TCIA Collections. cancerimagingarchive.net,” 2017. https://www.cancerimagingarchive.net/

[67] National Cancer Institute, “Quantitative Imaging Network (QIN),” 2017. https://imaging.cancer.gov/programs_ resources/specialized_initiatives/qin.htm

[68] Radiological Society of North America, “Quantitative Imaging Biomarkers Alliance® (QIBA®),” 2017. https://www.rsna.org/qiba/

[69] “www.medomics.ai.”

[70] “QuiC ConCePT,” 2017. http://www.quic-concept.eu/

[71] Z. Shi, A. Traverso, J. van Soest, A. Dekker, and L. Wee, “Technical Note: Ontology-guided radiomics analysis workflow (O-RAW),” *Med Phys*, vol. 46, no. 12, pp. 5677–5684, Dec. 2019, doi: https://doi.org/10.1002/mp.13844.

[72] “https://bioportal.bioontology.org/ontologies/RO.”

[73] H. Yang *et al.*, “Lead federated neuromorphic learning for wireless edge artificial intelligence,” *Nat Commun*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/S41467-022-32020-W.

[74] L. Zhang, D. V Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court, “IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics.,” *Med Phys*, vol. 42, no. 3, pp. 1341–1353, Mar. 2015, doi: 10.1118/1.4908210.

[75] E. Pfaehler, A. Zwanenburg, J. R. de Jong, and R. Boellaard, “RACAT: An open source and easy to use radiomics calculator tool,” *PLoS One*, vol. 14, no. 2, pp. 1–26, 2019, doi: 10.1371/journal.pone.0212223.

[76] C. Nioche *et al.*, “LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity,” *Cancer Res*, vol. 78, no. 16, pp. 4786 LP – 4789, Aug. 2018, doi: 10.1158/0008-5472.CAN-18-0125.

[77] A. Fedorov *et al.*, “3D Slicer as an image computing platform for the Quantitative Imaging Network,” *Magn Reson Imaging*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[78] FDA, “Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback,” *U.S Food & Drug Administration*, pp. 1–20, 2019.

SUMMARY

PART 1 – AI based diagnostic models for respiratory diseases

In **chapter 2,** we developed and externally validated a machine learning model that is able to discriminate between COVID-19 positive and negative patients, and which has been trained and validated using a regularized logistic regression model. The model showed an AUC of 0.882 (95% CI: 0.851–0.913) in the independent test dataset (641 patients). The optimal decision threshold, considering the cost of false negatives twice as high as the cost of false positives, resulted in an accuracy of 85.18%, a sensitivity of 69.52%, a specificity of 91.63%, a negative predictive value (NPV) of 94.46% and a positive predictive value (PPV) of 59.44%.

In **chapter 3**, We developed and externally validated a deep learning AI model for the classification of no-infection, COVID-19, or Influenza/CAP cases based upon CT imaging. The model showed a performance in the external validation set with an AUC of 0.90, 0.92 and 0.92 for COVID-19, Influenza/CAP and No infection respectively. The selection of the input slices based on automatic segmentation of the abnormalities in the lung reduces analysis time (56 second per scan) and computational burden of the model.

In **chapter 4,** We have developed and externally validated an AI model for classification of pulmonary embolism in CTPA images. The model showed an area under the curve (AUC) of 0.86 [0.800-0.919], a sensitivity of 82.68 % [75.16 - 88.27] and a specificity of 79.31 % [61.61 - 90.15] on the external validation set. The activation maps of the slices rightly predicted positive by the PE classifier showed good visual correspondence with areas of PE. This was also quantitatively confirmed as 79.2% of PE regions in the GT were highlighted in the activation maps and the percentage of activated regions corresponding to GT PE is 80.3%.

PART 2 – AI based models for treatment outcome prediction and detection of disease in oncological use cases

In **chapter 5,** we have externally validated the prognostic value of Signature-0 in a prospective cohort using a manual and an automated segmentation method for survival prediction. The results of this study showed that the original radiomics signature, Signature-0, developed in 2012, outperforms the contemporaneous standard of care (TNM 6th edition) producing superior stratification between survivors and non-survivors.

In **chapter 6,** we have developed and externally validated a DL based algorithm that is able to detect metastatic bone disease on Bone Scintigraphy images. The developed DL based algorithm is able to detect MBD on BSs, with high specificity and sensitivity (0.80 and 0.82 respectively on the external test set), in a shorter time compared to the nuclear medicine physicians (2.5 minutes for AI and 30 minutes for nuclear medicine physicians to classify 134 BSs), that could be applied to any BS regardless of the patient’s gender and history of cancer.

PART 3 – AI based model for diagnosis of a disorder in inner ear

In **chapter 7,** we have validated a machine learning model trained on radiomics features extracted from inner ear region on MRI for diagnosis of Menière’s disease. The classification accuracy of the model on the validation set was 82%, with a sensitivity of 83% and a specificity of 82%. The positive and negative predictive values were 71% and 90%, respectively.

In **chapter 8,** we have trained and externally validated an AI based model for auto-segmentation of inner ear on MRI images. The model showed precise Dice Similarity Coefficient scores (mean DSC- 0.8790) with a high True Positive Rate (91.5%) and low False Discovery Rate and False Negative Rates (14.8% and 8.49% respectively) across images from three different centers. The model proved to perform well with a DSC of 0.8768 on the clinical validation dataset.

SAMENVATTING

DEEL 1 – Op AI gebaseerde diagnostische modellen voor luchtwegaandoeningen

In **hoofdstuk 2** hebben we een machine learning-model ontwikkeld en extern gevalideerd dat onderscheid kan maken tussen COVID-19-positieve en negatieve patiënten, en dat is getraind en gevalideerd met behulp van een geregulariseerd logistisch regressiemodel. Het model vertoonde een AUC van 0,882 (95% BI: 0,851-0,913) in de onafhankelijke testdataset (641 patiënten). De optimale beslissingsdrempel, gezien de kosten van valse negatieven die twee keer zo hoog zijn als de kosten van valse positieven, resulteerde in een nauwkeurigheid van 85,18%, een gevoeligheid van 69,52%, een specificiteit van 91,63%, een negatief voorspellende waarde (NPV) van 94,46 % en een positief voorspellende waarde (PPV) van 59,44%.

In **hoofdstuk 3** hebben we een deep learning AI-model ontwikkeld en extern gevalideerd voor de classificatie van gevallen zonder infectie, COVID-19 of Influenza/CAP op basis van CT-beeldvorming. Het model liet een prestatie zien in de externe validatieset met een AUC van 0,90, 0,92 en 0,92 voor respectievelijk COVID-19, Influenza/CAP en Geen infectie. De selectie van de invoerschijfjes op basis van automatische segmentatie van de afwijkingen in de long vermindert de analysetijd (56 seconden per scan) en de rekenbelasting van het model.

In **hoofdstuk 4** hebben we een AI-model ontwikkeld en extern gevalideerd voor classificatie van longembolie in CTPA-beelden. Het model toonde een oppervlakte onder de curve (AUC) van 0,86 [0,800-0,919], een sensitiviteit van 82,68 % [75,16 - 88,27] en een specificiteit van 79,31% [61,61 - 90,15] op de externe validatieset. De activeringskaarten van de plakjes die terecht positief waren voorspeld door de PE-classificator, vertoonden een goede visuele overeenkomst met gebieden van PE. Dit werd ook kwantitatief bevestigd, aangezien 79,2% van de PE-regio's in de GT werden gemarkeerd in de activeringskaarten en het percentage geactiveerde regio's dat overeenkomt met GT PE is 80,3%.

DEEL 2 – Op AI gebaseerde modellen voor voorspelling van behandelresultaten en detectie van ziekte in oncologische gebruiksgevallen

In **hoofdstuk 5** hebben we de prognostische waarde van Signature-0 extern gevalideerd in een prospectief cohort met behulp van een handmatige en een geautomatiseerde segmentatiemethode voor overlevingsvoorspelling. De resultaten van deze studie toonden aan dat de originele radiomic-signatuur, Signature-0, ontwikkeld in 2012, beter presteert dan de huidige standaard van zorg (TNM 6e editie) en een superieure stratificatie produceert tussen overlevenden en niet-overlevenden.

In **hoofdstuk 6** hebben we een op DL gebaseerd algoritme ontwikkeld en extern gevalideerd dat metastatische botziekte kan detecteren op botscintigrafiebeelden. Het ontwikkelde op DL gebaseerde algoritme kan MBD detecteren op BS'en, met een hoge specificiteit en gevoeligheid (respectievelijk 0,80 en 0,82 op de externe testset), in een kortere tijd vergeleken met de nucleair geneeskundigen (2,5 minuten voor AI en 30 minuten voor nucleaire geneeskundigen om 134 BS'en te classificeren), die op elke BS kan worden toegepast, ongeacht het geslacht van de patiënt en de voorgeschiedenis van kanker.

DEEL 3 – AI-gebaseerd model voor diagnose van een stoornis in het binnenoor

In **hoofdstuk 7** hebben we een machinaal leermodel gevalideerd dat is getraind op radiomic-kenmerken die zijn geëxtraheerd uit het binnenoorgebied op MRI voor de diagnose van de ziekte van Menière. De classificatienauwkeurigheid van het model op de validatieset was 82%, met een sensitiviteit van 83% en een specificiteit van 82%. De positief en negatief voorspellende waarden waren respectievelijk 71% en 90%.

In **hoofdstuk 8** hebben we een op AI gebaseerd model voor autosegmentatie van het binnenoor op MRI-beelden getraind en extern gevalideerd. Het model toonde nauwkeurige Dice-overeenkomstcoëfficiëntscores (gemiddelde DSC-0,8790) met een hoge True Positive Rate (91,5%) en een lage False Discovery Rate en False Negative Rates (14,8% en 8,49% respectievelijk) over afbeeldingen van drie verschillende centra. Het model bleek goed te presteren met een DSC van 0.8768 op de dataset voor klinische validatie.

3 REFERENTIES

[1] W. Samek and K.-R. Müller, “Towards Explainable Artificial Intelligence,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11700 LNCS, pp. 5–22, Sep. 2019, doi: 10.1007/978-3-030-28954-6_1.

APPENDIX

3.1. TRIPOD CHECKLIST

Y=yes; N=no; R=referenced; NA=not applicable		**Development [D]**	**External validation [V]**	**Combined Development & External validation [D+V]**
Title and abstract				
1	**Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.**			**0**
i	The words developing/development, validation/validating, incremental/added value (or synonyms) are reported in the title	Y	Y	Y
ii	The words prediction, risk prediction, prediction model, risk models, prognostic models, prognostic indices, risk scores (or synonyms) are reported in the title	N	N	N
iii	The target population is reported in the title	N	N	N
iv	The outcome to be predicted is reported in the title	Y	Y	Y
2	**Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.**			**0**
i	The objectives are reported in the abstract	Y	Y	Y
ii	Sources of data are reported in the abstract E.g. Prospective cohort, registry data, RCT data.	Y	Y	Y
iii	The setting is reported in the abstract E.g. Primary care, secondary care, general population, adult care, or paediatric care. The setting should be reported for both the development and validation datasets, if applicable.	Y	Y	Y
iv	A general definition of the study participants is reported in the abstract E.g. patients with suspicion of certain disease, patients with a specific disease, or general eligibility criteria.	Y	Y	Y

v	The overall sample size is reported in the abstract	Y	Y	Y
vi	The number of events (or % outcome together with overall sample size) is reported in the abstract If a continuous outcome was studied, score Not applicable (NA).	Y	Y	Y
vii	Predictors included in the final model are reported in the abstract. For validation studies of well-known models, at least the name/acronym of the validated model is reported Broad descriptions are sufficient, e.g. ‘all information from patient history and physical examination’. Check in the main text whether all predictors of the final model are indeed reported in the abstract.	N	N	N
viii	The outcome is reported in the abstract	Y	Y	Y
ix	Statistical methods are described in the abstract For model development, at least the type of statistical model should be reported. For validation studies a quote like “model’s discrimination and calibration was assessed” is considered adequate. If done, methods of updating should be reported.	N	N	N
x	Results for model discrimination are reported in the abstract This should be reported separately for development and validation if a study includes both development and validation.	Y	Y	Y
xi	Results for model calibration are reported in the abstract This should be reported separately for development and validation if a study includes both development and validation.	N	N	N
xii	Conclusions are reported in the abstract In publications addressing both model development and validation, there is no need for separate conclusions for both; one conclusion is sufficient.	Y	Y	Y
3a	**Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.**			**1**
i	The background and rationale are presented	Y	Y	Y
ii	Reference to existing models is included (or stated that there are no existing models)	Y	Y	Y

3b	**Specify the objectives, including whether the study describes the development or validation of the model or both.**			**1**
i	It is stated whether the study describes development and/or validation and/or incremental (added) value	Y	Y	Y
Methods				
4a	**Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.**			**1**
i	The study design/source of data is described E.g. Prospectively designed, existing cohort, existing RCT, registry/medical records, case control, case series. This needs to be explicitly reported; reference to this information in another article alone is insufficient.	Y	Y	Y
4b	**Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.**			**1**
i	The starting date of accrual is reported	Y	Y	Y
ii	The end date of accrual is reported	Y	Y	Y
iii	The length of follow-up and prediction horizon/time frame are reported, if applicable E.g. “Patients were followed from baseline for 10 years“ and “10-year prediction of…”; notably for prognostic studies with long term follow-up. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable (NA).	NA	NA	NA
5a	**Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.**			**1**
i	The study setting is reported (e.g. primary care, secondary care, general population) E.g.: ‘surgery for endometrial cancer patients’ is considered to be enough information about the study setting.	R	R	R
ii	The number of centres involved is reported If the number is not reported explicitly, but can be concluded from the name of the centre/centres, or if clearly a single centre study, score Yes.	Y	Y	Y

iii	The geographical location (at least country) of centres involved is reported If no geographical location is specified, but the location can be concluded from the name of the centre(s), score Yes.	Y	Y	Y
5b	**Describe eligibility criteria for participants.**			**0**
i	In-/exclusion criteria are stated These should explicitly be stated. Reasons for exclusion only described in a patient flow is not sufficient.	N	N	N
5c	**Give details of treatments received, if relevant.** (i.e. notably for prognostic studies with long term follow-up)			**Not applicable**
i	Details of any treatments received are described This item is notably for prognostic modelling studies and is about treatment at baseline or during follow-up. The 'if relevant' judgment of treatment requires clinical knowledge and interpretation. If you are certain that treatment was not relevant, e.g. in some diagnostic model studies, score Not applicable.	NA	NA	NA
6a	**Clearly define the outcome that is predicted by the prediction model, including how and when assessed.**			**1**
i	The outcome definition is clearly presented This should be reported separately for development and validation if a publication includes both.	Y	Y	Y
ii	It is described how outcome was assessed (including all elements of any composite, for example CVD [e.g. MI, HF, stroke]).	R	R	R
iii	It is described when the outcome was assessed (time point(s) since T0)	R	R	R
6b	**Report any actions to blind assessment of the outcome to be predicted.**			**0**
i	Actions to blind assessment of outcome to be predicted are reported If it is clearly a non-issue (e.g. all-cause mortality or an outcome not requiring interpretation), score Yes. In all other instances, an explicit mention is expected.	N	N	N

7a	**Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.**			**0**
i	All predictors are reported For development, “all predictors” refers to all predictors that potentially could have been included in the ‘final’ model (including those considered in any univariable analyses). For validation, “all predictors” means the predictors in the model being evaluated.	N	N	N
ii	Predictor definitions are clearly presented	Y	Y	Y
iii	It is clearly described how the predictors were measured	Y	Y	Y
iv	It is clearly described when the predictors were measured	N	N	N
7b	**Report any actions to blind assessment of predictors for the outcome and other predictors.**			**0**
i	It is clearly described whether predictor assessments were blinded for outcome For predictors for which it is clearly a non-issue (e.g. automatic blood pressure measurement, age, sex) and for instances where the predictors were clearly assessed before outcome assessment, score Yes. For all other predictors an explicit mention is expected.	N	N	N
ii	It is clearly described whether predictor assessments were blinded for the other predictors	N	N	N
8	**Explain how the study size was arrived at.**			**1**
i	It is explained how the study size was arrived at Is there any mention of sample size, e.g. whether this was done on statistical grounds or practical/logistical grounds (e.g. an existing study cohort or data set of a RCT was used)?	Y	Y	Y
9	**Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.**			**0**

i	The method for handling missing data (predictors and outcome) is mentioned E.g. Complete case (explicit mention that individuals with missing values have been excluded), single imputation, multiple imputation, mean/median imputation. If there is no missing data, there should be an explicit mention that there is no missing data for all predictors and outcome. If so, score Yes. If it is unclear whether there is missing data (from e.g. the reported methods or results), score No. If it is clear there is missing data, but the method for handling missing data is unclear, score No.	N	N	N
ii	If missing data were imputed, details of the software used are given When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA
iii	If missing data were imputed, a description of which variables were included in the imputation procedure is given When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA
iv	If multiple imputation was used, the number of imputations is reported When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA
10a	**Describe how predictors were handled in the analyses.**			**1**
i	For continuous predictors it is described whether they were modelled as linear, nonlinear (type of transformation specified) or categorized A general statement is sufficient, no need to describe this for each predictor separately. If no continuous predictors were reported, score Not applicable.	NA	Not applicable	NA
ii	For categorical or categorized predictors, the cut-points were reported If no categorical or categorized predictors were reported, score Not applicable.	Y	Not applicable	Y

iii	For categorized predictors the method to choose the cut-points was clearly described If no categorized predictors, score Not applicable.	NA	Not applicable	NA
10b	**Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.**			**0**
i	The type of statistical model is reported E.g. Logistic, Cox, other regression model (e.g. Weibull, ordinal), other statistical modelling (e.g. neural network)	Y	Not applicable	Y
ii	The approach used for predictor selection before modelling is described 'Before modelling' means before any univariable or multivariable analysis of predictor-outcome associations. If no predictor selection before modelling is done, score Not applicable. If it is unclear whether predictor selection before modelling is done, score No. If it is clear there was predictor selection before modelling but the method was not described, score No.	NA	Not applicable	NA
iii	The approach used for predictor selection during modelling is described E.g. Univariable analysis, stepwise selection, bootstrap, Lasso. 'During modelling' includes both univariable or multivariable analysis of predictor-outcome associations. If no predictor selection during modelling is done (so-called full model approach), score Not applicable. If it is unclear whether predictor selection during modelling is done, score No. If it is clear there was predictor selection during modelling but the method was not described, score No.	N	Not applicable	N
iv	Testing of interaction terms is described If it is explicitly mentioned that interaction terms were not addressed in the prediction model, score Yes. If interaction terms were included in the prediction model, but the testing is not described, score No.	N	Not applicable	N
v	Testing of the proportionality of hazards in survival models is described	N	Not applicable	N

	If no proportional hazard model is used, score Not applicable.			
vi	Internal validation is reported E.g. Bootstrapping, cross validation, split sample. If the use of internal validation is clearly a non-issue (e.g. in case of very large data sets), score Yes. For all other situations an explicit mention is expected.	N	Not applicable	N
10c	**For validation, describe how the predictions were calculated.**			**0**
i.	It is described how predictions for individuals (in the validation set) were obtained from the model being validated E.g. Using the original reported model coefficients with or without the intercept, and/or using updated or refitted model coefficients, or using a nomogram, spreadsheet or web calculator.	Not applicable	N	N
10d	**Specify all measures used to assess model performance and, if relevant, to compare multiple models.** These should be described in methods section of the paper (item 16 addresses the reporting of the results for model performance).			**0**
i	Measures for model discrimination are described E.g. C-index / area under the ROC curve.	Y	Y	Y
ii	Measures for model calibration are described E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.	N	N	N
iii	Other performance measures are described E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.	Y	Y	Y
10e	**Describe any model updating (e.g., recalibration) arising from the validation, if done.**			**Not applicable**
i	A description of model-updating is given E.g. Intercept recalibration, regression coefficient recalibration, refitting the whole model, adding a new predictor If updating was done, it should be clear which updating method was applied to score Yes. If it is not explicitly mentioned that updating was	Not applicable	NA	NA

	applied in the study, score this item as ‘Not applicable’.			
11	**Provide details on how risk groups were created, if done.** If risk groups were not created, score this item as Yes.			**0**
i	If risk groups were created, risk group boundaries (risk thresholds) are specified Score this item separately for development and validation if a study includes both development and validation. If risk groups were not created, score this item as not applicable.	N	N	N
12	**For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.**			**0**
i	Differences or similarities in definitions with the development study are described Mentioning of any differences in all four (setting, eligibility criteria, predictors and outcome) is required to score Yes. If it is explicitly mentioned that there were no differences in setting, eligibility criteria, predictors and outcomes, score Yes.	Not applicable	N	N
Results				
13a	**Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.**			**0**
i	The flow of participants is reported	Y	Y	Y
ii	The number of participants with and without the outcome are reported If outcomes are continuous, score Not applicable.	N	N	N
iii	A summary of follow-up time is presented This notably applies to prognosis studies and diagnostic studies with follow-up as diagnostic outcome. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.	NA	NA	NA

13b	**Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.**			**0**
i	Basic demographics are reported	Y	Y	Y
ii	Summary information is provided for all predictors included in the final developed/validated model	Y	Y	Y
iii	The number of participants with missing data for predictors is reported	N	N	N
iv	The number of participants with missing data for the outcome is reported	N	N	N
13c	**For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).**			**0**
i	Demographic characteristics (at least age and gender) of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
ii	Distributions of predictors in the model of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
iii	Outcomes of the validation study participants are reported along with those of the original development study	Not applicable	N	N
14a	**Specify the number of participants and outcome events in each analysis.**			**1**
i	The number of participants in each analysis (e.g. in the analysis of each model if more than one model is developed) is specified	Y	Not applicable	Y
ii	The number of outcome events in each analysis is specified (e.g. in the analysis of each model if more than one model is developed) If outcomes are continuous, score Not applicable.	NA	Not applicable	NA
14b	**If done, report the unadjusted association between each candidate predictor and outcome.**			**Not applicable**
i	The unadjusted associations between each predictor and outcome are reported If any univariable analysis is mentioned in the methods but not in the results, score No.	NA	Not applicable	NA

	If nothing on univariable analysis (in methods or results) is reported, score this item as Not applicable.			
15a	**Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).**			**0**
i	The regression coefficient (or a derivative such as hazard ratio, odds ratio, risk ratio) for each predictor in the model is reported	N	Not applicable	N
ii	The intercept or the cumulative baseline hazard (or baseline survival) for at least one time point is reported	N	Not applicable	N
15b	**Explain how to use the prediction model.**			**1**
i	An explanation (e.g. a simplified scoring rule, chart, nomogram of the model, reference to online calculator, or worked example) is provided to explain how to use the model for individualised predictions.	Y	Not applicable	Y
16	**Report performance measures (with confidence intervals) for the prediction model.** These should be described in results section of the paper (item 10 addresses the reporting of the methods for model performance).			**0**
i	A discrimination measure is presented E.g. C-index / area under the ROC curve.	Y	Y	Y
ii	The confidence interval (or standard error) of the discrimination measure is presented	N	Y	N
iii	Measures for model calibration are described E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.	N	N	N
iv	Other model performance measures are presented E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.	Y	Y	Y
17	**If done, report the results from any model updating (i.e., model specification, model performance, recalibration).**			**Not applicable**

	If updating was not done, score this TRIPOD item as ‘Not applicable’.			
0	Model updating was done If "No", then answer 17i-17v with "Not applicable"	Not applicable	N	N
i	The updated regression coefficients for each predictor in the model are reported If model updating was described as ‘not needed’, score Yes.	Not applicable	NA	NA
ii	The updated intercept or cumulative baseline hazard or baseline survival (for at least one time point) is reported If model updating was described as ‘not needed’, score Yes.	Not applicable	NA	NA
iii	The discrimination of the updated model is reported	Not applicable	NA	NA
iv	The confidence interval (or standard error) of the discrimination measure of the updated model is reported	Not applicable	NA	NA
v	The calibration of the updated model is reported	Not applicable	NA	NA
Discussion				
18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).			1
i	Limitations of the study are discussed Stating any limitation is sufficient.	Y	Y	Y
19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.			1
i	Comparison of results to reported performance in development studies and/or other validation studies is given	Not applicable	Y	Y
19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.			1
i	An overall interpretation of the results is given	Y	Y	Y
20	Discuss the potential clinical use of the model and implications for future research.			1

i	The potential clinical use is discussed E.g. an explicit description of the context in which the prediction model is to be used (e.g. to identify high risk groups to help direct treatment, or to triage patients for referral to subsequent care).	Y	Y	Y
ii	Implications for future research are discussed E.g. a description of what the next stage of investigation of the prediction model should be, such as ”We suggest further external validation”.	Y	Y	Y
Other information				
21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.			
i	Information about supplementary resources is provided	Y	Y	Y
22	Give the source of funding and the role of the funders for the present study.			1
i	The source of funding is reported or there is explicit mention that there was no external funding involved	Y	Y	Y
ii	The role of funders is reported or there is explicit mention that there was no external funding	Y	Y	Y

Number of applicable TRIPOD items			**32**
Number of TRIPOD items adhered			**15**
OVERALL adherence to TRIPOD			**47%**

4.1. STATISTICAL ANALYSIS

The probability threshold for defining a patient positive for PE was determined on the internal validation set and was optimized for both sensitivity and specificity. To quantify the variability in the results, 95% DeLong confidence intervals [14] (CIs) were computed for the AUC metric and 95% Wilson score CIs for sensitivity and specificity [15]. Calibration plots were used to measure the model's ability to generate probabilities that are, on average, close to the average observed risk on both internal validation and external test set [16]. Briefly, the model predictions were sorted in ascending order according to their probabilities and the [0,1] interval is discretized into 10 bins in a way that each bin contains an equal number of samples. For each bin the predicted and observed risk is computed and a linear model is used to fit the points. The slope and the intercept of the resulting linear fit are used as metrics for the model calibration on a specific dataset.

4.2. TRIPOD - TRANSPARENT REPORTING OF A MULTIVARIABLE PREDICTION MODEL FOR INDIVIDUAL PROGNOSIS OR DIAGNOSIS

	Y=yes; N=no; R=referenced; NA=not applicable	Development [D]	External validation [V]	Combined Development & External validation [D+V]
Title and abstract				
1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.			0
i	The words developing/development, validation/validating, incremental/added value (or synonyms) are reported in the title	Y	Y	Y
ii	The words prediction, risk prediction, prediction model, risk models, prognostic models, prognostic indices, risk scores (or synonyms) are reported in the title	N	N	N
iii	The target population is reported in the title	N	N	N
iv	The outcome to be predicted is reported in the title	Y	Y	Y
2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.			0
i	The objectives are reported in the abstract	Y	Y	Y
ii	Sources of data are reported in the abstract E.g. Prospective cohort, registry data, RCT data.	Y	Y	Y
iii	The setting is reported in the abstract E.g. Primary care, secondary care, general population, adult care, or paediatric care. The setting should be reported for both the development and validation datasets, if applicable.	Y	Y	Y
iv	A general definition of the study participants is reported in the abstract E.g. patients with suspicion of certain disease, patients with a specific disease, or general eligibility criteria.	Y	Y	Y
v	The overall sample size is reported in the abstract	Y	Y	Y
vi	The number of events (or % outcome together with overall sample size) is reported in the abstract If a continuous outcome was studied, score Not applicable (NA).	N	N	N

vii	Predictors included in the final model are reported in the abstract. For validation studies of well-known models, at least the name/acronym of the validated model is reported Broad descriptions are sufficient, e.g. ‘all information from patient history and physical examination’. Check in the main text whether all predictors of the final model are indeed reported in the abstract.	Y	Y	Y
viii	The outcome is reported in the abstract	Y	Y	Y
ix	Statistical methods are described in the abstract For model development, at least the type of statistical model should be reported. For validation studies a quote like “model’s discrimination and calibration was assessed” is considered adequate. If done, methods of updating should be reported.	Y	Y	Y
x	Results for model discrimination are reported in the abstract This should be reported separately for development and validation if a study includes both development and validation.	Y	Y	Y
xi	Results for model calibration are reported in the abstract This should be reported separately for development and validation if a study includes both development and validation.	N	N	N
xii	Conclusions are reported in the abstract In publications addressing both model development and validation, there is no need for separate conclusions for both; one conclusion is sufficient.	Y	Y	Y
3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.			1
i	The background and rationale are presented	Y	Y	Y
ii	Reference to existing models is included (or stated that there are no existing models)	Y	Y	Y
3b	Specify the objectives, including whether the study describes the development or validation of the model or both.			1
i	It is stated whether the study describes development and/or validation and/or incremental (added) value	Y	Y	Y
Methods				

4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.			1
i	The study design/source of data is described E.g. Prospectively designed, existing cohort, existing RCT, registry/medical records, case control, case series. This needs to be explicitly reported; reference to this information in another article alone is insufficient.	Y	Y	Y
4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.			1
i	The starting date of accrual is reported	R	R	R
ii	The end date of accrual is reported	R	R	R
iii	The length of follow-up <u>and</u> prediction horizon/time frame are reported, if applicable E.g. “Patients were followed from baseline for 10 years“ and “10-year prediction of…”; notably for prognostic studies with long term follow-up. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable (NA).	NA	NA	NA
5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.			0
i	The study setting is reported (e.g. primary care, secondary care, general population) E.g.: ‘surgery for endometrial cancer patients’ is considered to be enough information about the study setting.	N	Y	N
ii	The number of centres involved is reported If the number is not reported explicitly, but can be concluded from the name of the centre/centres, or if clearly a single centre study, score Yes.	N	Y	N
iii	The geographical location (at least country) of centres involved is reported If no geographical location is specified, but the location can be concluded from the name of the centre(s), score Yes.	N	Y	N
5b	Describe eligibility criteria for participants.			0
i	In-/exclusion criteria are stated These should explicitly be stated. Reasons for exclusion only described in a patient flow is not sufficient.	N	N	N

5c	Give details of treatments received, if relevant. (i.e. notably for prognostic studies with long term follow-up)			Not applicable
i	Details of any treatments received are described This item is notably for prognostic modelling studies and is about treatment at baseline or during follow-up. The ‘if relevant’ judgment of treatment requires clinical knowledge and interpretation. If you are certain that treatment was not relevant, e.g. in some diagnostic model studies, score Not applicable.	NA	NA	NA
6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.			0
i	The outcome definition is clearly presented This should be reported separately for development and validation if a publication includes both.	Y	Y	Y
ii	It is described how outcome was assessed (including all elements of any composite, for example CVD [e.g. MI, HF, stroke]).	N	N	N
iii	It is described when the outcome was assessed (time point(s) since T0)	N	N	N
6b	Report any actions to blind assessment of the outcome to be predicted.			1
i	Actions to blind assessment of outcome to be predicted are reported If it is clearly a non-issue (e.g. all-cause mortality or an outcome not requiring interpretation), score Yes. In all other instances, an explicit mention is expected.	Y	Y	Y
7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.			0
i	All predictors are reported For development, “all predictors” refers to all predictors that potentially could have been included in the ‘final’ model (including those considered in any univariable analyses). For validation, “all predictors” means the predictors in the model being evaluated.	N	N	N
ii	Predictor definitions are clearly presented	Y	Y	Y
iii	It is clearly described how the predictors were measured	Y	Y	Y
iv	It is clearly described when the predictors were measured	R	R	R

7b	Report any actions to blind assessment of predictors for the outcome and other predictors.			0
i	It is clearly described whether predictor assessments were blinded for outcome For predictors for which it is clearly a non-issue (e.g. automatic blood pressure measurement, age, sex) and for instances where the predictors were clearly assessed before outcome assessment, score Yes. For all other predictors an explicit mention is expected.	N	N	N
ii	It is clearly described whether predictor assessments were blinded for the other predictors	N	N	N
8	Explain how the study size was arrived at.			0
i	It is explained how the study size was arrived at Is there any mention of sample size, e.g. whether this was done on statistical grounds or practical/logistical grounds (e.g. an existing study cohort or data set of a RCT was used)?	N	N	N
9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.			0
i	The method for handling missing data (predictors and outcome) is mentioned E.g. Complete case (explicit mention that individuals with missing values have been excluded), single imputation, multiple imputation, mean/median imputation. If there is no missing data, there should be an explicit mention that there is no missing data for all predictors and outcome. If so, score Yes. If it is unclear whether there is missing data (from e.g. the reported methods or results), score No. If it is clear there is missing data, but the method for handling missing data is unclear, score No.	N	N	N
ii	If missing data were imputed, details of the software used are given When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA
iii	If missing data were imputed, a description of which variables were included in the imputation procedure is given When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA

iv	If multiple imputation was used, the number of imputations is reported When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.	NA	NA	NA
10a	Describe how predictors were handled in the analyses.			1
i	For continuous predictors it is described whether they were modelled as linear, nonlinear (type of transformation specified) or categorized A general statement is sufficient, no need to describe this for each predictor separately. If no continuous predictors were reported, score Not applicable.	Y	Not applicable	Y
ii	For categorical or categorized predictors, the cut-points were reported If no categorical or categorized predictors were reported, score Not applicable.	NA	Not applicable	NA
iii	For categorized predictors the method to choose the cut-points was clearly described If no categorized predictors, score Not applicable.	NA	Not applicable	NA
10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.			1
i	The type of statistical model is reported E.g. Logistic, Cox, other regression model (e.g. Weibull, ordinal), other statistical modelling (e.g. neural network)	Y	Not applicable	Y
ii	The approach used for predictor selection before modelling is described 'Before modelling' means before any univariable or multivariable analysis of predictor-outcome associations. If no predictor selection before modelling is done, score Not applicable. If it is unclear whether predictor selection before modelling is done, score No. If it is clear there was predictor selection before modelling but the method was not described, score No.	NA	Not applicable	NA
iii	The approach used for predictor selection during modelling is described E.g. Univariable analysis, stepwise selection, bootstrap, Lasso. 'During modelling' includes both univariable or multivariable analysis of predictor-outcome associations. If no predictor selection during modelling is done (so-called full model approach), score Not	NA	Not applicable	NA

	applicable. If it is unclear whether predictor selection during modelling is done, score No. If it is clear there was predictor selection during modelling but the method was not described, score No.			
iv	Testing of interaction terms is described If it is explicitly mentioned that interaction terms were not addressed in the prediction model, score Yes. If interaction terms were included in the prediction model, but the testing is not described, score No.	Y	Not applicable	Y
v	Testing of the proportionality of hazards in survival models is described If no proportional hazard model is used, score Not applicable.	NA	Not applicable	NA
vi	Internal validation is reported E.g. Bootstrapping, cross validation, split sample. If the use of internal validation is clearly a non-issue (e.g. in case of very large data sets), score Yes. For all other situations an explicit mention is expected.	Y	Not applicable	Y
10c	For validation, describe how the predictions were calculated.			1
i.	It is described how predictions for individuals (in the validation set) were obtained from the model being validated E.g. Using the original reported model coefficients with or without the intercept, and/or using updated or refitted model coefficients, or using a nomogram, spreadsheet or web calculator.	Not applicable	Y	Y
10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models. These should be described in methods section of the paper (item 16 addresses the reporting of the results for model performance).			1
i	Measures for model discrimination are described E.g. C-index / area under the ROC curve.	Y	Y	Y
ii	Measures for model calibration are described E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.	Y	Y	Y

iii	Other performance measures are described E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.	Y	Y	Y
10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.			Not applicable
i	A description of model-updating is given E.g. Intercept recalibration, regression coefficient recalibration, refitting the whole model, adding a new predictor If updating was done, it should be clear which updating method was applied to score Yes. If it is not explicitly mentioned that updating was applied in the study, score this item as 'Not applicable'.	Not applicable	NA	NA
11	Provide details on how risk groups were created, if done. If risk groups were not created, score this item as Yes.			Not applicable
i	If risk groups were created, risk group boundaries (risk thresholds) are specified Score this item separately for development and validation if a study includes both development and validation. If risk groups were not created, score this item as not applicable.	NA	NA	NA
12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.			1
i	Differences or similarities in definitions with the development study are described Mentioning of any differences in all four (setting, eligibility criteria, predictors and outcome) is required to score Yes. If it is explicitly mentioned that there were no differences in setting, eligibility criteria, predictors and outcomes, score Yes.	Not applicable	Y	Y
Results				
13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.			1
i	The flow of participants is reported	Y	Y	Y
ii	The number of participants with and without the outcome are reported If outcomes are continuous, score Not applicable.	NA	NA	NA

iii	A summary of follow-up time is presented This notably applies to prognosis studies and diagnostic studies with follow-up as diagnostic outcome. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.	NA	NA	NA
13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.			0
i	Basic demographics are reported	Y	Y	Y
ii	Summary information is provided for all predictors included in the final developed/validated model	Y	Y	Y
iii	The number of participants with missing data for predictors is reported	N	N	N
iv	The number of participants with missing data for the outcome is reported	N	N	N
13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).			1
i	Demographic characteristics (at least age and gender) of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
ii	Distributions of predictors in the model of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
iii	Outcomes of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
14a	Specify the number of participants and outcome events in each analysis.			0
i	The number of participants in each analysis (e.g. in the analysis of each model if more than one model is developed) is specified	N	Not applicable	N
ii	The number of outcome events in each analysis is specified (e.g. in the analysis of each model if more than one model is developed) If outcomes are continuous, score Not applicable.	NA	Not applicable	NA
14b	If done, report the unadjusted association between each candidate predictor and outcome.			Not applicable

i	The unadjusted associations between each predictor and outcome are reported If any univariable analysis is mentioned in the methods but not in the results, score No. If nothing on univariable analysis (in methods or results) is reported, score this item as Not applicable.	NA	Not applicable	NA
15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).			0
i	The regression coefficient (or a derivative such as hazard ratio, odds ratio, risk ratio) for each predictor in the model is reported	N	Not applicable	N
ii	The intercept or the cumulative baseline hazard (or baseline survival) for at least one time point is reported	N	Not applicable	N
15b	Explain how to use the prediction model.			0
i	An explanation (e.g. a simplified scoring rule, chart, nomogram of the model, reference to online calculator, or worked example) is provided to explain how to use the model for individualised predictions.	N	Not applicable	N
16	Report performance measures (with confidence intervals) for the prediction model. These should be described in results section of the paper (item 10 addresses the reporting of the methods for model performance).			1
i	A discrimination measure is presented E.g. C-index / area under the ROC curve.	Y	Y	Y
ii	The confidence interval (or standard error) of the discrimination measure is presented	Y	Y	Y
iii	Measures for model calibration are described E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.	Y	Y	Y
iv	Other model performance measures are presented E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.	Y	Y	Y
17	If done, report the results from any model updating (i.e., model specification, model performance, recalibration). If updating was not done, score this TRIPOD item as ‘Not applicable’.			Not applicable
0	Model updating was done If "No", then answer 17i-17v with "Not applicable"	Not applicable	N	N

i	The updated regression coefficients for each predictor in the model are reported If model updating was described as ‘not needed’, score Yes.	Not applicable	NA	NA
ii	The updated intercept or cumulative baseline hazard or baseline survival (for at least one time point) is reported If model updating was described as ‘not needed’, score Yes.	Not applicable	NA	NA
iii	The discrimination of the updated model is reported	Not applicable	NA	NA
iv	The confidence interval (or standard error) of the discrimination measure of the updated model is reported	Not applicable	NA	NA
v	The calibration of the updated model is reported	Not applicable	NA	NA
Discussion				
18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).			1
i	Limitations of the study are discussed Stating any limitation is sufficient.	Y	Y	Y
19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.			1
i	Comparison of results to reported performance in development studies and/or other validation studies is given	Not applicable	Y	Y
19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.			1
i	An overall interpretation of the results is given	Y	Y	Y
20	Discuss the potential clinical use of the model and implications for future research.			1
i	The potential clinical use is discussed E.g. an explicit description of the context in which the prediction model is to be used (e.g. to identify high risk groups to help direct treatment, or to triage patients for referral to subsequent care).	Y	Y	Y
ii	Implications for future research are discussed E.g. a description of what the next stage of investigation of the prediction model should be, such as ”We suggest further external validation”.	Y	Y	Y
Other information				
21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.			

i	Information about supplementary resources is provided	Y	Y	Y
22	Give the source of funding and the role of the funders for the present study.			1
i	The source of funding is reported or there is explicit mention that there was no external funding involved	Y	Y	Y
ii	The role of funders is reported or there is explicit mention that there was no external funding	Y	Y	Y
	Number of applicable TRIPOD items			31
	Number of TRIPOD items adhered			18
	OVERALL adherence to TRIPOD			58%

4.3. CLAIM: CHECKLIST FOR ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING

Section / Topic	**No.**	**Item**	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	x
	2	Structured summary of study design, methods, results, and conclusions	x
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	x
	4	Study objectives and hypotheses	x
METHODS			
Study Design	5	Prospective or retrospective study	x
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	x
Data	7	Data sources	x
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	
	9	Data pre-processing steps	x
	10	Selection of data subsets, if applicable	
	11	Definitions of data elements, with references to Common Data Elements	

	12	De-identification methods	
	13	How missing data were handled	
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	x
	15	Rationale for choosing the reference standard (if alternatives exist)	
	16	Source of ground-truth annotations; qualifications and preparation of annotators	x
	17	Annotation tools	x
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	
Data Partitions	19	Intended sample size and how it was determined	
	20	How data were assigned to partitions; specify proportions	x
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	x
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections	x
	23	Software libraries, frameworks, and packages	x
	24	Initialization of model parameters (e.g., randomization, transfer learning)	
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained	
	26	Method of selecting the final model	
	27	Ensembling techniques, if applicable	
Evaluation	28	Metrics of model performance	x
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	x
	30	Robustness or sensitivity analysis	
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	x
	32	Validation or testing on external data	x
RESULTS			
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	x
	34	Demographic and clinical characteristics of cases in each partition	

Model performance	35	Performance metrics for optimal model(s) on all data partitions	x
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	x
	37	Failure analysis of incorrectly classified cases	
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	x
	39	Implications for practice, including the intended use and/or clinical role	x
OTHER INFORMATION			
	40	Registration number and name of registry	
	41	Where the full study protocol can be accessed	
	42	Sources of funding and other support; role of funders	x

5.1 ADDITIONAL RESULTS

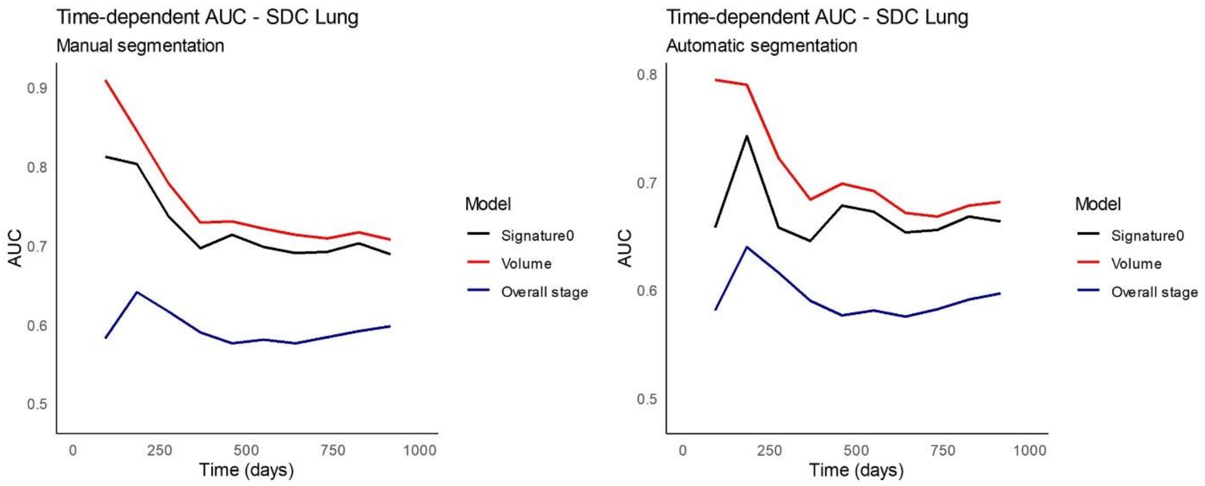


Fig. 1 Time-dependent AUC for Signature-0 (black) vs volume feature (red) vs Overall Stage (blue) on manual segmentation (left) and Signature-0 vs volume feature on automatic segmentation (right).

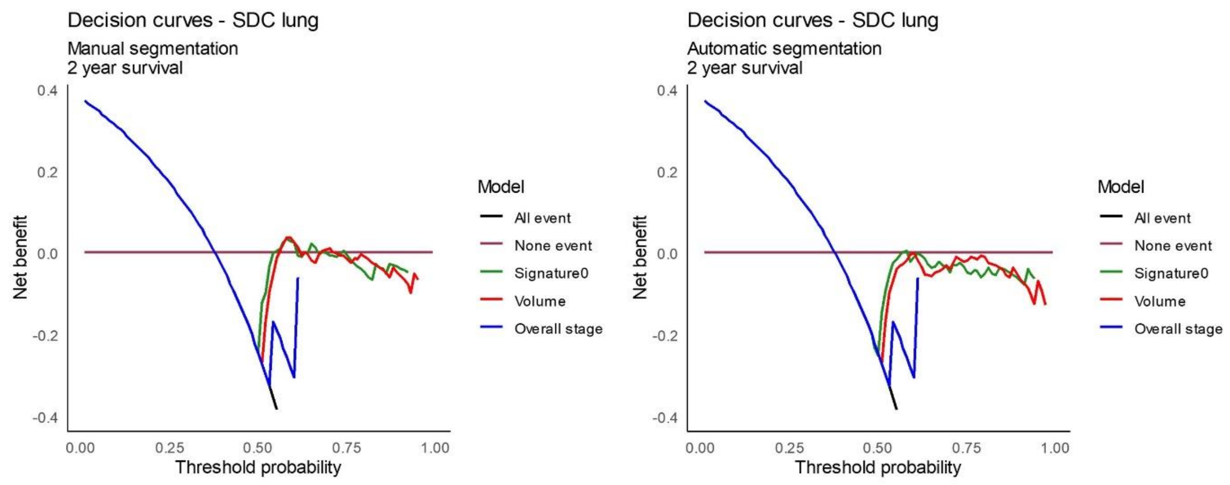


Fig. 2 Decision curves for Signature-0 (black) vs volume feature (red) vs Overall Stage (grey) on manual segmentation (left) and automatic segmentation (right) at 2 year survival for SDC Lung dataset

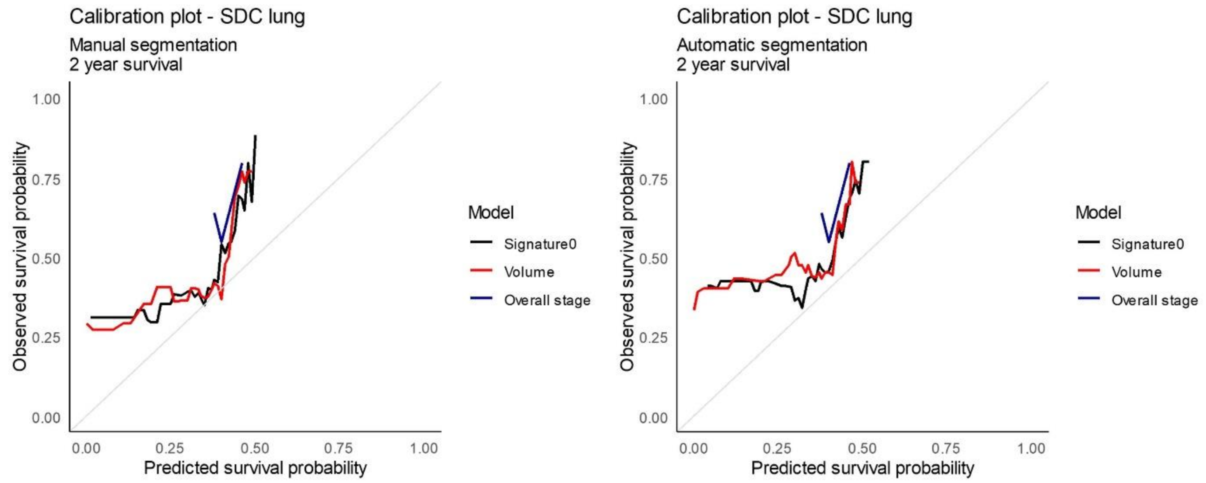


Fig. 3 Calibration curves of Signature-0 (black) vs volume feature (red) vs Overall Stage (grey) on manual segmentation (left) and automatic segmentation (right) for 2 year survival on SDC Lung dataset

8.1 ADDITIONAL RESULTS

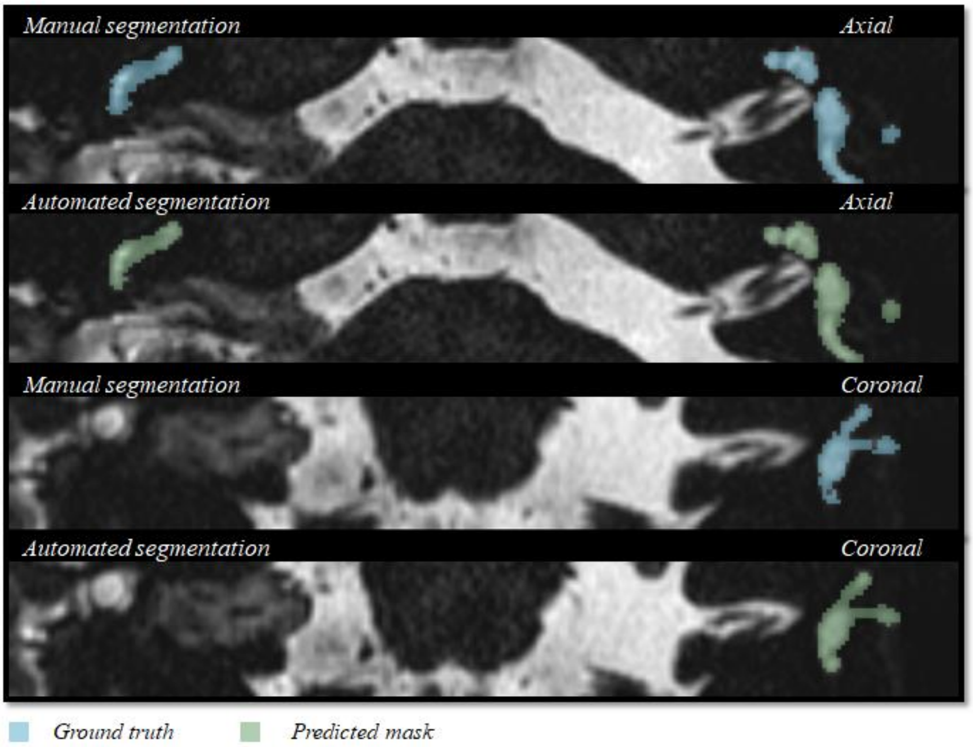


Figure 1a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows the presence of a vestibular schwannoma after a translabyrinthine resection on the right side. Therefore, the right semi-circular canals and vestibule are not segmented. DSC: 0.8973, Ground Truth Volume: 316.11 mm^3, True Positive Volume: 294.69mm^3, True Positive Rate: 93.22%, False Negative Rate: 6.77%. *False* Positive Rate: 0.0005%

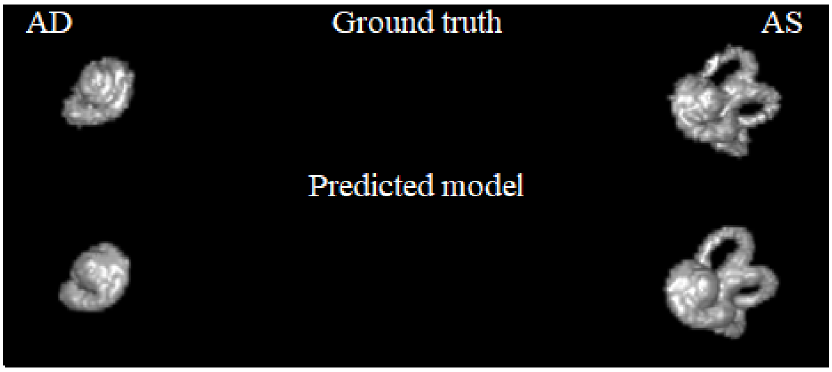


Figure 1b. The 3D volume rendering of the ground truth and the predicted mask. The semi-circular canals and the vestibule of the right inner ear were not fully displayed on MRI. The model has correctly not segmented the semi-circular canals and the vestibule of the right inner ear. AD= auriculum dextra, AS=auriculum sinistra

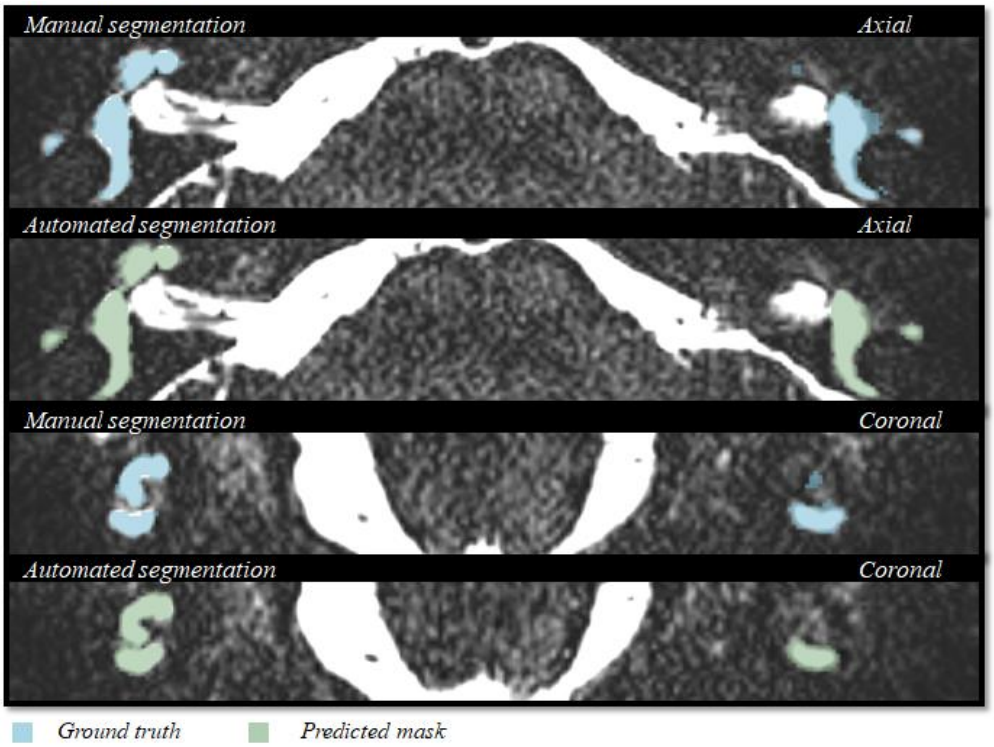


Figure 2a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows obliteration of the apical and middle turn of the left cochlea, indicating the presence of either labyrinthitis ossificans or a vestibular schwannoma. The left cochlea is, therefore, not fully segmented. DSC: 0.8691, Ground Truth Volume: 680.79 mm^3, True Positive Volume: 573.26 mm^3, True Positive Rate: 84.20%, False Negative Rate: 15.79%. *False* Positive Rate: 0.0007%

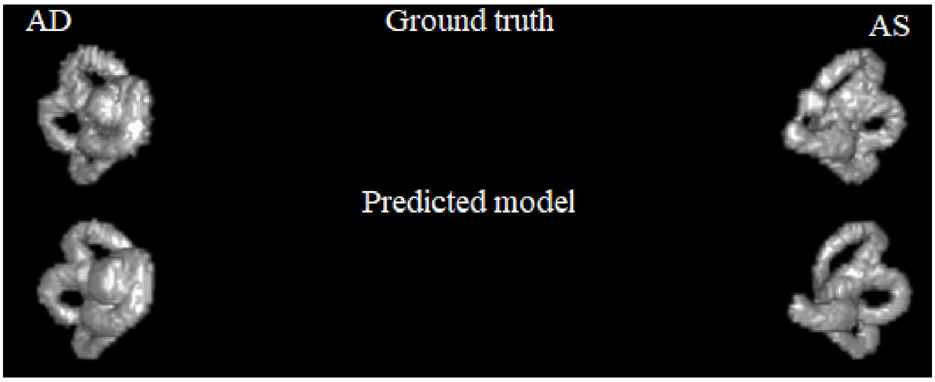


Figure 2b. The 3D volume rendering of the ground truth and the predicted mask. The cochlea of the left inner ear was not fully displayed on MRI. The model has correctly not segmented parts of the cochlea. AD= auriculum dextra, AS=auriculum sinistra

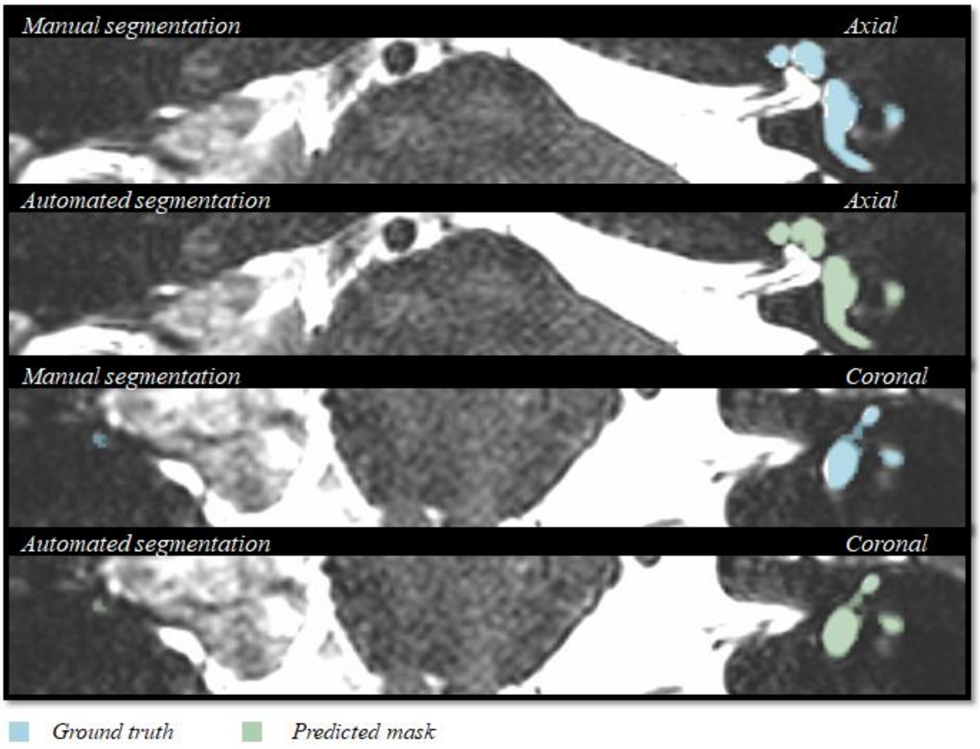


Figure 3a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows the presence of a vestibular schwannoma after a translabyrinthine resection on the right side. The left semi-circular canals and the vestibule are, therefore, not segmented. DSC: 0.8704, Ground Truth Volume: 359.94 mm^3, True Positive Volume: 314.92 mm^3, True Positive Rate: 87.49%, False Negative Rate: 12.50%, *False* Positive Rate: 0.0005%

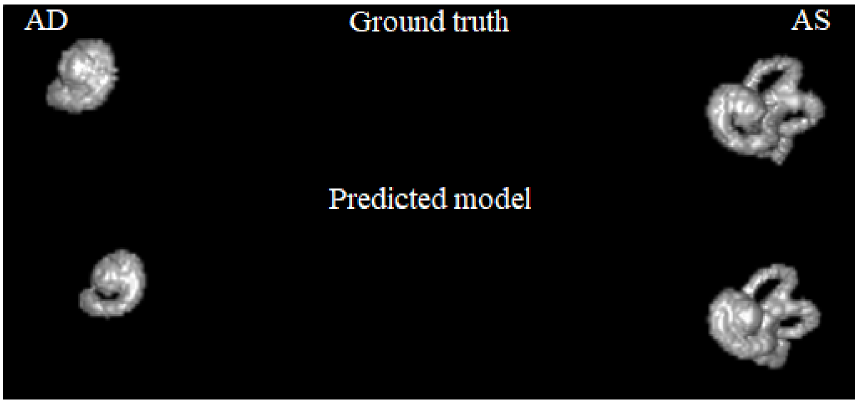


Figure 3b The 3D volume rendering of the ground truth and the predicted mask. The semi-circular canals and the vestibule of the right inner ear were not displayed on MRI. The model has correctly not segmented the semi-circular canals and the vestibule. AD= auriculum dextra, AS=auriculum sinistra

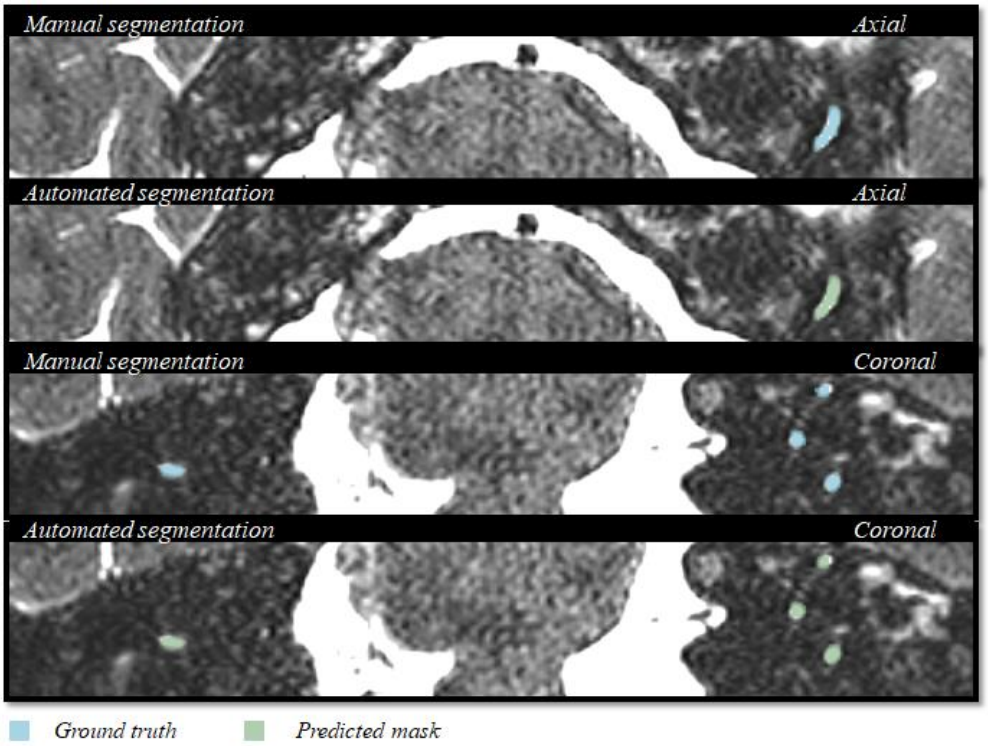


Figure 4a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows post-therapeutic fibrosis in the left inner ear. The right superior and inferior semi-circular canals are not segmented. DSC: 0.8916, Ground Truth Volume: 510.3 mm^3, True Positive Volume: 442.5 mm^3, True Positive Rate: 86.71%, False Negative Rate: 13.28%. *False* Positive Rate: 0.0004%

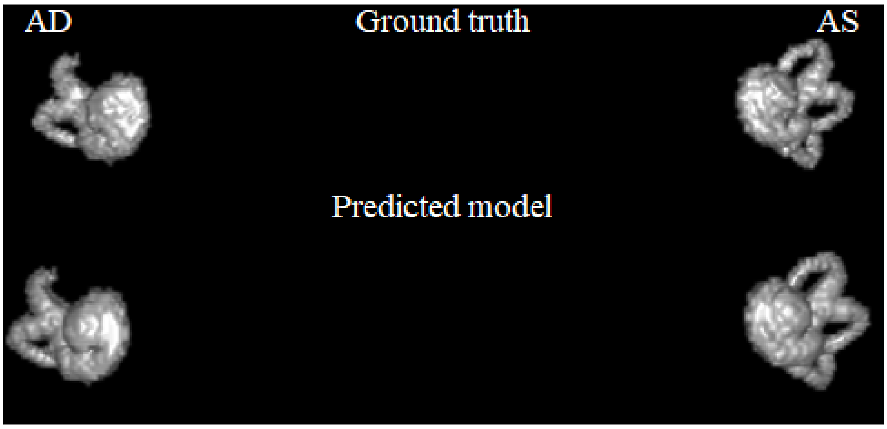


Figure 4b. The 3D volume rendering of the ground truth and the predicted mask. The superior and inferior semi-circular canals of the right inner ear were not displayed on MRI. The model has correctly not segmented these semi-circular canals AD= auriculum dextra, AS=auriculum sinistra

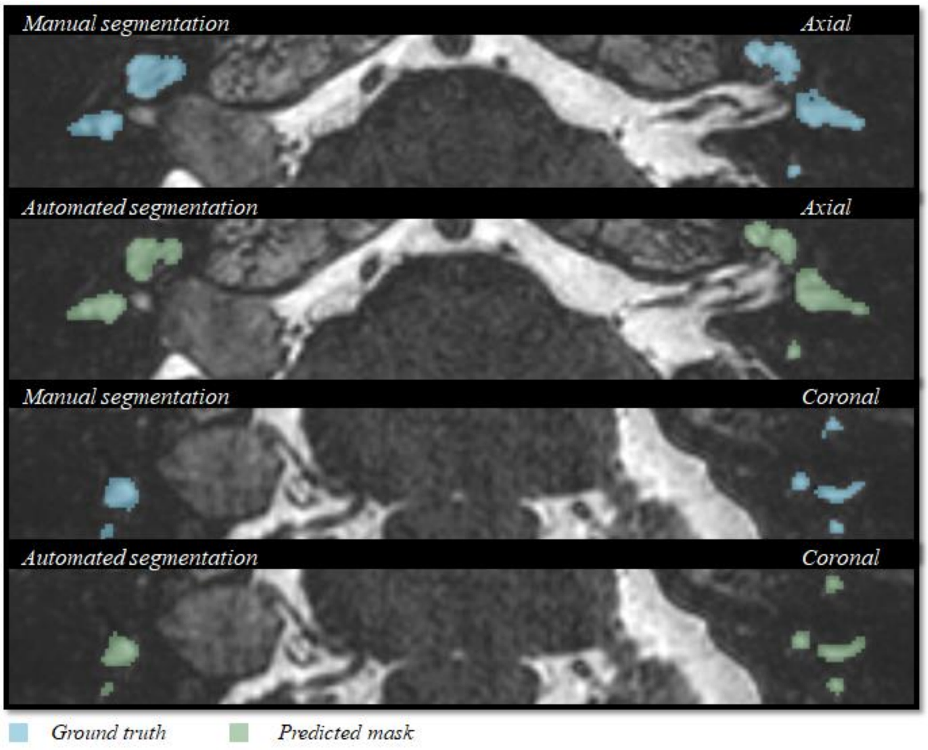


Figure 5a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows the presence of a vestibular schwannoma on the right side, with changes in the signal intensities the inner ear, indicating fibrosis. The right, superior, lateral and inferior semi-circular canals are not fully segmented. DSC: 0.8770, Ground Truth Volume: 486.78 mm^3, True Positive Volume: 434.21 mm^3 , True Positive Rate: 89.19%, False Negative Rate: 10.8% . False Positive Rate: 0.013%.

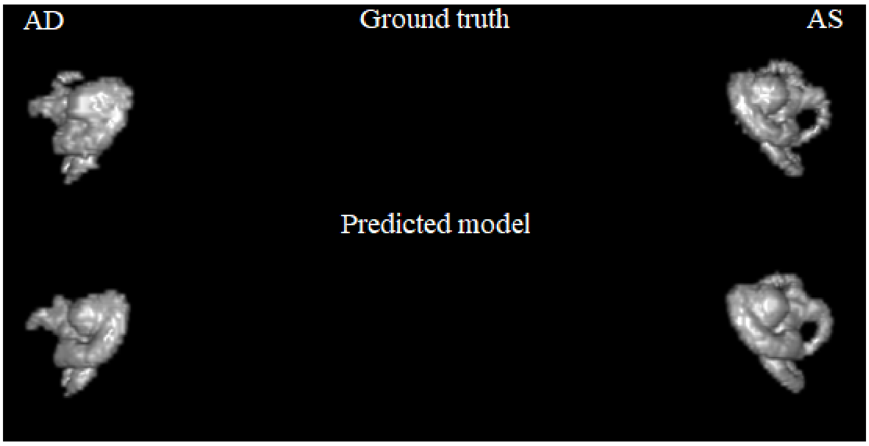


Figure 5b. The 3D volume rendering of the ground truth and the predicted mask. The superior, lateral and inferior semi-circular canals of the right inner ear were not properly displayed on MRI. The model has correctly not segmented these semi-circular canals AD= auriculum dextra, AS=auriculum sinistra.

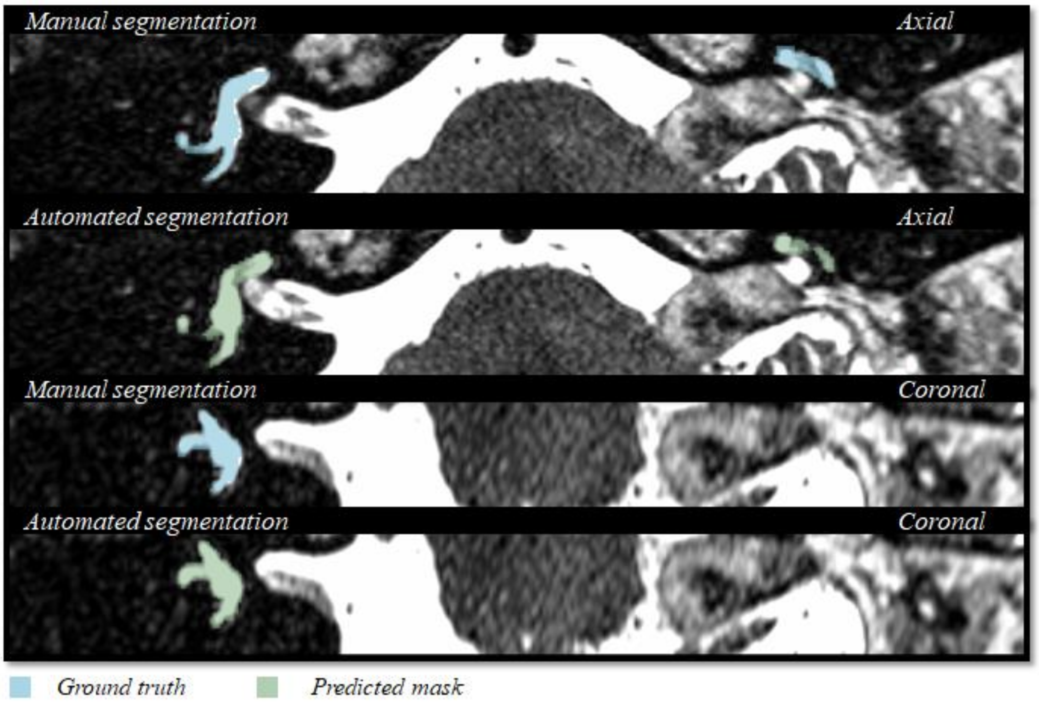


Figure 6a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows the presence of a vestibular schwannoma after a translabyrinthine resection on the left side. The left semi-circular canals and vestibule are not segmented. DSC: 0.8631, Ground Truth Volume: 395.39mm^3, True Positive Volume: 348.55 mm^3 , True Positive Rate: 88.15% , False Negative Rate: 11.84%, *False* Positive Rate: 0.0004%

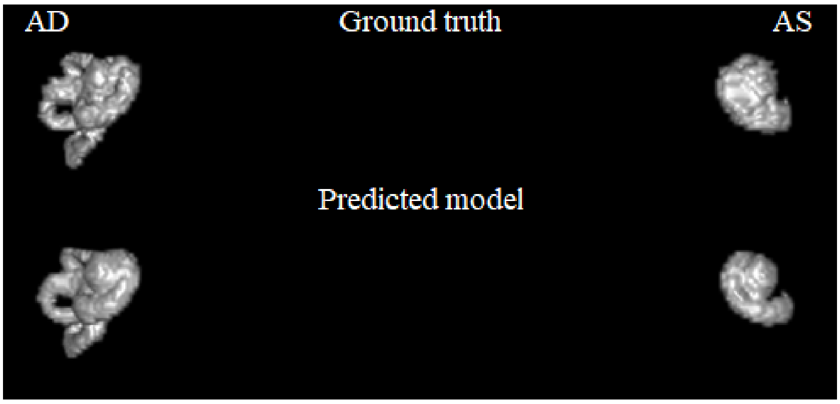


Figure 6b. The 3D volume rendering of the ground truth and the predicted mask. The semi-circular canals and the vestibule of the right inner ear were not displayed on MRI. The model has correctly not segmented the semi-circular canals and vestibule. AD= auriculum dextra, AS=auriculum sinistra.

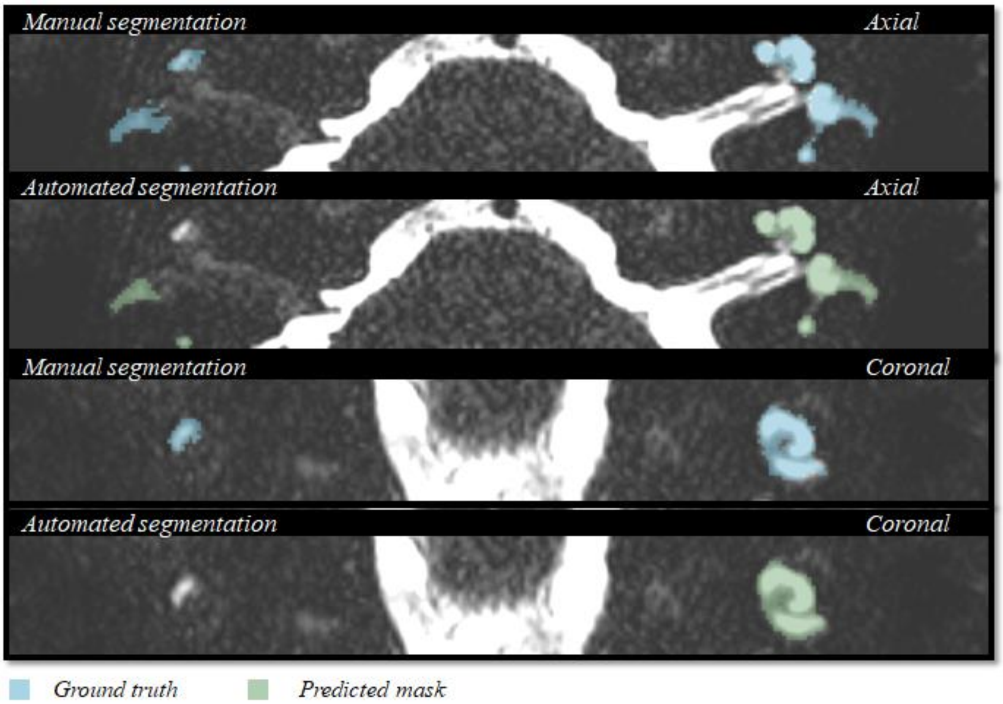


Figure 7a. Example of one of the clinical validation MRI scans in the axial and coronal plane This case shows a transmodiolar and macular schwannoma on the right side. The right vestibule and cochlea are not fully segmented. DSC: 0.8648, Ground Truth Volume: 472.39 mm^3, True Positive Volume:405.14 mm^3, True Positive Rate:85.76%, False Negative Rate:14.23% . False Positive Rate: 0.0006%.

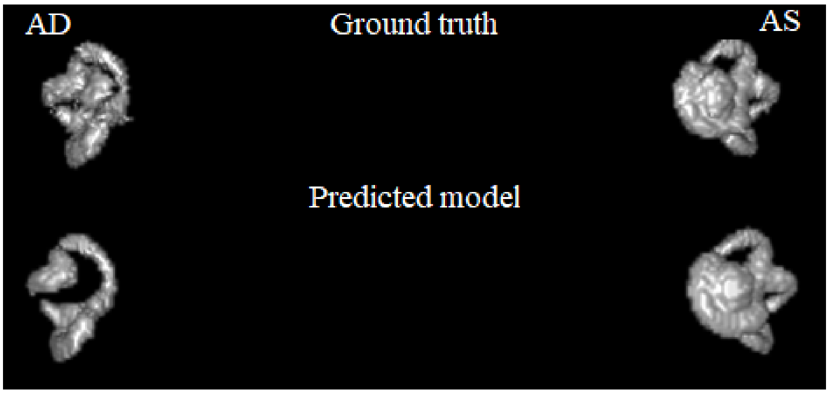


Figure 7b. The 3D volume rendering of the ground truth and the predicted mask. The vestibule and the cochlea of the right inner ear were not displayed on MRI. The model has correctly not segmented the vestibule and the cochlea. AD= auriculum dextra, AS=auriculum sinistra.

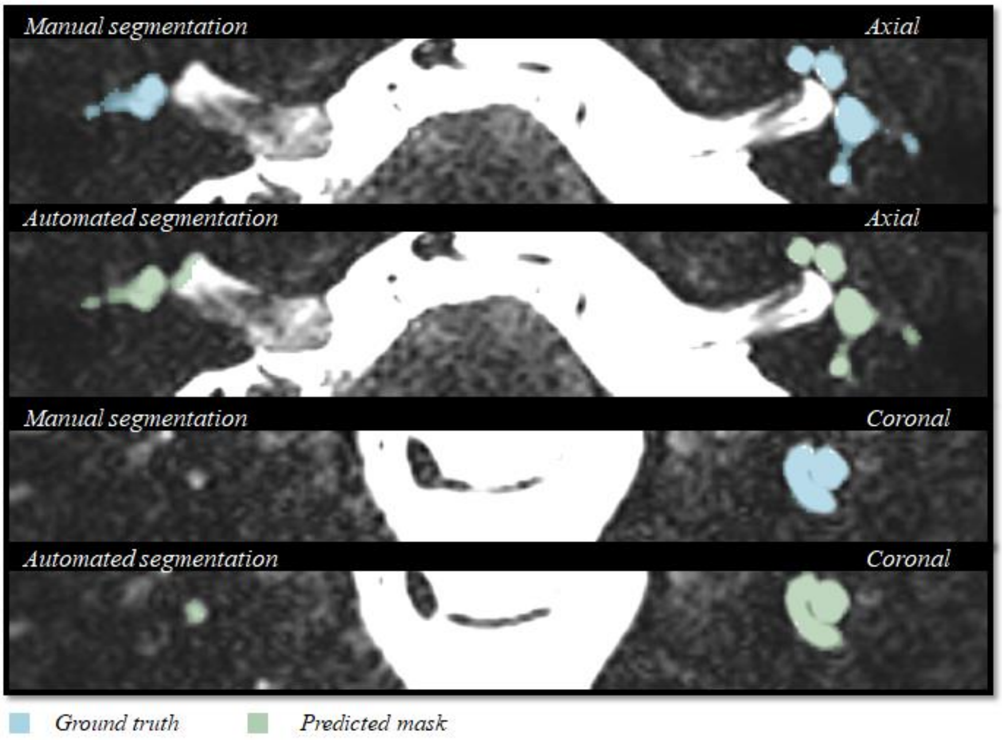


Figure 8a. Example of one of the clinical validation MRI scans in the axial and coronal plane. This case shows obliteration of the apical, middle and basal turn of the right cochlea, indicating the presence of either labyrinthitis ossificans or a vestibular schwannoma. The right superior and inferior semi-circular canals, the vestibule and the cochlea are not fully segmented. DSC: 0.8810, Ground Truth Volume: 369.603 mm^3, True Positive Volume: 325.95 mm^3 , True Positive Rate: 88.19%, False Negative Rate: 11.80%. *False* Positive Rate: 0.0005%.

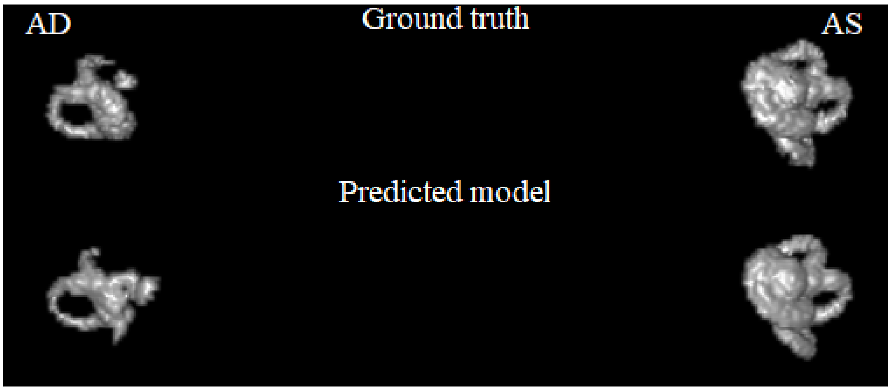


Figure 8b. The 3D volume rendering of the ground truth and the predicted mask. The superior and inferior semi-circular canals, the vestibule and the cochlea of the right inner ear were not displayed on MRI. The model has correctly not segmented these structures. AD= auriculum dextra, AS=auriculum sinistra.

IMPACT ADDENDUM

The primary goal of the work presented in this thesis is to provide Radiomics methodologies for disease detection, localization, quantification, diagnosis, prognosis, and treatment outcome prediction. The proposed methods have been tested on external validation cohorts, with different imaging parameters and morphological information, to assess their generalizability and robustness, and to pave the way for a possible application in a real clinical setting.

SCIENTIFIC IMPACTS

The combination of AI-based auto-segmentation model and radiomics features extracted from the segmented lung GTV region on CT images described in **chapter 5** can be leveraged for gaining more insights on clinical endpoints like genetic mutation status in a tumor, progressive response to treatment, determining automated RECIST scoring, etc. Similar work has been done to predict HPV status from standard CT images of anal and vulvar cancer patients. [1], automatic RECIST score evaluation using diffusion MRI [2]. Furthermore, radiomic features extracted using such methodologies, in combination with genomic, and proteomic data can lead to Biomarker discovery in the future.

The AI model and the automatic labeler methodology described in **chapter 6** can be adapted to any application that requires the recording of localized information on images in a text format, linked for example to Electronic Health Records (EHR). This would greatly benefit an optimal connection between the AI-based system and patient records documentation.

The deep learning methodology for segmentation described in **chapter 8** can be reproduced for any application that requires delineation of regions on images and such a model can be used as a support tool for detection, localization, and quantification purposes in radiology. An adaptation of the model mentioned in the chapter was trained and validated on liver lesions [3].

Furthermore, all our studies are published in medical and technical peer-reviewed international journals such as, Diagnostics, Journal of Neurology, IEEE Access, Medical Research Reviews, Journal of Personalized Medicine, La radiologia medica, European Respiratory Journal, and Journal of Clinical Oncology and the publications are available as open access.

SOCIETAL IMPACTS

All the tools that were developed in the context of my research can be efficiently deployed and usable in a clinical setting. In actual fact, The AI-based tools developed in **chapter 2** called COVIA have been implemented at CHU-Liege and were used as an additional diagnostic research tool when the pandemic was at its peak in 2020. The methodology in **chapters 2 & 3** could be adapted to support incidental findings or to provide a second independent verification of the occurrence of the disease, inclusive and beyond the emergency status of this pandemic. Especially, the method for localization of abnormalities in lungs can be leveraged to analyze new unseen abnormalities if in-case another pandemic occurs.

Models and methodologies presented in the thesis have already been used in a research context by both Biotech and major pharmaceutical companies to explore possible improvements in drug development. One of the models has been filed as a novel invention and is currently being evaluated. To be accepted as an international patent application [4].

Finally, leveraging, and improvement of our methods can greatly benefit society by providing increased accuracy and efficiency concerning diagnosis, prognosis, and appropriate treatment selection and ultimately superior understanding of disease or biology of any type of abnormality presenting in our body.

REFERENCES

[1] R. T. H. Leijenaar *et al.*, “External validation of a radiomic signature to predict HPV (p16) status from standard CT images of anal and vulvar cancer patients.,” *https://doi.org/10.1200/JCO.2021.39.15_suppl.e15502*, vol. 39, no. 15_suppl, pp. e15502–e15502, May 2021, doi: 10.1200/JCO.2021.39.15_SUPPL.E15502.

[2] E. Baidya Kayal, D. Kandasamy, R. Yadav, S. Bakhshi, R. Sharma, and A. Mehndiratta, “Automatic segmentation and RECIST score evaluation in osteosarcoma using diffusion MRI: A computer aided system process,” *Eur J Radiol*, vol. 133, Dec. 2020, doi: 10.1016/J.EJRAD.2020.109359.

[3] A. Vaidyanathan *et al.*, “PO-1710: A novel AI solution for auto-segmentation of multi-origin liver neoplasms,” *Radiotherapy and Oncology*, vol. 152, pp. S944–S945, Nov. 2020, doi: 10.1016/S0167-8140(21)01728-X.

[4] “BE1028836B1 - Methods and systems for biomedical image segmentation based on a combination of arterial and portal image information - Google Patents.” https://patents.google.com/patent/BE1028836B1/en?inventor=Akshayaa+Vaidyanathan&oq=Akshayaa+Vaidyanathan (accessed Sep. 06, 2022).

ACKNOWLEDGMENTS

First of all, I would like to thank prof. Philippe Lambin for providing me with numerous exciting opportunities throughout my research journey. Your inexhaustible enthusiasm and passion for the field of AI and precision medicine have inspired me throughout these years. Your feedback has been highly constructive and insightful. Thanks for that.

Furthermore, I would like to thank my industry supervisor Sean Walsh for persistently motivating me throughout my Ph.D. journey. The scientific advice I have received from you, during various projects I have worked on in the industry was very useful. Without your support, I wouldn't have reached my destination. A huge thanks to you for your trust in my research, to be used as part of commercial activities in the company. This gave me more confidence to further put forth my ideas.

I would like to thank Ralph Leijenaar for providing me with insights on image processing and radiomics. Your knowledgeable insights undeniably played a huge role in my research work.

Many thanks to Henry Woodruff. I had a chance to work under your guidance mostly during the first year of my Ph.D. You have been very supportive during the times I have doubted myself as a researcher. Thanks to your efforts in initiating research collaborations between me and researchers in other institutes (one of which I enjoyed the most is the collaboration on the MRI HCC project with Imperial college!).

I would like to thank the most my colleague Fabio Bottari, a very talented medical writer who helped me with scientific writing. Without your commendable writing skills and timely support, I wouldn't have published my research work while maintaining good scientific writing standards.

Many thanks to Wim Vos for providing all the resources and framework required for my research work. The scientific collaborations initiated by you were very useful for my research work. Thanks for that. Thanks to Mariaelena Occhipinti for providing insights into the clinical aspects of my research work.

Many thanks to my colleague Benjamin Miraglio for the timely support by facilitating various technical needs to accomplish my research tasks.

A huge thanks to Steven Lauwers for providing me the knowledge and support to organize efficiently my time and duties, balancing my Ph.D. research with the company projects.

Big thanks to all my colleagues at Radiomics and fellow researchers at D-Lab, Yousif Widaatalla, Abdalla Ibrahim, Sergey Primakov, and Manon Beuque for their interesting collaborations throughout these years.

A special thanks to Julien Guiot for our research collaboration on respiratory research during the COVID crisis and Marly Lubbe, Marc Hoof Van, and Raymond van den Berg for the most interesting collaborations I have had on our research on Inner ear diseases.

Finally, I would like to thank Arthur Jochems by whom I was first considered for an open Ph.D. position at the D-Lab, I was eventually introduced to Philippe, Henry, and Sean and that's where my Ph.D. journey started. I thank Arthur for his trust and giving me a chance.

AKSHAYAA VAIDYANATHAN

174, BLOCK D, BELLEVUE, DUBLIN, D08 DA48, IRELAND ☎ +353 899466108

DETAILS

174, Block D, Bellevue, Dublin, D08 DA48, Ireland
+353 899466108
akshayaavaidya@gmail.com

Date of birth
06/06/1992

EMPLOYMENT HISTORY

Head of R&D at Radiomics, Liege
June 2021 — June 2022

Lead AI scientist at Radiomics, Liege
June 2020 — June 2021

Artificial Intelligence Researcher at Radiomics, Liege
May 2019 — May 2020

Artificial Intelligence Researcher at Donders institute of Brain, Cognition and Behaviour, Nijmegen
September 2018 — December 2018

Student researcher at Center of Biorobotics, Tallinn, Estonia
July 2018 — August 2018

Machine learning engineer at Synchronoss technologies, Dublin
May 2017 — July 2018

IT Analyst at Tata Consultancy services, Chennai
December 2013 — July 2017

EDUCATION

PhD, Maastricht University, Maastricht
January 2019 — Present

Worked as AI researcher at the D-Lab. Collaborated with multiple research groups on projects related to application of AI on medical imaging data.

MSc. Computer Science (Cognitive Psychology major), University College Dublin, Dublin
January 2017 — September 2018

As part of masters' thesis, worked as researcher intern at Synchronoss technologies. Contributed in research and development of AI based image analytics platform for mobile application.

BE. Biomedical Sciences, Anna University, Chennai
June 2009 — June 2013

As part of bachelors' thesis, worked on a research project involving analysis of cardiac data for abnormalities detection based on mathematical transformations applied on ECG signals.

COURSES

Computational Psychiatry, University of Zurich
August 2020 — September 2020

CONFERENCES

BIR 2022, Oral presentation

ECR 2022, Poster presentation

ECR 2021, Oral presentation

ESMO 2020, Merit award holder, Oral presentation

ESTRO 2020, Poster presentation

GROW Science Day - 2020, Oral presentation

LIST OF MANUSCRIPTS

PUBLISHED ARTICLES

- Guiot, Julien, **Akshayaa Vaidyanathan**, Louis Deprez, Fadila Zerka, Denis Danthine, Anne-Noëlle Frix, Marie Thys, Monique Henket, Gregory Canivet, Stephane Mathieu, Evanthia Eftaxia, Philippe Lambin, Nathan Tsoutzidis, Benjamin Miraglio, Sean Walsh, Michel Moutschen, Renaud Louis, Paul Meunier, Wim Vos, Ralph T.H. Leijenaar, and Pierre Lovinfosse. 2021. "Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19" *Diagnostics* 11, no. 1: 41. https://doi.org/10.3390/diagnostics11010041
- **Akshayaa Vaidyanathan**, Julien Guiot, Fadila Zerka, Flore Belmans, Ingrid Van Peufflik, Louis Deprez, Denis Danthine, Gregory Canivet, Philippe Lambin, Sean Walsh, Mariaelena Occchipinti, Paul Meunier, Wim Vos, Pierre Lovinfosse & Ralph T.H. Leijenaar, *An externally validated fully automated deep learning algorithm to classify COVID-19 and other pneumonias on chest CT*, ERJ Open (2022)
- **Vaidyanathan, A.,** van der Lubbe, M.F.J.A., Leijenaar, R.T.H. et al. Deep learning for the fully automated segmentation of the inner ear on MRI. Sci Rep 11, 2885 (2021). https://doi.org/10.1038/s41598-021-82289-y
- van der Lubbe, M.F.J.A., **Vaidyanathan, A.**, Van Rompaey, V. et al. The "hype" of hydrops in classifying vestibular disorders: a narrative review. J Neurol 267, 197–211 (2020). https://doi.org/10.1007/s00415-020-10278-8
- Guiot, J, **Vaidyanathan, A**, Deprez, L, et al. A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev. 2022; 42: 426- 440. https://doi.org/10.1002/med.21846
- van der Lubbe, M.F.J.A., **Vaidyanathan, A**., de Wit, M. et al. A non-invasive, automated diagnosis of Menière's disease using radiomics and machine learning on conventional magnetic resonance imaging: A multicentric, case-controlled feasibility study. Radiol med (2021). https://doi.org/10.1007/s11547-021-01425-w
- Mohamed Yacin Sikkandar, **V. Akshayaa**, Acharya Divya Dinesh, and L. Dinikshaa Sree International Journal of Biomedical Engineering and Technology 2013 13:1, 69-86
- - F. Zerka, V. Urovi, F. Bottari, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, M. Gueuning, **A. Vaidyanathan**, W. Vos, M. Occhipinti, H.C. Woodruff, M. Dumontier, P. Lambin, Privacy preserving distributed learning classifiers – Sequential learning with small sets of data, Computers in Biology and Medicine. 136 (2021) 104716. https://doi.org/10.1016/j. compbiomed.2021.104716.
- - F. Zerka, V. Urovi, **A. Vaidyanathan**, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. GabraniJuma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ ACCESS.2020.3029445.
- Frix, A.-N.; Cousin, F.; Refaee, T.; Bottari, F.; **Vaidyanathan, A**.; Desir, C.; Vos, W.; Walsh, S.; Occhipinti, M.; Lovinfosse, P.; Leijenaar, R.T.H.; Hustinx, R.; Meunier, P.; Louis, R.; Lambin, P.; Guiot, J. Radiomics in Lung Diseases Imaging: State-of-the-Art for Clinicians. *J. Pers. Med.* **2021**, *11*, 602. https://doi.org/10.3390/jpm11070602

PUBLISHED ABSTRACTS

- 4MO A novel AI solution for auto-segmentation of multi-origin liver neoplasms **Vaidyanathan, A**. et al. Annals of Oncology, Volume 31, S246, https://doi.org/10.1016/j.annonc.2020.08.157
- Julien Guiot, **Akshayaa Vaidyanathan**, Fadila Zerka, Louis Deprez, Denis Danthine, Anne-Noëlle Frix, Fabio Bottari, Monique Henket, Stephane Mathieu, Philippe Lambin, Sean Walsh, Mariaelena Occhipinti, Benoit Misset, Louis Renard, Paul Meunier, Wim Vos, Ralph T.H. Leijenaar, Pierre Lovinfosse, European Respiratory Journal Sep 2021, 58 (suppl 65) PA361; DOI: 10.1183/13993003.congress-2021.PA361
- Bart Liefers, Johanna Maria Colijn, Cristina González-Gonzalo, **Akshayaa Vaidyanathan**, Harm van Zeeland, Paul Mitchell, Caroline C W Klaver, Clara I Sanchez; Prediction of areas at risk of developing geographic atrophy in color fundus images using deep learning. *Invest. Ophthalmol. Vis. Sci.* 2019;60(9):1455.

- Cristina González-Gonzalo, Bart Liefers, **Akshayaa Vaidyanathan**, Harm van Zeeland, Caroline C W Klaver, Clara I Sanchez; Opening the "black box" of deep learning in automated screening of eye diseases. *Invest. Ophthalmol. Vis. Sci.* 2019;60(9):1443.
- **Akshayaa Vaidyantahan**, Julien Guiot, Fadila Zerka, Primakov Sergey, L Deprez, D Danthine, G Canivet, P Lambin, S Walsh, P Meunier, Wim Vos, R.T.H Leijenaar, P Lovinfosse. European Respiratory Journal Sep 2020, 56 (suppl 64) 3587; DOI: 10.1183/13993003.congress-2020.3587
- Walsh, S., Leijenaar, R., Miraglio, B., Barakat, S., Zerka, F., **Vaidyanathan, A.**, & Lambin, P. (2020). OC-0587: Prospective Validation of a Radiomics Signature for Chemoradiotherapy Lung Cancer Patients. Radiotherapy and Oncology, 152, S330-S331.
- fadila Zerka, **Akshayaa Vaidyanathan**, Julien Guiot, Louis Deprez, Denis Danthine, Grégory Canivet, Mathieu Stéphane, E Eftaxia, Monique Henket, M Thys, Philippe Lambin, Nathan Tsoutzidis, Benjamin Miraglio, Sean Wlash, Paul Meunier, Wim Vos, Ralph Leijenaar, Pierre Lovinfosse. European Respiratory Journal Sep 2020, 56 (suppl 64) 4152; DOI: 10.1183/13993003.congress-2020.4152
- Zerka, F., **Vaidyanathan, A.**, Barakat, S., Benjamin, M., TH, L. R., Sean, W., & Philippe, L. (2020). PO-1744: Privacy preserving distributed liver tumor segmentation. Radiotherapy and Oncology, 152, S968-S969.
- Walsh, S., Leijenaar, R., Miraglio, B., Barakat, S., Zerka, F., **Vaidyanathan, A.**, & Lambin, P. (2020). OC-0587: Prospective Validation of a Radiomics Signature for Chemoradiotherapy Lung Cancer Patients. Radiotherapy and Oncology, 152, S330-S331.

PUBLICATIONS UNDER REVIEW

- **Akshayaa Vaidyanathan**, Flore Belmans, Fabio Bottari, François Blistein, Ingrid van Peufflik, Wim Vos, Mariaelena Occhipinti, Philippe Lambin, Julien Guiot, Sean Walsh, Externally validated deep learning model for the diagnosis and detection of pulmonary embolism on chest CTPA images
- Ralph T.H. Leijenaar, **Akshayaa Vaidyanathan**, Dirk De Ruysscher, Andre Dekker, Lizza E.L. Hendriks, Petros Kalendralis, Flore Belmans, Martin Gueuning, Fabio Bottari, Mariaelena Occhipinti, Julien Guiot, Pierre Lovinfosse, Wim Vos, Philippe Lambin, & Sean Walsh, Prospective validation of a prognostic radiomics signature for patients with non-metastatic NSCLC treated with standard of care non-surgical therapy
- Abdalla Ibrahim, **Akshayaa Vaidyanathan**, Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Refaee, Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Henry C. Woodruff, Helle D. Zacho, Sean Walsh, Wim Vos, Mariaelena Occhipinti, François-Xavier Hanin, Philippe Lambin, Felix M. Mottaghy, Roland Hustin., Deep learning based identification of bone scintigraphies containing metastatic bone disease foci