

Radiomics and artificial intelligence in radiology and nuclear medicine

Edited by

Giorgio Treglia and Salvatore Annunziata

Published in

Frontiers in Medicine



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2634-7
DOI 10.3389/978-2-8325-2634-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Radiomics and artificial intelligence in radiology and nuclear medicine

Topic editors

Giorgio Treglia — Ente Ospedaliero Cantonale (EOC), Switzerland

Salvatore Annunziata — Fondazione Policlinico Universitario A. Gemelli IRCCS, Italy

Citation

Treglia, G., Annunziata, S., eds. (2023). *Radiomics and artificial intelligence in radiology and nuclear medicine*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-2634-7

Table of contents

05	Editorial: Radiomics and artificial intelligence in radiology and nuclear medicine Salvatore Annunziata and Giorgio Treglia
07	Prediction of MYCN Amplification, 1p and 11q Aberrations in Pediatric Neuroblastoma via Pre-therapy 18F-FDG PET/CT Radiomics Luodan Qian, Shen Yang, Shuxin Zhang, Hong Qin, Wei Wang, Ying Kan, Lei Liu, Jixia Li, Hui Zhang and Jigang Yang
18	Application of 18F-FDG PET-CT Images Based Radiomics in Identifying Vertebral Multiple Myeloma and Bone Metastases Zhicheng Jin, Yongqing Wang, Yizhen Wang, Yangting Mao, Fang Zhang and Jing Yu
30	Chest L-Transformer: Local Features With Position Attention for Weakly Supervised Chest Radiograph Segmentation and Classification Hong Gu, Hongyu Wang, Pan Qin and Jia Wang
40	Healthy Organs Uptake on Baseline ¹⁸F-FDG PET/CT as an Alternative to Total Metabolic Tumor Volume to Predict Event-Free Survival in Classical Hodgkin's Lymphoma David Morland, Elizabeth Katherine Anna Triumbari, Elena Maiolo, Annarosa Cuccaro, Giorgio Treglia, Stefan Hohaus and Salvatore Annunziata
48	Radiomics and Its Applications and Progress in Pancreatitis: A Current State of the Art Review Gaowu Yan, Gaowen Yan, Hongwei Li, Hongwei Liang, Chen Peng, Anup Bhetuwal, Morgan A. McClure, Yongmei Li, Guoqing Yang, Yong Li, Linwei Zhao and Xiaoping Fan
61	Non-contrast and contrast enhanced computed tomography radiomics in preoperative discrimination of lung invasive and non-invasive adenocarcinoma Yingli Sun, Wei Zhao, Kaiming Kuang, Liang Jin, Pan Gao, Shaofeng Duan, Yi Xiao, Jun Liu and Ming Li
71	PET image enhancement using artificial intelligence for better characterization of epilepsy lesions Anthime Flaus, Tahya Deddah, Anthonin Reilhac, Nicolas De Leiris, Marc Janier, Ines Merida, Thomas Grenier, Colm J. McGinnity, Alexander Hammers, Carole Lartizien and Nicolas Costes
85	Exploratory analysis of radiomic as prognostic biomarkers in ¹⁸F-FDG PET/CT scan in uterine cervical cancer Nadja Rolim Gonçalves de Alencar, Marcos Antônio Dórea Machado, Felipe Alves Mourato, Mércia Liane de Oliveira, Thauan Fernandes Moraes, Luiz Alberto Reis Mattos Junior, Tien-Man Cabral Chang, Carla Rameri Alexandre Silva de Azevedo and Simone Cristina Soares Brandão

- 93 **MRI-derived radiomics to guide post-operative management of glioblastoma: Implication for personalized radiation treatment volume delineation**
S. Chiesa, R. Russo, F. Beghella Bartoli, I. Palumbo, G. Sabatino, M. C. Cannatà, R. Gigli, S. Longo, H. E. Tran, L. Boldrini, N. Dinapoli, C. Votta, D. Cusumano, F. Pignotti, M. Lupattelli, F. Camilli, G. M. Della Pepa, G. Q. D'Alessandris, A. Olivi, M. Balducci, C. Colosimo, M. A. Gambacorta, V. Valentini, C. Aristei and S. Gaudino
- 102 **External validation of a convolutional neural network for the automatic segmentation of intraprostatic tumor lesions on ^{68}Ga -PSMA PET images**
Samuele Ghezzi, Sofia Mongardi, Carolina Bezzi, Ana Maria Samanes Gajate, Erik Preza, Irene Gotuzzo, Francesco Baldassi, Lorenzo Jonghi-Lavarini, Ilaria Neri, Tommaso Russo, Giorgio Brembilla, Francesco De Cobelli, Paola Scifo, Paola Mapelli and Maria Picchio
- 109 **Artificial intelligence-based ^{68}Ga -DOTATOC PET denoising for optimizing $^{68}\text{Ge}/^{68}\text{Ga}$ generator use throughout its lifetime**
Elske Quak, Kathleen Weyts, Cyril Jaudet, Anaïs Prigent, Gauthier Foucras and Charline Lasnon
- 118 **Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review**
Bart M. de Vries, Gerben J. C. Zwezerijnen, George L. Burchell, Floris H. P. van Velden, Catharina Willemien Menke-van der Houven van Oordt and Ronald Boellaard



OPEN ACCESS

EDITED AND REVIEWED BY
Francesco Cicone,
Magna Græcia University, Italy

*CORRESPONDENCE
Giorgio Treglia
✉ giorgio.treglia@eoc.ch

RECEIVED 03 May 2023
ACCEPTED 10 May 2023
PUBLISHED 23 May 2023

CITATION
Annunziata S and Treglia G (2023) Editorial:
Radiomics and artificial intelligence in radiology
and nuclear medicine. *Front. Med.* 10:1216434.
doi: 10.3389/fmed.2023.1216434

COPYRIGHT
© 2023 Annunziata and Treglia. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Editorial: Radiomics and artificial intelligence in radiology and nuclear medicine

Salvatore Annunziata¹ and Giorgio Treglia^{2,3,4*}

¹Unità di Medicina Nucleare, GSTeP Radiopharmacy - TracerGLab, Dipartimento di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Rome, Italy, ²Division of Nuclear Medicine, Imaging Institute of Southern Switzerland, Ente Ospedaliero Cantonale, Bellinzona, Switzerland, ³Faculty of Biomedical Sciences, Università della Svizzera italiana, Lugano, Switzerland, ⁴Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

KEYWORDS

radiology, artificial intelligence, radiomics, machine learning, nuclear medicine, imaging, deep learning

Editorial on the Research Topic

Radiomics and artificial intelligence in radiology and nuclear medicine

Artificial intelligence (AI) and radiomics algorithms in radiology and nuclear medicine have demonstrated a good performance as diagnostic, predictive or prognostic markers for several diseases with a high potential to be used as clinical tools. However, these algorithms should be further validated in clinical practice to spread their routine use worldwide.

After the successful publication of a previous Research Topic about Radiomics in Positron Emission Tomography (PET) on Frontiers in Medicine in 2021 (<https://www.frontiersin.org/research-topics/15427/artificial-intelligence-in-positron-emission-tomography>) a wider and more comprehensive Research Topic on the role of AI and radiomics in radiology and nuclear medicine was launched in 2022. This Research Topic comprises 12 articles.

For successful employment of AI and deep learning algorithms in the clinical practice, explainable artificial intelligence (XAI) could be introduced for several imaging modalities as clearly discussed in the comprehensive review of De Vries et al. including 75 articles (De Vries et al.). However, the authors demonstrated that there is currently no clear consensus on how XAI should be used in order to close the gap between medical professionals and deep learning algorithms for clinical implementation. Furthermore, De Vries et al. also suggested a systematic technical and clinical quality assessment of XAI methods.

Seven articles included in this Research topic are dedicated to the use of AI in oncological imaging.

Radiomic features could be very useful for a better prognostic stratification in patients with glioblastoma. The study of Chiesa et al. including 90 patients with glioblastoma applied a radiomic analysis focusing on healthy tissue ring around the surgical cavity on post-operative magnetic resonance imaging. This study provided a preliminary model for a decision support tool for a customization of the radiation target volume in glioblastoma patients to achieve a margin reduction strategy (Chiesa et al.).

Sun et al. assessed the value of radiomics based on computed tomography (CT) images in the preoperative discrimination between lung invasive adenocarcinomas and non-invasive adenocarcinomas among 1,185 pulmonary nodules. The authors found that radiomics based on CT images showed good predictive performance in discriminating between these tumoral entities, especially in part solid nodule group. Furthermore, radiomics based on contrast enhanced CT images provided no additional value compared to non-contrast enhanced CT images (Sun et al.).

A brief research report evaluated the performance of fluorine-18 fluorodeoxyglucose (^{18}F FDG) PET/CT radiomic features to predict overall survival in 50 patients with locally advanced uterine cervical carcinoma. The authors found that standardized uptake value peak (SUV_{peak}) and the textural feature gray-level run-length matrix (GLRLM) presented the best performance to predict overall survival in patients with cervical cancer undergoing chemotherapy and brachytherapy (Goncalves de Alencar et al.).

A retrospective study assessed the predictive ability of ^{18}F FDG PET/CT radiomic features for MYCN, 1p and 11q abnormalities in 122 pediatric patients with neuroblastoma. The authors clearly demonstrated that baseline ^{18}F FDG PET/CT radiomics is able to predict MYCN amplification and 1p and 11 aberrations in patients with neuroblastoma, thus aiding tumor staging, risk stratification and disease management (Qian et al.).

Another retrospective study on 131 patients explored the application of ^{18}F FDG PET/CT radiomics in the identification and correct classification of spine multiple myeloma lesions and bone metastases. The radiomics model constructed based on ^{18}F FDG PET/CT images achieved satisfactory diagnostic performance for the classification of multiple myeloma and bone metastases. In addition, the radiomics model showed significant improvement in diagnostic performance compared to human experts and PET conventional parameters (Jin et al.).

A retrospective study from Morland et al. estimated the ability of a new index, uptake formula, including healthy organs standardized uptake values on ^{18}F FDG PET/CT to predict event free survival in 163 patients with Hodgkin lymphoma. The Uptake Formula showed a similar performance to total metabolic tumor volume in predicting event free survival in Hodgkin lymphoma (Morland et al.).

Ghezzi et al. tested on a cohort of 85 prostate cancer patients a recently proposed convolutional neural network for the automatic segmentation of intraprostatic cancer lesions on prostate specific membrane antigen PET images. The authors demonstrated that the AI model could be used to automatically segment intraprostatic cancer lesions to define the volume of interest for radiomics or deep learning analysis. However, more robust performance is needed for the generation of AI-based decision support technologies to be proposed in clinical practice (Ghezzi et al.).

Beyond oncological indications of imaging methods, AI may be also applied for other indications. For instance, the review article of Yan et al. have summarized the application of radiomics for predicting recurrent pancreatitis, evaluating the clinical

severity of pancreatitis, differentiating pancreatitis from pancreatic adenocarcinoma, and functional abdominal pain from pancreatitis, identifying pancreatitis, its risk factors and complications (Yan et al.).

Flaus et al. developed a deep learning-based ^{18}F FDG PET image enhancement method using simulated brain PET to improve visualization of epileptogenic lesions. However, the authors recommended further evaluation to generalize their method and to assess its clinical performance in a larger cohort (Flaus et al.).

Weakly supervised deep learning models have gained increasing popularity in medical image segmentation. However, these models are not suitable for the critical characteristics presented in chest radiographs: the global symmetry of chest radiographs and dependencies between lesions and their positions. In their study, Gu et al. proposed a weakly supervised model, Chest L-Transformer, to take these characteristics into account. The authors demonstrated a significant segmentation performance improvement over the current state-of-the-art while achieving competitive classification performance (Gu et al.).

Lastly, an original article by Quak et al. including 67 patients demonstrated that the degradation of image quality on PET due to a reduction in injected activity at the end of the $^{68}\text{Ge}/^{68}\text{Ga}$ generator lifespan can be effectively counterbalanced by using AI-based PET denoising (Quak et al.).

Finally, we would like to underline that AI and radiomics tools are widely used for research purpose in the fields of radiology and nuclear medicine. Nevertheless, large validation protocols and real-life experience are needed to allow an increasing use of these tools in clinical practice, with possible benefit for patients' treatments and outcomes.

Author contributions

SA and GT drafted the manuscript and revised the final version. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



Prediction of MYCN Amplification, 1p and 11q Aberrations in Pediatric Neuroblastoma via Pre-therapy 18F-FDG PET/CT Radiomics

Luodan Qian^{1†}, Shen Yang^{2†}, Shuxin Zhang¹, Hong Qin², Wei Wang¹, Ying Kan¹, Lei Liu³, Jixia Li^{4,5*†}, Hui Zhang⁶ and Jigang Yang^{1*†}

OPEN ACCESS

Edited by:

Giorgio Treglia,
Ente Ospedaliero Cantonale
(EOC), Switzerland

Reviewed by:

Bi Cong Yan,
Shanghai Sixth People's
Hospital, China
Salvatore Annunziata,
Fondazione Policlinico Universitario A.
Gemelli IRCCS, Italy

*Correspondence:

Jixia Li
j.li@auckland.ac.nz
Jigang Yang
yangjigang@ccmu.edu.cn

[†]These authors have contributed
equally to this work and share first
authorship

[‡]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

Received: 21 December 2021

Accepted: 13 January 2022

Published: 18 March 2022

Citation:

Qian L, Yang S, Zhang S, Qin H,
Wang W, Kan Y, Liu L, Li J, Zhang H
and Yang J (2022) Prediction of MYCN
Amplification, 1p and 11q Aberrations
in Pediatric Neuroblastoma via
Pre-therapy 18F-FDG PET/CT
Radiomics. *Front. Med.* 9:840777.
doi: 10.3389/fmed.2022.840777

¹ Department of Nuclear Medicine, Beijing Friendship Hospital, Capital Medical University, Beijing, China, ² Department of Surgical Oncology, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China, ³ Sinounion Medical Technology (Beijing) Co., Ltd., Beijing, China, ⁴ Department of Molecular Medicine and Pathology, School of Medical Science, The University of Auckland, Auckland, New Zealand, ⁵ Department of Laboratory Medicine of Medical School, Foshan University, Foshan, China, ⁶ Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, China

Purpose: This study aimed to assess the predictive ability of 18F-FDG PET/CT radiomic features for MYCN, 1p and 11q abnormalities in NB.

Method: One hundred and twenty-two pediatric patients (median age 3.2 years, range, 0.2–9.8 years) with NB were retrospectively enrolled. Significant features by multivariable logistic regression were retained to establish a clinical model (C_model), which included clinical characteristics. 18F-FDG PET/CT radiomic features were extracted by Computational Environment for Radiological Research. The least absolute shrinkage and selection operator (LASSO) regression was used to select radiomic features and build models (R_model). The predictive performance of models constructed by clinical characteristic (C_model), radiomic signature (R_model), and their combinations (CR_model) were compared using receiver operating curves (ROCs). Nomograms based on the radiomic score (rad-score) and clinical parameters were developed.

Results: The patients were classified into a training set ($n = 86$) and a test set ($n = 36$). Accordingly, 6, 8, and 7 radiomic features were selected to establish R_models for predicting MYCN, 1p and 11q status. The R_models showed a strong power for identifying these aberrations, with area under ROC curves (AUCs) of 0.96, 0.89, and 0.89 in the training set and 0.92, 0.85, and 0.84 in the test set. When combining clinical characteristics and radiomic signature, the AUCs increased to 0.98, 0.91, and 0.93 in the training set and 0.96, 0.88, and 0.89 in the test set. The CR_models had the greatest performance for MYCN, 1p and 11q predictions ($P < 0.05$).

Conclusions: The pre-therapy 18F-FDG PET/CT radiomics is able to predict MYCN amplification and 1p and 11q aberrations in pediatric NB, thus aiding tumor stage, risk stratification and disease management in the clinical practice.

Keywords: 18F-FDG PET/CT, radiomics, neuroblastoma, MYCN amplification, 1p aberration, 11q aberration

INTRODUCTION

Neuroblastoma (NB), the most common extracranial solid pediatric tumor, accounts for about 8–10% of all childhood cancer and 12–15% of childhood cancer mortality (1). Using selected clinical, pathologic, and genetic factors, patients diagnosed with NB can be classified into different risk groups for treatment (2). Previous studies have shown that patient outcomes of NB are highly correlated with risk stratification, with more than 90% cure in non-high risk patients and <50% event-free survival rate in high risk patients (3). It is therefore very important to obtain a better understanding of risk factors so that treatment strategies for children with NB can be tailored accordingly. Previous studies have demonstrated the value of prognostic factors such as patients age, tumor stage using the International Neuroblastoma Staging System (INSS), tumor histopathology using the International Neuroblastoma Pathology Classification (INPC) system, DNA ploidy, cytogenetics such as MYCN amplification status and chromosome aberrations of 1p and 11q (1, 4, 5). In addition, CT or MR image-defined risk factors (IDRFs) were used to distinguish low-risk tumors from high-risk tumors (6, 7). However, the predictive value of nuclear medicine functional imaging techniques on tumor biology has been less studied.

Nuclear medicine functional imaging plays an important role in the assessment of NB. Currently, ^{123}I -Metaiodobenzylguanidine (^{123}I -MIBG) scintigraphy is a standard practice in the diagnosis of NB (6), with ~90% of patients having MIBG avid tumors. However, in some countries, including China, ^{123}I -MIBG has not been approved for clinical use and cannot be included in the standard clinical protocols for NB patients. In our practice, we have utilized ^{18}F -fluorodeoxyglucose positron emission tomography/computer tomography (^{18}F -FDG PET/CT) in the diagnosis and follow-up of NB patients. ^{18}F -FDG PET imaging has been reported to be equal or superior to ^{123}I -MIBG scan for delineating NB disease extent in the chest, abdomen, and pelvis (8). In case the tumor is not MIBG avid, ^{18}F -FDG PET is also recommended as a complementary option to ^{123}I -MIBG scintigraphy (9).

The purpose of this study aims to evaluate whether diagnostic ^{18}F -FDG PET/CT imaging plays a role in risk stratification prediction in children with NB. The relationship between diagnostic ^{18}F -FDG PET/CT image features and the tumor biology of NB were investigated to answer this question. Specifically, cytogenetic factors, MYCN amplification status and chromosome aberrations of 1p and 11q, are chosen as representative indicators of tumor biology. It was well-documented that MYCN amplification and chromosome aberrations of 1p and 11q are powerful prognostic markers and have a strong association with worse outcome in NB (5). Amplification of MYCN can be detected in 20% of cases with NB and is closely linked with high-risk disease and poorer outcome (10). Loss of heterozygosity on chromosome 1p and 11q are correlated with increased disease severity (2, 11). For the PET/CT image analysis method, radiomic analysis was chosen in this study. In contrast to conventional visual image features, radiomics is expected to provide more comprehensive

description of tissues, with the potential to aid clinical care in several aspects including diagnosis, prognosis and treatment selection (12, 13). Currently, a number of studies demonstrated the value of ^{18}F -FDG PET/CT-based radiomics in predicting the histological subtypes of lung cancer (14) and distinguishing breast carcinoma from breast lymphoma (15). So far, there is little study to investigate the predictive value of ^{18}F -FDG PET/CT on the status of MYCN, 1p and 11q in pediatric NB. Therefore, this study was designed to evaluate whether ^{18}F -FDG PET/CT-based radiomics can predict the status of MYCN, 1p and 11q, which in turn, can be used in risk stratification prediction in children with NB.

METHODS

Patients

The records of 139 pediatric patients with newly diagnosed NB were reviewed retrospectively between March 2018 and November 2019 in our hospital. The inclusion criteria were as follows: (1) pathologically confirmed NB; (2) age ≤ 18 years at diagnosis; (3) complete PET/CT imaging data; (4) complete clinical information; (5) no cancer therapy before PET/CT imaging; (6) complete MYCN amplification and 1p and 11q aberrations data. Subsequently, 17 cases were excluded because of unavailable MYCN, 1p and 11q information, and 122 patients were included in this study. These patients were randomly divided into training set and test set with a ratio of 7:3. This retrospective study was approved by Institutional Review Board of our hospital and the requirement of written informed consent was waived.

Determination of MYCN Amplification and 1p and 11q Aberrations by FISH

MYCN amplification and 1p and 11q aberrations were determined using FISH from paraffin-embedded tissue obtained by biopsy or surgery at initial diagnosis according to the previously published method (16). According to the recommendations of the European Neuroblastoma Quality Assessment group (17, 18), MYCN amplification was defined as a > four-fold increase of signals.

Clinical Data and ^{18}F -FDG PET/CT Imaging Clinical Characteristics

Patient gender, age, neuron-specific enolase (NSE), serum ferritin (SF), lactate dehydrogenase (LDH), vanillylmandelic acid (VMA), homovanillic acid (HVA), maximum tumor diameter (MTD) in Ultrasound, and MTD in CT and/or MRI.

All patients underwent whole body scan on the PET/CT scanner (Biograph mCT-64 PET/CT; Siemens, Knoxville, Tenn) in accordance with EANM guidelines (19, 20) and a biopsy/surgery for pathological diagnosis of NB was performed within 3 months. The PET scan was carried out with 3 min per bed position immediately after the whole body CT scan. PET images were reconstructed using the ordered subsets-expectation maximization algorithm with time-of-flight. The regions-of-interest (ROIs) of primary tumor were manually drawn by an experienced nuclear medicine physician using the longitudinal

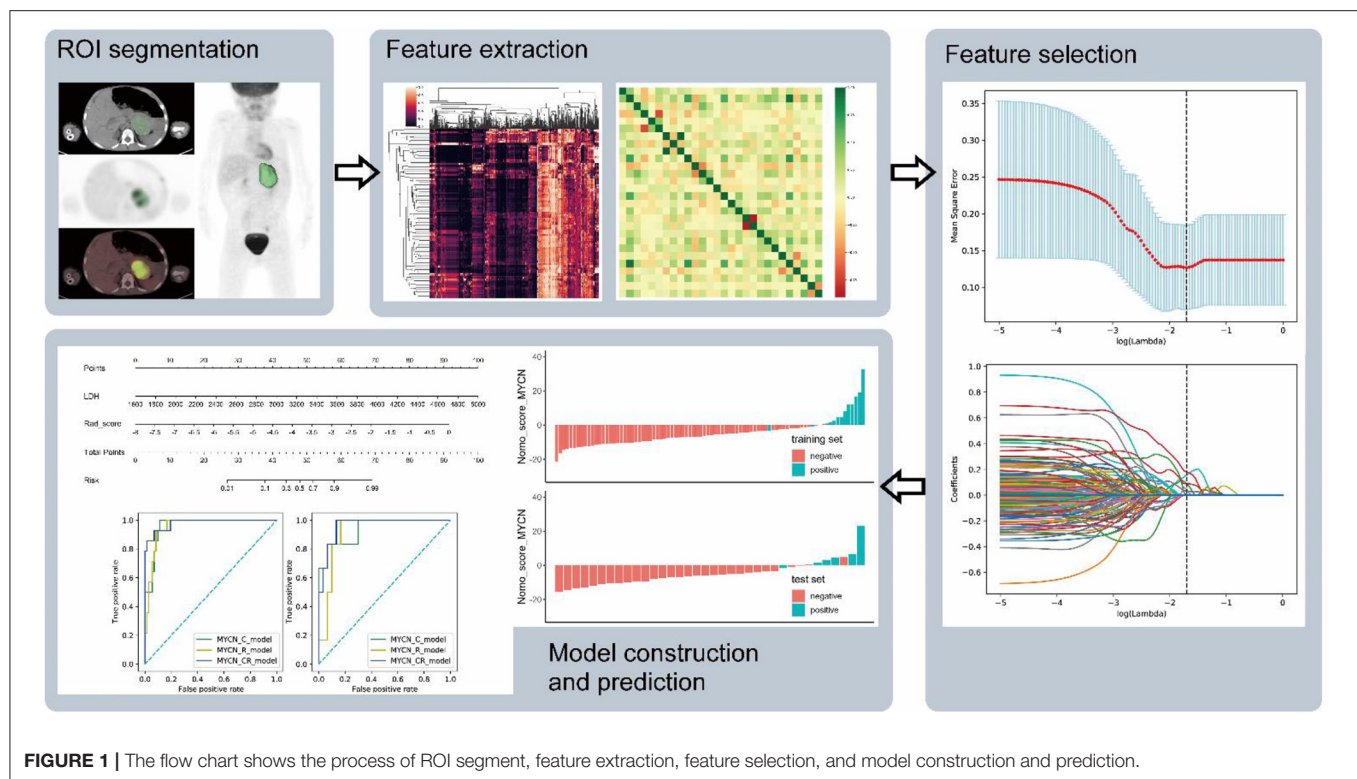


FIGURE 1 | The flow chart shows the process of ROI segment, feature extraction, feature selection, and model construction and prediction.

PET/CT module in 3D Slicer (version 4.10.1). ROIs were delineated along the edge of NB on CT images, which included the entire tumor, metastatic lesions and unclear demarcations between the primary tumor and its surrounding metastasis. In order to map to the PET image, the ROIs were resampled based on B-spline interpolation to ensure that it had the same pixel spacing as the PET image.

Feature Extraction and Selection and Model Construction

Univariate analysis was performed to compare the differences in clinical characteristics. Based on the selected characteristics, a clinical model (C-model) was established.

Radiomic features from CT and PET images were computed separately using pyradiomics, an open-source python package for the extraction of radiomic features from medical imaging (21). First order features ($n = 18$), shape features ($n = 14$), gray level co-occurrence matrix (GLCM) features ($n = 24$), gray level run length matrix (GLRLM) features ($n = 16$), gray level size zone matrix (GLSZM) features ($n = 16$), neighboring gray tone difference matrix (NGTDM) features ($n = 5$), and gray level dependence matrix (GLDM) features ($n = 14$) were extracted from the original and the pre-processed images. The following methods were used in the imaging processing: wavelet filtering, square, square root, logarithm, exponential and gradient filtering (Figure 1).

Intraclass correlation coefficients (ICC) were obtained to assess the reliability of variables using the features extracted from the two sets of ROIs portrayed separately by two different nuclear

medicine physicians in 24 out of the 122 patients with NB after 2 months. Because of imbalanced datasets, synthetic minority oversampling technique (SMOTE) was used to improve random oversampling in the training set. Least absolute shrinkage and selection operator (LASSO) was applied for variable selection and regularization in the training set. Predictive R_models were built by logistic regression and the radiomic score (rad-score) for each patient was computed based on the selected radiomic features. Additionally, the selected clinical characteristics combined with radiomics features were used to construct the combination model (CR_model). All models were built and trained in the training set, and the prediction performance was evaluated in the training and test sets. Ten-fold cross-validation was applied to prevent model overfitting in the training process. Receiver operating characteristic (ROC) curve and area under curve (AUC) were employed for the evaluation of the diagnostic performance in the training and test sets.

Statistical Analysis

Statistical analyses were performed with Python (ver. 3.7.8, www.python.org) and R (ver. 4.0.3, www.r-project.org). The Python packages of “sklearn,” “numpy,” and “pandas” were used for LASSO binary logistic regression and ROC curve; the “scipy” was for analyzing statistical properties; the “imblearn” was for SMOTE. The R package “rms” was employed to create nomograms. The *t*-test or Mann-Whitney *U*-test was applied for univariate analysis, and $p < 0.05$ with a 95% confidence interval was considered as statistical significance. AUC-ROC curve was calculated for evaluating the diagnostic performance of models.

TABLE 1 | Clinical features of NB patients.

Clinical features	Total	MYCN			1p			11q		
		Positive	Negative	p-value	Positive	Negative	p-value	Positive	Negative	p-value
Number	122	20	102		47	75		48	74	
Gender				0.224			0.062			0.345
Male	52	11	41		25	27		23	29	
Female	70	9	61		22	48		25	45	
Age (year)	3.2 (0.2–9.8)	2.5	3.4	0.1082	3.4	2.8	0.0885	4.0	2.3	0.0002
NSE (ng/ml)	219.1 (14.7–2627.1)	666.5	152.6	0.0046	370.0	129.1	0.0004	336.2	128.8	0.2977
SF (ng/ml)	210.2 (8.1–1807.0)	216.6	202.0	0.0744	220.1	189.5	0.0929	247.8	117.8	0.0019
LDH (U/L)	553 (177–6029)	2261	427	0.0001	936	386	<0.0001	596	411	0.0460
VMA	236.2 (5.2–5975.0)	28.6	364.8	<0.0001	164.2	396.9	0.0055	507.6	98.3	0.0080
HVA	54.7 (1.5–1532.0)	42.5	69.3	0.1169	51.1	61.8	0.0526	108.6	33.4	0.0141
MTD Ultra (cm)	9.1 (2.2–20.0)	11.3	9.0	0.0820	10.5	8.4	0.0161	9.6	8.7	0.0882
MTD CT/MRI (cm)	9.3 (2.1–17.4)	11.4	9.1	0.0382	11.1	9.0	0.0044	10.1	9.1	0.1196

Each feature was expressed as median (minimum–maximum) except for gender.

NSE, neuron-specific enolase; SF, serum ferritin; LDH, lactate dehydrogenase; VMA, Vanillylmandelic Acid; HVA, homovanillic acid; MTD Ultra, maximum tumor diameter (MTD) in ultrasound; MTD CT/MRI, MTD in CT/MRI.

AUC ranging from 0.5 to 1.0 is commonly used as a measure of classifier performance. A value of 0.5 is equal to random guessing, while 1.0 means a perfect classifier.

RESULTS

Clinical Characteristics of Patients

According to the inclusion criteria, 122 out of 139 patients with NB were enrolled in this study. Eighty six patients were assigned to the training set and 36 patients were assigned to the test set. All clinical characteristics are summarized in **Table 1**, including gender, age, neuron-specific enolase (NSE), serum ferritin (SF), lactate dehydrogenase (LDH), vanillylmandelic acid (VMA), homovanillic acid (HVA), maximum tumor diameter (MTD) in Ultrasound, and MTD in CT and/or MRI. The percentages of MYCN-, 1p- and 11q-positive cases were 16.4% (20/122), 38.5% (47/122), and 39.3% (48/122), respectively. Among these variables, NSE, LDH, VMA, and MTD in CT/MRI were significantly different between MYCN-positive and negative groups (All $p < 0.05$). Between 1p-positive and negative cases, NSE, LDH, VMA, MTD in Ultrasound and MTD in CT/MRI were distinct (All $p < 0.05$). Between 11q-positive and negative cases, age, SF, LDH, VMA, and HVA were distinct (All $p < 0.05$) (**Table 1**).

Predictive Model Construction

The total of 2,632 radiomic features were extracted from PET/CT images using pyradiomics. After assessing the robustness, 1,623 out of 2,632 features retained for model building, with intraclass correlation coefficients (ICC) > 0.75 . In respect of C-model (clinical variables) constructed by logistic regression and trained in the training set, 4 clinical characteristics (LDH, NSE, VMA, and SF) were selected for MYCN prediction, with 3 characteristics (LDH, NSE and age) for 1p prediction and

3 characteristics (LDH, SF and HVA) for 11q prediction. As for R_model (radiomics signature) establishment, 6 radiomic features were chosen for MYCN prediction, with 8 features for 1p prediction and 7 features for 11q prediction (**Table 2** and **Supplementary Table 1**).

In regard to CR_model (combinations of clinical and radiomic features) construction, eight features were chosen for MYCN prediction, which included 4 clinical characteristics (NSE, LDH, VMA, and MTD in CT/MRI) and 2 PET, 2 CT features (**Tables 1, 3**). Eleven features were selected for 1p prediction, which included 5 clinical characteristics (NSE, LDH, VMA, MTD in Ultrasound and MTD in CT/MRI) and 5 PET, 1 CT features (**Tables 1, 3**). Eleven features were picked up for 11q prediction, which included 5 clinical characteristics (age, SF, LDH, VMA, and HVA) and 1 PET, 5 CT features (**Tables 1, 3**).

Rad-scores were calculated by the following formula:

$$\begin{aligned}
 \text{Rad_score_MYCN} &= -2.6446 \\
 &+ 0.17750 \times \text{PET_wavelet-LLH_glszm_GrayLevelNonUniformity} \\
 &+ 0.88251 \times \text{PET_wavelet-HHH_glszm_SizeZoneNonUniformity} \\
 &- 0.00069 \times \text{CT_exponential_glrlm_LongRunEmphasis} \\
 &- 0.02217 \times \text{CT_wavelet-HHL_firstorder_Maximum} \\
 \text{Rad_score_1p} &= 2.9612 \\
 &- 115.24 \times \text{PET_squareroot_ngtdm_Contrast} \\
 &- 0.29673 \times \text{PET_logarithm_firstorder_Minimum} \\
 &+ 0.04218 \times \text{PET_wavelet-LLH_glrlm_LongRunLowGrayLevelEmphasis} \\
 &+ 2.1217 \times \text{PET_wavelet-HHH_glszm_SmallAreaHighGrayLevelEmphasis} \\
 &- 5.5262 \times \text{PET_wavelet-HHH_glszm_LowGrayLevelZoneEmphasis} \\
 &- 5.1213 \times \text{CT_exponential_glszm_SmallAreaEmphasis}
 \end{aligned}$$

TABLE 2 | Comparison of the radiomic features between positive and negative in training sets of R_model.

Radiomic feature	<i>p</i> -value
MYCN	
PET_squareroot_gldm_HighGrayLevelEmphasis	0.0234
PET_wavelet-LHL_gldm_DependenceNonUniformity	0.0233
PET_wavelet-HHH_glszm_SizeZoneNonUniformity	0.0361
CT_logarithm_firstorder_Skewness	0.0001
CT_wavelet-LLL_gldm_DependenceVariance	0.0009
CT_wavelet-HLL_glszm_LargeAreaHighGrayLevelEmphasis	0.0156
1p	
PET_squareroot_gldm_Idmn	0.0009
PET_logarithm_firstorder_Minimum	0.0940
PET_wavelet-LLL_gldm_InverseVariance	0.0061
PET_wavelet-HHL_gldm_DependenceVariance	0.0436
PET_wavelet-HHH_glszm_SmallAreaHighGrayLevelEmphasis	<0.0001
PET_wavelet-HHH_glszm_LowGrayLevelZoneEmphasis	0.0002
CT_exponential_glszm_SmallAreaEmphasis	0.0554
CT_wavelet-HHH_glszm_SizeZoneNonUniformityNormalized	0.0885
11q	
PET_original_glszm_GrayLevelNonUniformity	0.0108
PET_wavelet-LHL_gldm_DependenceNonUniformityNormalized	0.0271
CT_original_shape_Flatness	0.0043
CT_wavelet-LLL_gldm_RunVariance	0.0006
CT_wavelet-LHL_firstorder_Median	0.0613
CT_wavelet-LHL_gldm_Imc1	0.0166
CT_wavelet-HHH_firstorder_Entropy	0.0291

$$\text{Rad_score_11q} = -2217.3$$

$$\begin{aligned} &- 147.63 \times \text{PET_wavelet-LHL_gldm_DependenceNonUniformityNormalized} \\ &- 0.41560 \times \text{CT_wavelet-LLL_gldm_RunVariance} \\ &- 0.59915 \times \text{CT_wavelet-LHL_firstorder_Median} \\ &+ 58.736 \times \text{CT_wavelet-LHL_gldm_Imc1} \\ &- 14.536 \times \text{CT_wavelet-HLL_gldm_LowGrayLevelRunEmphasis} \\ &+ 2232.9 \times \text{CT_wavelet-HHH_firstorder_Entropy.} \end{aligned}$$

The *p*-values of radiomic features are shown in **Table 3**. Rad-scores presented significant difference between positive and negative groups in the training and test sets (*p* < 0.001). NB with MYCN, 1p and 11q positive had higher Rad-score than those with negative in both the training and test sets.

Nomogram score (Nomo_score) was calculated by the following formula (**Figure 2**):

$$\begin{aligned} \text{Nomo_score_MYCN} &= -0.7569 + 0.0064 \times \text{LDH} + 2.4857 \times \text{Rad_score_MYCN} \\ \text{Nomo_score_1p} &= -0.5175 + 0.0017 \times \text{LDH} + 1.0476 \times \text{Rad_score_1p} \\ \text{Nomo_score_11q} &= -0.3897 - 0.0020 \times \text{LDH} + 0.0088 \times \text{SF} + 1.6657 \times \text{Rad_score_11q} \end{aligned}$$

TABLE 3 | Comparison of the radiomic features between positive and negative in training sets of CR_model.

Radiomic feature	<i>p</i> -value
MYCN	
PET_wavelet-LLH_glszm_GrayLevelNonUniformity	0.0125
PET_wavelet-HHH_glszm_SizeZoneNonUniformity	0.0361
CT_exponential_gldm_LongRunEmphasis	0.0224
CT_wavelet-HHL_firstorder_Maximum	0.0832
1p	
PET_squareroot_ngldm_Contrast	0.0286
PET_logarithm_firstorder_Minimum	0.0940
PET_wavelet-LLH_gldm_LongRunLowGrayLevelEmphasis	0.0105
PET_wavelet-HHH_glszm_SmallAreaHighGrayLevelEmphasis	<0.0001
PET_wavelet-HHH_glszm_LowGrayLevelZoneEmphasis	0.0002
CT_exponential_glszm_SmallAreaEmphasis	0.0554
11q	
PET_wavelet-LHL_gldm_DependenceNonUniformityNormalized	0.0271
CT_wavelet-LLL_gldm_RunVariance	0.0006
CT_wavelet-LHL_firstorder_Median	0.0613
CT_wavelet-LHL_gldm_Imc1	0.0166
CT_wavelet-HLL_gldm_LowGrayLevelRunEmphasis	0.0037
CT_wavelet-HHH_firstorder_Entropy	0.0291

The nomogram was created based on the training set, which represented individualized prediction and visualized proportion of each factor (**Figure 3**).

Model Performance

To evaluate the performance in predicting MYCN, 1p and 11q status, C_model, R_model and CR_model were compared. The predictive abilities of models (sensitivity, specificity, and AUC) were shown in **Table 4**, and ROC curves were displayed in **Figure 4**. Obviously, the CR_models were the best predictive models for MYCN, 1p and 11q abnormalities, with AUCs of 0.98 (sensitivity, 0.93; specificity, 0.93), 0.91 (sensitivity, 0.85; specificity, 0.83), and 0.93 (sensitivity, 0.82; specificity, 0.90) in the training set, respectively. In the test set, their AUCs were 0.96 (sensitivity, 0.83; specificity, 0.87), 0.88 (sensitivity, 0.79; specificity, 0.77), and 0.89 (sensitivity, 0.86; specificity, 0.72), sequentially. The CR_model for MYCN prediction had the greatest performance in the training and test sets compared to the CR_models for 1p and 11q prediction. In addition, the R_models for predicting 1p and 11q performed better than the C_models in the test set (AUCs = 0.85 vs. 0.77 for 1p; AUCs = 0.84 vs. 0.74 for 11q). In contrast, the C_model for MYCN prediction was better than the R_model in the test set (AUCs = 0.94 vs. 0.92).

DISCUSSION

Considering the well-established role of MYCN, 1p and 11q abnormalities in the prognosis of NB, identifying these events are crucial for risk stratification. This study provided three distinct forms of predictive models (clinical variables,

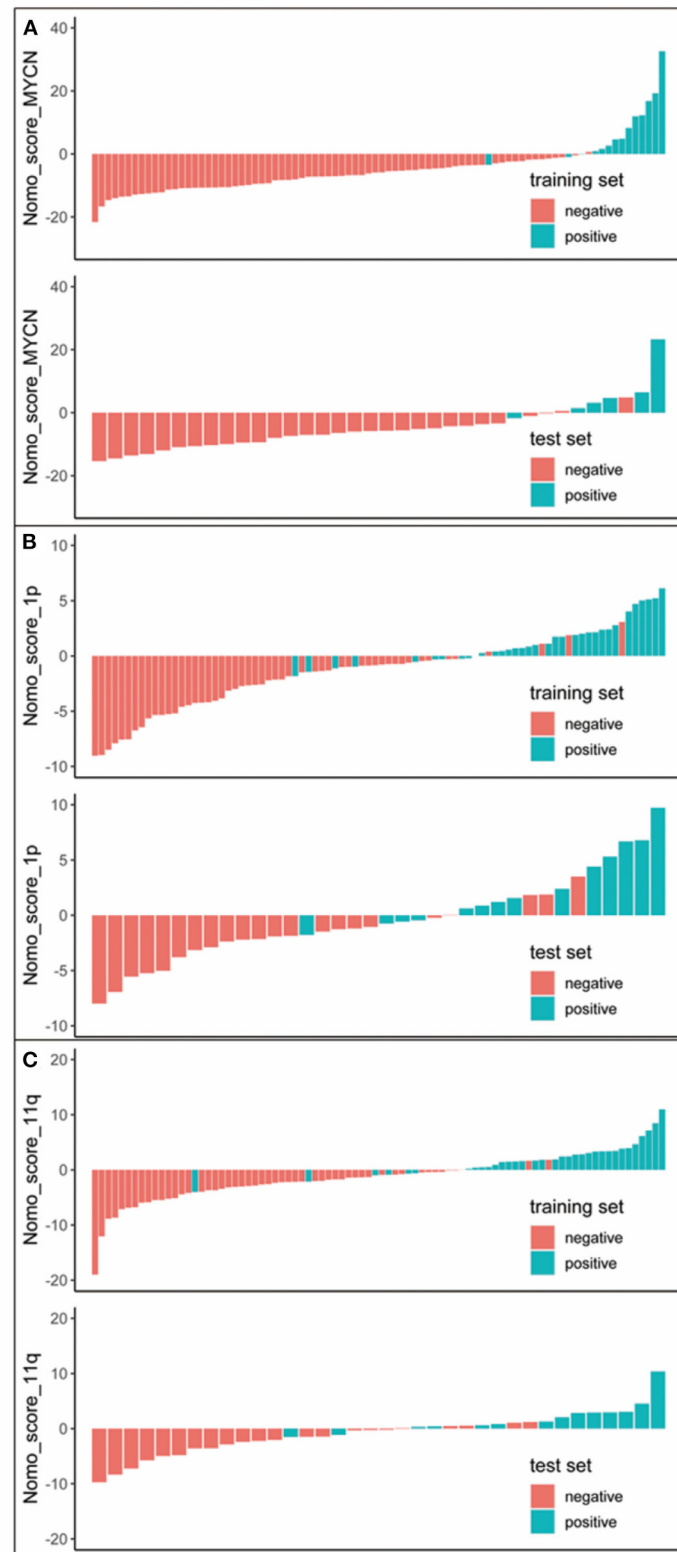


FIGURE 2 | Nomo_score for every patient in each set. The red marks indicate negative samples, while the blue marks indicate the positive samples. **(A)** Nomo_score of MYCN status prediction. **(B)** Nomo_score of 1p status prediction. **(C)** Nomo_score of 11q status prediction.

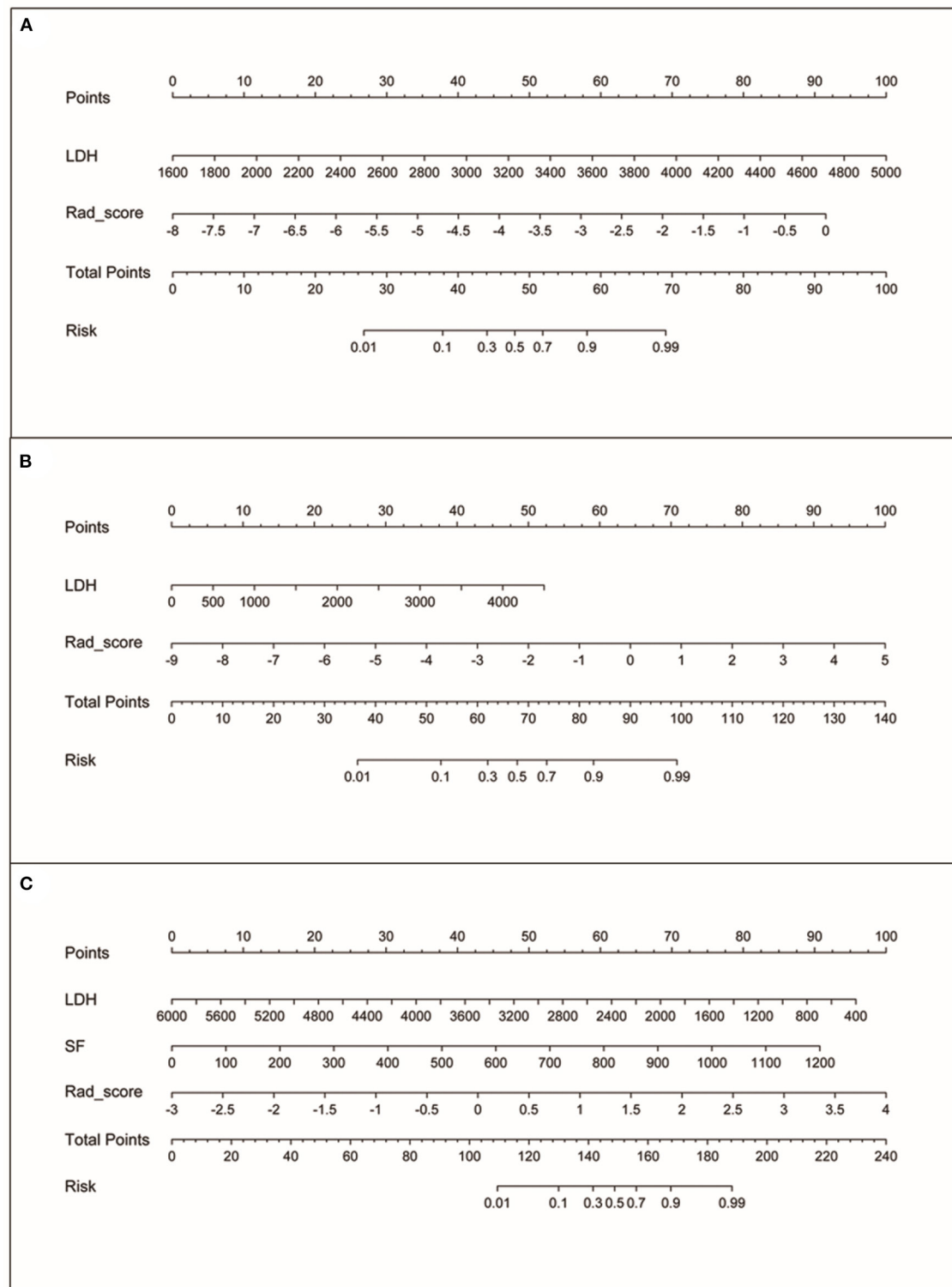


FIGURE 3 | The nomograms. **(A)** Nomogram based on rad-score and LDH for MYCN status prediction. **(B)** Nomogram based on rad-score and LDH for 1p status prediction. **(C)** Nomogram based on rad-score, LDH and SF for 11q status prediction.

TABLE 4 | The predictive value of the models in MYCN, 1p and 11q.

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	AUC (95%CI)	Sensitivity	Specificity	Accuracy	AUC (95%CI)
MYCN								
C_model	1.00	0.88	0.90	0.96 (0.93–0.99)	0.83	0.93	0.92	0.94 (0.85–1.00)
R_model	0.86	0.92	0.91	0.96 (0.93–0.99)	0.83	0.90	0.89	0.92 (0.82–1.00)
CR_model	0.93	0.93	0.93	0.98 (0.96–0.99)	0.83	0.87	0.86	0.96 (0.90–1.00)
1p								
C_model	0.64	0.71	0.68	0.79 (0.73–0.85)	0.79	0.59	0.67	0.77 (0.62–0.91)
R_model	0.73	0.75	0.74	0.89 (0.85–0.93)	0.93	0.64	0.75	0.85 (0.73–0.97)
CR_model	0.85	0.83	0.84	0.91 (0.87–0.95)	0.79	0.77	0.78	0.88 (0.78–0.98)
11q								
C_model	0.71	0.73	0.72	0.77 (0.71–0.83)	0.64	0.64	0.64	0.74 (0.60–0.88)
R_model	0.76	0.83	0.80	0.89 (0.85–0.93)	0.79	0.68	0.72	0.84 (0.73–0.95)
CR_model	0.82	0.90	0.87	0.93 (0.90–0.96)	0.86	0.72	0.77	0.89 (0.79–0.99)

radiomic signature and their combinations) for identifying MYCN and chromosomal abnormalities in a non-invasive way, demonstrating that pre-therapy ^{18}F -FDG PET/CT-based radiomics had an extremely important role in predicting MYCN amplification and 1p and 11q aberrations. In particular, CR_model was suggested to be the best model for the prediction of MYCN, 1p and 11q status with the largest AUCs in the training and test sets.

Recently, clinical variables (such as LDH and SF) have been demonstrated to be prognostic biomarkers in large-scale studies, which suggested to reconsider utilizing LDH and SF as NB risk stratification factors (22, 23). In the present study, LDH and SF were also predictors of MYCN, 1p and 11q abnormalities. The radiomics models had a power to predict these aberrations, but models integrating PET and CT features with clinical variables led to higher predictive performance for training and test cohorts, in comparison with models with radiomic features or clinical parameters alone (Table 2). In line with other studies (24), the integration of radiomic features with clinical parameters has a complementary and added impact in abnormal genetic and/or molecular prediction.

In this study, radiomic features were selected to construct CR_model for predicting MYCN, 1p and 11q abnormalities, including: PET_wavelet-LLH_glszm_GrayLevelNonUniformity, PET_wavelet-HHH_glszm_SizeZoneNonUniformity, CT_exponential_glrlm_LongRunEmphasis, CT_wavelet-HHL_firstorder_Maximum, PET_squareroot_ngtdm_Contrast, PET_logarithm_firstorder_Minimum, PET_wavelet-LLH_glrlm_LongRunLowGrayLevelEmphasis, PET_wavelet-HHH_glszm_SmallAreaHighGrayLevelEmphasis, PET_wavelet-HHH_glszm_LowGrayLevelZoneEmphasis, CT_exponential_glszm_SmallAreaEmphasis, PET_wavelet-LHL_gldm_DependenceNonUniformityNormalized, CT_wavelet-LLL_glrlm_RunVariance, CT_wavelet-LHL_firstorder_Median, CT_wavelet-LHL_gldm_Imc1, CT_wavelet-HLL_glrlm_LowGrayLevelRunEmphasis, and CT_wavelet-HHH_firstorder_Entropy. The majority of these

features (12/16) were not derived from the primary image but from wavelet decomposition images, possibly because wavelet transformed features contained high-order information that may be more helpful for MYCN, 1p and 11q prediction. Previous studies have revealed the potential value of wavelet features in histologic subtype prediction and prognostic assessment (25, 26). In agreement with that, our data also indicated that wavelet features possess remarkable abilities in MYCN, 1p and 11q prediction models. In addition, approximately half of the selected features were extracted from GLRLM (4/16) and GLSZM (5/16). Long run emphasis (LRE) in GLRLM quantifies the distribution of long run lengths, with a larger value representing longer run lengths and more coarse structural textures. Size-zone non-uniformity (SZN) in GLSZM quantifies the variability of size zone volumes in the image, with a smaller value representing more homogeneity in size zone volumes. Our results showed that the greater value of LRE or SZN was correlated with the higher possibility of MYCN amplification and 1p and 11q aberrations.

Currently, ^{123}I -MIBG scan is the most frequently used imaging modality and is regarded as standard of care in patients with NB. In comparison with ^{18}F -FDG PET/CT, ^{123}I -MIBG scan is carried out over 2 days and the image quality is less ideal that could post a challenge to inexperienced physicians (27). At many centers, planar I-MIBG imaging scans are performed, but radiomics based on these images was very limited. Moreover, false-negative MIBG scans were reported as early as 1990, which may result in incorrect down-staging (9). In about 8% of NB patients, false-negative scans at diagnosis occurred despite the solid evidence of disease. ^{18}F -FDG PET/CT describes the metabolic state of cancer cells and provides information about malignancy (28). The value of ^{18}F -FDG PET/CT in NB has been investigated in many studies. For example, Shulkin et al. demonstrated that ^{18}F -FDG uptake was increased in the most of lesions, with about 94% of NB showing elevated ^{18}F -FDG activity (28). Melzer et al. reported that ^{123}I -MIBG SPECT/CT and ^{18}F -FDG PET/CT had significant differences in their uptake

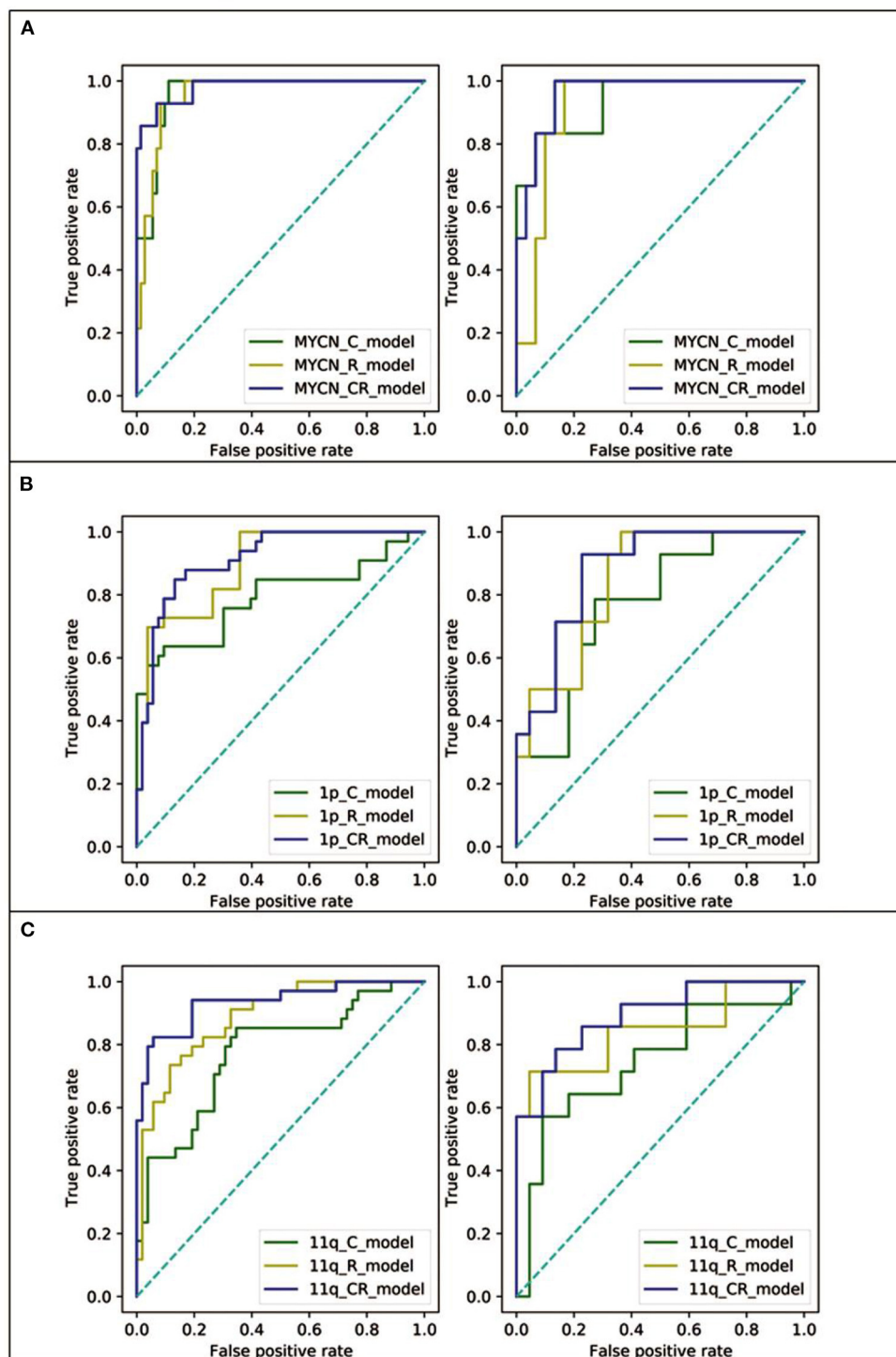


FIGURE 4 | The ROC curves of the C_model (green line), R_model (yellow line), and CR_model (blue line) in both training (left) and test (right) set. **(A)** The ROC curves of MYCN status prediction. **(B)** The ROC curves of 1p status prediction. **(C)** The ROC curves of 11q status prediction.

patterns. In NB patients, ^{18}F -FDG PET/CT had higher sensitivity and specificity for the detection of lesions (9), and showed more extensive primary and/or residual lesions in stage 1 and

2 (8). Overall, ^{18}F -FDG PET/CT was superior in depicting NB, although ^{123}I -MIBG might be needed to exclude higher-stage (8). Interestingly, the FDG-avid but MIBG-negative and

MIBG-avid but FDG-negative NB can coexist in the same tumor (28).

The potential clinical significance of the present study included: (1) radiomics based on pre-therapy ^{18}F -FDG PET/CT provides a relatively accurate method in a non-invasive way for predicting MYCN, 1p and 11q, which can be applicable to pediatric NB patients; (2) the status of MYCN, 1p and 11q can be used for risk stratification, therapy selection, therapy response monitor and prognosis prediction.

This study had limitations. Small size cohort from single center may influence the generalized ability, sensitivity and specificity of the predictive models. Therefore, prospective larger cohort from multi-center is necessary to validate the results and improve the reliability of models for MYCN, 1p and 11q predictions in NB.

CONCLUSION

The models developed by the pre-therapy ^{18}F -FDG PET/CT radiomic signature and clinical parameters are able to predict MYCN amplification and 1p and 11 aberrations in pediatric NB, thus risk stratification, disease management and guiding personalized malignancy therapy in the clinical practice.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Beijing Friendship Hospital, Capital Medical

University. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LQ, SY, and SZ made substantial contributions to study design, image acquisition, data analysis and interpretation, and new software creation in this work. SZ, HQ, WW, YK, LL, JL, and HZ contributed writing and/or revising the manuscript. JY and JL approved all versions to be published and were responsible for all aspects of this study. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Capital Health Development Research Project (No. 2020-2-2025), National Natural Science Foundation of China (Nos. 81971642, 82001861, and 82102088), and National Key Research and Development Plan (No. 2020YFC0122000).

ACKNOWLEDGMENTS

We would like to thank Dr Dehui Sun for helping us in imaging analysis of this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2022.840777/full#supplementary-material>

REFERENCES

- Dzian J, Rodriguez Garcia A, Westermarck UK, Henley AB, Eyre Sánchez E, Träger C, et al. MYCN-amplified neuroblastoma maintains an aggressive and undifferentiated phenotype by deregulation of estrogen and NGF signaling. *Proc Natl Acad Sci USA*. (2018) 115:E1229–38. doi: 10.1073/pnas.1710901115
- Park JR, Bagatell R, London WB, Maris JM, Cohn SL, Mattay KK, et al. Children's Oncology Group's 2013 blueprint for research: neuroblastoma. *Pediatr Blood Cancer*. (2013) 60:985–93. doi: 10.1002/pbc.24433
- Matthay KK, George RE, Yu AL. Promising therapeutic targets in neuroblastoma. *Clin Cancer Res*. (2012) 18:2740–53. doi: 10.1158/1078-0432.CCR-11-1939
- Huang M, Weiss WA. Neuroblastoma and MYCN. *Cold Spring Harb Perspect Med*. (2013) 3:a014415. doi: 10.1101/cshperspect.a014415
- Irwin MS, Park JR. Neuroblastoma: paradigm for precision medicine. *Pediatr Clin North Am*. (2015) 62:225–56. doi: 10.1016/j.pcl.2014.09.015
- Bar-Sever Z, Biassoni L, Shulkin B, Kong G, Hofman MS, Lopci E, et al. Guidelines on nuclear medicine imaging in neuroblastoma. *Eur J Nucl Med Mol Imaging*. (2018) 45:2009–24. doi: 10.1007/s00259-018-4070-8
- Phelps HM, Ndolo JM, Van Arendonk KJ, Chen H, Dietrich HL, Watson KD, et al. Association between image-defined risk factors and neuroblastoma outcomes. *J Pediatr Surg*. (2019) 54:1184–91. doi: 10.1016/j.jpedsurg.2019.02.040
- Sharp SE, Shulkin BL, Gelfand MJ, Salisbury S, Furman WL. 123I-MIBG scintigraphy and ^{18}F -FDG PET in neuroblastoma. *J Nucl Med*. (2009) 50:1237–43. doi: 10.2967/jnumed.108.060467
- Melzer HI, Coppenrath E, Schmid I, Albert MH, von Schweinitz D, Tudball C, et al.¹²³I-MIBG scintigraphy/SPECT versus ^{18}F -FDG PET in paediatric neuroblastoma. *Eur J Nucl Med Mol Imaging*. (2011) 38:1648–58. doi: 10.1007/s00259-011-1843-8
- Maris JM, Hogarty MD, Bagatell R, Cohn SL. Neuroblastoma. *Lancet*. (2007) 369:2106–20. doi: 10.1016/S0140-6736(07)60983-0
- Bosse KR, Maris JM. Advances in the translational genomics of neuroblastoma: from improving risk stratification and revealing novel biology to identifying actionable genomic alterations. *Cancer*. (2016) 122:20–33. doi: 10.1002/cncr.29706
- Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. (2016) 278:563–77. doi: 10.1148/radiol.2015151169
- Ligero M, Garcia-Ruiz A, Viaplana C, Villacampa G, Raciti MV, Landa J, et al. A CT-based radiomics signature is associated with response to immune checkpoint inhibitors in advanced solid tumors. *Radiology*. (2021) 299:109–19. doi: 10.1148/radiol.2021200928
- Hyun SH, Ahn MS, Koh YW, Lee SJ. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clin Nucl Med*. (2019) 44:956–60. doi: 10.1097/RLU.00000000000002810

15. Ou X, Zhang J, Wang J, Pang F, Wang Y, Wei X, et al. Radiomics based on (18)F-FDG PET/CT could differentiate breast carcinoma from breast lymphoma using machine-learning approach: a preliminary study. *Cancer Med.* (2020) 9:496–506. doi: 10.1002/cam4.2711
16. Yue ZX, Huang C, Gao C, Xing TY, Liu SG, Li XJ, et al. MYCN amplification predicts poor prognosis based on interphase fluorescence *in situ* hybridization analysis of bone marrow cells in bone marrow metastases of neuroblastoma. *Cancer Cell Int.* (2017) 17:43. doi: 10.1186/s12935-017-0412-z
17. Theissen J, Boensch M, Spitz R, Betts D, Stegmaier S, Christiansen H, et al. Heterogeneity of the MYCN oncogene in neuroblastoma. *Clin Cancer Res.* (2009) 15:2085–90. doi: 10.1158/1078-0432.CCR-08-1648
18. Villamon E, Berbegall AP, Piqueras M, Tadeo I, Castel V, Djos A, et al. Genetic instability and intratumoral heterogeneity in neuroblastoma with MYCN amplification plus 11q deletion. *PLoS ONE.* (2013) 8:e53740. doi: 10.1371/journal.pone.0053740
19. Stauss J, Franzius C, Pfluger T, Juergens KU, Biassoni L, Begent J, et al. Guidelines for 18F-FDG PET and PET-CT imaging in paediatric oncology. *Eur J Nucl Med Mol Imaging.* (2008) 35:1581–8. doi: 10.1007/s00259-008-0826-x
20. Delbeke D, Coleman RE, Guiberteau MJ, Brown ML, Royal HD, Siegel BA, et al. Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0. *J Nucl Med.* (2006) 47:885–95.
21. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472.Can-17-0339
22. Morgenstern DA, London WB, Stephens D, Volchenboum SL, Hero B, Di Cataldo A, et al. Metastatic neuroblastoma confined to distant lymph nodes (stage 4N) predicts outcome in patients with stage 4 disease: a study from the International Neuroblastoma Risk Group Database. *J Clin Oncol.* (2014) 32:1228–35. doi: 10.1200/jco.2013.53.6342
23. Moroz V, Machin D, Hero B, Ladenstein R, Berthold F, Kao P, et al. The prognostic strength of serum LDH and serum ferritin in children with neuroblastoma: a report from the International Neuroblastoma Risk Group (INRG) project. *Pediatr Blood Cancer.* (2020) 67:e28359. doi: 10.1002/pbc.28359
24. Zhang J, Zhao X, Zhao Y, Zhang J, Zhang Z, Wang J, et al. Value of pre-therapy (18)F-FDG PET/CT radiomics in predicting EGFR mutation status in patients with non-small cell lung cancer. *Eur J Nucl Med Mol Imaging.* (2020) 47:1137–46. doi: 10.1007/s00259-019-04592-1
25. Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, et al. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother Oncol.* (2016) 120:258–66. doi: 10.1016/j.radonc.2016.05.024
26. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol.* (2016) 6:71. doi: 10.3389/fonc.2016.00071
27. Wen Z, Zhang L, Zhuang H. Roles of PET/computed tomography in the evaluation of neuroblastoma. *PET Clin.* (2020) 15:321–31. doi: 10.1016/j.cpet.2020.03.003
28. Shulkin BL, Hutchinson RJ, Castle VP, Yanik GA, Shapiro B, Sisson JC. Neuroblastoma: positron emission tomography with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose compared with metaiodobenzylguanidine scintigraphy. *Radiology.* (1996) 199:743–50. doi: 10.1148/radiology.199.3.8637999

Conflict of Interest: LL was employed by the company Sinounion Medical Technology (Beijing) Co., Ltd.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Qian, Yang, Zhang, Qin, Wang, Kan, Liu, Li, Zhang and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Application of 18F-FDG PET-CT Images Based Radiomics in Identifying Vertebral Multiple Myeloma and Bone Metastases

Zhicheng Jin^{1†}, Yongqing Wang^{2†}, Yizhen Wang^{1†}, Yangting Mao¹, Fang Zhang^{1*} and Jing Yu^{1*}

¹ Department of Nuclear Medicine, Second Affiliated Hospital, Dalian Medical University, Dalian, China, ² School of Geophysics and Information Technology, China University of Geosciences, Beijing, China

OPEN ACCESS

Edited by:

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC),
Switzerland

Reviewed by:

Sharjeel Usmani,
Kuwait Cancer Control Center, Kuwait
Salvatore Annunziata,
Fondazione Policlinico Universitario A.
Gemelli (IRCCS), Italy

*Correspondence:

Fang Zhang
m18810016346@163.com
Jing Yu
yujing_2020@dmu.edu.cn

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Frontiers in Medicine,
a section of the journal
Frontiers in Medicine

Received: 13 February 2022

Accepted: 17 March 2022

Published: 18 April 2022

Citation:

Jin Z, Wang Y, Wang Y, Mao Y,
Zhang F and Yu J (2022) Application
of 18F-FDG PET-CT Images Based
Radiomics in Identifying Vertebral
Multiple Myeloma and Bone
Metastases. *Front. Med.* 9:874847.
doi: 10.3389/fmed.2022.874847

Purpose: The purpose of this study was to explore the application of 18F-fluorodeoxyglucose positron emission tomography/computed tomography (18F-FDG PET/CT) image radiomics in the identification of spine multiple myeloma (MM) and bone metastasis (BM), and whether this method could improve the classification diagnosis performance compared with traditional methods.

Methods: This retrospective study collected a total of 184 lesions from 131 patients between January 2017 and January 2021. All images were visually evaluated independently by two physicians with 20 years of experience through the double-blind method, while the maximum standardized uptake value (SUVmax) of each lesion was recorded. A total of 279 radiomics features were extracted from the region of interest (ROI) of CT and PET images of each lesion separately by manual method. After the reliability test, the least absolute shrinkage and selection operator (LASSO) regression and 10-fold cross-validation were used to perform dimensionality reduction and screening of features. Two classification models of CT and PET were derived from CT images and PET images, respectively and constructed using the multivariate logistic regression algorithm. In addition, the ComModel was constructed by combining the PET model and the conventional parameter SUVmax. The performance of the three classification diagnostic models, as well as the human experts and SUVmax, were evaluated and compared, respectively.

Results: A total of 8 and 10 features were selected from CT and PET images for the construction of radiomics models, respectively. Satisfactory performance of the three radiomics models was achieved in both the training and the validation groups (Training: AUC: CT: 0.909, PET: 0.949, ComModel: 0.973; Validation: AUC: CT: 0.897, PET: 0.929, ComModel: 0.948). Moreover, the PET model and ComModel showed significant improvement in diagnostic performance between the two groups compared to the human expert (Training: $P = 0.01$ and $P = 0.001$; Validation: $P = 0.018$ and $P = 0.033$), and no statistical difference was observed between the CT model and human experts ($P = 0.187$ and $P = 0.229$, respectively).

Conclusion: The radiomics model constructed based on 18F-FDG PET/CT images achieved satisfactory diagnostic performance for the classification of MM and bone metastases. In addition, the radiomics model showed significant improvement in diagnostic performance compared to human experts and PET conventional parameter SUVmax.

Keywords: radiomics, multiple myeloma, bone metastases, 18F-FDG PET-CT, SUVmax

1. INTRODUCTION

Multiple myeloma (MM) was a malignant clonal cell tumor that originated from bone marrow plasma cells. MM extensively invades bone marrow, bones, and extramedullary organs, leading to prime syndromes such as bone pain, anemia, infection, fractures, and kidney damage (1). Bone metastasis (BM) was a common event in tumor progression. The common primary tumors were lung cancer, breast cancer, and prostate cancer (2). The spine contained a rich blood supply and was also the most frequent site to be involved. MM and BM had different pathogenesis, but the site of occurrence, clinical manifestations, and imaging features were similar, which makes it difficult to distinguish. Lesions that were difficult to characterize were often misdiagnosed as other orthopedic diseases, especially for MM and bone metastases with unknown primary lesions. Misclassifications will significantly affect the quality of patient survival and survival rates due to the variability of treatment options (3, 4). Therefore, it was particularly important to improve the diagnostic accuracy of MM and BM.

Previous studies had considered serologic markers such as serum creatinine, serum globulin, and serum alkaline phosphatase as crucial information for differentiating MM from BM. However, some patients with light chain secretory, low, and non-secretory myeloma may have low or normal levels of these serologic markers, and such examinations were often used for preliminary screening (5, 6). 18F-fluorodeoxyglucose positron emission tomography/computed tomography (18F-FDG PET/CT) images combined anatomical and metabolic information to provide relatively high sensitivity and specificity to assess bone damage and detect extramedullary lesions (7, 8). In patients with early MM, up to 40% of patients could detect additional lesions by PET/CT examination to guide individualized treatment plans (9). The International Myeloma Working Group had reached a consensus and recommended 18F-FDG PET/CT as one of the best imaging methods for the examination of MM and other plasma cell diseases (10). However, there still exist lesions that were difficult to identify even for experienced physicians in clinical work, especially osteolytic lesions (11, 12).

Radiomics converted texture, intensity, density, and other features extracted from medical images into mineable high-dimensional data through automated or semi-automated methods, which could be used as a non-invasive assessment of spatial heterogeneity of tumors and facilitate personalized patient treatment (13, 14). The performance of radiomics analysis had been demonstrated in previous studies to identify cancer types,

predict treatment efficacy, and predict disease progression (15–17). In addition, radiomics had shown unique advantages in molecular areas such as prediction of cancer gene expression and lymph node metastasis (18, 19). However, previous radiomics mostly focused on CT and MRI, and the diagnostic value of radiomics combined with 18F-FDG PET/CT for MM and BM was still unclear (20, 21).

The purpose of this study was to explore the feasibility of radiomics based on 18F-FDG PET/CT images in the identification of MM and BM and whether it could improve the diagnostic performance of these two diseases.

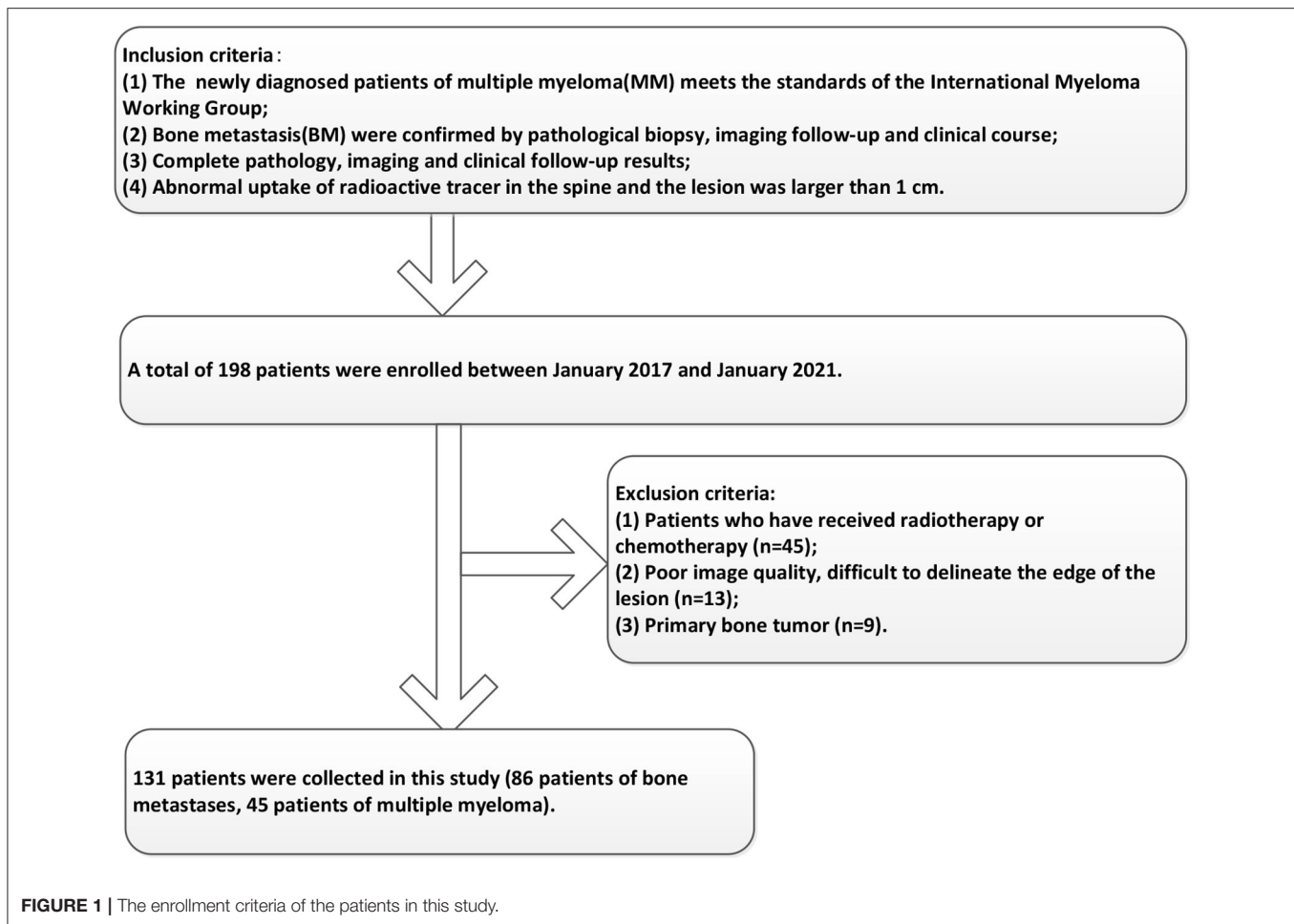
2. MATERIALS AND METHODS

2.1. Patients

Participants between January 2017 and January 2021 were enrolled in this study according to the following inclusion criteria: (1) The diagnosis of MM meets the standards of the International Myeloma Working Group (22); (2) BM were confirmed by pathological biopsy, imaging follow-up, and clinical course; (3) Complete pathology, imaging, and clinical follow-up results; (4) Abnormal uptake of radioactive tracer in the spine and the lesion was larger than 1 cm. In addition, the exclusion criteria included the following: (1) Patients who have received radiotherapy or chemotherapy; (2) Poor image quality, difficulty to delineate the edge of the lesion; (3) Primary bone tumor. The enrolled patients were randomly divided into training groups and validation groups according to the ratio of 7:3 for subsequent model construction. This retrospective study was approved by the hospital's ethical review, and the patient's informed consent requirement was waived. The enrollment criteria of the patients in this study were shown in Figure 1.

2.2. Image Protocol

All the image acquisition procedures in this study were completed in the PET/CT (Philips Ingenuity TF). The radioactive tracer 18F-FDG was automatically synthesized by the cyclotron (Sumitomo, Japan) and the 18F-FDG chemical synthesis module (Sumitomo, Japan), and the radioactive tracer purity was guaranteed to be >95%. The patient fasted for at least 6 h before the examination, and the glucose level lower than 11.1 mmol/L was ensured by routine measurement of the blood samples before the PET/CT examination. Patients were injected with 18F-FDG (5.55 MBq/kg) intravenously in a quiet state and were placed in a room with dim light for 40–60 min,



and then underwent PET/CT after emptying the bladder. The scanning process includes a low-dose CT scan and PET scan from the top of the skull to the upper thigh. CT acquisition parameters were as follows: tube current tube voltage was automatically generated according to the positioning image, tube rotation time: 0.35 s, output voltage: 70–140 KV, output current: 20–450 mA, layer thickness: 0.7 mm, reconstruction time: 40 frames/s, reconstruction matrix: 512×512, number of detector rows: 64, pitch: 0.15–1.5. After standardizing all parameters of the patient's PET/CT images, the window width and window level of the CT images were set to 350 and 50, respectively, and the PET data were reconstructed by attenuation correction and iterative method (Ordered Subsets Expectation Maximization, OSEM), and then transmitted to the MedEx workstation together with the CT images for fusion imaging. The maximum standardized uptake value (SUVmax) was automatically generated by the workstation based on the information of the subject's weight, injection dose, and time. The region of interest (ROI) was outlined along with the extent of the lesion at the level where the concentration of the radioactive tracer was most obvious, and the workstation automatically calculates the SUVmax.

2.3. Confirmation of Lesions and Huamn Expert's Qualitative Classification

Considering that detailed pathological examination was not available in all patients, we determined the diagnosis of the lesions on the basis of pathological biopsy, imaging follow-up, and clinical course of the disease. Independent visual analysis of lesions was evaluated by two physicians (TAJ and JY) with 20 years of diagnostic experience using the double-blind method, physicians were not informed of the patient's clinical information and pathology but were told that the lesion was either MM or BM. The weighted kappa analysis was used to determine the interobserver agreement. Kappa coefficients of 0–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1 were considered to be slight, fair, moderate, good, and almost perfect agreement, respectively (23).

2.4. Segmentation and Feature Extraction

Segmentation of the lesions was also performed by double-blind methods with physicians who had 10 years of experience (CMY and ZJN) in diagnostic work. All features were extracted in MaZda software, which has been proven in previous studies

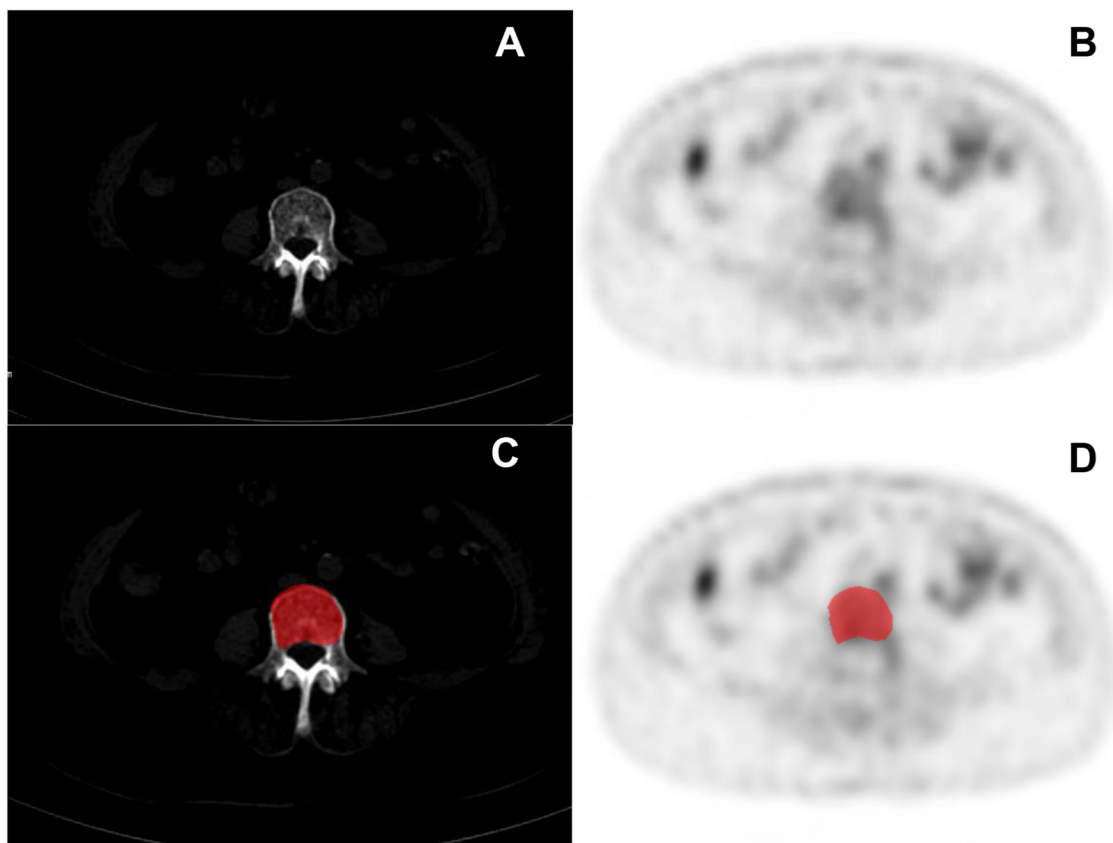


FIGURE 2 | Clinical cases PET/CT images of multiple myeloma (MM) (A,B) and the delineation of the region of interest (ROI) (C,D).

for radiomics, and all radiomics features extracted were in accordance with Image Biomarker Standardization Initiative (IBSI) standards (24, 25). The source images were extracted from the hospital PACS workstation and saved in BMP format. The object of this study was the largest cross-sectional area of the vertebral lesion, and the features were extracted by the 2D mode of Mazda software. Before the image was applied to MaZda software for feature extraction, uniform and standardized pre-processing of the image was performed by the method of $\mu \pm 3\sigma$ to make the features more reproducible and verifiable. The abnormal uptake of radioactive tracer on the image was used as the initial ROI, and the physician carefully identified the edges of the lesion and progressively outlined the ROI on the PET and CT images along the edge of the lesion. Because of the long examination time, the physician could make minor adjustments to determine the lesion of interest outlined and eliminate the effects of patient movement or expiratory motion. An example of the ROI outline was illustrated in **Figure 2**. A total of 279 features were extracted for each ROI, which were included in the following six common categories: gray-level histogram (HSLM), gray-level absolute gradient (GRM), gray-level run-length matrix (GLRLM), gray-level co-occurrence matrix (GLCM), autoregressive model (ARM), and wavelet. The interpretation of the features is described in detail in the previous study (26).

2.5. Reliability Analysis

To ensure the stability and reproducibility of the extracted features, a reliability test was performed. Another physician with 10 years of diagnostic experience repeated the outlining of the above ROI by randomly selecting 30 lesions. The reliability of the ROI outlined by the two physicians was assessed by the class correlation coefficient. Class correlation coefficients greater than 0.75 for radiomic features were considered to have good stability and reproducibility and were used for subsequent feature screening and model construction.

2.6. Dimensionality Reduction and Model Establishment

Before the feature screening, the normalization of the features was performed by the Z-score method, which aims to avoid the training of the model with too small weights, causing numerical instability, and to improve the comparability of the data, while enabling the parameter optimization to converge at a faster rate. After the reliability test, the training group was subjected to the least absolute shrinkage and selection operator (LASSO) regression for further data selection. LASSO regression was performed by fitting a generalized linear model with variable selection and complexity adjustment regularization. The filtering features were validated by 10-fold cross-validation based on the bias minimization criterion. Finally, for the final selected

TABLE 1 | Basic information for patients in the training and validation cohorts.

	The training cohort		<i>P</i>	The validation cohort		<i>P</i>
	BM	MM		BM	MM	
Gender			0.171			0.079
Female	22	16		11	8	
Male	38	15		15	6	
Age	63.58 ± 12.07	58.71 ± 10.08	0.470	60.88 ± 11.15	57.79 ± 13.20	0.521
Range	33–90	43–75		37–86	34–77	
Lesion form			0.057			0.101
Osteolytic	52	37		22	18	
Osteoblastic	15	2		6	0	
Mixed	13	10		6	3	
ISS stage						
I	-	10		-	3	
II	-	23		-	13	
III	-	16		-	5	
Extramedullary mass	27	19	0.604	11	7	0.778
SUVmax	6.84 ± 3.32	4.06 ± 1.61	<0.001	6.79 ± 3.31	4.38 ± 1.60	0.001
Osteoporosis			0.001			0.007
Positive	17	33		8	13	
Negative	63	16		24	8	
Confirmation			0.001			0.001
Biopsy	29	49		14	21	
Follow-up	51	0		18	0	

P < 0.05 was considered to be statistically significant; BM, bone metastases; MM, multiple myeloma; Extramedullary mass, Extramedullary soft tissue mass; ISS stage, International Staging System classification.

non-zero features, a classification model was built by multivariate logistic regression. CT models and PET models were constructed based on the final selected features (features were derived from CT and PET images, respectively). In order to better evaluate the performance of radiomics, a combined model (ComModel) was constructed by adding the PET conventional parameter SUVmax combined with PET radiomics features.

2.7. Model Comparison

The performance of all classification diagnostic models was evaluated by comparing the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV), while 95% CI of AUC were calculated. The DeLong test was used to compare the diagnostic effects between the models, and *P* < 0.05 was considered to be statistically different. In addition, calibration curves and Brier scores were used to evaluate the predictive ability and goodness of fit of the classification models to observe the agreement between the actual and predicted probabilities of the models. Decision curve analysis (DCA) was used to visualize and evaluate the clinical net benefit and clinical utility of the classification prediction model by the graphical presentation.

2.8. Statistical Analysis

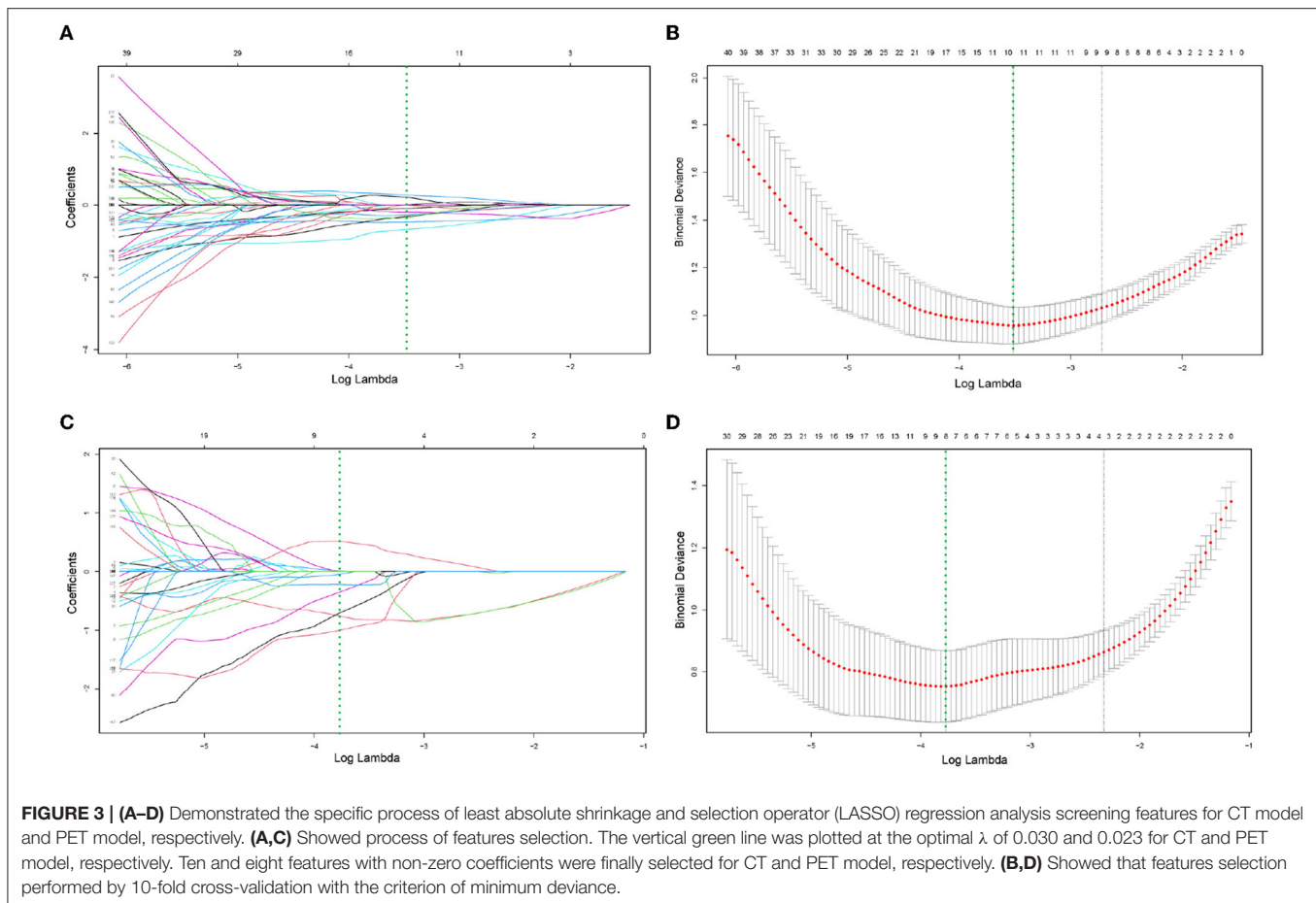
Independent samples *t*-test or Mann–Whitney *U*-test was used to compare continuous variables with normal or non-normal

distribution in the MM and BM groups. Categorical variables between the two groups were assessed using the chi-square test or Fisher test and weighted Kappa statistics were used to assess interobserver agreement. The processing of features screen, model construction, and comparison of the diagnostic performance of the models were performed in R software (version 4.1.1) and Python (version 3.8.1). IBM SPSS (version 21.0) and MedCalc software (version 20.0) were used for other clinical data analysis and ROC curve plotting. Probability values of *P* < 0.05 were considered statistically significant.

3. RESULTS

3.1. Basic Patient Information

According to the inclusion and exclusion criteria, a total of 131 patients were enrolled, including 86 patients who were diagnosed with bone metastases (BM), and the remaining 45 patients were confirmed as MM. The statistics and comparison of basic information of the patients were shown in **Table 1**. According to the diagnostic criteria of the International Myeloma Working Group, the stages of MM(ISS standard) were as follows: stage I: *n* = 10; stage II: *n* = 19; stage III: *n* = 16. The primary tumor of patients diagnosed with BM were as follows: lung cancer: *n* = 24; breast cancer: *n* = 17; prostate cancer: *n* = 14; pancreatic cancer: *n* = 4; kidney cancer: *n* = 3; ureteral cancer: *n* = 3; stomach cancer: *n* = 3; thyroid cancer: *n* = 3; liver cancer: *n* = 2; bladder cancer: *n* = 2; colon cancer: *n* = 2; parotid cancer: *n* = 2; gallbladder



cancer: $n = 2$; fallopian tube cancer: $n = 2$; uroepithelial cancer: $n = 1$; esophagus cancer: $n = 1$; cervical cancer: $n = 1$. A total of 184 lesions were obtained and randomly divided into training and validation groups according to the ratio of 7:3 (The training group: BM: $n = 80$, MM: $n = 49$; The validation group: BM: $n = 34$, MM: $n = 21$).

3.2. Feature Selection, Model Establishment, and Validation

After reliability testing and excluding features with ICC coefficients less than 0.75, 223 and 234 radiomics features were extracted from CT and PET in the training group, respectively. Then, 10 and 8 texture features were obtained from CT and PET after LASSO regression and 10-fold cross-validation, respectively. The LASSO regression screening process was described in detail in **Figure 3**, and the final filtered feature information in the training and validation groups of the MM group and BM group was illustrated in the heat map in **Supplementary Figures S1–S4**, which the differences in feature expression between the MM and BM groups were clearly seen. Furthermore, the selected features of CT and PET coefficients were also described in **Supplementary Table A**. In the training group, all models achieved very high AUC values and the classification diagnostic performance of the ComModel (AUC:0.973; CI95%:0.928–0.993) was significantly

improved compared to the CT (AUC:0.909; CI95%:0.846–0.952) and PET models (AUC:0.949; CI95%:0.896–0.980) ($P = 0.013$ and $P = 0.024$, respectively), while the PET model did not show a statistical difference in the DeLong test although it had a higher diagnostic performance compared to the CT model ($P = 0.131$). In the validation group, the ComModel (AUC: 0.948; CI95%: 0.853–0.990) and the PET model (AUC:0.929; CI95%: 0.826–0.981) achieved similar diagnostic performance and outperformed the CT model (AUC:0.897; CI95%: 0.785–0.963), and the DeLong test suggested no statistical difference between the three models ($P = 0.309$, $P = 0.466$, and $P = 0.496$, respectively).

3.3. Diagnostic Performance Between the CT Model, PET Model, ComModel, Human Experts, and SUVmax

Human experts' classification diagnostic of lesions was estimated by the kappa coefficient, and in this study, the weighted k -value for the interobserver agreement was 0.832, which indicates a relatively reliable agreement. In the training and validation groups, the AUC values of human experts for the classification and diagnosis performance of MM and BM were 0.835 (CI95%:0.760–0.895) and 0.840 (CI95%:0.717–0.925), respectively. while the AUC values of SUVmax between the two groups were 0.802 (CI95%:0.723–0.867) and 0.810

(CI95%:0.681–0.903), respectively. Both the ComModel and the PET model showed significant differences in the classification diagnosis of MM and BM compared to human experts in both the training ($P = 0.001$ and $P = 0.01$, respectively) and validation groups ($P = 0.033$ and $P = 0.018$, respectively). The CT model was not statistically different between the two groups compared to the human experts ($P = 0.187$ and $P = 0.299$, respectively). The ComModel and the PET model also showed great superiority compared to SUVmax between the two groups (Training group: $P < 0.001$ and $P = 0.001$; Validation group: $P = 0.019$, $P = 0.045$). No statistical difference was observed that the human expert compared to SUVmax between the two groups ($P = 0.036$ and $P = 0.732$). The classification diagnostic performance of all models was described and illustrated in **Table 2**, and the ROC curves of all classification models were illustrated in **Figure 4**, in addition, the detailed results of the DeLong test were also recorded in **Figure 4**.

3.4. Clinical Use and Calibration

According to the calibration curves, all the radiomics models were closer to the ideal curve, implying a good categorical diagnostic performance. In addition, the ComModel had a better fitness compared to the PET and CT models because of the smaller Brier scores (Brier scores were 0.070, 0.088, and 0.119, respectively), the calibration curve was shown in **Figure 5**. In terms of the net clinical gain of the models, both the ComModel and the PET model achieved good net clinical gain and outperformed the other models, the decision curve was shown in **Figure 6**.

4. DISCUSSION

Both myeloma and metastases were common malignant lesions of the spine. When patients only present with lumbar pain and no previous history of tumor, the clinical symptoms and imaging

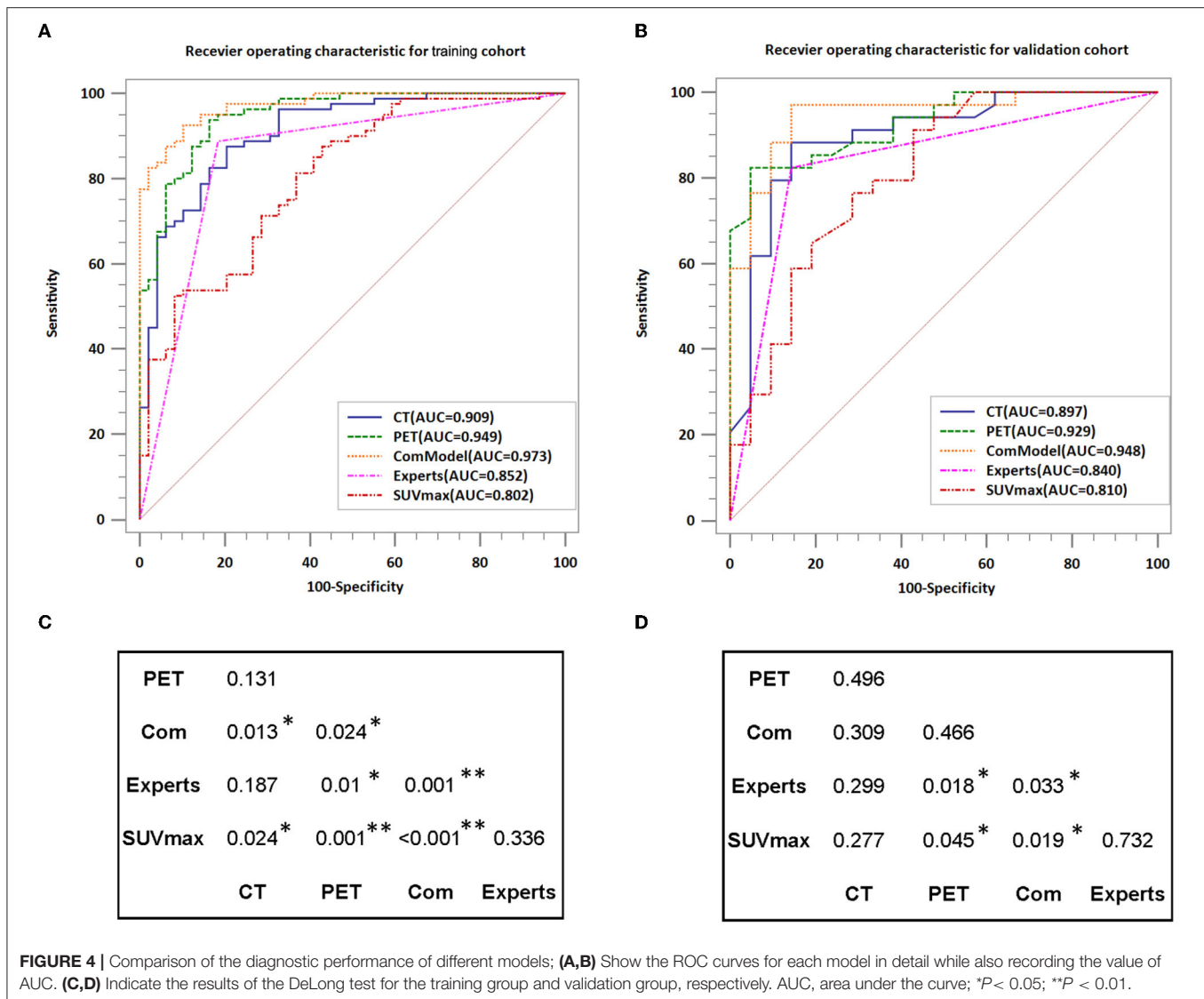
manifestations of both were similar. However, the treatment and prognosis of them were significantly different. Although bone biopsy was the gold standard for identifying benign and malignant lesions, there were limitations in clinical diagnosis due to its invasive examination. Radiomics provided a non-invasive assessment of the lesion and its microenvironment and allow quantification of the spatial heterogeneity of the lesion, which allows identification and evaluates the prognosis. In this study, we constructed and validated a radiomics model based on 18F-FDG PET-CT images and achieved excellent performance in classifying and diagnosing BM and MM. Furthermore, the radiomics model showed unique superiority and clinical utility compared to human experts as well as PET conventional parameter SUVmax. This will play a decisive role as a non-invasive and easy-to-use method in the diagnosis, staging, and re-staging of diseases and even in the selection of treatment strategies for diseases.

Previous studies had pointed out that the traditional imaging features of bone destruction in myeloma involved a series of small focal-like, worm-like, and broad bone destruction, especially for the imaging features such as chisel-like changes in the skull and broad bone destruction in the ribs were specific (27, 28). In addition, most patients with MM showed different degrees of osteoporosis and rarely osteoblastic bone changes. The imaging of bone metastases was characterized by the tendency to invade the pedicle rather than the vertebral body and lack of involvement of extremity bones (29). Mutlu et al. suggest that features such as more sclerotic margins around BM lesions and sharper margins around MM lesions may also be used to differentiate among them (12). Although these studies indicate that these features may be crucial information to identify them, there were still exist similar imaging presentations of bone metastases or atypical lesions in clinical work. Moreover, MM and BM cannot be discriminated simply from bone destruction. In this study, physicians successfully identified all osteoblastic

TABLE 2 | The diagnostic ability of each model for discriminating vertebral multiple myeloma (MM) from bone metastasis (BM).

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
CT model						
training cohort	0.909(0.846–0.952)	0.829	0.875	0.796	0.889	0.735
validation cohort	0.897(0.785–0.963)	0.836	0.882	0.857	0.882	0.762
PET model						
training cohort	0.949(0.896–0.980)	0.884	0.937	0.837	0.900	0.857
validation cohort	0.929(0.826–0.981)	0.873	0.824	0.952	0.882	0.857
ComModel						
training cohort	0.973(0.928–0.993)	0.915	0.925	0.898	0.925	0.898
validation cohort	0.948(0.853–0.990)	0.891	0.971	0.857	0.912	0.857
Human experts						
training cohort	0.835(0.760–0.895)	0.845	0.875	0.796	0.875	0.796
validation cohort	0.840(0.717–0.925)	0.836	0.824	0.857	0.824	0.857
SUVmax						
training cohort	0.802(0.723–0.867)	0.729	0.875	0.571	0.775	0.653
validation cohort	0.810(0.681–0.903)	0.745	0.912	0.571	0.794	0.667

AUC, the area under the ROC curve; PPV, positive predictive value; NPV, negative predictive value.



lesions in both the training and validation cohort of the BM group. However, 17.1% (22/129) of the lesions were incorrectly identified in the classification of osteolytic lesions. Physician identification of lesions on conventional imaging mainly was attributed to subjective visual assessment as well as diagnostic experience. Still, this approach was undoubtedly challenging for younger physicians with less diagnostic experience. Accurate classification of BM and MM was crucial as it relates to the plan of individualized treatment, reduction of complications, and improvement of prognosis.

Maximum standardized uptake value as the conventional parameter of PET/CT was used in past studies to determine the treatment sensitivity and prognostic value of malignant lymphoma in the early and intermediate stages (30, 31). On the other hand, the study of Polat et al. confirmed the predictive value of SUVmax for grading and staging of renal clear cell carcinoma and the risk of stratification (32, 33). In our study, SUVmax in the MM group (SUVmax: 4.06 ± 1.61) compared

to the BM (SUVmax: 6.84 ± 3.22) group showed a significant decrease in both the training and validation groups, and similar results had been reported several times in the past studies (34, 35). However, the SUVmax for MM and BM in study Li et al. was (1.6 ± 0.7 and 5.5 ± 2.7), respectively, and this variability may be due to the subjects of that study being derived from 334 patients with 8,896 lesions throughout the body, whereas our study focused on the spine and maximum of two lesions per case (5). There are no clear diagnostic thresholds for SUVmax to identify MM and BM. Furthermore, the AUC values of SUVmax for discriminating the BM and MM in the training and validation groups in our study were 0.802 and 0.810, and the accuracy were 0.729 and 0.745, respectively, which achieved only moderate diagnostic efficacy and were not sufficient to make accurate predictions for the classification of lesions. Furthermore, it was difficult for SUVmax to provide a comprehensive description of the heterogeneity and spatial consistency of lesions.

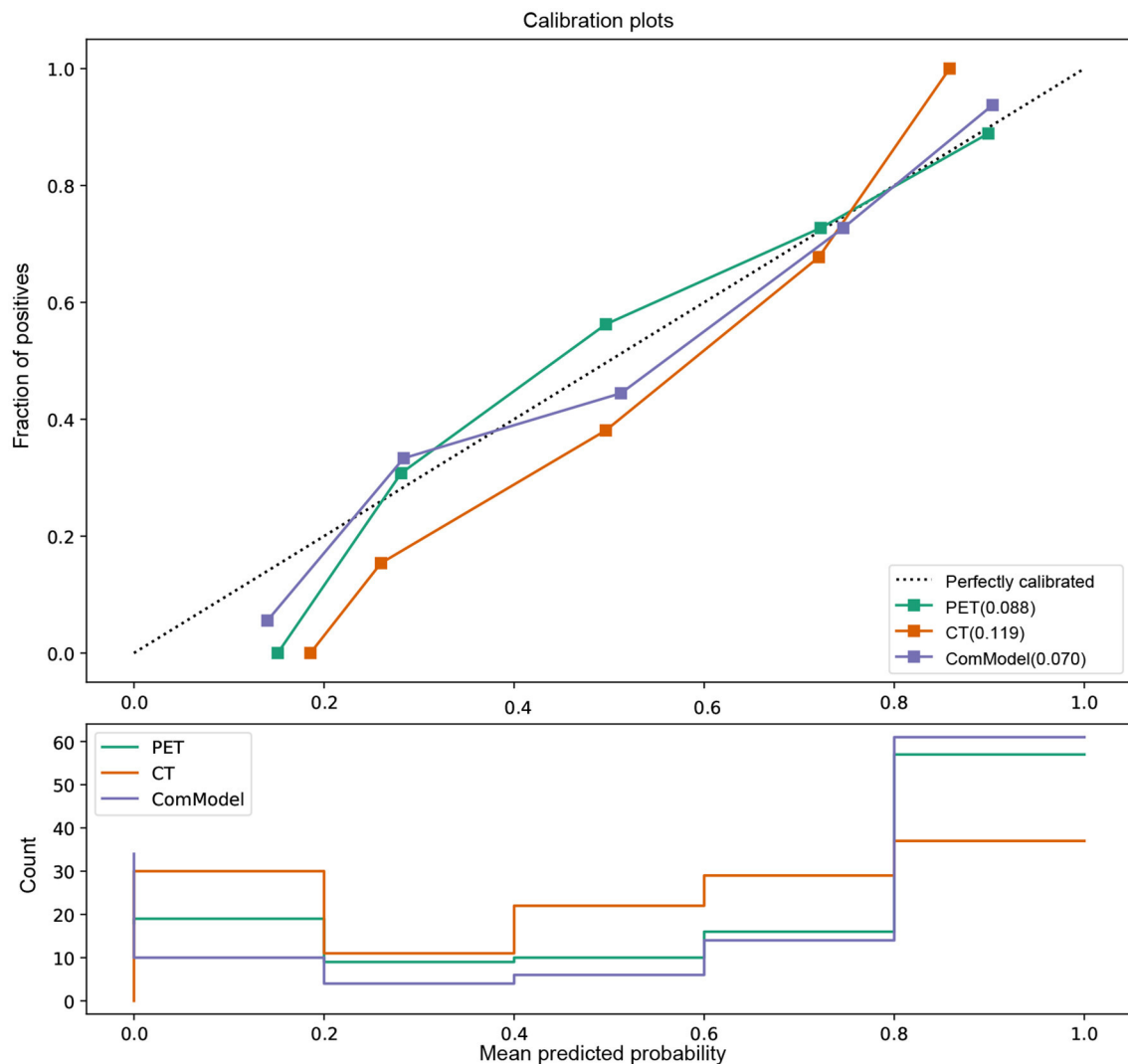


FIGURE 5 | Comparison of the calibration curve and Brier score of different models. All three model's calibration curves were closed to ideal curves, indicating that the models had good fitness and predictive ability. The ComModel had better goodness of fit compared to the PET and CT models because of the smaller Brier scores (Brier scores were 0.070, 0.088, and 0.119, respectively). The following figure shows the distribution of the probability of diagnosis for different models.

Radiomics could transcend subjective visual assessment to provide an objective evaluation of lesion and tissue heterogeneity, which served as a new tool to provide valuable information about the microenvironment of lesions that cannot be observed by the human eyes. PET/CT radiomics was demonstrated several times in past studies to play an essential role in the diagnosis and prognosis of diseases and performing assessment of therapeutic efficacy. In our research, the radiomics models constructed based on PET/CT images had high diagnostic efficacy in discriminating MM and BM not only in the training group, with AUC values of 0.909, 0.949, and 0.973 for the CT model, PET model, and ComModel, respectively, but also in the validation group, with AUC values of 0.897, 0.929, and 0.948 for the CT model, PET model, and ComModel, respectively. In addition, the diagnostic performance and clinical utility of the radiomics model were superior to those of

the human expert and SUVmax, with incremental value for differential and diagnostic purposes, especially the PET and COM models. It should note that a proportion of patients incorrectly staged by human experts (10.9%) and SUVmax (23.9%) were correctly diagnosed by our radiomics model, indicating that the radiomics model could complement the current staging scheme. More importantly, our findings suggest that the PET model had a higher value than conventional CT radiomics in discriminating MM from BM. Although there was no statistical difference in the Delong test, the AUC, Accuracy, Sensitivity, and Specificity of the PET model were significantly improved compared with the CT model. This may be due to the fact that PET images represent radioactive tracer uptake and metabolic information of the lesion. At the same time, PET/CT radiomics reflected the quantification of tumor uptake heterogeneity and earlier detection of lesions compared to conventional imaging,

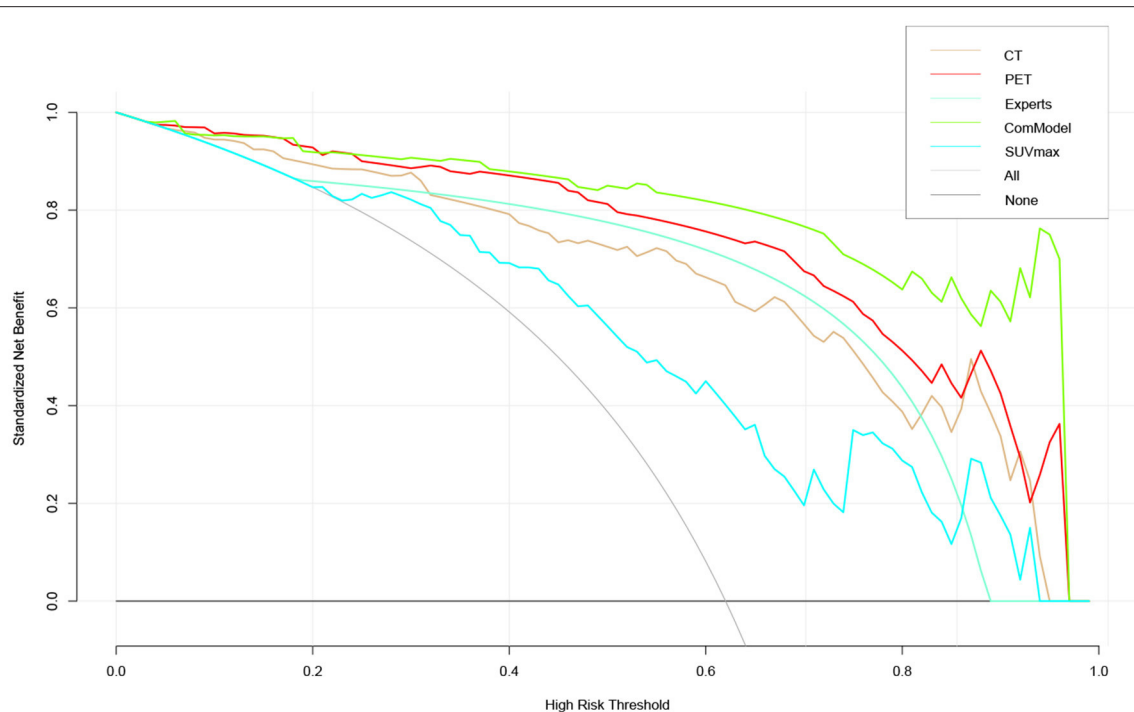


FIGURE 6 | The clinical practicability of the models in this study was evaluated and compared, which indicated that the PET model and the ComModel had better net clinical benefit than the other models.

which brings additional value for lesion and tissue specificity identification.

Our results show that Perc.01% and Vertl_RLNonUni were the most representative features in identifying MM and BM, as they appeared in both CT and PET models. The radiomics parameter Perc.01% derived from HSLM reflected the brightness value of the area where it was located and the number of pixels, which further confirmed the excellent performance of the radiomic model since these parameters were closely related to bone density in CT images and radioactive tracer uptake in PET images, as well as to the osteoporosis exhibited by patients with myeloma (36). The radiomics parameter Vertl_RLNonUni derived from the GLRLM reflects the heterogeneity of the images in different directions, and the Vertl_RLNonUni had been demonstrated in previous studies as a reliable indicator that could be used to predict the grading and staging of clear cell renal cell carcinoma and to perform risk stratification (37). In our study, it was hypothesized that this may be due to the different pathological mechanisms of MM and BM, as well as the fact that BM was an infiltrative lesion while MM was a diffuse lesion (38).

Our findings were highly reproducible because we applied rigorous subject screening and reliability testing of lesion segmentation during the study. The LASSO regression algorithm has been applied and validated many times in the past to have good utility in screening the feature parameters. Our radiomics model has been validated by methods such as the Calibration curve and DCA curve, and has good fitness and is very close to the ideal curve. It was worth noting that we performed the

outline of ROI in 2D mode rather than 3D mode, and the radiomic features generated by different mode outlines may be different. Still, past studies had demonstrated that the models constructed in 2D or 3D mode achieved similar classification diagnostic performance (39). In this study, we compared the bone metastases of different malignant with MM, it is unclear whether the characteristic parameters of bone metastases caused by different malignancies are the same, Xiong et al. have tried to distinguish the BM characteristics of lung cancer and other cancers, and their results achieved only moderate diagnostic results in distinguishing them (20). In addition, patients with BM and MM may have altered images after chemotherapy or radiotherapy, such as focal radiotracer uptake and SUV measurements. Post-treatment lesions exhibiting flare phenomena and osteogenic-type responses may also introduce changes to the features extraction. Therefore, we strictly screened the enrolled patients to eliminate the effect of treatment on the images and ensure the rigor of this study. Finally, the use of pre-treatment images to construct radiomic models had the potential to help clinicians physicians to determine the sensitivity of patients to radiotherapy or chemotherapy and, thus, better stratify patients to determine more appropriate individualized treatment plans.

There were still exist some limitations in our study. First, as a single-center study, this study may be biased in terms of patient selection, and thus, the results may hardly represent generalizable findings. In addition, there was a lack of external validation data from the multicenter association. Therefore, all aspects still

need to be adjusted and optimized before applying to the clinic. Second, retrospective studies may have selection bias related to study data collection. Third, not all patients enrolled had pathological findings, and we combined pathological findings and follow-up results to determine the classification of lesions under strict adherence to inclusion and exclusion criteria.

5. CONCLUSION

The radiomics model constructed based on 18F-FDG PET/CT images achieved satisfactory diagnostic performance for the classification of MM and bone metastases. In addition, the radiomics model showed a significant improvement in diagnostic performance compared to human experts and PET conventional parameter SUVmax. This non-invasive method could be used as a complement to traditional diagnostic methods. Furthermore, it had the potential to help clinicians physicians to develop individualized treatment plans, avoid adverse risks, and improve treatment outcomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the

local legislation and institutional requirements. Written informed consent from the patients/participants their next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ZJ and JY contributed to the conception and design of the study. FZ and YIW carried out data statistics and analysis. ZJ wrote the manuscript. JY revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the Nature Science Foundation of Liaoning Province of China (No. 2019-ZD-0620).

ACKNOWLEDGMENTS

We would like to thank Aijuan Tian, Meiyan Chen, and Jianan Zhang from the Department of Nuclear Medicine of the Second Hospital of Dalian Medical University for their help in image processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2022.874847/full#supplementary-material>

REFERENCES

- Edgardo, J, C, Angtuaco, Athanasios, B. Multiple myeloma: clinical review and diagnostic imaging. *Radiology*. (2004) 231:11–23. doi: 10.1148/radiol.2311020452
- Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. *Cancer Treat Rev*. (2001) 27:165–76. doi: 10.1053/ctrv.2000.0210
- Weber KL. Evaluation of the adult patient (aged >40 years) with a destructive bone lesion. *J Am Acad Orthop Surg*. (2010) 18:169–79. doi: 10.5435/00124635-201003000-00006
- Peabody T. The rodged metastasis is a sarcoma: strategies to prevent inadvertent surgical procedures on primary bone malignancies. *Instr Course Lect*. (2004) 53:657–661.
- Li X, Wu N, Zhang W, Liu Y, Ming Y. Differential diagnostic value of 18F-FDG PET/CT in osteolytic lesions. *J Bone Oncol*. (2020) 24:100302. doi: 10.1016/j.jbo.2020.100302
- Wang QT, Li Y, Liang Y, Hu C, Zhai Y, Zhao G, et al. Construction of a multiple myeloma diagnostic model by magnetic bead-based MALDI-TOF mass spectrometry of serum and pattern recognition software. *Anatom Record*. (2009) 292:604–10. doi: 10.1002/ar.20871
- Usmani, Saad Z. Prognostic implications of serial 18-fluoro-deoxyglucose emission tomography in multiple myeloma treated with total therapy 3. *Blood*. (2013) 121:1819–23. doi: 10.1182/blood-2012-08-451690
- Moreau P, Attal M, Caillot D, Macro M, Karlin L, Garderet L, et al. Prospective evaluation of magnetic resonance imaging and [(18)F]Fluorodeoxyglucose positron emission tomography-computed tomography at diagnosis and before maintenance therapy in symptomatic patients with multiple myeloma included in the IFM/DFCI 2009. *Trial*. (2017) 35:2911–2918. doi: 10.1200/JCO.2017.72.2975
- Agarwal A, Chirindel A, Shah BA, Subramaniam RM. Evolving role of FDG PET/CT in multiple myeloma imaging and management. *Am J Roentgenol*. (2013) 200:884–90. doi: 10.2214/AJR.12.9653
- Hur J, Yoon CS, Ryu YH, Yun MJ, Suh JS. Efficacy of multidetector row computed tomography of the spine in patients with multiple myeloma: comparison with magnetic resonance imaging and fluorodeoxyglucose-positron emission tomography. *J Comput Assist Tomogr*. (2007) 31:342. doi: 10.1097/01.rct.0000237820.41549.c9
- Yildirim M, Baykara M. Differentiation of multiple myeloma and lytic bone metastases: histogram analysis. *J Comput Assist Tomogr*. (2020) 44:953–955. doi: 10.1097/RCT.0000000000001086
- Mutlu U, Balci A, zsan GH, Zkal S, zgl HA. Computed tomography characteristics of multiple myeloma and other osteolytic metastatic bone lesions. *Acta radiol*. (2020) 62:1639–47. doi: 10.1177/0284185120977035
- Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. (2016) 278:563–77. doi: 10.1148/radiol.2015151169
- Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGP, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036
- Cy A, Mh A, Sl B, Jc B, Yao YB, Na QB, et al. Radiomics model of magnetic resonance imaging for predicting pathological grading and lymph node metastases of extrahepatic cholangiocarcinoma. *Cancer Lett*. (2020) 470:1–7. doi: 10.1016/j.canlet.2019.11.036
- Marino MA, Pinker K, Leithner D, Sung J, Jochelson M. Contrast-Enhanced mammography and radiomics analysis for noninvasive breast

- cancer characterization: initial results. *Mol Imaging Biol.* (2019) 22:780–7. doi: 10.1007/s11307-019-01423-5
17. Ripani D, Caldarella C, Za T, Rossi E, Stefano VD, Giordano A. Progression to symptomatic multiple myeloma predicted by texture analysis-derived parameters in patients without focal disease at 18 F-FDG PET/CT. *Clin Lymphoma Myeloma Leuk.* (2021) 21:536–44. doi: 10.1016/j.clml.2021.03.014
 18. Min Y, Xiang K, Feng Y, Chen H, Chen J, Wei X, et al. Development and validation of a population-based model for predicting the regional lymph node metastasis in adolescent differentiated thyroid carcinoma. *Oral Oncol.* (2021) 121:105507. doi: 10.1016/j.oraloncology.2021.105507
 19. Bi Z, Chen JJ, Liu PC, Chen P, Wang YS. Candidates of genomic tests in HR+/HER2- breast cancer patients with 1-2 positive sentinel lymph node without axillary lymph node dissection: analysis from multicentric cohorts. *Front Oncol.* (2021) 11:722325. doi: 10.3389/fonc.2021.722325
 20. Xiong X, Wang J, Hu S, Dai Y, Hu C. Differentiating between multiple myeloma and metastasis subtypes of lumbar vertebra lesions using machine learning based radiomics. *Front Oncol.* (2021) 11:601699. doi: 10.3389/fonc.2021.601699
 21. Tagliafico AS, Cea M, Rossi F, Valdora F, Dominietto A. Differentiating diffuse from focal pattern on computed tomography in multiple myeloma: added value of a radiomics approach. *Eur J Radiol.* (2019) 121:108739. doi: 10.1016/j.ejrad.2019.108739
 22. Greipp PR, San Miguel J, Durie BGM, Crowley JJ, Barlogie B, Blad J, et al. International staging system for multiple myeloma. *J Clin Oncol.* (2020) 23:3412–20. doi: 10.1200/JCO.2005.04.242
 23. Chen Y, Xi W, Yao W, Wang L, Zhang H. Dual-Energy computed tomography-based radiomics to predict peritoneal metastasis in gastric cancer. *Front Oncol.* (2021) 11:659981. doi: 10.3389/fonc.2021.659981
 24. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Lck S. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* (2020) 295:191145. doi: 10.1148/radiol.2020191145
 25. Korte JC, Cardenas C, Hardcastle N, Kron T, Wang J, Bahig H, et al. Radiomics feature stability of open-source software evaluated on apparent diffusion coefficient maps in head and neck cancer. *Sci Rep.* (2021) 11:17633. doi: 10.1038/s41598-021-96600-4
 26. Li Y, Xu X, Weng S, Yan C, Chen J, Ye R. CT image-based texture analysis to predict microvascular invasion in primary hepatocellular carcinoma. *J Digit Imaging.* (2020) 33:1365–75. doi: 10.1007/s10278-020-00386-2
 27. Chang MC, Chen JH, Liang JA, Lin CC, Kao CH. Meta-analysis: comparison of F-18 fluorodeoxyglucose-positron emission tomography and bone scintigraphy in the detection of bone metastasis in patients with lung cancer. *Acad Radiol.* (2011) 19:349–57. doi: 10.1016/j.acra.2011.10.018
 28. D'Anastasi M, Grandl S, Reiser ME, Baur-Melnyk A. Radiological diagnostics of multiple myeloma. *Radiologe.* (2014) 54:556–63. doi: 10.1007/s00117-013-2628-9
 29. Even-Sapir E, Martin RH, Barnes DC, Pringle CR, Iles SE, Mitchell MJ. Role of SPECT in differentiating malignant from benign lesions in the lower thoracic and lumbar vertebrae. *Radiology.* (1993) 187:193. doi: 10.1148/radiology.187.1.8451412
 30. Rossi C, Kanoun S, Berriolo-Riedinger A, Dygai-Cochet I, Humbert O, Legouge C, et al. Interim 18F-FDG PET SUVmax reduction is superior to visual analysis in predicting outcome early in hodgkin lymphoma patients. *J Nuclear Med.* (2014) 55:569–73. doi: 10.2967/jnumed.113.130609
 31. Itti E, Meignan M, Berriolo-Riedinger A, Biggi A, Cashen AF, Vra P, et al. An international confirmatory study of the prognostic value of early PET/CT in diffuse large B-cell lymphoma: comparison between Deauville criteria and SUVmax. *Eur J Nucl Med.* (2013) 40:1312–20. doi: 10.1007/s00259-013-2435-6
 32. Polat EC, Otunctemur A, Ozbek E, Besiroglu H, Horsanali MO. Standardized uptake values highly correlate with tumor size and fuhrman grade in patients with clear cell renal cell carcinoma. *Asian Pacific J Cancer Prevent.* (2014) 15:7821–4. doi: 10.7314/APJCP.2014.15.18.7821
 33. Nakaigawa N, Kondo K, Tateishi U, Minamimoto R, Kaneta T, Namura K, et al. FDG PET/CT as a prognostic biomarker in the era of molecular-targeting therapies: max SUVmax predicts survival of patients with advanced renal cell carcinoma. *BMC Cancer.* (2016) 16:67. doi: 10.1186/s12885-016-2097-4
 34. Dai D, Xu WG, Wang Q. The value of SUV of FDG-PET/CT in differentiation of myeloma and metastasis in patients with malignant skeletal diseases of unknown origin. *Chin J Orthopaedics.* (2009) 29:1127–30. doi: 10.1016/j.ceramint.2007.09.109
 35. erao T, Machida Y, Hirata K, Kuzume A, Tabata R, Tsushima T, et al. Prognostic impact of metabolic heterogeneity in patients with newly diagnosed multiple myeloma using 18F-FDG PET/CT. *Clin Nucl Med.* (2021) 46:790–6. doi: 10.1097/RLU.0000000000003773
 36. Ito K, Muraoka H, Hirahara N, Sawada E, Kaneda T. Quantitative assessment of normal submandibular glands and submandibular sialadenitis using CT texture analysis: a retrospective study. *Oral Surg Oral Med Oral Pathol Oral Radiol.* (2020) 132:112–17. doi: 10.1016/j.oooo.2020.10.007
 37. Demirjian NL, Varghese BA, Cen SY, Hwang DH, Aron M, Siddiqui I, et al. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *Eur Radiol.* (2021) 32:2552–63. doi: 10.1007/s00330-021-08344-4
 38. Haznedar R, Ak SZ, zgr U Akdemir, zkurt ZN, zcan eneli, Yac M, et al. Value of 18F-fluorodeoxyglucose uptake in positron emission tomography/computed tomography in predicting survival in multiple myeloma. *Eur J Nucl Med Mol Imaging.* (2011) 38:1046–1053. doi: 10.1007/s00259-011-1738-8
 39. Meng L, Dong D, Chen X, Fang M, Tian J. 2D and 3D CT radiomic features performance comparison in characterization of gastric cancer: a multi-center study. *IEEE J Biomed Health Inf.* (2020) 25:755–63. doi: 10.1109/JBHI.2020.3002805

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Jin, Wang, Wang, Mao, Zhang and Yu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Chest L-Transformer: Local Features With Position Attention for Weakly Supervised Chest Radiograph Segmentation and Classification

Hong Gu¹, Hongyu Wang¹, Pan Qin^{1*} and Jia Wang^{2*}

¹ Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, ² Department of Surgery, The Second Hospital of Dalian Medical University, Dalian, China

OPEN ACCESS

Edited by:

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC),
Switzerland

Reviewed by:

Hongxiang Lin,
University College London,
United Kingdom
Salvatore Annunziata,
Fondazione Policlinico Universitario A.
Gemelli IRCCS, Italy

*Correspondence:

Pan Qin
qp112cn@dlut.edu.cn
Jia Wang
wangjia77@hotmail.com

Specialty section:

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

Received: 19 April 2022

Accepted: 12 May 2022

Published: 02 June 2022

Citation:

Gu H, Wang H, Qin P and Wang J
(2022) Chest L-Transformer: Local
Features With Position Attention for
Weakly Supervised Chest Radiograph
Segmentation and Classification.
Front. Med. 9:923456.
doi: 10.3389/fmed.2022.923456

We consider the problem of weakly supervised segmentation on chest radiographs. The chest radiograph is the most common means of screening and diagnosing thoracic diseases. Weakly supervised deep learning models have gained increasing popularity in medical image segmentation. However, these models are not suitable for the critical characteristics presented in chest radiographs: the global symmetry of chest radiographs and dependencies between lesions and their positions. These models extract global features from the whole image to make the image-level decision. The global symmetry can lead these models to misclassification of symmetrical positions of the lesions. Thoracic diseases often have special disease prone areas in chest radiographs. There is a relationship between the lesions and their positions. In this study, we propose a weakly supervised model, called Chest L-Transformer, to take these characteristics into account. Chest L-Transformer classifies an image based on local features to avoid the misclassification caused by the global symmetry. Moreover, associated with Transformer attention mechanism, Chest L-Transformer models the dependencies between the lesions and their positions and pays more attention to the disease prone areas. Chest L-Transformer is only trained with image-level annotations for lesion segmentation. Thus, Log-Sum-Exp voting and its variant are proposed to unify the pixel-level prediction with the image-level prediction. We demonstrate a significant segmentation performance improvement over the current state-of-the-art while achieving competitive classification performance.

Keywords: weakly supervised, lesion segmentation, transformer, local feature, chest radiograph

1. INTRODUCTION

The chest radiograph is widely applied for the diagnosis of thoracic diseases. Diagnostic imaging often requires the classification of findings, as well as their geometrical information. Segmentation of lesions is an indispensable part of clinical diagnosis (1). Deep learning models have achieved considerable success in chest radiograph segmentation (2–4). Unfortunately, these supervised models require substantial pixel-level annotated data to locate the lesions (3–5). The pixel-level annotated medical data are prohibitively expensive to acquire

with long working hours of expert radiologists. On the contrary, image-level annotations can be relatively easy to access with the text analysis techniques on radiological reports (6, 7). Thus, a good alternative to supervised learning is weakly supervised learning, which leverages image-level annotations to search the segmentation prediction (8). Existing deep learning models for weakly supervised medical segmentation class the images with features extracted with convolutions (9–12). The pixel-level and image-level predictions are unified with algorithms based on Multiple-instance learning (MIL) (9, 10, 13) or class activation map (CAM) (11, 12, 14). Moreover, the attention mechanism is adopted to promote their performances (9–12). However, these weakly supervised models do not consider the critical characteristics of chest radiographs: the global symmetry of lungs and dependencies between lesions and their positions.

There is an imperfect symmetry between the left and right lungs (15), which the existing weakly supervised models don't take into account. They extract global features from the whole image and it is unclear how the latent feature space is related to the pixel space (9–12). The global symmetry of the lungs can lead these models to contrast symmetrical positions in the left and right lungs to classify the lesions (9). As a result, features of lesions appear at the symmetrical positions of the lesions in the feature space, and the symmetrical positions are misclassified as lesions (9).

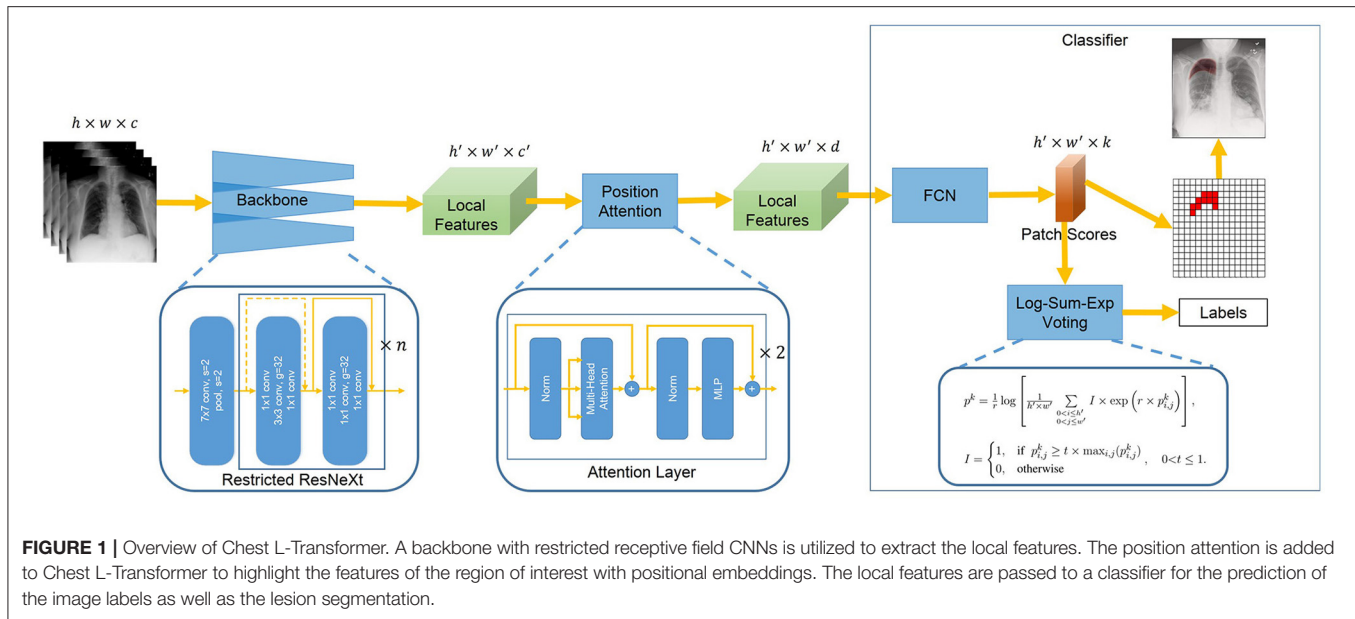
Convolutional neural networks (CNNs) with restricted receptive fields have been applied to relate the feature space and the pixel space exactly (16–18). In these models, the images are sliced into patches and the features are extracted within local patches (16–18). The class evidences produced by local features are averaged across all patches to infer the image-level labels with the softmax activation (16–18). However, the selection of the patch size is a hard problem for CNNs with restricted receptive fields to apply in weakly supervised segmentation. Increasing the patch size expands the receptive field and leads to better local features for classification, but coarsens the segmentation output (16–18). Another problem is the way to aggregate pixel-level evidences to the image-level decision. Unlike the images used in (16–18), all of which contain objects, the medical image datasets contain an extra class: no lesion. Averaging the class evidences, patches have the same weight to infer the image-level class. In the abnormal images, the patches with no lesion are more than those with lesions. To assign the right label to the images with lesions, many patches with no lesion may be classed as lesions. The patch with more evidence of lesion should have larger weights in the aggregation. There is another common function for aggregation: the max function, which encourages the model to just consider the most-likely lesion patch (13). But training with just one patch of the whole image, the model is hard to converge (19). Moreover, chest radiographs contain special areas, like the muscle and the black background, which are unrelated to thoracic diseases. It is necessary to filter them out in the aggregation of patches. Moreover, the softmax activation is designed for mutual exclusion. But different diseases can appear in one chest radiograph and may even have an overlapping region.

Another characteristic of chest radiographs is the dependencies between lesions and their positions. Thoracic diseases often have special disease prone areas in chest radiographs. This fact implies a relationship between the lesions and their positions. Weakly supervised deep learning models highlight salient parts of feature maps and separate redundant information with CNN attention modules to promote their performance (9–12). These CNN attention modules treat areas of the whole image equally, with the same convolution and pooling operations (9–12). But the salient parts are more likely located in the disease prone areas and extra attention should be paid to these areas. These models lack the ability to model the position information present in chest radiographs.

To tackle the aforementioned problems, we propose a weakly supervised deep learning model, called Chest L-Transformer, for lesion segmentation and disease classification on chest radiographs. Chest L-Transformer completes these two tasks only using image-level annotations. We present a new restricted receptive field CNN, called Restricted ResNeXt, as the backbone of Chest L-Transformer. Restricted ResNeXt extracts local features with a restricted receptive field and relates the feature space and the pixel space exactly. Hence, the features of lesions only appear at nearby positions of themselves, and the misclassification caused by the symmetry is avoided. Furthermore, Restricted ResNeXt extracts the local features not only from image patches but also from a limited nearby area around them. It can expand the receptive field while maintaining the fine scale of the segmentation output. A particular voting function, called Log-Sum-Exp voting, is proposed to aggregate pixel-level evidences. With this function, patches with differential evidences will have different weights to infer the image-level classes. Furthermore, a variant of Log-Sum-Exp voting is proposed to filter the unrelated areas. To ensure that multiple diseases can be detected simultaneously, the sigmoid activation takes place of the softmax one. Finally, Transformer attention mechanism (20) is introduced into the attention block of Chest L-Transformer to utilize the dependencies between the lesions and their positions. The attention block focuses on the disease prone areas with additional learnable positional embeddings (20, 21). We demonstrate a significant segmentation performance improvement over the current state of the art with competitive classification performance.

2. METHODS

With image-level annotated images, we aim to design a deep learning model that simultaneously produces disease classification and lesion segmentation. The proposed architecture is shown in **Figure 1**. It consists of three components: backbone, position attention block, and classifier. The backbone extracts the local features with Restricted ResNeXt. The local feature maps are downsampled and each pixel of the feature maps represents a small patch in the original image. The features of the region of interest are highlighted by the position attention block, which is mainly realized by two attention layers. The classifier first assigns



each patch a probability of the lesion for the segmentation task by the fully convolutional network (FCN). Then, Log-Sum-Exp voting allocating patches with differential evidences differential weights are used by the classifier in inferring the image-level classes with the probabilities of patches.

2.1. Backbone

We propose a variant of ResNeXt architecture as the backbone given its dominant performance in image analysis (22). Our backbone, Restricted ResNeXt, differs from ResNeXt (22) mainly in the replacement of many 3×3 by 1×1 convolutions for a restricted receptive field (see **Figure 2**). Restricted ResNeXt addresses the gradient vanish problem with the residual learning (23) and reduces the model complexity with the split-transform-merge strategy (24). After removing the final classification and pooling layers, an input image with shape $h \times w \times c$ produces a local feature tensor with shape $h' \times w' \times c'$. Here, h , w , and c are the height, width, and number of channels of the input image respectively while $h' = h/16$, $w' = w/16$, and $c' = 2,048$. The output of this network encodes the images into a set of abstracted feature maps. Each pixel of the feature maps represents a small patch (size 16×16) in chest radiographs. The receptive field size of the topmost convolutional layer of Restricted ResNeXt is limited to 39×39 pixels. The size of the receptive field can be increased by reducing the number of replaced 3×3 convolutions, while the scale of the output remains unchanged.

2.2. Position Attention

The position attention block (see **Figure 3**) highlights local features of the region of interest with Transformer attention mechanism (20). In the position attention block, the local features \mathbf{x} are mapped into a d -dimensional ($d = 1,024$) embeddings \mathbf{z}_0 with position information (Equation 1). The local features $\mathbf{x} \in \mathbb{R}^{h' \times w' \times c'}$ are reshaped into a sequence of flattened 2D features

$\mathbf{x}_p \in \mathbb{R}^{(h' \cdot w') \times c'}$. The flattened features \mathbf{x}_p are mapped into a latent d -dimensional embedding space using a trainable linear projection. To use position information, learnable positional embeddings (25) are added to the feature embeddings to retain position information as follows:

$$\mathbf{z}_0 = \mathbf{x}_p \times \mathbf{E} + \mathbf{E}_{pos}, \quad (1)$$

where $\mathbf{E} \in \mathbb{R}^{c' \times d}$ denotes the patch embedding projection and $\mathbf{E}_{pos} \in \mathbb{R}^{(h' \cdot w') \times d}$ denotes the positional embeddings. Then, d -dimensional embeddings \mathbf{z}_0 are put into a stack of $L = 2$ identical attention layers. Each layer has two sub-layers including a multi-head self-attention (MSA) mechanism and a small multi-layer perceptron (MLP) with one hidden layer. The MSA is an extension of “Scaled Dot-Product Attention” (20). We run $M = 12$ “Scaled Dot-Product Attention” operations and project their concatenated outputs in the MSA. We employ a residual connection (23) around each of the two sub-layers, followed by layer normalization (26). Therefore the output features of the l -th layer can be written as follows:

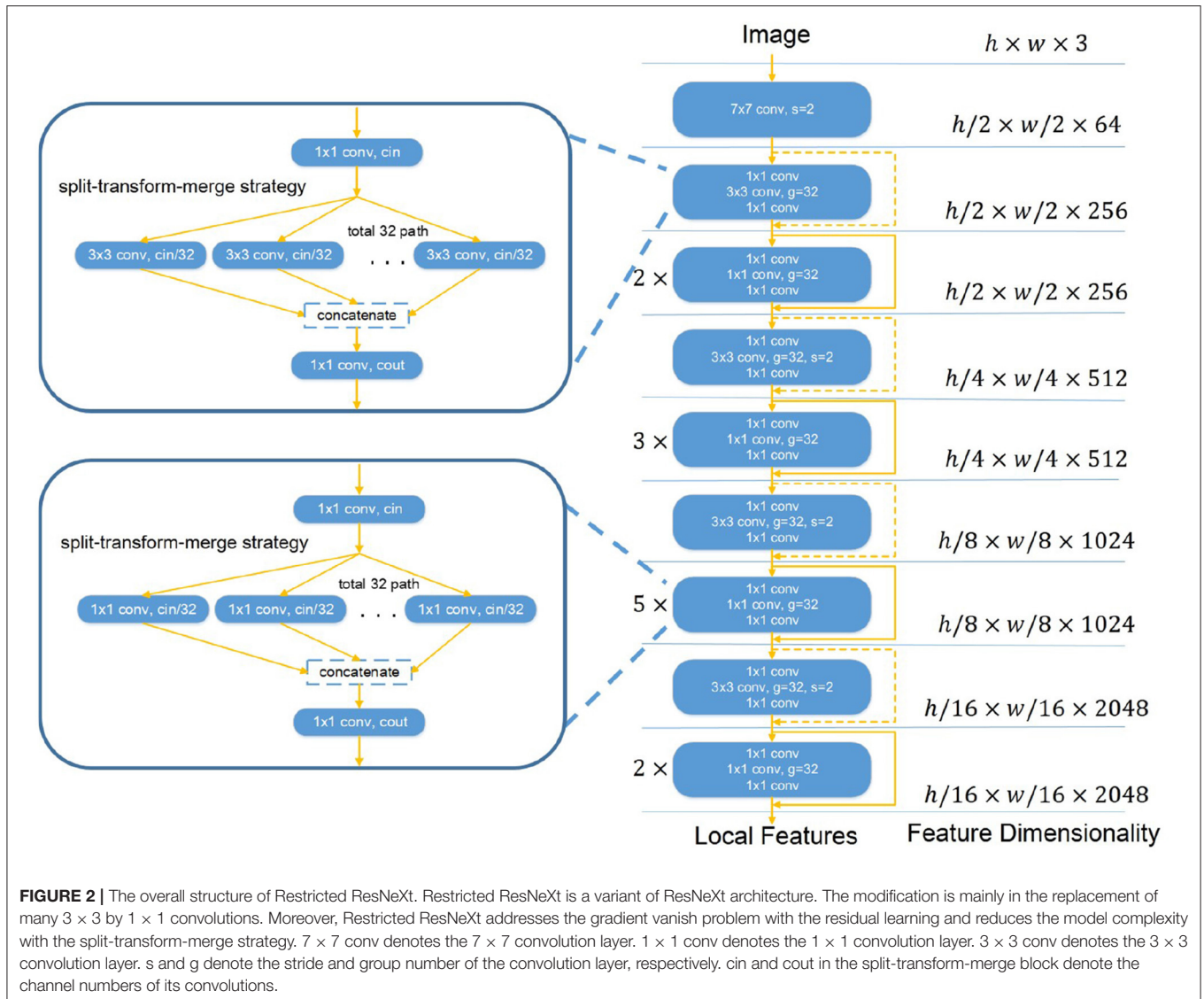
$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad (2)$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad (3)$$

where $l \in \{1, 2\}$ is the layer number, \mathbf{z}_l denotes the output by the l -th layer, and LN denotes the layer normalization operator. At last, the 2D features \mathbf{z}_2 are reshaped back into 3D features $\mathbf{x}' \in \mathbb{R}^{h' \times w' \times d}$.

2.3. Segmentation and Classification

Our model divides the input image into $h' \times w'$ patch grid. Each patch is assigned a probability of the diseases by a small FCN (27) with features $\mathbf{x}' \in \mathbb{R}^{h' \times w' \times d}$ as the segmentation result. The small



FCN consists of two pointwise convolution layers and sigmoid activation.

Chest L-Transformer is only trained with image-level annotations. To aggregate the pixel-level evidences to an image-level decision, a smooth and convex approximation of the max and average functions (28) is chosen to build Log-Sum-Exp voting as follows:

$$p^k = \frac{1}{r} \log \frac{1}{h' \times w'} \sum_{\substack{0 < i \leq h' \\ 0 < j \leq w'}} \exp(r \times p_{ij}^k), \quad (4)$$

where p^k is the probability of the k -th class for an image and p_{ij}^k is the probability of the k -th class for the patch at location (i, j) . r is a positive hyper-parameter controlling the smoothness. Log-Sum-Exp voting will be a max function for $r \rightarrow \infty$ and be an average function for $r \rightarrow 0$. With r , the voting function assigns larger weights to the more important patches.

In chest radiographs, not all the areas are related to thoracic diseases. Although increasing r can decrease the weight of these unrelated areas in the voting process, the weight of less important areas of lesions will also be turned to a small value. The model may just focus on the more related areas of the lesions and ignore the less related ones. Moreover, a big value of r may lead to an overflow in the calculation. To ignore the unrelated areas, we propose adaptive Log-Sum-Exp voting as follows:

$$p^k = \frac{1}{r} \log \left[\frac{1}{h' \times w'} \sum_{\substack{0 < i \leq h' \\ 0 < j \leq w'}} I \times \exp(r \times p_{ij}^k) \right], \quad (5)$$

$$I = \begin{cases} 1, & \text{if } p_{ij}^k \geq t \times \max_{ij}(p_{ij}^k), \\ 0, & \text{otherwise} \end{cases}, \quad 0 < t \leq 1.$$

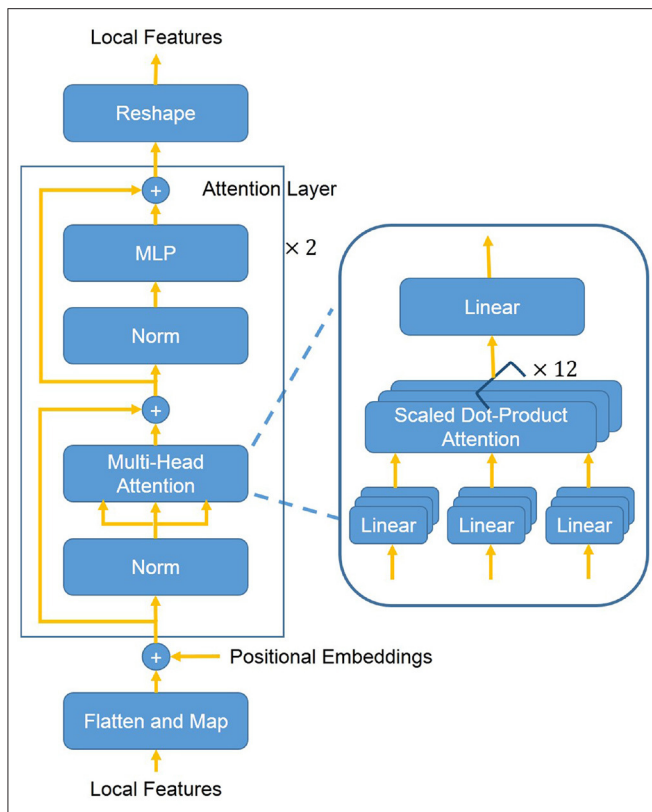


FIGURE 3 | The overall structure of the position attention block. First, the 3D local features are flattened and mapped into a latent d -dimensional embedding space. Moreover, learnable positional embeddings are added to the feature embeddings. Then added embeddings are put into a stack of $L = 2$ identical attention layers. The attention layer mainly consists of a MSA mechanism and a MLP. The MSA consists of $M = 12$ concatenated “Scaled Dot-Product Attention”. At last, the flattened features are reshaped back into 3D features.

We filter the unrelated areas with an adaptive threshold $t \times \max_{i,j}(p_{i,j}^k)$. The patches with similar evidences will have similar probabilities. With the threshold $t \times \max_{i,j}(p_{i,j}^k)$, only the patches similar to the most likely abnormal patch participate in the voting. Adaptive Log-Sum-Exp voting adapts the range of voting patches according to their class evidences automatically. t controls how similar the voting patches should be to the most likely abnormal patch. Adaptive Log-Sum-Exp voting guarantees only the patches related to diseases involve in the production of image-level probability p^k . For the images of diseases, the model will ignore the unrelated areas with this voting function. For the images of normal persons, the model will take more attention to assigning the areas, which are easier to misclassify as lesions, a correct label.

At last, we combine Log-Sum-Exp voting (including adaptive Log-Sum-Exp voting) with the α -balanced focal loss (29) as the weakly supervised loss:

$$L = \sum_k [-\alpha y^k (1 - p^k)^\gamma \log(p^k) - (1 - \alpha) (1 - y^k) (p^k)^\gamma \log(1 - p^k)], \quad (6)$$

where y^k is the binary label of the k -th class. The focal loss is initially applied in the object detection task to deal with the foreground-background imbalance. Here, we introduce it to the weakly supervised loss of Chest L-Transformer. Parameter γ is used to down-weight easy cases and focus training on hard-classified cases. Parameter α balances the importance of positive/negative cases.

3. EXPERIMENTS

3.1. Datasets

We utilize the SIIM-ACR Pneumothorax Segmentation dataset (30) to verify the proposed method. The dataset contains 12,047 frontal-view chest radiographs with pixel-level annotations, in which 2,669 chest radiographs contain lung pneumothorax and 9,378 chest radiographs have no pneumothorax. The chest radiographs were directly extracted from the DICOM file and resized as $1,024 \times 1,024$ bitmap images. Six board-certified radiologists participated in the annotation process. All annotations were then independently reviewed by 12 thoracic radiologists followed by adjudication by an additional thoracic radiologist.

3.2. Metrics

To assess the classification performance of Chest L-Transformer, we compute the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and F1 score on the testing set. Intersection over union (IoU) is computed to assess the segmentation performance.

Sensitivity and specificity are statistical measures of the performance of a binary classification test. The F1 score is used to measure the test accuracy. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

$$\text{sensitivity} = \frac{TP}{TP+FN}, \quad (7)$$

$$\text{specificity} = \frac{TN}{TN+FP}, \quad (8)$$

$$F1 = \frac{2TP}{2TP+FP+FN}, \quad (9)$$

where true positive, false positive, true negative, and false negative are denoted as TP, FP, TN, and FN, respectively.

IoU, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets. IoU can be used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \quad (10)$$

A is the predicted set of pixels and B is the ground truth.

TABLE 1 | Comparison of Chest L-Transformer with the state-of-the-art models (classification).

Model	Main method	AUC	F1	Sensitivity	Specificity
Mask R-CNN	Supervised	0.84	0.60	0.63	0.87
U-net	Supervised	0.85	0.54	0.43	0.85
ResNeXt	Classification	0.84	0.53	0.43	0.95
Chest L-Transformer	Weakly Supervised	0.81	0.57	0.67	0.79

3.3. Experimental Settings

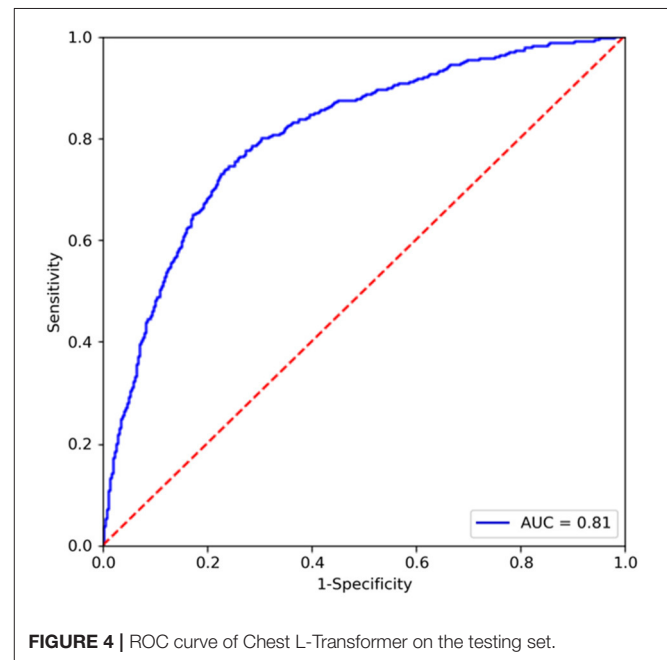
The SIIM-ACR Pneumothorax dataset is used to evaluate the classification and segmentation performance of the proposed Chest L-Transformer with 7:1:2 training:validation:test set split with no intersection. We performed an ablation study to show the effects of different blocks of Chest L-Transformer. First, we train a model with ResNeXt-50 as the backbone without position attention. The second model is Restricted ResNeXt without position attention. The third model is Restricted ResNeXt with position attention. Adaptive Log-Sum-Exp voting is utilized for the three models. The models are named as RNX50-LVT, rRNX50-LVT, and rRNX50-LVT-PA, respectively. Finally, we perform an ablation study for different versions of voting. We compare four voting functions: rRNX50-LVT-PA (adaptive Log-Sum-Exp voting), rRNX50-LV-PA (Log-Sum-Exp voting), rRNX50-AV-PA (average voting), and rRNX50-MV-PA (max voting). As shown in (9), we also train Chest L-Transformer with 400 radiographs with pixel-level annotations and the rest of the dataset with image-level annotations. The binary cross-entropy loss and Dice loss are used for the pixel-level annotated data (4).

The stochastic gradient descent (SGD) optimizer with momentum (0.9) (31) is used to train 500 epochs with an initial learning rate of 0.001. The learning rate is reduced by 0.3 when the training loss stops. We train our model with a batch size of 8 and resize the original images to 512×512 as the input. The parameters r , t , α , and γ are set to 8, 0.6, 0.6, and 2, respectively. In our experiments, we determine them with a search on 10% of the training and validation set. Chest L-Transformer is implemented in Python using PyTorch framework. Referring to the experiment in (3), we initialize the backbones with pre-trained weights.

4. RESULTS

4.1. Classification

We conduct an experiment to evaluate the performance on the classification task and compare it to the state-of-the-art segmentation models on the SIIM-ACR Pneumothorax dataset. As few weakly supervised segmentation models on chest radiographs are available, we compare Chest L-Transformer with some supervised models: Mask R-CNN (2, 32) and U-net (2, 3, 33). Chest L-Transformer is trained only with image-level annotations in a weakly supervised manner. The supervised segmentation methods are trained with pixel-level annotations in a supervised manner. We used the maximum probability of lesion areas in a radiograph as the classification probability of supervised segmentation models (2). Moreover,

**FIGURE 4** | ROC curve of Chest L-Transformer on the testing set.

Chest L-Transformer is compared with the classification model ResNeXt (22). The classification performance of Chest L-Transformer is shown in **Table 1**. Chest L-Transformer achieve an AUC of 0.81, slightly worse than supervised segmentation models (Mask R-CNN AUC = 0.84, U-net AUC = 0.85) and classification model (ResNeXt AUC = 0.84). The receiver operating characteristic (ROC) curve of Chest L-Transformer is illustrated in **Figure 4**. The results validate the classification effectiveness of Chest L-Transformer.

4.2. Segmentation

To evaluate the performance of Chest L-Transformer for segmentation, we computed IoU on the testing set, compared with Mask R-CNN (2, 32), U-net (2, 3, 33), which are trained with pixel-level annotations, and Tiramisu with CNN attention (9), which is trained with image-level annotations, shown in **Table 2**. Chest L-Transformer achieves an effective result (IoU of 0.70). It performs slightly worse than Mask R-CNN (IoU = 0.75) and U-net (IoU = 0.76) with supervised training. Moreover, Chest L-Transformer outperforms the state-of-the-art weakly supervised model (9) (Tiramisu IoU = 0.13). After added pixel-level annotations, Chest L-Transformer outperforms the state-of-the-art weakly supervised model (9) with IoU increased by

10.4%. **Figure 5** shows a few examples of the weakly supervised predictions output by Chest L-Transformer.

4.3. Ablation Study

For the ablation study, we study the effectiveness of our modified backbone, position attention block, and proposed voting function.

Table 3 shows the classification results of the ablation study of the architecture of Chest L-Transformer (backbone and position attention block) with the AUC, F1 score, sensitivity, and specificity, while segmentation results of

IoU are shown in **Table 4**. Compared with RNX50-LVT (AUC = 0.80, IoU = 0.62), the classification result of rRNX50-LVT (AUC = 0.74) is worse, but the segmentation result is significantly improved (IoU = 0.69). Although the classification performance decreases, a remarkable improvement in segmentation is achieved by applying Restricted ResNeXt to extract the local features. Compared with rRNX50-LVT, rRNX50-LVT-PA achieves improvements in both classification (AUC = 0.81) and segmentation (IoU = 0.70) with the addition of position attention by 9.5% and 1.4%, respectively. Moreover, rRNX50-LVT-PA outperforms RNX50-LVT in both classification and segmentation.

Table 5 shows the classification results of the ablation study of voting functions of Chest L-Transformer with the AUC, F1 score, sensitivity, and specificity, while segmentation results of IoU are shown in **Table 6**. Among the compared models, rRNX50-MV-PA achieves the worst AUC of 0.66 and IoU of 0.61. rRNX50-AV-PA achieves an AUC of 0.78 and an IoU of 0.66. With Log-Sum-Exp voting, rRNX50-LV-PA (AUC = 0.78, IoU = 0.68) performs better than rRNX50-AV-PA and rRNX50-AV-PA. rRNX50-LVT-PA achieved the best result (AUC = 0.81, IoU = 0.70).

5. DISCUSSION

We propose Chest L-Transformer for the weakly chest radiograph segmentation and classification. Chest L-Transformer is designed with a restricted receptive field

TABLE 2 | Comparison of Chest L-Transformer with the state-of-the-art segmentation models (segmentation).

Model	Main method	IoU
Mask R-CNN	Supervised	0.75
U-net	Supervised	0.76
Tiramisu	Weakly supervised	0.13
Tiramisu	Weakly supervised + 400 pixel-level annotated radiographs	0.67
Chest L-Transformer	Weakly supervised	0.70
Chest L-Transformer	Weakly supervised + 400 pixel-level annotated radiographs	0.74

“+ 400 pixel-level annotated radiographs” means that the model is trained with 400 radiographs with pixel-level annotations and the rest of the dataset with image-level annotations.

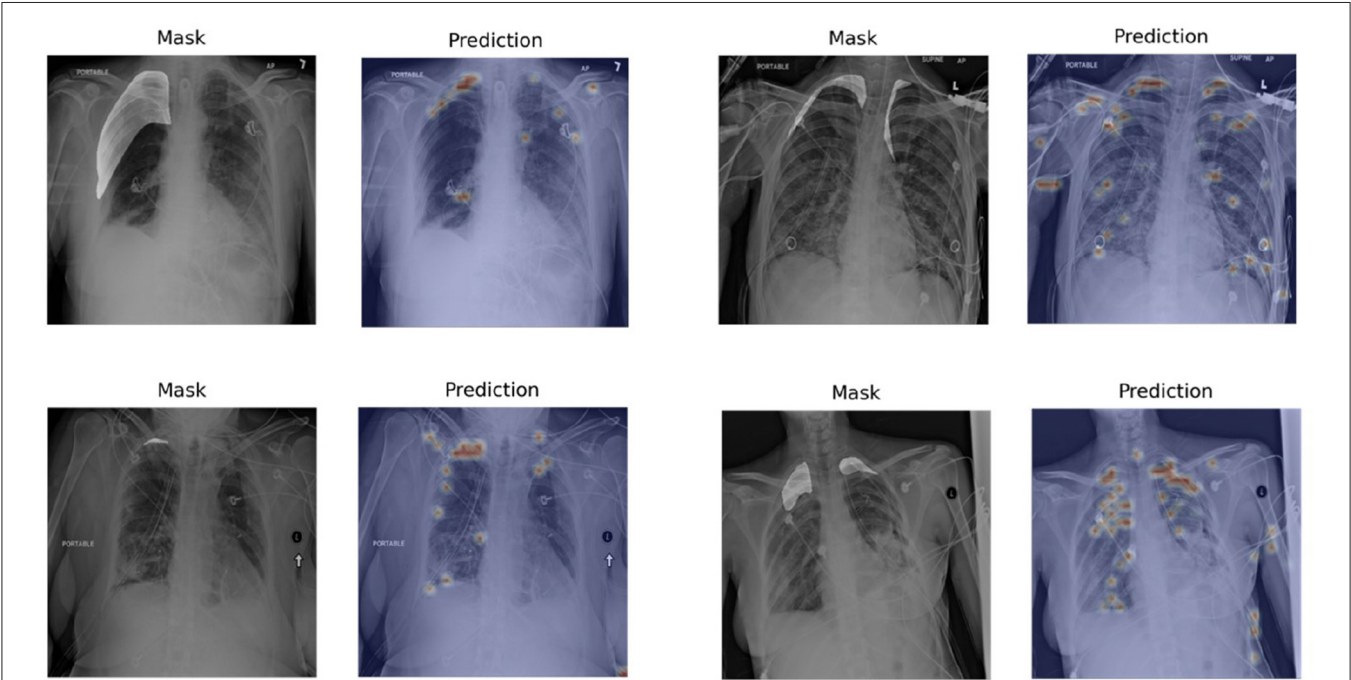


FIGURE 5 | Examples of segmentation visualization on the testing set. The visualization is generated by rendering the pixel-level outputs as heatmaps and overlapping on the original images. The left image in each pair is the original chest radiograph with highlighted masks and the right one is the segmentation visualization.

TABLE 3 | Analyzing different architectures of Chest L-Transformer (classification).

Model	AUC	F1	Sensitivity	Specificity
RNX50-LVT	0.80	0.60	0.62	0.80
rRNX50-LVT	0.74	0.41	0.35	0.90
rRNX50-LVT-PA	0.81	0.57	0.67	0.79

Numbers in bold indicate the best result among the models.

TABLE 4 | Analyzing different architectures of Chest L-Transformer (segmentation).

Model	IoU
RNX50-LVT	0.62
rRNX50-LVT	0.69
rRNX50-LVT-PA	0.70

Numbers in bold indicate the best result among the models.

backbone to analyze the contribution of each patch to the final image-level decision. Furthermore, Chest L-Transformer focuses on disease prone areas and highlights salient features useful for the diagnostic task by adding Transformer attention mechanism. Log-Sum-Exp voting and its variant are proposed to aggregate the pixel-level evidences to an image-level decision. Chest L-Transformer outperforms the state-of-the-art weakly supervised model and is comparable to the supervised segmentation and classification models (Tables 1, 2).

Extracting features from the whole image makes the pixel assignments difficult (16). The weakly supervised segmentation accuracy is depressed by the misclassification of the symmetrical positions of the lesions (9). Thus, we propose Restricted ResNeXt to extract local features with a simple modification of ResNeXt. Compared with RNX50-LVT, although the classification performance of rRNX50-LVT decreases (Table 3), it achieves remarkable improvement in segmentation (Table 4). Given the simplicity modification, the architecture of Restricted ResNeXt can be easily generalized to other deep learning models to trade a bit of classification accuracy for better weakly supervised segmentation.

The attention mechanism is an effective feature learning technique shown to be helpful in promoting the performances of image analysis models. The diseases often have special disease prone areas. But CNN attention modules treat areas of the whole image equally and fail to model the relationship between the lesions and their position (9–12). To make use of the position information, we introduce Transformer attention mechanism into our model for the position attention block. Learned positional embeddings are added to the feature embeddings to make the position attention block sensitive to certain positions. The prediction ability of Chest L-Transformer is enhanced with additional position attention. This is demonstrated in the comparison of the rRNX50-LVT and rRNX50-LVT-PA (Tables 3, 4). Moreover, the enhanced prediction of Chest L-Transformer

TABLE 5 | Analyzing different voting functions of Chest L-Transformer (classification).

Model	AUC	F1	Sensitivity	Specificity
rRNX50-AV-PA	0.78	0.53	0.55	0.85
rRNX50-MV-PA	0.66	0.38	0.40	0.79
rRNX50-LV-PA	0.78	0.51	0.49	0.88
rRNX50-LVT-PA	0.81	0.57	0.67	0.79

Numbers in bold indicate the best result among the models.

TABLE 6 | Analyzing different voting functions of Chest L-Transformer (segmentation).

Model	IoU
rRNX50-AV-PA	0.66
rRNX50-MV-PA	0.61
rRNX50-LV-PA	0.68
rRNX50-LVT-PA	0.70

Numbers in bold indicate the best result among the models.

outperforms the model with global features, RNX50-LVT (Tables 3, 4). The classification accuracy depressed by local features is offset by position attention. Chest L-Transformer can serve physicians in thoracic disease diagnosis with the effective classification and position information of findings.

To unify classification and segmentation into the same underlying prediction model, we proposed Log-Sum-Exp voting and its variant. In the ablation study, we compare the performance of different voting functions. The average voting used by the previous models achieves high accuracy in classification (Table 5) but low segmentation results (Table 6). It assigns the same weight to all patches of the image in the voting. This may lead to the misclassification of no lesion patches in the abnormal image. The model with the maximum voting is difficult to converge and achieves disappointing results in both classification and segmentation (Tables 5, 6). Log-Sum-Exp voting is proposed to take the place of the two frequently-used functions. It assigns more important patches larger weights than the less important ones. The Log-Sum-Exp voting outperforms these two functions in both classification and segmentation (Tables 5, 6). Chest radiographs contain some patches which are unrelated to the disease. To ignore the unrelated areas, we proposed adaptive Log-Sum-Exp voting, which adapts the range of voting patches with their class evidences automatically. With an adaptive threshold, Chest L-Transformer achieves further improvement in the two prediction tasks (Tables 5, 6).

Chest L-Transformer predicts rough areas of the lesions automatically. The mistakes are mainly led by therapeutic equipment, such as catheters and lines (see Figure 5). Because most of the radiographs with lesions contain therapeutic equipment, this kind of mistake can hardly be avoided with only image-level annotations. Most of the mistakes caused by

equipment would be checked out by radiologists quickly. Chest L-Transformer provides good initial areas for the pixel-level annotation and thus reduces the workload of radiologists on this work (30). Chest L-Transformer can speed up the progress of the diagnosis and treatment planning. Moreover, Chest L-Transformer will contribute to the development of medical image data for segmentation, because it reduces the cost of pixel-level annotation.

6. CONCLUSIONS

In this study, Chest L-Transformer is proposed for weakly supervised segmentation and classification on chest radiographs. The proposed backbone, Restricted ResNeXt, circumvents the misclassification of the symmetrical positions of the lesions. The position attention block embedded into Chest L-Transformer can model the position information and further provide improvement for predictions. Moreover, the Log-Sum-Exp voting and its variant aggregate the pixel-level evidences effectively. We have shown that Chest L-Transformer obtains accurate segmentation and classification predictions with image-level annotations. Therefore, Chest L-Transformer can contribute to the auxiliary diagnosis of thoracic diseases and the development of chest radiograph segmentation datasets. Moreover, the architecture of Chest L-Transformer can be easily generalized to other deep learning models for weakly supervised segmentation.

REFERENCES

- Masood S, Sharif M, Masood A, Yasmin M, Raza M. A survey on medical image segmentation. *Curr Med Imaging*. (2015) 11:3–14. doi: 10.2174/157340561101150423103441
- Wang H, Gu H, Qin P, Wang J. CheXLocNet: Automatic localization of pneumothorax in chest radiographs using deep convolutional neural networks. *PLoS One*. (2020) 15:e0242013. doi: 10.1371/journal.pone.0242013
- Tolkachev A, Sirazitdinov I, Kholiavchenko M, Mustafaev T, Ibragimov B. Deep learning for diagnosis and segmentation of pneumothorax: the results on the Kaggle competition and validation against radiologists. *IEEE J Biomed Health Inform*. (2020) 25:1660–72. doi: 10.1109/JBHI.2020.3023476
- Wang Y, Wang K, Peng X, Shi L, Sun J, Zheng S, et al. DeepSDM: boundary-aware pneumothorax segmentation in chest X-ray images. *Neurocomputing*. (2021) 454:201–11. doi: 10.1016/j.neucom.2021.05.029
- Wang H, Gu H, Qin P, Wang J. U-shaped GAN for semi-supervised learning and unsupervised domain adaptation in high resolution chest radiograph segmentation. *Front Med*. (2021) 8:782664. doi: 10.3389/fmed.2021.782664
- Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI (2017). p. 2097–106. doi: 10.1109/CVPR.2017.369
- Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, HI (2019). p. 590–7. doi: 10.1609/aaai.v33i01.3301590
- Zeng Y, Zhuge Y, Lu H, Zhang L. Joint learning of saliency detection and weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul (2019). p. 7223–33.
- Ouyang X, Xue Z, Zhan Y, Zhou XS, Wang Q, Zhou Y, et al. Weakly supervised segmentation framework with uncertainty: a study on pneumothorax segmentation in chest x-ray. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2019). p. 613–21. doi: 10.1007/978-3-030-32226-7_68
- Chikontwe P, Luna M, Kang M, Hong KS, Ahn JH, Park SH. Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening. *Med Image Anal*. (2021) 72:102105. doi: 10.1016/j.media.2021.102105
- Gadgil SU, Endo M, Wen E, Ng AY, Rajpurkar P. Chexseg: Combining expert annotations with DNN-generated saliency maps for x-ray segmentation. In: *Medical Imaging with Deep Learning*. Lübeck: PMLR (2021). p. 190–204.
- Patel G, Dolz J. Weakly supervised segmentation with cross-modality equivariant constraints. *Med Image Anal*. (2022) 2022:102374. doi: 10.1016/j.media.2022.102374
- Babenko B. *Multiple Instance Learning: Algorithms and Applications*. NCBI Google Scholar. [Preprint] (2008). Available online at: http://ailab.jbnu.ac.kr/seminar_board/pds1_files/bbabenko_re.pdf (accessed May 21, 2022).
- Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 2921–9. doi: 10.1109/CVPR.2016.319
- Ahdi Rezaeieh S, Zamani A, Bialkowski K, Abbosh A. Novel microwave torso scanner for thoracic fluid accumulation diagnosis and monitoring. *Sci Rep*. (2017) 7:1–10. doi: 10.1038/s41598-017-00436-w
- Brendel W, Bethge M. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. *arXiv[Preprint].arXiv:190400760*. (2019). doi: 10.48550/arXiv.1904.00760
- Theodorus A, Nauta M, Seifert C. Evaluating CNN interpretability on sketch classification. In: *Twelfth International Conference on Machine Vision (ICMV 2019)*. Amsterdam: International Society for Optics Photonics (2020). p. 114331. doi: 10.1117/12.2559536

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

HG and HW conceived the idea for this study. HW worked on the end-to-end implementation of the study. JW provided relevant insights on the clinical impact of the research work and handled the redaction of the manuscript. PQ managed the project. PQ and JW provided the funding for the research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Numbers 61633006 and 81872247), the Fundamental Research Funds for the Central Universities, China (Grant Number DUT21YG118), and “1+X” program for Clinical Competency enhancement-Clinical Research Incubation Project, The Second Hospital of Dalian Medical University (Grant Number 2022JCXYB07).

18. Ilanchezian I, Kobak D, Faber H, Ziemssen F, Berens P, Ayhan MS. Interpretable gender classification from retinal fundus images using BagNets. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2021). p. 477–87. doi: 10.1007/978-3-030-87199-4_45
19. Pinheiro PO, Collobert R. From image-level to pixel-level labeling with convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA (2015). p. 1713–21. doi: 10.1109/CVPR.2015.7298780
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in Neural Information Processing Systems, Vol. 30*. (Long Beach, CA: Curran Associates, Inc.) (2017). p. 1–11.
21. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. (2020). doi: 10.48550/arXiv.2010.11929
22. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI (2017). p. 1492–500. doi: 10.1109/CVPR.2017.634
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
24. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA (2015). p. 1–9. doi: 10.1109/CVPR.2015.7298594
25. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: *International Conference on Machine Learning*. Sydney: PMLR (2017). p. 1243–52.
26. Ba JL, Kiros JR, Hinton GE. Layer normalization. *arXiv[Preprint].arXiv:1607.06450*. (2016). doi: 10.48550/arXiv.1607.06450
27. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA (2015). p. 3431–40. doi: 10.1109/CVPR.2015.7298965
28. Boyd S, Boyd SP, Vandenberghe L. *Convex Optimization*. Cambridge, MA: Cambridge University Press (2004). doi: 10.1017/CBO9780511804441
29. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. (2018) 42:318–27. doi: 10.1109/TPAMI.2018.2858826
30. Filice RW, Stein A, Wu CC, Arteaga VA, Borstelmann S, Gaddikeri R, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. *J Digit Imaging*. (2020) 33:490. doi: 10.1007/s10278-019-00299-9
31. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: *International Conference on Machine Learning*. Atlanta: PMLR (2013). p. 1139–47.
32. He K, Gkioxari G, Dollár P, Girshick R. Mask r-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice (2017). p. 2961–9. doi: 10.1109/ICCV.2017.322
33. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gu, Wang, Qin and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Healthy Organs Uptake on Baseline ^{18}F -FDG PET/CT as an Alternative to Total Metabolic Tumor Volume to Predict Event-Free Survival in Classical Hodgkin's Lymphoma

David Morland^{1,2,3,4**†}, Elizabeth Katherine Anna Triumbari^{1†}, Elena Maiolo⁵, Annarosa Cuccaro⁶, Giorgio Treglia^{7,8,9}, Stefan Hohaus^{5,10} and Salvatore Annunziata¹

¹ Unità di Medicina Nucleare, TracerGLab, Dipartimento di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Roma, Italy, ² Service de Médecine Nucléaire, Institut Godinot, Reims, France, ³ Laboratoire de Biophysique, UFR de Médecine, Université de Reims Champagne-Ardenne, Reims, France, ⁴ CReSTIC (Centre de Recherche en Sciences et Technologies de l'Information et de la Communication), EA 3804, Université de Reims Champagne-Ardenne, Reims, France, ⁵ Unità di Ematologia, Dipartimento di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Roma, Italy, ⁶ Unità di Ematologia, ASL Toscana N/O Spedali Riuniti Livorno, Livorno, Italy, ⁷ Clinic of Nuclear Medicine, Imaging Institute of Southern Switzerland, Ente Ospedaliero Cantonale, Bellinzona, Switzerland, ⁸ Faculty of Biomedical Sciences, Università della Svizzera italiana, Lugano, Switzerland, ⁹ Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland, ¹⁰ Section of Hematology, Department of Radiological Sciences, Radiotherapy and Hematology, Università Cattolica del Sacro Cuore, Roma, Italy

OPEN ACCESS

Edited by:

Ronald Boellaard,
AmsterdamUMC, Netherlands

Reviewed by:

Domenico Albano,
University of Brescia, Italy
Francesco Dondi,
Università degli Studi di Brescia, Italy

*Correspondence:

David Morland
david.morland@reims.unicancer.fr

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

Received: 06 April 2022

Accepted: 31 May 2022

Published: 22 June 2022

Citation:

Morland D, Triumbari EKA, Maiolo E, Cuccaro A, Treglia G, Hohaus S and Annunziata S (2022) Healthy Organs Uptake on Baseline ^{18}F -FDG PET/CT as an Alternative to Total Metabolic Tumor Volume to Predict Event-Free Survival in Classical Hodgkin's Lymphoma. *Front. Med.* 9:913866. doi: 10.3389/fmed.2022.913866

Purpose: Healthy organs uptake, including cerebellar and liver SUVs have been reported to be inversely correlated to total metabolic tumor volume (TMTV), a controversial predictor of event-free survival (EFS) in classical Hodgkin's Lymphoma (cHL). The objective of this study was to estimate TMTV by using healthy organs SUV measurements and assess the performance of this new index (UF, Uptake Formula) to predict EFS in cHL.

Methods: Patients with cHL were retrospectively included. SUV values and TMTV derived from baseline ^{18}F -FDG PET/CT were harmonized using ComBat algorithm across PET/CT systems. UF was estimated using ANOVA analysis. Optimal thresholds of TMTV and UF were calculated and tested using Cox models.

Results: 163 patients were included. Optimal UF model of TMTV included age, lymphoma maximum SUVmax, hepatic SUVmean and cerebellar SUVmax (R^2 14.0% - $p < 0.001$). $\text{UF} > 236.8$ was a significant predictor of EFS (HR: 2.458 [1.201–5.030], $p = 0.01$) and was not significantly different from $\text{TMTV} > 271.0$ (HR: 2.761 [1.183–5.140], $p = 0.001$). $\text{UF} > 236.8$ remained significant in a bivariate model including IPS score ($p = 0.02$) and determined two populations with different EFS (63.7 vs. 84.9%, $p = 0.01$).

Conclusion: The Uptake Formula, a new index including healthy organ SUV values, shows similar performance to TMTV in predicting EFS in Hodgkin's Lymphoma. Validation cohorts will be needed to confirm this new prognostic parameter.

Keywords: cerebellum, liver, metabolic tumor volume, Hodgkin's Lymphoma, prognosis, prediction

INTRODUCTION

Hodgkin's Lymphoma (HL) affects young adults and represents about 2.3 cases per 100,000 people per year, with an associated mortality of 0.4 cases per 100,000 per year (1). Despite treatment, about 20% of HL patients still relapse (1). ^{18}F -Fluorodeoxyglucose (^{18}F -FDG) Positron Emission Tomography (PET) coupled with Computed Tomography (CT) plays a central role in HL patients management, whether in staging or response assessment settings (1).

PET-derived parameters, volumetric ones above all, have been proposed to refine prognosis prediction of HL (2). The role of Total Metabolic Tumor Volume (TMTV) is debated in Hodgkin's Lymphoma. It has been reported as a negative prognostic factor in early-stage HL treated with ABVD (Adriamycin, Bleomycin, Vinblastine, Dacarbazine) regimen (3–5) and HIV-associated HL (6). However, some publications reported no association between TMTV and Progression-Free Survival (PFS) in advanced-stage HL when treated with escalated BEACOPP (7). Furthermore, TMTV threshold varies from one study to another (3–5, 7, 8).

The drawbacks of TMTV calculation (results depending on the segmentation method (9), time required for delineation (10), difficulty in evaluating bone involvement) led to ponder other prognostic markers, such as healthy organ ^{18}F -FDG uptake. In 2010, Hanaoka et al. (11) reported an inverse correlation between cerebellar uptake and total lesion glycolysis (TLG) in a population with aggressive lymphoma. The mechanism underlying this phenomenon is poorly understood but could correspond to a metabolic theft of ^{18}F -FDG by the tumor mass. Because TLG is correlated with TMTV, Godard et al. (12) speculated that cerebellar metabolism might have a prognostic value and suggested to normalize cerebellar ^{18}F -FDG uptake to hepatic ^{18}F -FDG uptake to account for differences between PET/CT systems. This index has been shown to be a prognostic parameter for PFS prediction in diffuse large-B-cell lymphoma (10) and follicular lymphoma (12). Normalization to liver was not optimal: liver ^{18}F -FDG uptake was also negatively correlated with TMTV as was cerebellar ^{18}F -FDG uptake ($r = -0.34$ and $r = -0.42$, respectively) (10). These two healthy organs could thus prove useful in predicting prognosis.

The objective of this study was to model TMTV by integrating healthy organ uptake data in classical HL (cHL). The resulting estimate was then tested for EFS prediction.

MATERIALS AND METHODS

Study Population

This study was approved by the Ethical Committee of Fondazione Policlinico Universitario A. Gemelli IRCCS (study code: 3834). All included subjects signed an informed consent form. All procedures performed were in accordance with the ethical standards defined by the 1964 Helsinki Declaration and its later amendments.

All patients with HL referred to our Institution for their baseline ^{18}F -FDG PET/CT between September 2010 and January 2020 were retrospectively screened. Inclusion criteria were as follows: histologically proven cHL; baseline PET/CT performed

within 4 weeks prior to treatment. Exclusion criteria were: Recent history of other cancer <1 year; Nodular lymphocyte-predominant Hodgkin lymphoma histology [slow growing LH subtype with completely different prognosis (13)]; any factor interfering with measurement of cerebellar uptake, liver uptake or TMTV: cerebellum not fully included in field of view, movement artifacts, extensive surgically resected disease before staging PET/CT, diffuse lymphomatous involvement of liver or brain lymphoma; nonobservance of a fasting period of at least 6h before ^{18}F -FDG administration; glycemia > 2.0 g/l; no follow-up available after staging PET/CT; nonstandard treatment regimen.

The following clinicobiological data were collected: date of birth, date of diagnosis, date of last observation, HL subtype, Eastern Cooperative Oncology Group (ECOG) performance status, International Prognostic Score (IPS) items (age; sex; Ann Arbor stage; serum albumin levels, white blood cell count, lymphocyte count, hemoglobinemia at baseline), first-line treatment, event-free survival (EFS: time interval between date of diagnosis and the event of progression, recurrence, change of therapy or death) and overall survival (OS: time interval between date of diagnosis and death). Imaging data collected included: date of staging PET/CT, PET/CT system, administered ^{18}F -FDG activity, glycemia levels.

^{18}F -FDG PET/CT Acquisition and Measurements

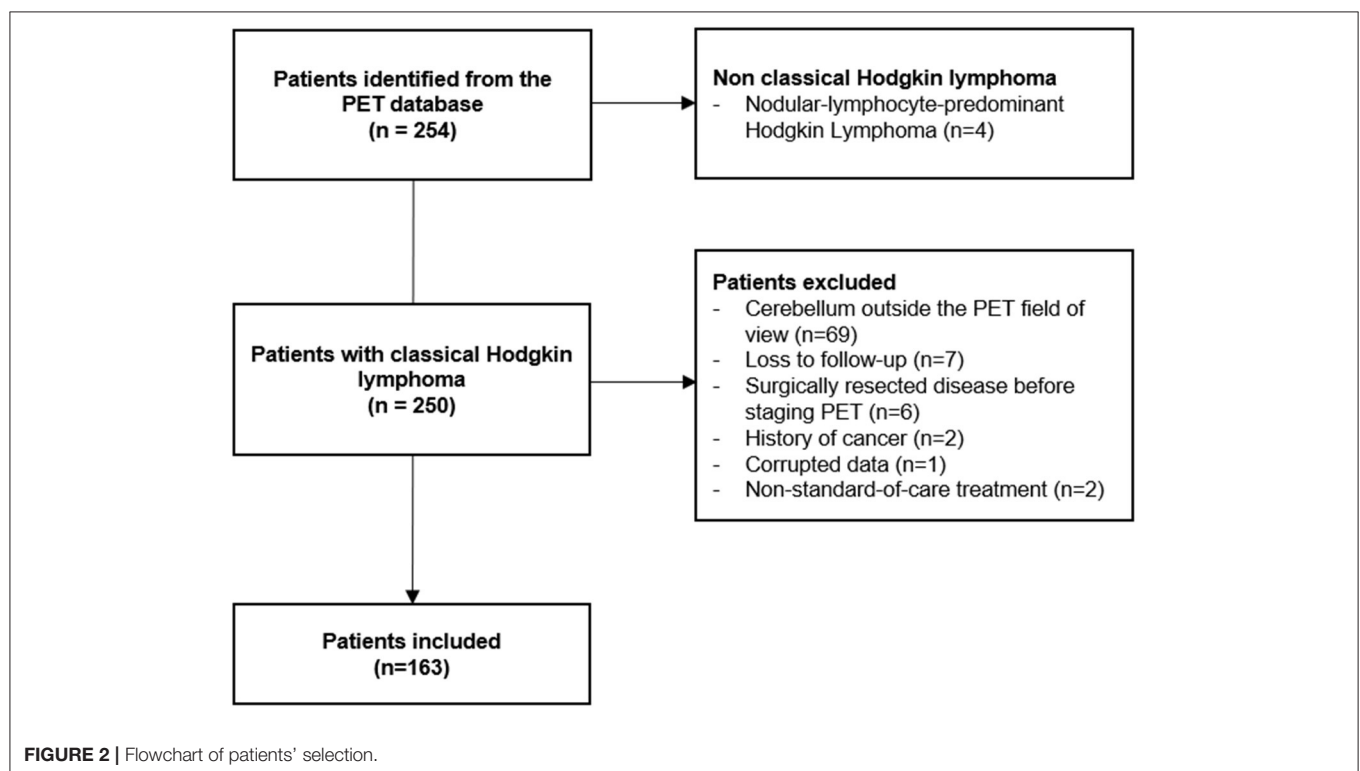
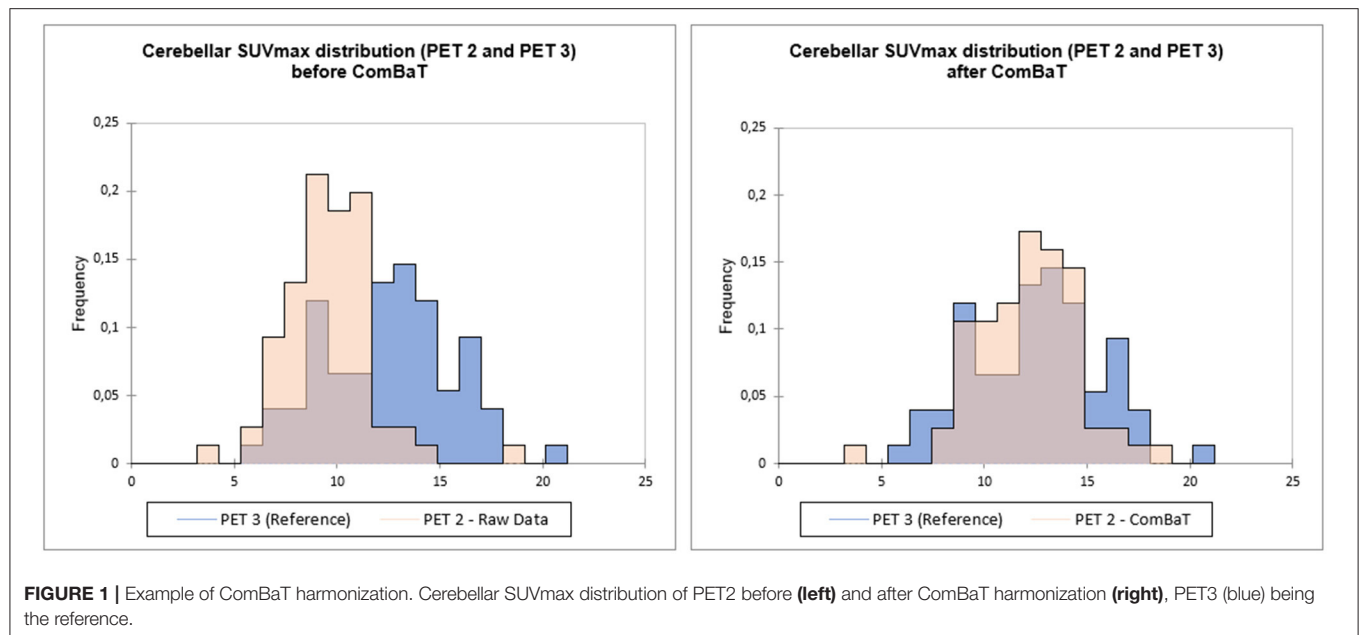
After verification of patients' blood glucose levels, baseline ^{18}F -FDG PET/CT was performed at 60 ± 10 min after intravenous injection of mean 236.34 MBq (range 137–366) of ^{18}F -FDG. Due to the long-time span of the inclusion period, images were acquired using 3 different PET/CT integrated systems denoted as PET1, PET2 and PET3 in chronological order.

PET1 corresponded to a Gemini Dual GS PET/CT scanner (Philips Healthcare): images were acquired in three-dimensional mode with an acquisition time of 3 min per bed position and reconstructed on a 128×128 matrix using Row-Action Maximum Likelihood Algorithm (RAMLA, 2 iterations, blob size of 2.5 pixels, voxel size: $4 \times 6 \times 20 \text{ mm}^3$) without Point Spread Function (PSF) or Time of Flight (TOF).

PET2 corresponded to a Gemini GXL PET/CT scanner (Philips Healthcare): images were acquired with a 3 min per bed position acquisition time and reconstructed on a 128×128 matrix using 3D-Line Of Response RAMLA (3D-LOR-RAMLA, 3 iterations, 33 subsets, voxel size: $4 \times 4 \times 4 \text{ mm}^3$) without PSF or TOF.

PET3 corresponded to a Biograph mCT PET/CT scanner (Siemens Healthineers): images were acquired in 2.5 min per bed position and reconstructed on a 400×400 matrix using 3D Ordered Subset Expectation Maximization algorithm (3D-OSEM, 2 iterations, 21 subsets, voxel size: $3.2 \times 3.2 \times 5 \text{ mm}^3$) with PSF and TOF. A gaussian filter was also applied (3D isotropic Gaussian kernel of 2 mm full width at half-maximum).

CT acquisition protocol was the same for the three machines: 120 kV, 50 mAs, reconstruction on a 512×512 matrix with a voxel size of $0.97 \times 0.97 \times 3 \text{ mm}^3$. PET/CT images were acquired at least from skull base to proximal thighs.



PET/CT were displayed on dedicated interpretation consoles (Syngo.via for SUV measurements and version 7.0.5 of MIM Encore Software for volumetric parameters). The following data were collected: (1) cerebellar SUV_{max} , (2) hepatic SUV_{mean} , (3) TMTV, (4) lymphoma maximal SUV_{max} (lesion SUV_{max} , henceforth denoted as L).

After a first visual check using a rainbow 10 point-color scale, enclosing region of interest (ROI) were drawn on areas with highest visual uptake, excluding any voxel of the neighboring brain hemispheres. The highest SUV_{max} of all these ROI corresponded to cerebellar SUV_{max} . A default spherical ROI (2-cm diameter) was positioned on the right liver to measure its SUV_{mean} . The lymphoma maximal SUV_{max} was determined

TABLE 1 | Patients' characteristics.

	Included patients (n=163)
Hodgkin's Lymphoma subtype	
Nodular sclerosis	125 (76.7%)
Mixed cellularity	9 (5.5%)
Lymphocyte-rich	2 (1.2 %)
Lymphocyte-depleted	7 (4.3%)
Not specified	23 (14.1%)
ECOG performance status	
0	65 (39.9%)
1	41 (25.2%)
2	7 (4.3%)
3	2 (1.2%)
Not available	48 (29.4%)
International Prognosis Score items	
Age \geq 45 years	51 (31.3%)
Male sex	77 (47.2%)
Ann Arbor stage IV	56 (34.4%)
Serum albumin < 4 g/dl	78 (47.9%)
White Cell count \geq 15,000/mm ³	23 (14.1%)
Lymphocyte count < 600/mm ³	9 (5.5%)
Hemoglobin < 10.5 g/dl	30 (18.4%)
First-line chemotherapy treatment	
ABVD	127 (77.9%)
ABVD + BEACOPP	18 (11.0%)
MBVD	10 (6.1%)
BEACOPP	5 (3.1%)
Not available	3 (1.8%)
Radiotherapy	134 (82.2%)
Number of EFS events	40 (24.5%)
Number of OS events	9 (5.5%)
PET/CT systems	
PET1	9 (5.5%)
PET2	70 (42.9%)
PET3	84 (51.5%)

ABVD, Adriamycin, Bleomycin, Vinblastine, Dacarbazine; MBVD, nonpegylated liposomal doxorubicin (Myocet), Bleomycin, Vinblastine, Dacarbazine; BEACOPP, Bleomycin, Etoposide, Adriamycin, Cyclophosphamide, vincristine (Oncovin), Procarbazine, Prednisolone.

manually by an experienced nuclear medicine physician and was defined as the SUV_{max} of the hottest nodal lesion. TMTV was measured using a PET segmentation tool (LesionID, version 7.0.5 of MIM Encore Software Inc., Cleveland, OH). As previously described (14), the software proceeds in 4 steps: first, a PET Response Criteria in Solid Tumors (PERCIST)-based background threshold (liver) thresholding was applied; second, a VOI encompassing all detected lesions (above the threshold) was automatically drawn. The detected lesions could thus include bone and spleen, depending on their uptake; third, a second thresholding at 41% of the SUV_{max} of the detected lesions was applied to determine lesions' boundaries; finally, physicians were required to reject false positive lesions before computation of TMTV.

TABLE 2 | ANOVA analysis and derived model for TMTV prediction.

	Coefficient (95%CI)	p-value
Constant	382.150 [181.543, 582.757]	<0.001**
Age (A)	-2.449 [-4.675, -0.223]	0.031*
Lesion SUV _{max} (L)	9.145 [4.263-14.026]	<0.001**
Cerebellar SUV _{max} (C)	-13.674 [-26.652, -0.695]	0.039*
Hepatic SUV _{mean} (H)	-20.008 [-42.541, 2.526]	0.081
Glycemia (G)	Rejected	Rejected

* $p < 0.05$; ** $p < 0.001$.

ComBat Harmonization

The "batch effect" introduced by the use of 3 different PET/CT machines was compensated using a validated statistical harmonization method (15) implemented on RStudio (16). ComBat was applied on log transformed data, followed by exponentiation to improve the algorithm effectiveness (15), and ensure positive values. TMTV, hepatic SUV_{mean} (H), cerebellar SUV_{max} (C) and lesion SUV_{max} (L) were harmonized. Reference batch was set to PET3. An example is presented in **Figure 1**.

Statistical Analyses

Statistical Software

If not stated otherwise, the following statistical analyses were performed on Xlstat (2020, Addinsoft, New York, USA). p -value threshold for significance was set at 0.05.

TMTV Modelization

An ANOVA analysis was used to model the TMTV from the following 5 clinicobiological data: age, blood glucose, H, C, L. The selection of the optimal model was based on the R^2 value with a number of allowed parameters ranging from 2 to 5. The resulting formula is hereafter referred to as the Uptake Formula (UF). Significance was assessed by F-statistic.

Analysis

Optimal cut-off for TMTV, UF and IPS were calculated using CutoffFinder (17) with respect to EFS using the survival analysis method. This method fits Cox proportional hazard models to the dichotomized variable and the survival variable: optimal cutoff is defined as the point with the most significant (log-rank test) split. Missing IPS values were replaced by mean-values. Derived Hazard Ratios were compared based on their 95% Confidence Intervals. Bivariate analysis was performed using TMTV+IPS and UF+IPS. TMTV and UF were not combined for collinearity issues. Survival curves were drawn for UF and TMTV.

RESULTS

Two-hundred and fifty-four patients were retrieved from the database (**Figure 2**). Among them, 4 had a nodular-lymphocyte-predominant HL (1.6%) and were excluded. Among the remaining 250 patients, 77 were excluded due to the impossibility of measuring the needed parameters (cerebellum outside the field of view, surgically resected disease, corrupted data). Seven

TABLE 3 | Univariate and bivariate analyses for Event-Free Survival (EFS) based on Cox model.

	Hazard Ratios (95% CI)	p-value
UF > 236.8	2.458 [1.201–5.030]	0.014*
TMTV > 271.0	2.761 [1.183–5.140]	0.001*
IPS >= 2	2.050 [1.023, 4.106]	0.043*
UF + IPS Model	2.320 [1.131–4.760]	0.022*
- UF		
- IPS	1.903 [0.947–3.822]	0.071
TMTV + IPS Model	2.507 [1.333–4.715]	0.004*
- TMTV		
- IPS	1.732 [0.854–3.513]	0.128

* $p < 0.05$.

patients were lost at follow-up just after the baseline PET. Two patients had a history of recent cancer. Two patients were treated with non-standard-of-care chemotherapy. Finally, 163 patients were included in the analysis. Their main characteristics are presented in **Table 1**. The median follow-up was 51 months (range 3–127 months). Overall, 9 patients died during follow-up (5.5%) and 40 EFS events were recorded (24.5%).

TMTV Modelization

ANOVA analysis selected 1 constant and 4 parameters to model TMTV (R^2 : 14.0%— $p < 0.001$): age, lesion SUV_{max} (L), cerebellar SUV_{max} (C), and hepatic SUV_{mean} (H). Coefficient values are presented in **Table 2**. Lesion SUV_{max} was a positive coefficient while healthy organs SUV values corresponded to negative coefficients. Glycemia was excluded.

The resultant UF was: $TMTV = 382.150 - 2.449 \text{ Age} + 9.145 L - 13.674 C - 20.008 H$.

Event-Free Survival Analysis

Optimal threshold for UF, TMTV and IPS were 236.8, 271.0, and 2.0, respectively (**Table 3**). The three parameters were significant predictors of EFS with HR between 2.050 and 2.761. When pooled with IPS, both UF and TMTV remained significant predictors of EFS ($p = 0.022$ and $p = 0.004$, respectively).

EFS survival curves based on UF are presented in **Figure 3**.

DISCUSSION

Healthy Organs ^{18}F -FDG Uptake Values and Derived Formula

No significant differences in hazard ratio were found between UF and TMTV. UF remained significant at bivariate analysis when adding IPS score with an HR of 2.3 (derived EFS of 84.9 vs. 63.7%).

The metabolic theft hypothesis for which FDG-avid tumor masses would deprive healthy organs of ^{18}F -FDG was investigated in two papers on diffuse large-cell B-cell and follicular lymphomas (10, 12). Cerebellar and hepatic ^{18}F -FDG uptake values were reported to be inversely correlated to TMTV. The optimal model we found to estimate TMTV is coherent

with these findings: both liver and cerebellum coefficients are negative, meaning an inverse correlation with TMTV estimate. The addition of these two parameters contributed to the significance of the model, which, however, solves only part of the variability in TMTV (R^2 of 14.0%), explaining the slight difference between TMTV and UF optimal thresholds.

Besides healthy organ uptake, age and tumoral ^{18}F -FDG uptake were also selected in the model. Apart from the study by Angelopoulou et al. (18), who reported that SUV_{max} was predictive of PFS in a study of 162 patients with HL, other studies reported no significance (3). The lack of harmonization is probably one of the overriding factors for those results.

Both SUV_{mean} and SUV_{max} were used. SUV_{mean} , which is less sensitive to noise, was preferred for the measurement of hepatic uptake to promote reproducibility: as the liver is a homogeneous organ, variations in the positioning of the ROI have little impact on SUV_{mean} measurement as already noted in a previously published study (12). The cerebellum has on the other hand a heterogeneous uptake mainly concentrated in the gray matter. The measurement of SUV_{mean} would have required a precise contouring of this structure that could have introduced contouring bias (19). SUV_{max} , which relies on only a single pixel was then chosen to ensure reproducibility, as already demonstrated in another study (12).

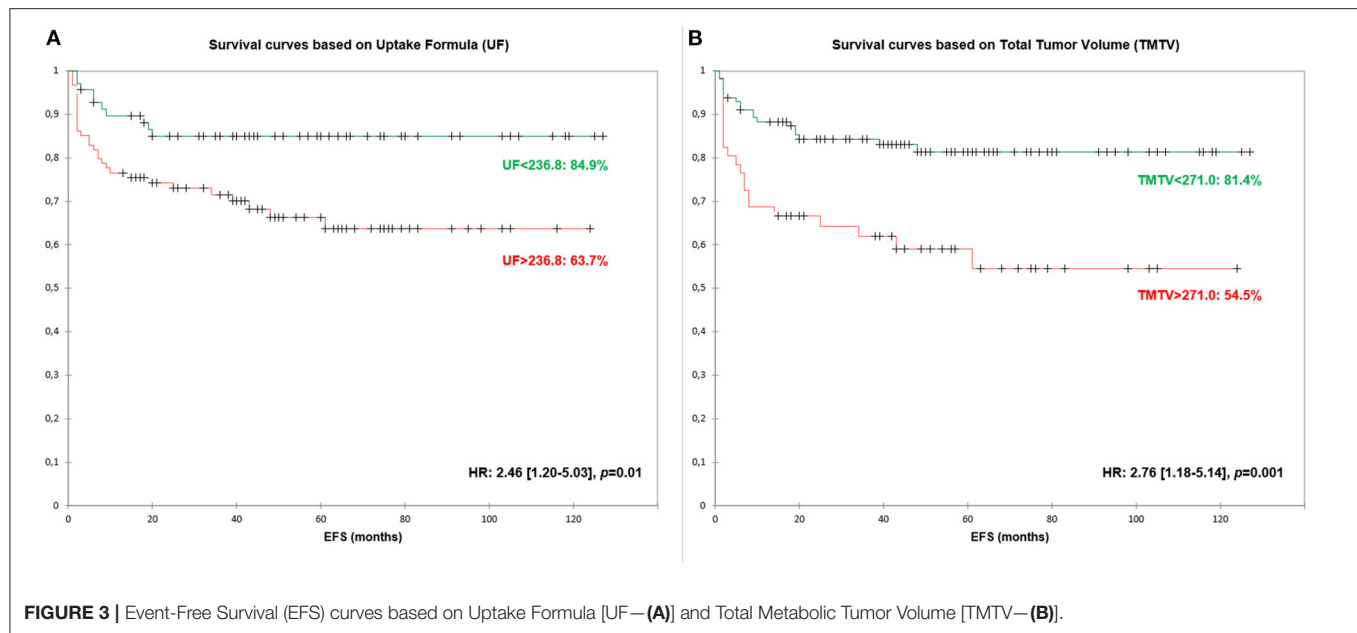
UF thus had the advantage of speed of calculation, requiring only 3 measurements of SUV values vs. several minutes for TMTV [6.2 min on average, ranging from 0.4 to 21.6 min in the study by Ilyas et al. (20)].

Metabolic Tumor Volume as Prognostic Factor

Some studies have already investigated the prognostic value of TMTV in HL (3–7, 18, 21–24) with conflicting results, presumably related to the difference in patients' therapeutic management and the low number of events encountered in HL (25). Most studies reported a significant ability of TMTV to predict PFS (3, 5, 22), with an overall HR calculated by Frood et al. (25) of 2.13 (CI 95% 1.53–2.96). These results were however associated with high levels of heterogeneity. Segmentation methods and cut-offs varied greatly [TMTV cut-off from 89 ml (22) to 225 ml (21)] and no test-retest of these thresholds were performed. Moreover, the cut-off determination method is another aspect that needs to be addressed. As pointed out by Schöder and Moskowitz (26), most studies rely on Receiver Operating Curves to determine variables' cut-offs, neglecting censored data and leading to inappropriate results (27–29). To overcome this issue, a survival-based cut-off method was used in this study. Even if our TMTV threshold (271 cm^3) was higher than the previously mentioned ones, we found a similar HR to previously reported ones.

Harmonization

Pooling images from different scanners is not simple, as many quantitative biomarkers (SUV, TMTV) are sensitive to a scanner effect (15, 30). Although procedures were proposed to harmonize image quality (31), a dedicated reconstruction requiring raw data storage would be needed and



would mostly be not feasible in a retrospective setting (15). To counteract this batch effect, the ComBat harmonization method, initially introduced in the field of genomics (32), has been proposed (15) and used (33). ComBat is a data-driven method that does not require phantom acquisitions to estimate the scanner effect but requires data from the different sites with sufficient sample size. It always theoretically improves the alignment of the mean and standard deviation of the distributions, given the criterion optimized by the method (15).

In our study, we chose to harmonize SUV and TMTV values using PET3 scanner as a reference for several reasons: it is the most recent machine among the three, corresponding to currently available technology in PET scanners; furthermore, the majority of patients was scanned on the PET3, so TMTV and SUV values were not modified for the majority of patients. Harmonization allowed us to study cerebellar and hepatic uptake independently, without having to use a ratio for normalization purposes. The use of a ratio disturbed the correlation between healthy organ and TMTV in the article by Morland et al. (10), but was still necessary to ensure good inter-machine agreement. The ComBat harmonization allowed us to overcome this problem.

Limitations

Some limitations of this study can be pointed out. This exploratory retrospective study, lacks an external and/or prospective validation cohort which would be desirable to confirm our findings. Moreover, the cohort is heterogeneous due to different stages at diagnosis that may have interfered with the performance of the parameters tested. Nevertheless, large retrospective studies are commonly designed to evaluate prognostic parameters in lymphoma, needing

long-term follow-up to register a significant number of adverse events.

CONCLUSION

The Uptake Formula, a new index including healthy organ uptake values, shows similar performance to TMTV in predicting EFS in Hodgkin's Lymphoma. Validation cohorts will be needed to confirm this new prognostic parameter.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Committee of Fondazione Policlinico Universitario A. Gemelli IRCCS (Study Code: 3834). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

SA and ET were supported by the Italian Ministry of Health (GR-2019-12370372).

REFERENCES

- Eichenauer DA, Aleman BMP, André M, Federico M, Hutchings M, Illidge T, et al. Hodgkin lymphoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* (2018) 29:iv19–29. doi: 10.1093/annonc/mdy080
- Bertagna F, Giubbini R, Albano D. Evidence-Based PET for Haematological Tumours. In: Treglia G, Giovanella L, éditeurs. *Evidence-based Positron Emission Tomography*. Cham: Springer International Publishing (2020). p. 79–88. Available online at: http://link.springer.com/10.1007/978-3-030-47701-1_9 (cité 17 mai 2022)
- Cottreau AS, Versari A, Loft A, Casasnovas O, Bellei M, Ricci R, et al. Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. *Blood.* (2018) 131:1456–63. doi: 10.1182/blood-2017-07-795476
- Song MK, Chung JS, Lee JJ, Jeong SY, Lee SM, Hong JS, et al. Metabolic tumor volume by positron emission tomography/computed tomography as a clinical parameter to determine therapeutic modality for early stage Hodgkin's lymphoma. *Cancer Sci.* (2013) 104:1656–61. doi: 10.1111/cas.12282
- Akhari M, Milgrom SA, Pinnix CC, Reddy JP, Dong W, Smith GL, et al. Reclassifying patients with early-stage Hodgkin lymphoma based on functional radiographic markers at presentation. *Blood.* (2018) 131:84–94. doi: 10.1182/blood-2017-04-773838
- Louarn N, Galicier L, Bertinchamp R, Lussato D, Montravers F, Oksenhendler E, et al. First Extensive Analysis of ¹⁸F-Labelled Fluorodeoxyglucose Positron Emission Tomography-Computed Tomography in a Large Cohort of Patients With HIV-Associated Hodgkin Lymphoma: Baseline Total Metabolic Tumor Volume Affects Prognosis. *JCO.* (2022) JCO.21.01228. doi: 10.1200/JCO.21.01228
- Mettler J, Müller H, Voltin CA, Baues C, Klaeser B, Moccia A, et al. Metabolic Tumor Volume for Response Prediction in Advanced-Stage Hodgkin Lymphoma. *J Nucl Med.* (2019) 60:207–11. doi: 10.2967/jnumed.118.210047
- Pinochet P, Texte E, Stamatoullas-Bastard A, Vera P, Mihailescu SD, Becker S. Prognostic value of baseline metabolic tumour volume in advanced-stage Hodgkin's lymphoma. *Sci Rep.* (2021) 11:23195. doi: 10.1038/s41598-021-02734-w
- Driessen J, Zwezerijnen GJ, Schöder H, Drees EE, Kersten MJ, Moskowitz AJ, et al. The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in (18)F-FDG PET scans of patients with classical Hodgkin lymphoma. *J Nucl Med.* (2022) 6:jnumed.121.263067doi: 10.2967/jnumed.121.263067
- Morland D, Zizi G, Godard F, Gauchy AC, Durot C, Hoeffel C, et al. 18F-FDG cerebellum/liver index as a prognostic factor for progression-free survival in diffuse large B-cell lymphoma. *Ann Nucl Med.* (2021) 35:785–93. doi: 10.1007/s12149-021-01609-4
- Hanaoka K, Hosono M, Shimono T, Usami K, Komeya Y, Tsuchiya N, et al. Decreased brain FDG uptake in patients with extensive non-Hodgkin's lymphoma lesions. *Ann Nucl Med.* (2010) 24:707–11. doi: 10.1007/s12149-010-0415-5
- Godard F, Durot E, Durot C, Hoeffel C, Delmer A, Morland D. Cerebellum/liver index in pretherapeutic 18F-FDG PET/CT as a predictive marker of progression-free survival in follicular lymphoma treated by immunochemotherapy and rituximab maintenance. *Medicine.* (2022) 101:e28791. doi: 10.1097/MD.00000000000028791
- Shankar A, Hall GW, McKay P, Gallop-Evans E, Fielding P, Collins GP. Management of children and adults with all stages of nodular lymphocyte predominant Hodgkin lymphoma—All St AGE s : A consensus-based position paper from the Hodgkin lymphoma subgroup of the UK National Cancer Research Institute. *Br J Haematol.* (2022) bjh.18169. doi: 10.1111/bjh.18169
- Dean EA, Mhaskar RS, Lu H, Mousa MS, Krivenko GS, Lazaryan A, et al. High metabolic tumor volume is associated with decreased efficacy of axicabtagene ciloleucel in large B-cell lymphoma. *Blood Adv.* (2020) 4:3268–76. doi: 10.1182/bloodadvances.2020001900
- Orlhac F, Eertink JJ, Cottreau AS, Zijlstra JM, Thieblemont C, Meignan M, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med.* (2022) 63:172–9. doi: 10.2967/jnumed.121.262464
- RStudio Team (2022). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. Available online at: <http://www.rstudio.com/>
- Budczies J, Klauschen F, Sinn BV, Gyorffy B, Schmitt WD, Darb-Esfahani S, et al. Cutoff Finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. van Diest P, éditeur. *PLoS ONE.* (2012) 7:e51862. doi: 10.1371/journal.pone.0051862
- Angelopoulou MK, Mosa E, Pangalis GA, Rondogianni P, Chatziioannou S, Prassopoulos V, et al. The Significance of PET/CT in the Initial Staging of Hodgkin Lymphoma: Experience Outside Clinical Trials. *Anticancer Res.* (2017) 37:5727–36. doi: 10.21873/anticancer.12011
- Berghmans T, Dusart M, Paesmans M, Hossein-Foucher C, Buvat I, Castaigne C, et al. Primary Tumor Standardized Uptake Value (SUVmax) Measured on Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) is of Prognostic Value for Survival in Non-small Cell Lung Cancer (NSCLC): a Systematic Review and Meta-Analysis (MA) by the European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project. *J. Thor. Oncol.* (2008) 3:6–12. doi: 10.1097/JTO.0b013e31815e6d6b
- Ilyas H, Mikhael NG, Dunn JT, Rahman F, Möller H, Smith D, et al. Defining the optimal method for measuring baseline metabolic tumour volume in diffuse large B cell lymphoma. *Eur J Nucl Med Mol Imaging.* (2018) 45:1142–54. doi: 10.1007/s00259-018-3953-z
- Kanoun S, Tal I, Berriolo-Riedinger A, Rossi C, Riedinger JM, Vigneaude JM, et al. Influence of software tool and methodological aspects of total metabolic tumor volume calculation on baseline [18F]FDG PET to predict survival in Hodgkin Lymphoma. Chen CT, éditeur. *PLoS ONE.* (2015) 10:e0140830. doi: 10.1371/journal.pone.0140830
- Albano D, Mazzeletti A, Spallino M, Muzi C, Zilioli VR, Pagani C, et al. Prognostic role of baseline 18F-FDG PET/CT metabolic parameters in elderly HL: a two-center experience in 123 patients. *Ann Hematol.* (2020) 99:1321–30. doi: 10.1007/s00277-020-04039-w
- Lue KH, Wu YF, Liu SH, Hsieh TC, Chuang KS, Lin HH, et al. Prognostic Value of Pretreatment Radiomic Features of 18F-FDG PET in Patients With Hodgkin Lymphoma. *Clin Nucl Med.* (2019) 44:e559–65. doi: 10.1097/RLU.0000000000002732
- Tseng D, Rachakonda LP, Su Z, Advani R, Horning S, Hoppe RT, et al. Interim-treatment quantitative PET parameters predict progression and death among patients with hodgkin's disease. *Radiat Oncol.* (2012) 7:5. doi: 10.1186/1748-717X-7-5
- Frood R, Burton C, Tsoumpas C, Frangi AF, Gleeson F, Patel C, et al. Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review. *Eur J Nucl Med Mol Imag.* (2021) 48:3198–220. doi: 10.1007/s00259-021-05233-2
- Schöder H, Moskowitz C. Metabolic Tumor Volume in Lymphoma: Hype or Hope? *JCO.* (2016) 34:3591–4. doi: 10.1200/JCO.2016.69.3747
- Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res.* (2004) 10:7252–9. doi: 10.1158/1078-0432.CCR-04-0713
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* (2000) 56:337–44. doi: 10.1111/j.0006-341X.2000.00337.x
- Zheng Y, Cai T, Feng Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics.* (2006) 62:279–87. doi: 10.1111/j.1541-0420.2005.00441.x
- Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol.* (2017) 27:4498–509. doi: 10.1007/s00330-017-4859-z
- Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging.* (2015) 42:328–54. doi: 10.1007/s00259-014-2961-x
- Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* (2007) 8:118–27. doi: 10.1093/biostatistics/kjx037

33. Da-Ano R, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, et al. A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. *PLoS ONE*. (2021) 16:e0253653. doi: 10.1371/journal.pone.0253653

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Morland, Triumbari, Maiolo, Cuccaro, Treglia, Hohaus and Annunziata. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Radiomics and Its Applications and Progress in Pancreatitis: A Current State of the Art Review

Gaowu Yan^{1†}, Gaowen Yan^{2†}, Hongwei Li^{3†}, Hongwei Liang^{4†}, Chen Peng^{5†}, Anup Bhetuwal⁶, Morgan A. McClure⁷, Yongmei Li^{4*}, Guoqing Yang^{1*}, Yong Li^{1*}, Linwei Zhao¹ and Xiaoping Fan¹

¹ Department of Radiology, Suining Central Hospital, Suining, China, ² Department of Radiology, The First Hospital of Suining, Suining, China, ³ Department of Radiology, The Third Hospital of Mianyang and Sichuan Mental Health Center, Mianyang, China, ⁴ Department of Radiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, ⁵ Department of Gastroenterology, The First Hospital of Suining, Suining, China, ⁶ Sichuan Key Laboratory of Medical Imaging, Department of Radiology, Affiliated Hospital of North Sichuan Medical College, Nanchong, China, ⁷ Department of Radiology and Imaging, Institute of Rehabilitation and Development of Brain Function, The Second Clinical Medical College of North Sichuan Medical College, Nanchong Central Hospital, Nanchong, China

OPEN ACCESS

Edited by:

Alessandro Granito,
University of Bologna Department of
Medical and Surgical Sciences, Italy

Reviewed by:

Zubair Khan,
University of Texas Health Science
Center at Houston, United States
Linda Beenet,
University of California, Los Angeles,
United States

*Correspondence:

Yongmei Li
lymzhang70@aliyun.com
Guoqing Yang
13890893057@163.com
Yong Li
lmy2008hy@163.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Gastroenterology,
a section of the journal
Frontiers in Medicine

Received: 17 April 2022

Accepted: 31 May 2022

Published: 23 June 2022

Citation:

Yan G, Yan G, Li H, Liang H, Peng C,
Bhetuwal A, McClure MA, Li Y,
Yang G, Li Y, Zhao L and Fan X (2022)
Radiomics and Its Applications and
Progress in Pancreatitis: A Current
State of the Art Review.
Front. Med. 9:922299.
doi: 10.3389/fmed.2022.922299

Radiomics involves high-throughput extraction and analysis of quantitative information from medical images. Since it was proposed in 2012, there are some publications on the application of radiomics for (1) predicting recurrent acute pancreatitis (RAP), clinical severity of acute pancreatitis (AP), and extrapancreatic necrosis in AP; (2) differentiating mass-forming chronic pancreatitis (MFPC) from pancreatic ductal adenocarcinoma (PDAC), focal autoimmune pancreatitis (AIP) from PDAC, and functional abdominal pain (functional gastrointestinal diseases) from RAP and chronic pancreatitis (CP); and (3) identifying CP and normal pancreas, and CP risk factors and complications. In this review, we aim to systematically summarize the applications and progress of radiomics in pancreatitis and its associated situations, so as to provide reference for related research.

Keywords: radiomics, acute pancreatitis, chronic pancreatitis, autoimmune pancreatitis, pancreatic ductal adenocarcinoma, computed tomography, magnetic resonance imaging, positron emission tomography/computed tomography

INTRODUCTION

Radiomics and Its Process

Inspired by the knowledge systems and research fields of such as genomics, proteomics, radiogenomics, etc., Lambin et al. first proposed the concept of radiomics in 2012 (1–6). Radiomics refers to high-throughput extraction and analysis of a large number of advanced quantitative imaging features from medical images obtained by computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography (PET) (2). The workflow of radiomics mainly includes the following steps (1–6). (1) *Image acquisition* is the first step of radiomics. The images may come from CT, MRI, PET, as well as X-ray radiography and ultrasonography (US), etc. (7–10). Because the distribution of images features may be affected by many factors such as equipment vendors, scanning protocols, imaging parameters, reconstruction algorithms, etc., it is of great importance to establish standards and consensus imaging protocols. (2) *Image segmentation* uses dedicated software to draw two dimensions (2-D) or three dimensions (3-D) of regions of interest (ROIs) of lesions or organs by means of manual, semi-automatic, or automatic segmentations.

(3) *Image preprocessing* is to homogenize the data before extracting radiomics features which mainly includes two methods: image resampling and gray-level discretization (4). *Features extraction* uses dedicated software or software packages to extract morphological features, first-order statistical features, second-order statistical features, and high-order statistical features from 2-D or 3-D ROIs after segmentation. Morphological features ($n = 16$) are used to describe the 3-D shape and size of a ROI including asphericity, compactness, maximum diameter, sphericity, surface area, surface to volume ratio, volume, etc. The first-order statistical feature ($n = 18$) represents the histogram of voxel intensity values contained within a ROI to include mean, median, maximum, minimum, standard deviation, percentile, skewness, kurtosis, uniformity, energy, entropy, etc. Second order statistical features are used to describe the spatial distribution of voxel intensities within a ROI to include gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), gray-level distance-zone matrix (GLDZM), neighborhood gray tone difference matrix (NGTDM), and neighboring gray level dependence (NGLDM). After applying filters or mathematical transformations to the images, the higher-order statistics features can be obtained (5). *Feature selection* is the process of removing redundant features and selecting the most relevant features according to specific research tasks. Common methods are univariate analysis, logistic regression analysis, least absolute shrinkage and selection operator (LASSO), minimum redundancy maximum relevance (MRMR), etc. (6). *Modelization and validation* is after a classification or prediction model is established, it needs to be tested internally and externally to evaluate the robustness and repeatability of the model.

In recent years, due to the progress and rapid developments of various hardware and software technologies, radiomics has gradually developed into a relatively mature discipline or medical image analysis method (1). There are more and more publications on the application of radiomics for the diagnosis, differential diagnosis, treatment options, and prognosis evaluation of many human diseases (11–16). Among them, Hong et al. (13) extracted 10 radiomics features from the contrast-enhanced CT (CECT) images of 241 patients with a bone island or osteoblastic metastasis to establish a random forest (RF) prediction model. The results showed that the RF model based on CT was helpful to differentiate bone islands from osteoblastic metastases, and its diagnostic performance was higher than that of inexperienced radiologists but equivalent to that of experienced radiologists. In another study, Tian et al. (16) reported the diagnostic value of preoperative evaluation of microvascular invasion of solitary small hepatocellular carcinoma (HCC) based on nomogram of gadolinium ethoxybenzyl diethylenetriamine pentaacetic acid (Gd-EOB-DTPA) enhanced MRI. The results indicated that the clinical-radiological-radiomics model achieved the highest diagnostic performance with area under the receiver operating characteristic curves (AUCs) of 0.934, 0.889 and 0.875 for the training, internal and external validation sets, respectively.

In this review, we aim to systematically summarize the applications and progress of radiomics in pancreatitis and associated situations (Table 1) so as to provide reference for related research.

CLINICAL APPLICATIONS

Predicting Recurrent Acute Pancreatitis

Acute pancreatitis (AP) is a common disease in clinical practice and meta-analysis showed that the annual incidence rate of AP in the world is about 33.74/100 000, along with an annual mortality rate of about 1.16/100 000 (36). With the increase in population aging, biliary calculus, hyperlipidemia, obesity, and many other AP risk factors, the incidence of AP is also gradually increasing (37–39). Recurrent acute pancreatitis (RAP) is a special type of pancreatitis, and it is different from AP and chronic pancreatitis (CP). The definition of RAP is that patients should experience at least two separate episodes of AP at least 3 months apart, and there are no abnormalities in pancreatic tissue structure or function in remission (40). It is reported that the recurrence rate of AP is about 10–30% (17). About 10% of patients with first-episode of AP and 36% of patients with RAP may progress to CP, and the risk is higher among men, smokers, and alcoholics (41). Another study also reported that CP may increase the risk of pancreatic cancer (PC) in patients (42). After 5 and 9 years of the diagnosis of CP, the risk of PC in CP patients increased by eight times and three times, respectively. Therefore, early prediction of RAP and appropriate management measures can not only decrease the recurrence of AP, but it also prevents or delays its progression to CP and even PC.

Chen et al. (17) included 389 first-episode AP patients. On the CT images of arterial and venous phases, 412 radiomics features were extracted from the ROIs of the whole pancreatic parenchyma, and 10 features were finally selected to establish the prediction model. In the training cohort ($n = 271$, including 145 patients with AP and 126 patients with RAP), the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and AUC of the radiomics model in predicting patients with RAP were 86.7%, 87.6%, 89.7%, 84.1%, 87.1%, and 0.941%, respectively. In the validation cohort ($n = 118$, including 63 patients with AP and 55 patients with RAP), the same diagnostic indexes of the radiomics model in predicting patients with RAP were 83.8%, 97.7%, 98.4%, 78.2%, 89.0%, and 0.929%, respectively. The results in the training and validation cohorts were all significantly higher than those of the clinical model (all P -values < 0.05).

Quantitative investigation on predicting RAP is still in a paucity at present. Previous studies mostly focused on the risk factors of RAP after the first attack of AP such as demography (like gender, age, etc.), and clinical characteristics (like etiology, local complications, etc.) (43–45). Chen et al. (17) first showed that the radiomics model based on CECT exhibits promising value in the early prediction of RAP. In another similar study, Hu et al. (18) constructed a multivariate logistic regression radiomics model, radiomics,

TABLE 1 | Characteristics of the included publications on radiomics in pancreatitis.

Study ID	Year	Country	Design	Sample size	Objective (s)	Reference standard	Imaging modality	Imaging phases (slice thickness)	Segmentation method	Segmentation software	Feature extraction software	Feature type
Chen et al. (17)	2019	China	Retrospective	389	Predicting the recurrence of AP	Follow-up	Somatom Definition AS and Somatom Definition Flash (Siemens Healthineers), and LightSpeed VCT (GE Healthcare)	Arterial phase and venous phase images (5.0 mm)	Manual	IBEX	IBEX	S and Q
Hu et al. (18)	2022	China	Retrospective	190	Predicting the recurrence of AP	Follow-up	3.0 T MRI (Discovery 750, GE Healthcare)	T2WI (5.0 mm)	Manual	IBEX	IBEX	S and Q
Lin et al. (19)	2020	China	Retrospective	259	Predicting severity of AP	2012 revised Atlanta classification of AP	3.0 T MRI (Discovery 750, GE Healthcare)	Portal venous phase images (5.2 mm)	Manual	IBEX	IBEX	S and Q
Zhou et al. (20)	2021	China	Retrospective	135	Predicting EXPN in AP	Pathology and follow-up	3.0 T MRI (Discovery 750, GE Healthcare)	T2WI images of extra pancreatic collections and late arterial phase images of the pancreatic parenchyma (6.0 mm)	Manual	IBEX	IBEX	S and Q
Zhang et al. (21)	2022	China	Retrospective	138	Differentiating MFCP from PDAC	Pathology and CP consensus	Brilliance-16P (Philips Healthcare) and Aquilion ONE (Canon Medical Systems)	Portal venous phase images	Manual	3D Slicer	Pyradiomics	S and Q
Liu et al. (22)	2022	China	Retrospective	102	Distinguishing PC from MFCP	Pathology and follow-up	3.0 T MRI (MAGNETOM Skyra, Siemens Healthineers)	Axial T1WI, T2WI, DWI (b=800 s/mm ²), and ADC images	Manual	ITK-Snap	Pyradiomics	S and Q
Ma et al. (23)	2022	China	Retrospective	175	Differentiating between PC and CP (AIP and MFCP)	Including pathology and follow-up	Discovery CT 750 HD, Revolution CT, and Optima CT660 (GE Healthcare)	Arterial phase and venous phase images	Manual	MITK	Pyradiomics	S and Q
Deng et al. (24)	2021	China	Retrospective	119	Distinguishing PDAC from MFCP	Pathology	3.0 T MRI (Discovery 750, GE Healthcare)	Axial T1WI, T2WI, and the arterial phase and portal venous phase images	Manual	IBEX	IBEX	S and Q
Ren et al. (25)	2020	China	Retrospective	109	Differentiating MFCP from PDAC	Pathology	Brilliance 64 (Philips Healthcare) and Optima 670 (GE Healthcare)	Unenhanced CT images (3.0 mm)	Manual	ITK-SNAP	Analysis Kit	Q only
Ren et al. (26)	2019	China	Retrospective	109	Differentiating MFCP from PDAC	Pathology	Brilliance 64 (Philips Healthcare) and Optima 670 (GE Healthcare)	Arterial and portal phase CT images (3.0 mm)	Manual	ITK-SNAP	Analysis Kit	S and Q
Zhang et al. (27)	2019	China	Retrospective	109	Differentiating MFCP from PDAC	Pathology	Brilliance 64 (Philips Healthcare), Light speed VCT and Discovery HD750 (GE Healthcare)	Parenchymal phase images (5.0 mm)	Manual	ITK-SNAP	Analysis Kit	S and Q
Li et al. (28)	2022	China	Retrospective	97	Differentiating AIP from PDAC	Pathology and follow-up	Brilliance-16P (Philips Healthcare); Aquilion ONE (Canon Medical Systems)	Portal venous phase images (0.8/1.0 mm)	Manual	3D Slicer	Pyradiomics	S and Q

(Continued)

TABLE 1 | Continued

Study ID	Year	Country	Design	Sample size	Objective (s)	Reference standard	Imaging modality	Imaging phases (slice thickness)	Segmentation method	Segmentation software	Feature extraction software	Feature type
Liu et al. (29)	2021	China	Retrospective	112	Differentiating AIP and PDAC	Pathology and follow-up	PET/CT (Biograph64, Siemens Healthineers)	early and delayed imaging (3.0 mm)	Manual	3D Slicer	MATLAB R2018a	S and Q
Linning et al. (30)	2020	China	Retrospective	96	Differentiating AIP and PDAC	Pathology and follow-up	A range of helical multidetector (16, 64, 128, and 256 slices)	Non-contrast, arterial, and venous phases (1.0-5.0 mm)	Manual	In-house imaging platform	In-house MATLAB 2016b program	S and Q
Park et al. (31)	2020	USA	Retrospective	182	Differentiating AIP from PDAC	Pathology and follow-up	Somatom Definition, Definition Flash, or Force, and Somatom Sensation (Siemens Healthineers)	Arterial phase and venous phase images (0.75/3.0 mm)	Manual	Velocity AI	Velocity AI	S and Q
Zhang et al. (32)	2019	China	Retrospective	111	Differentiating AIP and PDAC	Pathology and follow-up	PET/CT (Biograph64, Siemens Healthineers)	- (0.98 mm)	Manual	3D Slicer	MATLAB R2017a	Q only
Zhang et al. (33)	2019	China	Retrospective	111	Differentiating AIP and PDAC	Pathology and follow-up	PET/CT (Biograph64, Siemens Healthineers)	- (0.6 mm)	Manual	3D Slicer	MATLAB R2017a	S and Q
Mashayekhi et al. (34)	2020	USA	Retrospective	56	Differentiating FAP, RAP, and CP	Clinical criteria	Including Sensation 64 (Siemens Healthineers)	Portal venous phase images (3 mm)	Manual	In-house MATLAB program	In-house MATLAB program	Q only
Frøkjær et al. (35)	2020	Denmark	Retrospective	99	Differentiating CP from healthy pancreas; classification of CP based on two risk factors and two complications	Lüneburg criteria	1.5T MRI (Signa HDxt, GE Healthcare)	DWI (b = 0 s/mm ²) (2.6 mm)	Manual	3D Slicer	Pyradiomics	Q only
Study ID	Type of extracted features			Number of extracted features	Number of statistically significant features	Feature reduction and classification method	Modeling method	Evaluation index	Main conclusions		%RQS (points)	
Chen et al. (17)	Shape features; First-order texture features; Second-order texture features			412	10 (five from arterial phase and five from portal phase)	Independent samples <i>t</i> -test, Mann-Whitney <i>U</i> test, LASSO regression, and Spearman correlation	Multivariable logistic regression analysis and SVM	ROC curve analysis for radiomics and clinical models	The radiomics model based on CECT performed well in predicting AP recurrence		16 (44%)	
Hu et al. (18)	Shape features; First-order texture features; Second-order texture features			513	4	LASSO	Multivariable logistic regression analysis	ROC curve analysis for radiomics, clinical, and combined models	Radiomics features based on MRI-T2WI could be used as biomarkers to predict the recurrence of AP		12 (33%)	
Lin et al. (19)	Shape features; First-order texture features; Second-order texture features			353	11	Independent sample <i>t</i> -test, Mann-Whitney <i>U</i> test, and Boruta algorithm	SVM	ROC curve analysis for radiomics model, and scoring systems of APACHE II, BISAP and MRSI	CEMRI based radiomics model had good performance in the early prediction of AP severity		15 (42%)	

(Continued)

TABLE 1 | Continued

Study ID	Type of extracted features	Number of extracted features	Number of statistically significant features	Feature reduction and classification method	Modeling method	Evaluation index	Main conclusions	%RQS (points)
Zhou et al. (20)	Shape features; First-order texture features; Second-order texture features	350	22 (12 from the extrapancreatic collection images and 10 from the pancreatic parenchyma images)	Independent sample <i>t</i> -test, Mann–Whitney <i>U</i> test, and LASSO	SVM	ROC curve analysis for radiomics models, clinical model, and scoring systems of EPIM and MRSI	The MRI-based radiomics models of both the extrapancreatic collections and the pancreatic parenchyma had excellent predictive performance for early EXPN	16 (44%)
Zhang et al. (21)	Shape features; First-order texture features; Second-order texture features	1,409	8	Variance analysis, Spearman's correlation analysis, and LASSO	Multivariable logistic regression analysis	ROC curve analysis for the CT model and radiomics models	The CT and radiomics models both were shown to be reasonably accurate in their differentiation of MFCP from PDAC in patients with CP	15 (42%)
Liu et al. (22)	Shape features; First-order texture features; Second-order texture features	960	6 (1 from T1WI, 2 from T2WI, 1 from DWI, and 2 from ADC maps)	MRMR and LASSO algorithms	Nomogram of the mixed model incorporating the radiomic signature, the CA19–9 level, and the CEA level	Individual T1WI, T2WI, DWI, and ADC models; clinical model; multiparametric MRI model; mixed-prediction model	A comprehensive model based on multiparametric MRI and clinically independent risk factors displayed the best evaluation performance	16 (44%)
Ma et al. (23)	Shape features; First-order texture features; Second-order texture features	1,037	2 (both from venous phase CT images)	Preserve features with good consistence, univariate Wilcoxon rank–sum test, correlation analysis, LASSO	Multivariable logistic regression analysis	ROC curve analysis for the arterial phase, venous phase, and arterial phase combined with venous phase radiomics model; clinical feature model; radiomics combined with clinical feature comprehensive model	The radiomics combined with clinical feature model could be a potential tool to distinguish PC from CP	16 (44%)
Deng et al. (24)	First-order texture features; Second-order texture features	410	28 (the number of included features in the T1WI, T2WI, arterial phase and portal venous phase feature subsets were 5, 7, 7, and 9, respectively)	Independent sample <i>t</i> -test, Mann–Whitney <i>U</i> test, LASSO	SVM	ROC curve analysis for T1WI, T2WI, and the arterial phase and portal venous phase radiomics models, and a clinical model	Radiomic models based on multiparametric MRI have the potential to distinguish PDAC from MFCP	17 (47%)
Ren et al. (25)	Shape features; First-order texture features; Second-order texture features	396	10	Mann–Whitney <i>U</i> test and MRMR	RF	ROC curve analysis for radiomics model	Unenhanced CT texture analysis can be a promising non-invasive method in discriminating MFCP from PDAC	10 (28%)
Ren et al. (26)	Shape features; First-order texture features; Second-order texture features	396	9 (five were arterial phase texture parameters and four portal phase texture parameters)	Mann–Whitney <i>U</i> test and MRMR	Multivariate logistic regression analysis	ROC curve analysis for imaging feature-based, texture feature-based models in arterial phase, and portal phase, and the combined model	CT texture analysis demonstrates great potential to differentiate MFCP from PDAC	10 (28%)

(Continued)

TABLE 1 | Continued

Study ID	Type of extracted features	Number of extracted features	Number of statistically significant features	Feature reduction and classification method	Modeling method	Evaluation index	Main conclusions	%RQS (points)
Zhang et al. (27)	First-order texture features; Second-order texture features	160	4	LASSO	Multivariate logistic regression analysis	ROC curve analysis for imaging feature-based, texture feature-based models in parenchymal phase, and the combined model	The CECT combined with texture analysis model has the best diagnostic efficiency for differentiating MFCP from PDAC	10 (28%)
Li et al. (28)	Shape features; First-order texture features; Second-order texture features	1,409	4 (from portal venous phase CT images)	Variance analysis, Spearman's correlation analysis, and LASSO	Radiomics score	ROC curve analysis for radiomics score	The portal rad-score can accurately and non-invasively differentiate fAIP from PDAC	10 (28%)
Liu et al. (29)	Shape features; First-order texture features; Second-order texture features; MIP features	514	10 (three from CT, four from PET-early, and three from PET-delay)	SVM-RFE	SVM-LKF	ROC curve analysis for fusion feature based model, on ¹⁸ F-FDG PET/CT dual-time PET/CT images	The radiomics model based on ¹⁸ F-FDG PET/CT dual-time PET/CT images provided promising clinical performance for diagnostic indicators based discriminating AIP from PDAC model	15 (42%)
Linning et al. (30)	Shape features; First-order texture features; Second-order texture features	1,160	18 (six from non-contrast, arterial, and venous phases, respectively)	Unsupervised hierarchical clustering, MRMR, and IFS	RF	ROC curve analysis for the non-contrast, arterial phase, venous phase, and hybrid of three phases radiomics models	Radiomics is helpful for a differential diagnosis of AIP in clinical practice as a non-invasive and quantitative method	9 (25%)
Park et al. (31)	Shape features; First-order texture features; Second-order texture features; Filtered image features	431	35	MRMR	RF	ROC curve analysis for the arterial phase and venous phase radiomics features	Radiomic features help differentiate AIP from PDAC	8 (22%)
Zhang et al. (32)	First-order texture features; Second-order texture features; Filtered image features	418	8	Fisher's criterion >0.01 and SFS		ROC curve analysis for different feature selection and classification methods	The results proved that texture analysis of lesions helps to achieve accurate differentiation of AIP and PDAC	13 (36%)
Zhang et al. (33)	Shape features; First-order texture features; Second-order texture features	251	10	Spearman correlation, MRMR, and SVM	RF, adaptive boosting, and SVM	ROC curve analysis for different feature selection and classification methods	Radiomics could aid the non-invasive differentiation of AIP and PDAC in ¹⁸ F-FDG PET/CT images and the integration of multi-domain features is beneficial for the differentiation	15 (42%)
Mashayekhi et al. (34)	Shape features; First-order texture features; Second-order texture features	54	11	Wilcoxon rank-sum test	Isomap and SVM	ROC curve analysis for radiomic features	Certain radiomic features on CT imaging can differentiate patients with FAP, RAP, and CP	10 (28%)

(Continued)

TABLE 1 | Continued

Study ID	Type of extracted features	Number of extracted features	Number of statistically significant features	Feature reduction and classification method	Modeling method	Evaluation index	Main conclusions	%RQS (points)
Frokjaer et al. (35)	Shape features; First-order texture features; Second-order texture features; Filtered image features	851	5 (for differentiation between healthy pancreas and CP)	10-fold cross-validation forward selection procedure	Naive Bayes classifier	The average m-fold performance metrics for five demonstrated to be feasible in classifiers	Pancreatic texture analysis patients with CP and discriminate clinically relevant subgroups based on etiological risk factors and complications	8 (22%)
AP, acute pancreatitis; RAP, recurrent acute pancreatitis; CP, chronic pancreatitis; MFCP, mass-forming chronic pancreatitis; AIP, autoimmune pancreatitis; fAIP, focal type autoimmune pancreatitis; PC, pancreatic cancer; PDAC, pancreatic ductal adenocarcinoma; EPXN, extrapancreatic necrosis; FAP, functional abdominal pain; CT, computed tomography; CECT, contrast-enhanced computed tomography; MRI, magnetic resonance imaging; CEMRI, contrast-enhanced MRI; T1WI, T1-weighted imaging; T2WI, T2-weighted imaging; DWI, diffusion weighted imaging; ADC, apparent diffusion coefficient; EUS, endoscopic ultrasound; ¹⁸ F-FDG PET/CT, ¹⁸ F-fluorodeoxyglucose positron emission tomography/computed tomography; CAD, computer-aided diagnosis; MIP, maximum intensity projection; MRMR, minimum-redundancy maximum-relevance; LASSO, least absolute shrinkage and selection operator; SFS, sequential forward selection; IFS, incremental forward search; RF, random forest; SVM, support vector machine; SVM-RFE, support vector machine recursive feature elimination; SVM-LKF, support vector machine with a linear kernel function; MITK, medical imaging interaction toolkit; S, semantic; Q, quantitative; RQS, radiomics quality score; ROC, receiver operating characteristic curve; NA, not available; EPIM, extrapancreatic inflammation on MRI; MRSI, magnetic resonance severity index; APACHE II, acute physiology and chronic health evaluation II; BISAP, bedside index for severity in acute pancreatitis.								

and clinical characteristics combined model based on MRI-T2WI, and their results were consistent with those of Chen et al. (17).

Predicting Clinical Severity of AP

Based on the 2012 revised Atlanta classification and definition (2012-RACD) by international consensus, AP can be divided into three categories stratified by its clinical severity: mild acute pancreatitis (MAP), moderately severe acute pancreatitis (MSAP), and severe acute pancreatitis (SAP) (46). MAP is characterized by no organ failure and local or systemic complications. It can return to normal within 1–2 weeks. Usually, there is no need for an imaging examination of the pancreas, and the mortality rate is very low. MSAP is characterized by transient organ failure (<48 h), or accompanied by local or systemic complications, while no persistent organ failure (more than 48 h) exists. MSAP can be cured without intervention or may require long-term specialist care. The mortality rate of MSAP is much lower than that of SAP. SAP is characterized by persistent single or multiple organ failure (more than 48 h). Patients with persistent organ failure usually have one or more local complications. In the first few days after AP onset, patients with persistent organ failure have an increased risk of death, and the mortality reported in the literature is as high as 36–50% (46), and the mortality rate of patients with persistent organ failure complicated with infectious necrosis is very high (46). Therefore, early prediction of the clinical severity of AP is of utmost importance, which is not only good for the early diagnosis and treatment of MSAP and SAP patients, and also in favor of the early diversion or referral of MSAP and SAP patients.

Currently, methods of early predicting the clinical severity of AP mainly depend on clinical characteristics [such as scoring systems of acute physiology and chronic health evaluation II (APACHE II, ≥eight points), bedside index for severity in acute pancreatitis (BISAP, ≥three points), Ranson (≥three points) and modified Marshall score (≥two points)], laboratory tests [such as C-reactive protein concentration (≥150 mg/l), serum procalcitonin (>0.5 ng/ml), interleukin-6 (>50 pg/l) and neutrophil/lymphocyte ratio (>10)] as well as findings on imaging examinations [such as computed tomography severity index (CTSI, ≥four points), modified computed tomography severity index (mCTSI, ≥four points), and extrapancreatic inflammation on computed tomography (EPIC, ≥four points)] (47–50).

Lin et al. (19) first reported a contrast-enhanced MRI (CEMRI) based radiomics model to predict the clinical severity of AP (MAP vs. MSAP and SAP). In their study, they included 259 AP patients into the training (*n* = 180, with 99 MAP and 81 MSAP and SAP patients) and validation cohorts (*n* = 79, with 43 MAP and 36 MSAP and SAP patients). From the portal vein phase images, Lin et al. (19) extracted 353 radiomics features from the ROIs that contained the whole pancreatic parenchyma, and finally they selected 11 features to establish the support vector machine (SVM) model. In the training cohort, the sensitivity, specificity, PPV, NPV, accuracy, and AUC of the radiomics model to distinguish MAP from MSAP or SAP patients were 77.8%,

91.9%, 88.7%, 83.5%, 85.6%, and 0.917%, respectively. In the validation cohort, the corresponding diagnostic indexes of the radiomics model in distinguishing MAP from MSAP or SAP patients were 75.0%, 86.0%, 81.8%, 80.4%, 81.0%, and 0.848%, respectively. The both AUCs were significantly higher than that of APACHE II, BISAP, and MRSI scoring systems (all P -values were <0.05). This study showed that when compared with some existing clinical and radiological scoring systems, the portal phase MRI radiomics model may be more accurate in early predicting the clinical severity of AP.

Predicting Extrapancreatic Necrosis in AP

Based on the 2012-RACD (46), AP can be divided into two categories according to its morphological manifestations on imaging examination: (1) interstitial edematous pancreatitis (IEP; about 85%); and (2) necrotizing pancreatitis (NP; about 15%). Based on the distribution and location of necrosis, NP can be further subdivided into three subtypes (46): (1) combined pancreatic and peripancreatic necrosis (about 75.0%); (2) peripancreatic necrosis only (about 20.0%); and (3) pancreatic necrosis only (about 5.0%). The literature indicates that compared with NP, the mortality rate of IEP is about 3.0% while the mortality rate of NP is about 17%; and if combined with infection, the mortality rate of NP can rise to about 30% (46, 51). Consequently, it is of great clinical significance to distinguish IEP from NP for predicting the prognosis of AP patients. In the international structured reporting template of AP based on CECT published in 2020, experts also highlighted the importance of radiologists to clarify the morphologic subtypes of AP, the degree and anatomic area involvement of NP, the type and location of peripancreatic collections, and some other key points in the CT reports (51).

Zhou et al. (20) used an MRI based radiomics model to predict early extrapancreatic necrosis (EXPAN) in patients with AP. They enrolled 135 AP patients who were divided into the training ($n = 94$, with 47 EXPAN and 47 APFC patients) and validation cohorts ($n = 41$, with 20 EXPAN and 21 APFC patients). On the T2WI and late arterial phase images, Zhou et al. (20) extracted 350 image radiomics features from ROI of the peripancreatic collections (T2WI) and entire pancreatic parenchyma (late arterial phase). After dimension reduction and feature selection, 22 features (12 from the T2WI and 10 from the late arterial phase images) were selected for establishing SVM model. In the training cohort, the sensitivity, specificity, PPV, NPV, accuracy, and AUC of the T2WI peripancreatic collections and late arterial phase pancreatic parenchyma radiomics models for predicting EXPAN were 97.9% and 87.2%, 85.1% and 87.2%, 86.8% and 87.2%, 97.6% and 87.2%, 91.5% and 87.2%, 0.969% and 0.931%, respectively. In the validation cohort, the corresponding diagnostic parameters of the T2WI peripancreatic collections and late arterial phase pancreatic parenchyma radiomics models for predicting EXPAN were 95.0% and 75.0%, 90.5% and 90.5%, 90.5% and 88.2%, 95.0% and 79.2%, 92.7% and 82.9%, 0.976 and 0.921%, respectively. Both of the AUCs were significantly higher than those of clinical model, EPIM and MRSI scoring systems (all P -values < 0.05). This investigation showed that when compared with some existing clinical model and radiological scoring

systems, the MRI radiomics model based on T2WI peripancreatic collections and late arterial phase pancreatic parenchyma may be able to accurately predict EXPAN in AP patients at an early stage.

Differentiating Mass-Forming Chronic Pancreatitis From Pancreatic Ductal Adenocarcinoma

Pancreatic ductal adenocarcinoma (PDAC) is a malignant tumor that originating from pancreatic ductal epithelial cells, accounting for about 80–90% of all the pancreatic cancer (PC) patients with about 60–70% of the PDACs occur in the pancreatic heads (52, 53). The prognosis of PDAC is very poor ($<10\%$) and surgery has always been considered the first choice for the treatment of PDAC (52, 53). The mass-forming chronic pancreatitis (MFCP) is a special type of CP. Documents reported that MFCP accounts for about 27–50% of CP, and the vast majority of MFCP is located in the pancreatic heads (about 71%) (54–56). MFCP and PDAC share significant overlaps in the clinical manifestations (such as upper abdominal pain, nausea, weight loss, jaundice, diabetes, etc.), risk factors (such as alcohol, smoking, etc.), laboratory tests (such as elevated carbohydrate antigen 199 (CA199) and carcinoembryonic antigen (CEA) levels), and imaging findings (such as delayed enhancement) (57, 58). CT and endoscopic ultrasonography guided fine needle aspiration biopsy (EUS-FNA) can be used to improve the differential diagnosis accuracy of MFCP and PDAC, but both modalities are invasive examinations, which not only have sampling error, and also carry the risks of needle tract tumor seeding, bleeding, pancreatic juice leakage, etc. (59, 60). As a result, it is very difficult to accurately distinguish MFCP from PDAC prior to operation, yet it has very important clinical significance. Because accurate preoperative diagnosis of early PDAC can prevent it from being resectable to unresectable, and accurate diagnosis of MFCP can avoid unnecessary surgery.

With the rapid development of medical imaging technologies, radiomics has begun to be used in the differential diagnosis of MFCP and PDAC (21–27). For example, Deng et al. (24) studied 96 patients with PDAC and 23 patients with MFCP. They extracted four sets of radiomics features from T1WI, T2WI, as well as arterial and portal phase images of MRI to establish SVM models. When compared with the clinical model based on clinical characteristics and the evaluation results of two radiologists, the results demonstrated that in the primary cohort ($n = 64$, with 51 PDAC and 13 MFCP patients), the sensitivity, specificity and AUC of T1WI, T2WI, arterial phase and portal phase radiomics models, and the clinical model were 0.961, 0.769, and 0.893; 0.941, 0.769, and 0.911; 0.961, 0.923, and 0.958; 0.980, 1.000, and 0.997; 0.529, 0.692, and 0.516, respectively. In the testing cohort ($n = 55$, with 45 PDAC and 10 MFCP patients), the corresponding diagnostic data were 1.000, 0.733, and 0.882; 0.844, 0.900, and 0.902; 0.956, 0.900, and 0.920; 0.978, 0.900, and 0.962; 0.422, 0.900, and 0.649, respectively. There were no significant differences in the diagnostic performances between the four radiomics models (all P -values > 0.05), but they were all better than that of the clinical model and the radiologists' evaluation (all P -values < 0.05). This study demonstrated that

radiomics may be used to improve the differential diagnosis accuracy of MFCP and PDAC.

Differentiating Focal Autoimmune Pancreatitis From PDAC

Autoimmune pancreatitis (AIP) is a special type of CP. Yoshida et al. first proposed the concept of AIP in 1995; and the annual incidence rate of AIP is about 3.1/100 000, accounting for about 1.9%–6.6% of CP (61, 62). Pathologically, AIP is classified into two subtypes: (1) Type I, lymphoplasmacytic sclerosing pancreatitis (LPSP); and (2) Type II, idiopathic duct-centric chronic pancreatitis (IDCP) (61, 63). At present, Type I AIP has been considered as the pancreatic manifestation of a systemic disease named IgG4-related disease (IgG4-RD) and there are now dedicated criteria for IgG4-RD and some specific organs (like pancreas, biliary tract, kidney, ophthalmic tissues, and chest) (64–67). On imaging, AIP can be manifested as diffuse AIP and focal AIP, and about 40% of Type I AIP and 85% of Type II AIP are localized (64, 65). Focal AIP overlaps obviously with PDAC in clinical manifestations (such as obstructive jaundice, epigastric pain or discomfort, weight loss, etc.) and imaging findings (focal mass in the pancreas), and an accurate differential diagnosis is very challenging. However, the treatment methods after the establishment of diagnosis are very different because AIP responds well to glucocorticoid drugs while PDAC mainly needs comprehensive treatment methods such as surgery, chemotherapy and radiotherapy. Therefore, the accurate differential diagnosis of focal AIP and PDAC before the treatments has very important clinical value. Once focal AIP is misdiagnosed as PDAC, it will lead to unnecessary surgery, and once PDAC is misdiagnosed as focal AIP, it may delay the effective treatments of PDAC.

Radiomics may play a positive role in the differential diagnosis of focal AIP and PDAC (28–33). Among the studies, Linning et al. (30) studied 45 patients with focal AIP and 51 patients with PDAC to evaluate the value of radiomics model based on multi-phase CECT for the differential diagnosis of focal AIP from PDAC. The results showed that the sensitivity, specificity, PPV, NPV and accuracy of unenhanced, arterial phase, portal phase, and hybrid radiomics models were 71.11%, 86.27%, 77.19%, 82.05%, and 79.17%; 82.22%, 90.20%, 85.19%, 88.10%, and 86.46%; 93.33%, 96.08%, 92.00%, 89.13%, and 90.63%; 93.33%, 96.08%, 94.23%, 95.45%, and 94.80%, respectively. The AUCs were 0.827, 0.890, 0.953, and 0.977, respectively. The diagnostic performances were higher than those of the two radiologists ($P < 0.05$). In another study, Li et al. (28) used propensity score matching (PSM) in 45 patients with focal AIP and 51 patients with PDAC who were matched in gender, age, body mass index (BMI), and CT characteristics. They evaluated the diagnostic performance of radiomics model based on portal phase CECT images in the differential diagnosis of focal AIP and PDAC. Their results were consistent with the research of Linning et al. (30) The above two studies have shown that radiomics may play a positive role in the differential diagnosis of focal AIP and PDAC.

Differentiating Functional Abdominal Pain, RAP, and CP

Abdominal pain is a common clinical symptom and one of the most important reasons for patients to see a doctor. Its etiologies may come from abdominal solid organs, gastrointestinal tract, biliary system, urinary system, reproductive system, chest diseases, or systemic diseases. Because abdominal pain is a non-specific clinical symptom, early identifying the causes of abdominal pain helps the clinicians and patients to choose the appropriate treatment methods. In a study, Mashayekhi et al. (34) studied 19 patients with functional abdominal pain (functional gastrointestinal diseases, FGD), 20 patients with RAP, and 17 patients with CP and explored the value of a SVM classifier based on venous phase images of CECT in distinguishing FGD, RAP, and CP. The results showed that the overall predictive accuracy of the SVM classifier was 82.1%. In the one-to-one comparison of the three groups, the sensitivity, specificity, and AUC of the FGD group were 79%, 100%, and 0.91%, respectively; the same diagnostic parameters of the RAP group were 95%, 78%, and 0.88%, respectively; while the sensitivity, specificity and AUC of the CP group were 71%, 95%, and 0.90%, respectively. The results suggested that some radiomics features may be an effective method for radiologists and gastroenterologists to distinguish FGD, RAP, and CP.

Identifying CP and Normal Pancreas, CP Risk Factors, and Complications

Frøkjær et al. (35) studied 77 CP patients and 22 healthy controls, extracted 851 MRI texture features from diffusion-weighted imaging (DWI) images, and finally constructed five classifier models to address the potential use of MRI texture analysis of the pancreas in CP patients. The five radiomics classifiers were: (1) CP vs. healthy controls (with five selected radiomics features), (2) alcoholic vs. non-alcoholic etiology of CP (with nine selected radiomics features), (3) use of tobacco vs. no use of tobacco (with 10 selected radiomics features), (4) diabetes vs. no diabetes (with four selected radiomics features), and (5) pancreatic exocrine insufficiency vs. normal exocrine function (with three selected radiomics features). The results showed that the sensitivity, specificity, PPV, and accuracy of the above five radiomics classifiers were 0.71–0.97, 0.84–1.00, 0.71–1.00, and 0.82–0.98, respectively. These results implied that radiomics may be a potentially promising tool used to depict early-stage CP and monitor disease progression.

LIMITATIONS AND SOLUTIONS

Since it was proposed in 2012, due to the progress and rapid developments of various hardware and software technologies, radiomics has gradually developed into a relatively mature research field and knowledge system (1–6, 68). The authors performed a literature search in the PubMed database with the strategy of “(Radiomics [Title/Abstract]) OR (Radiomic [Title/Abstract]).” There were no restrictions on the publication time, language or research object. As of April 17, 2022, a total of 5,580 relevant publications were retrieved. This

search result has proved the degree of attention paid by researchers and related fields to radiomics in the past 10 years. However, the vast majority of radiomics models reported in the current literature are still in the stage of developing research, and their clinical applications have not really been implemented. The authors believe that this phenomenon is mainly caused by the limitations of radiomics. The current radiomics research and clinical applications still have the following limitations and difficulties (69): (1) the standardization of medical imaging data is insufficient; (2) the generalization ability of the models is not good enough; (3) poor biological interpretability; and (4) the clinical utility of the models needing to be improved.

Standardization of Medical Image Data

Standardized, homogeneous, and high-quality training data is an important cornerstone of radiomics research and clinical applications. Radiomics may refer to the FAIR guiding principles for scientific data management and stewardship that were proposed by the international community named Force 11 (The Future of Research Communications and e-Scholarship 2011) in 2016 (70). This international community emphasizes that scientific data management and stewardship should follow the principles of Findable (F), Accessible (a), Interoperable (I), and Reusable (R).

Generalization of the Models

The performance of a radiomics model in similar and different distribution of datasets (such as various times, treatment plans, geographical locations, etc.) is called the generalization of a radiomics model. That is to say the reproducibility and transferability of a radiomics model (71) which is an important premise for the clinical applications of radiomics. It is also an important problem that needs to be solved urgently in radiomics (72, 73). In addition to increasing the data sample size and data diversity, full-automatic and semi-automatic image segmentation methods need to be advocated, and reasonable features selection and dimensionality reduction methods also need to be adopted (69, 74). Federated machine learning is also expected to provide effective solutions to the above difficulties (75, 76).

Biological Interpretability

Radiomics researchers hope to explore the relationships between certain features and some diseases or clinical endpoints (such as the diagnosis and differential diagnosis of diseases, options of treatment schemes, predictions of treatment effects, pathological classification and grading, gene and protein phenotypes, etc.) by quantitatively extracting and analyzing image information (features) that cannot be recognized by human naked eye. This will provide more help for clinicians and patients for disease diagnosis and treatments. However, the biological interpretability of radiomics is still lacking, and the potential biological significance of each features is still unclear, which seriously

hinders its clinical applications (77–79). Therefore, how can we improve the biological interpretability of radiomics is an important problem to be faced in this field.

Clinical Utility

Radiomics models or systems with characteristics of easy to operate, short learning curve, good user experience, fast running speed, and broad use scenarios are often more in line with clinicians' work habits (80). Applications developed for mobile phones and internet users may become an effective carrier for the clinical applications of radiomics models or systems in the future.

CONCLUSIONS AND FUTURE PERSPECTIVES

Since it was proposed in 2012, radiomics has begun to demonstrate a promising potential both in scientific research and in clinical applications, such as predicting RAP, clinical severity of AP and EXPN of AP, and differentiating MFCP and focal AIP from PDAC (Table 1). However, most of the published studies hold the limitations of a single-center, retrospective, limited sample size, and low radiomics quality score (RQS) (4). In looking forward to the future, researchers may successively report some multicenter, prospective, large sample size, and high RQS studies. In addition to these, predicting AP clinical outcomes of organ failure, infection, death, hospitalization, admission to intensive care unit (ICU) and invasive intervention; quantifying pancreatic exocrine or (and) endocrine insufficiency; predicting the possibility of AP progress to CP or CP progress to PC; and effectively combining deep learning or some other technologies with radiomics may become the potential directions (81–87).

AUTHOR CONTRIBUTIONS

GaowuY, GaoweY, HLi, HLia, and CP designed the study and the drafting of the paper. GaowuY, GuY, YL, and YML revised the paper critically for intellectual content. All authors participated in the literature search and data collection. All authors approved the final version of the paper to be published.

FUNDING

This project was supported by grants from the Sichuan Provincial Commission of Health (Grant Nos. 18PJ138, 19PJ283, 19PJ284, and 20PJ284), Sichuan Provincial Department of Science and Technology (Grant No. 2019YFQ0028), and Science and Technology Association of Suining City (Grant Nos. 6 and 10).

ACKNOWLEDGMENTS

The authors would like to thank co-authors AB and MM for their help in proofreading the article.

REFERENCES

- Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036
- Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. (2012) 30:1234–48. doi: 10.1016/j.mri.2012.06.010
- Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures. They are data. *Radiology*. (2016) 278:563–77. doi: 10.1148/radiol.2015151169
- 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. doi: 10.1038/nrclinonc.2017.141
- Bartoli M, Barat M, Dohan A, Gaujoux S, Coriat R, Hoeffel C, et al. CT and MRI of pancreatic tumors: an update in the era of radiomics. *Jpn J Radiol*. (2020) 38:1111–24. doi: 10.1007/s11604-020-01057-6
- Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB, et al. Radiomics in oncology: a practical guide. *Radiographics*. (2021) 41:1717–32. doi: 10.1148/rg.2021210037
- Sivengphanom S, Gandomkar Z, Lewis SJ, Brennan PC. Mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs. *Acad Radiol*. (2021) S1076–6332(21)00468-2. doi: 10.1016/j.acra.2021.09.025
- Hu Z, Yang Z, Lafata KJ, Yin FF, Wang C. A radiomics-boosted deep-learning model for COVID-19 and non-COVID-19 pneumonia classification using chest x-ray images. *Med Phys*. (2022) 49:3213–22. doi: 10.1002/mp.15582
- Li MD, Cheng MQ, Chen LD, Hu HT, Zhang JC, Ruan SM, et al. Reproducibility of radiomics features from ultrasound images: influence of image acquisition and processing. *Eur Radiol*. (2022). doi: 10.1007/s00330-022-08662-1. [Epub ahead of print].
- Li C, Qiao G, Li J, Qi L, Wei X, Zhang T, et al. An Ultrasonic-based radiomics nomogram for distinguishing between benign and malignant solid renal masses. *Front Oncol*. (2022) 12:847805. doi: 10.3389/fonc.2022.847805
- Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal Cancer. *J Clin Oncol*. (2016) 34:2157–64. doi: 10.1200/JCO.2015.65.9128
- Rigirli F, Hoyer J, Lerebours R, Lafata KJ, Li C, Meyer M, et al. CT Radiomic features of superior mesenteric artery involvement in pancreatic ductal adenocarcinoma: a pilot study. *Radiology*. (2021) 301:610–22. doi: 10.1148/radiol.2021210699
- Hong JH, Jung JY, Jo A, Nam Y, Pak S, Lee SY, et al. Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT. *Radiology*. (2021) 299:626–32. doi: 10.1148/radiol.2021203783
- Guo W, She D, Xing Z, Lin X, Wang F, Song Y, et al. Multiparametric MRI-based radiomics model for predicting H3 K27M mutant status in diffuse midline glioma: a comparative study across different sequences and machine learning techniques. *Front Oncol*. (2022) 12:796583. doi: 10.3389/fonc.2022.796583
- Jimenez JE, Abdelhazef A, Mittendorf EA, Elshafee N, Yung JP, Litton JK, et al. A model combining pretreatment MRI radiomic features and tumor-infiltrating lymphocytes to predict response to neoadjuvant systemic therapy in triple-negative breast cancer. *Eur J Radiol*. (2022) 149:110220. doi: 10.1016/j.ejrad.2022.110220
- Tian Y, Hua H, Peng Q, Zhang Z, Wang X, Han J, et al. Preoperative evaluation of Gd-EOB-DTPA-enhanced MRI radiomics-based nomogram in small solitary hepatocellular carcinoma (≤ 3.0 cm) with microvascular invasion: a two-center study. *J Magn Reson Imaging*. (2022). doi: 10.1002/jmri.28157. [Epub ahead of print].
- Chen Y, Chen TW, Wu CQ, Lin Q, Hu R, Xie CL, et al. Radiomics model of contrast-enhanced computed tomography for predicting the recurrence of acute pancreatitis. *Eur Radiol*. (2019) 29:4408–17. doi: 10.1007/s00330-018-5824-1
- Hu Y, Liu N, Tang L, Liu Q, Pan K, Lei L, et al. Three-dimensional radiomics features of magnetic resonance T2-weighted imaging combined with clinical characteristics to predict the recurrence of acute pancreatitis. *Front Med*. (2022) 9:777368. doi: 10.3389/fmed.2022.777368
- Lin Q, Ji YF, Chen Y, Sun H, Yang DD, Chen AL, et al. Radiomics model of contrast-enhanced MRI for early prediction of acute pancreatitis severity. *J Magn Reson Imaging*. (2020) 51:397–406. doi: 10.1002/jmri.26798
- Zhou T, Xie CL, Chen Y, Deng Y, Wu JL, Liang R, et al. Magnetic resonance imaging-based radiomics models to predict early extra pancreatic necrosis in acute pancreatitis. *Pancreas*. (2021) 50:1368–75. doi: 10.1097/MPA.0000000000001935
- Zhang H, Meng Y, Li Q, Yu J, Liu F, Fang X, et al. Two nomograms for differentiating mass-forming chronic pancreatitis from pancreatic ductal adenocarcinoma in patients with chronic pancreatitis. *Eur Radiol*. (2022). doi: 10.1007/s00330-022-08698-3. [Epub ahead of print].
- Liu J, Hu L, Zhou B, Wu C, Cheng Y. Development and validation of a novel model incorporating MRI-based radiomics signature with clinical biomarkers for distinguishing pancreatic carcinoma from mass-forming chronic pancreatitis. *Transl Oncol*. (2022) 18:101357. doi: 10.1016/j.tranon.2022.101357
- Ma X, Wang YR, Zhuo LY, Yin XP, Ren JL, Li CY, et al. Retrospective analysis of the value of enhanced CT radiomics analysis in the differential diagnosis between pancreatic cancer and chronic pancreatitis. *Int J Gen Med*. (2022) 15:233–41. doi: 10.2147/IJGM.S337455
- Deng Y, Ming B, Zhou T, Wu JL, Chen Y, Liu P, et al. Radiomics model based on MR images to discriminate pancreatic ductal adenocarcinoma and mass-forming chronic pancreatitis lesions. *Front Oncol*. (2021) 11:620981. doi: 10.3389/fonc.2021.620981
- Ren S, Zhao R, Zhang J, Guo K, Gu X, Duan S, et al. Diagnostic accuracy of unenhanced CT texture analysis to differentiate mass-forming pancreatitis from pancreatic ductal adenocarcinoma. *Abdom Radiol*. (2020) 45:1524–33. doi: 10.1007/s00261-020-02506-6
- Ren S, Zhang J, Chen J, Cui W, Zhao R, Qiu W, et al. Evaluation of texture analysis for the differential diagnosis of mass-forming pancreatitis from pancreatic ductal adenocarcinoma on contrast-enhanced CT images. *Front Oncol*. (2019) 9:1171. doi: 10.3389/fonc.2019.01171
- Zhang JJ, Li QZ, Wang JH, Chen X, Ren S, Ye DD, et al. [Contrast-enhanced CT and texture analysis of mass-forming pancreatitis and cancer in the pancreatic head]. *Zhonghua Yi Xue Za Zhi*. (2019) 99:2575–80. doi: 10.3760/cma.j.issn.0376-2491.2019.33.004
- Li J, Liu F, Fang X, Cao K, Meng Y, Zhang H, et al. CT Radiomics features in differentiation of focal-type autoimmune pancreatitis from pancreatic ductal adenocarcinoma: a propensity score analysis. *Acad Radiol*. (2022) 29:358–66. doi: 10.1016/j.acra.2021.04.014
- Liu Z, Li M, Zuo C, Yang X, Ren S, et al. Radiomics model of dual-time 2-[18F]FDG PET/CT imaging to distinguish between pancreatic ductal adenocarcinoma and autoimmune pancreatitis. *Eur Radiol*. (2021) 31:6983–91. doi: 10.1007/s00330-021-07778-0
- Linning E, Xu Y, Wu Z, Li L, Zhang N, Yang H, et al. Differentiation of focal-type autoimmune pancreatitis from pancreatic ductal adenocarcinoma using radiomics based on multiphasic computed tomography. *J Comput Assist Tomogr*. (2020) 44:511–8. doi: 10.1097/RCT.0000000000001049
- Park S, Chu LC, Hruban RH, Vogelstein B, Kinzler KW, Yuille AL, et al. Differentiating autoimmune pancreatitis from pancreatic ductal adenocarcinoma with CT radiomics features. *Diagn Interv Imaging*. (2020) 101:555–64. doi: 10.1016/j.diii.2020.03.002
- Zhang Y, Cheng C, Liu Z, Wang L, Pan G, Sun G, et al. Radiomics analysis for the differentiation of autoimmune pancreatitis and pancreatic ductal adenocarcinoma in 18 F-FDG PET/CT. *Med Phys*. (2019) 46:4520–30. doi: 10.1002/mp.13733
- Zhang Y, Cheng C, Liu Z, Pan G, Sun G, Yang X, et al. [Differentiation of autoimmune pancreatitis and pancreatic ductal adenocarcinoma based on multi-modality texture features in 18F-FDG PET/CT]. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*. (2019) 36:755–62. doi: 10.7507/1001-5515.201807012
- Mashayekhi R, Parekh VS, Faghih M, Singh VK, Jacobs MA, Zaheer A. Radiomic features of the pancreas on CT imaging accurately differentiate functional abdominal pain, recurrent acute pancreatitis, and chronic pancreatitis. *Eur J Radiol*. (2020) 123:108778. doi: 10.1016/j.ejrad.2019.108778
- Frokjær JB, Lisitskaya MV, Jørgensen AS, Østergaard LR, Hansen TM, Drewes AM, et al. Pancreatic magnetic resonance imaging texture analysis in chronic

- pancreatitis: a feasibility and validation study. *Abdom Radiol.* (2020) 45:1497–506. doi: 10.1007/s00261-020-02512-8
36. Xiao AY, Tan ML, Wu LM, Asrani VM, Windsor JA, Yadav D, et al. Global incidence and mortality of pancreatic diseases: a systematic review, meta-analysis, and meta-regression of population-based cohort studies. *Lancet Gastroenterol Hepatol.* (2016) 1:45–55. doi: 10.1016/S2468-1253(16)30004-8
 37. Li CL, Jiang M, Pan CQ, Li J, Xu LG. The global, regional, and national burden of acute pancreatitis in 204 countries and territories, 1990–2019. *BMC Gastroenterol.* (2021) 21:332. doi: 10.1186/s12876-021-01906-2
 38. Roberts SE, Morrison-Rees S, John A, Williams JG, Brown TH, Samuel DG. The incidence and aetiology of acute pancreatitis across Europe. *Pancreatol.* (2017) 17:155–65. doi: 10.1016/j.pan.2017.01.005
 39. Bai X, Jin M, Zhang H, Lu B, Yang H, Qian J. Evaluation of Chinese updated guideline for acute pancreatitis on management of moderately severe and severe acute pancreatitis. *Pancreatol.* (2020) 20:1582–6. doi: 10.1016/j.pan.2020.09.013
 40. Guda NM, Muddana V, Whitcomb DC, Levy P, Garg P, Cote G, et al. Recurrent acute pancreatitis: international state-of-the-science conference with recommendations. *Pancreas.* (2018) 47:653–66. doi: 10.1097/MPA.0000000000001053
 41. Sankaran SJ, Xiao AY, Wu LM, Windsor JA, Forsmark CE, Petrov MS. Frequency of progression from acute to chronic pancreatitis and risk factors: a meta-analysis. *Gastroenterology.* (2015) 149:1490–500.e1. doi: 10.1053/j.gastro.2015.07.066
 42. Kirkegård J, Mortensen FV, Cronin-Fenton D. Chronic pancreatitis and pancreatic cancer risk: a systematic review and meta-analysis. *Am J Gastroenterol.* (2017) 112:1366–72. doi: 10.1038/ajg.2017.218
 43. Magnusdottir BA, Baldursdottir MB, Kalaitzakis E, Björnsson ES. Risk factors for chronic and recurrent pancreatitis after first attack of acute pancreatitis. *Scand J Gastroenterol.* (2019) 54:87–94. doi: 10.1080/00365521.2018.1550670
 44. Yu B, Li J, Li N, Zhu Y, Chen Y, He W, et al. Progression to recurrent acute pancreatitis after a first attack of acute pancreatitis in adults. *Pancreatol.* (2020) 20: 1340–6. doi: 10.1016/j.pan.2020.09.006
 45. Sun Y, Jin J, Zhu A, Hu H, Lu Y, Zeng Y, et al. Risk factors for recurrent pancreatitis after first episode of acute pancreatitis. *Int J Gen Med.* (2022) 15:1319–28. doi: 10.2147/IJGM.S344863
 46. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis-2012: revision of the Atlanta classification and definitions by international consensus. *Gut.* (2013) 62:102–11. doi: 10.1136/gutjnl-2012-302779
 47. Kuo DC, Rider AC, Estrada P, Kim D, Pillow MT. Acute pancreatitis: what's the score? *J Emerg Med.* (2015) 48:762–70. doi: 10.1016/j.jemermed.2015.02.018
 48. Van den Berg FF, de Bruijn AC, van Santvoort HC, Issa Y, Boermeester MA. Early laboratory biomarkers for severity in acute pancreatitis: a systematic review and meta-analysis. *Pancreatol.* (2020) 20:1302–11. doi: 10.1016/j.pan.2020.09.007
 49. Yan G, Li H, Bhethuwal A, McClure MA, Li Y, Yang G, et al. Pleural effusion volume in patients with acute pancreatitis: a retrospective study from three acute pancreatitis centers. *Ann Med.* (2021) 53:2003–18. doi: 10.1080/07853890.2021.1998594
 50. Zhou T, Chen Y, Wu JL, Deng Y, Zhang J, Sun H, et al. Extrapneumonic inflammation on magnetic resonance imaging for the early prediction of acute pancreatitis severity. *Pancreas.* (2020) 49:46–52. doi: 10.1097/MPA.0000000000001425
 51. Khurana A, Nelson LW, Myers CB, Akisik F, Jeffrey BR, Miller FH, et al. Reporting of acute pancreatitis by radiologists-time for a systematic change with structured reporting template. *Abdom Radiol.* (2020) 45:1277–89. doi: 10.1007/s00261-020-02468-9
 52. Zaky AM, Wolfgang CL, Weiss MJ, Javed AA, Fishman EK, Zaheer A. Tumor-vessel relationships in pancreatic ductal adenocarcinoma at multi detector CT: different classification systems and their influence on treatment planning. *Radiographics.* (2017) 37:93–112. doi: 10.1148/rg.2017160054
 53. Schawkat K, Manning MA, Glickman JN, Morteale KJ. Pancreatic ductal adenocarcinoma and its variants: pearls and perils. *Radiographics.* (2020) 40:1219–39. doi: 10.1148/rg.2020190184
 54. Schima W, Böhm G, Rösch CS, Klaus A, Függer R, Kopf H. Mass-forming pancreatitis versus pancreatic ductal adenocarcinoma: CT and MR imaging for differentiation. *Cancer Imaging.* (2020) 20:52. doi: 10.1186/s40644-020-00324-z
 55. Kothari K, Lopes Vendrami C, Kelahan LC, Shin JS, Mittal P, Miller FH. Inflammatory mimickers of pancreatic adenocarcinoma. *Abdom Radiol.* (2020) 45:1387–96. doi: 10.1007/s00261-019-02233-7
 56. Wolske KM, Ponnatapura J, Kolokythas O, Burke LMB, Tappouni R, Lalwani N. Chronic pancreatitis or pancreatic tumor? A problem-solving approach. *Radiographics.* (2019) 39:1965–82. doi: 10.1148/rg.2019190011
 57. Elsherif SB, Virarkar M, Javadi S, Ibarra-Rovira JJ, Tamm EP, Bhosale PR. Pancreatitis and PDAC: association and differentiation. *Abdom Radiol.* (2020) 45:1324–37. doi: 10.1007/s00261-019-02292-w
 58. Jia H, Li J, Huang W, Lin G. Multimodal magnetic resonance imaging of mass-forming autoimmune pancreatitis: differential diagnosis with pancreatic ductal adenocarcinoma. *BMC Med Imaging.* (2021) 21:149. doi: 10.1186/s12880-021-00679-0
 59. Tanaka H, Matsusaki S. The Utility of endoscopic-ultrasonography-guided tissue acquisition for solid pancreatic lesions. *Diagnostics.* (2022) 12:753. doi: 10.3390/diagnostics12030753
 60. DelMaschio A, Vanzulli A, Sironi S, Castrucci M, Mellone R, Staudacher C, et al. Pancreatic cancer versus chronic pancreatitis: diagnosis with CA 19-9 assessment, US, CT, and CT-guided fine-needle biopsy. *Radiology.* (1991) 178:95–9. doi: 10.1148/radiology.178.1.1984331
 61. Shimosegawa T, Chari ST, Frulloni L, Kamisawa T, Kawa S, Mino-Kenudson M, et al. International consensus diagnostic criteria for autoimmune pancreatitis: guidelines of the International Association of Pancreatolgy. *Pancreas.* (2011) 40:352–8. doi: 10.1097/MPA.0b013e3182142fd2
 62. Masamune A, Kikuta K, Hamada S, Tsuji I, Takeyama Y, Shimosegawa T, et al. Nationwide epidemiological survey of autoimmune pancreatitis in Japan in 2016. *J Gastroenterol.* (2020) 55:462–70. doi: 10.1007/s00535-019-01658-7
 63. Okazaki K, Kawa S, Kamisawa T, Ikeura T, Itoi T, Ito T, et al. Amendment of the Japanese consensus guidelines for autoimmune pancreatitis, 2020. *J Gastroenterol.* (2022) 57:225–45. doi: 10.1007/s00535-022-01857-9
 64. Vlachou PA, Khalili K, Jang HJ, Fischer S, Hirschfield GM, Kim TK. IgG4-related sclerosing disease: autoimmune pancreatitis and extrapancreatic manifestations. *Radiographics.* (2011) 31:1379–402. doi: 10.1148/rg.315105735
 65. Martínez-de-Alegría A, Baleato-González S, García-Figueiras R, Bermúdez-Naveira A, Abdulkader-Nallib I, Díaz-Peromingo JA, et al. IgG4-related disease from head to toe. *Radiographics.* (2015) 35:2007–25. doi: 10.1148/rg.357150066
 66. Umehara H, Okazaki K, Nakamura T, Satoh-Nakamura T, Nakajima A, Kawano M, et al. Current approach to the diagnosis of IgG4-related disease - combination of comprehensive diagnostic and organ-specific criteria. *Mod Rheumatol.* (2017) 27:381–91. doi: 10.1080/14397595.2017.1290911
 67. Nour E, Hammami A, Missaoui N, Bdioui A, Dahmani W, Ameur BW, et al. Multi-organ involvement of immunoglobulin g4-related disease. *Gastroenterol. Insights.* (2021) 12:350–7. doi: 10.3390/gastroent12030033
 68. Rogers W, Thulasi Seetha S, Refaee TAG, Lieverse RY, Granzier RY, Ibrahim A, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol.* (2020) 93:20190948. doi: 10.1259/bjr.20190948
 69. Avery E, Sanelli PC, Aboian M, Payabvash S. Radiomics: a primer on processing workflow and analysis. *Semin Ultrasound CT MR.* (2022) 43:142–6. doi: 10.1053/j.sult.2022.02.003
 70. Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, et al. Implementing the FAIR data principles in precision oncology: review of supporting initiatives. *Brief Bioinform.* (2020) 21:936–45. doi: 10.1093/bib/bbz044
 71. Van Soest J, Meldolesi E, van Stiphout R, Gatta R, Damiani A, Valentini V, et al. Prospective validation of pathologic complete response models in rectal cancer: transferability and reproducibility. *Med Phys.* (2017) 44:4961–7. doi: 10.1002/mp.12423
 72. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5
 73. Lee SH, Park H, Ko ES. Radiomics in breast imaging from techniques to clinical applications: a review. *Korean J Radiol.* (2020) 21:779–92. doi: 10.3348/kjr.2019.0855

74. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys.* (2018) 102:1143–58. doi: 10.1016/j.ijrobp.2018.05.053
75. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intellig.* (2020) 2:305–11. doi: 10.1038/s42256-020-0186-1
76. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med.* (2020) 3:119. doi: 10.1038/s41746-020-00323-1
77. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrta A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin.* (2019) 69:127–57. doi: 10.3322/caac.21552
78. Tomaszewski MR, Gillies RJ. The biological meaning of radiomic features. *Radiology.* (2021) 298:505–16. doi: 10.1148/radiol.2021202553
79. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? *Ann Intern Med.* (2020) 172:59–60. doi: 10.7326/M19-2548
80. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. *Cancer Cell.* (2021) 39:916–27. doi: 10.1016/j.ccell.2021.04.002
81. Ghandili S, Shayesteh S, Fouladi DF, Blanco A, Chu LC. Emerging imaging techniques for acute pancreatitis. *Abdom Radiol.* (2020) 45:1299–307. doi: 10.1007/s00261-019-02192-z
82. Parakh A, Tirkes T. Advanced imaging techniques for chronic pancreatitis. *Abdom Radiol.* (2020) 45:1420–38. doi: 10.1007/s00261-019-02191-0
83. Gorris M, Hoogenboom SA, Wallace MB, van Hooft JE. Artificial intelligence for the management of pancreatic diseases. *Dig Endosc.* (2021) 33:231–41. doi: 10.1111/den.13875
84. Goyal H, Mann R, Gandhi Z, Perisetti A, Zhang Z, Sharma N, et al. Application of artificial intelligence in pancreaticobiliary diseases. *Ther Adv Gastrointest Endosc.* (2021) 14:2631774521993059. doi: 10.1177/2631774521993059
85. Tong T, Gu J, Xu D, Song L, Zhao Q, Cheng F, et al. Deep learning radiomics based on contrast-enhanced ultrasound images for assisted diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis. *BMC Med.* (2022) 20:74. doi: 10.1186/s12916-022-02258-8
86. Ziegelmayer S, Kaissis G, Harder F, Jungmann F, Müller T, Makowski M, et al. Deep convolutional neural network-assisted feature extraction for diagnostic discrimination and feature visualization in pancreatic ductal adenocarcinoma (PDAC) versus autoimmune pancreatitis (AIP). *J Clin Med.* (2020) 9:4013. doi: 10.3390/jcm9124013
87. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* (2020) 295:328–38. doi: 10.1148/radiol.2020191145

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yan, Yan, Li, Liang, Peng, Bhetuwal, McClure, Li, Yang, Li, Zhao and Fan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Chunxue Bai,
Fudan University, China

REVIEWED BY

Jinghao Duan,
Shandong First Medical University
and Shandong Academy of Medical
Sciences, China
Mei Yuan,
Nanjing Medical University, China
Xingguan Yang,
Guilin People's Hospital, China

*CORRESPONDENCE

Jun Liu
junliu123@csu.edu.cn
Ming Li
ming_li@fudan.edu.cn

†These authors have contributed
equally to this work and share first
authorship

SPECIALTY SECTION

This article was submitted to
Pulmonary Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 09 May 2022

ACCEPTED 13 October 2022

PUBLISHED 04 November 2022

CITATION

Sun Y, Zhao W, Kuang K, Jin L, Gao P,
Duan S, Xiao Y, Liu J and Li M (2022)
Non-contrast and contrast enhanced
computed tomography radiomics
in preoperative discrimination of lung
invasive and non-invasive
adenocarcinoma.
Front. Med. 9:939434.
doi: 10.3389/fmed.2022.939434

COPYRIGHT

© 2022 Sun, Zhao, Kuang, Jin, Gao,
Duan, Xiao, Liu and Li. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Non-contrast and contrast enhanced computed tomography radiomics in preoperative discrimination of lung invasive and non-invasive adenocarcinoma

Yingli Sun^{1†}, Wei Zhao^{2†}, Kaiming Kuang³, Liang Jin¹,
Pan Gao¹, Shaofeng Duan⁴, Yi Xiao⁵, Jun Liu^{2*} and Ming Li^{1*}

¹Department of Radiology, Huadong Hospital Affiliated to Fudan University, Shanghai, China,

²Department of Radiology, Second Xiangya Hospital, Central South University, Changsha, China,

³Dianei Technology, Shanghai, China, ⁴GE Healthcare, Shanghai, China, ⁵Department of Radiology, Changzheng Hospital, Second Military Medical University, Shanghai, China

Objective: This study aimed to assess the value of radiomics based on non-contrast computed tomography (NCCT) and contrast-enhanced computed tomography (CECT) images in the preoperative discrimination between lung invasive adenocarcinomas (IAC) and non-invasive adenocarcinomas (non-IAC).

Methods: We enrolled 1,185 pulmonary nodules (478 non-IACs and 707 IACs) to build and validate radiomics models. An external testing set comprising 63 pulmonary nodules was collected to verify the generalization of the models. Radiomic features were extracted from both NCCT and CECT images. The predictive performance of radiomics models in the validation and external testing sets were evaluated and compared with radiologists' evaluations. The predictive performances of the radiomics models were also compared between three subgroups in the validation set (Group 1: solid nodules, Group 2: part-solid nodules, and Group 3: pure ground-glass nodules).

Results: The NCCT, CECT, and combined models showed good ability to discriminate between IAC and non-IAC [respective areas under the curve (AUCs): validation set = 0.91, 0.90, and 0.91; Group 1 = 0.82, 0.79, and 0.81; Group 2 = 0.93, 0.92, and 0.93; and Group 3 = 0.90, 0.90, and 0.89]. In the external testing set, the AUC of the three models were 0.89, 0.91, and 0.89, respectively. The accuracies of these three models were comparable to those of the senior radiologist and better than those of the junior radiologist.

Conclusion: Radiomic models based on CT images showed good predictive performance in discriminating between lung IAC and non-IAC, especially in part solid nodule group. However, radiomics based on CECT images provided no additional value compared to NCCT images.

KEYWORDS

adenocarcinoma, lung, radiomics, solitary pulmonary nodule, X-ray computed tomography

Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths worldwide (1). Despite the recent development of targeted therapies for selected sub-types of lung adenocarcinoma, the overall cure and survival rates for this cancer remain relatively low (2). Adenocarcinoma is the most common form of lung cancer and has recently been classified into pre-invasive adenocarcinoma [atypical adenocarcinoma hyperplasia (AAH), adenocarcinoma *in situ* (AIS)], minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC) (3). The 5-year disease-free survival rates in AIS and MIA are 100% or close to 100%, which are significantly higher than that in IAC (38–86%, depending on the predominant histological subtypes) (4, 5). Therefore, the accurate preoperative diagnosis of lung adenocarcinoma is critical for clinical decision-making processes and the assessment of prognoses.

Due to the diversity and overlap of radiographic features of these lesions, diagnosing and differentiating lung IAC is challenging for radiologists. Radiomics is an emerging method that can extract many features to facilitate the precision medicine (6). Many studies have explored the value of radiomics in the detection, characterization, and monitoring of lung nodules, resulting in promising performance (7–9). However, those studies focused on the radiomic features extracted from non-contrast CT (NCCT) images. The National Comprehensive Cancer Network (NCCN) recommends contrast-enhanced CT (CECT) examinations for some lung nodules: solid nodules > 15 mm on initial screening, part solid nodules with solid components > 8 mm in initial screening, new or increased solid nodules \geq 8 mm during the follow-up, new or increased part-solid nodules with solid components > 1.5 mm during the follow-up (10). The CECT images can yield better vascular information and improve the accuracy of the diagnoses. Several studies have assessed the value of radiomics based on CECT images in the diagnosis of pulmonary nodules (9, 11–14), but their conclusions are inconsistent. Moreover, whether

the radiomics extracted from CECT images can provide **Supplementary information** for differentiation of IAC from non-IAC remains unknown, especially for different types nodules (i.e., solid nodules, part-solid nodules, and pure ground-glass nodules).

Therefore, this study assessed the value of radiomics based on NCCT and CECT images to discriminate between IAC and non-IAC and compared the performances of models of different nodule subtypes (solid nodules, part-solid nodules and pure ground-glass nodules).

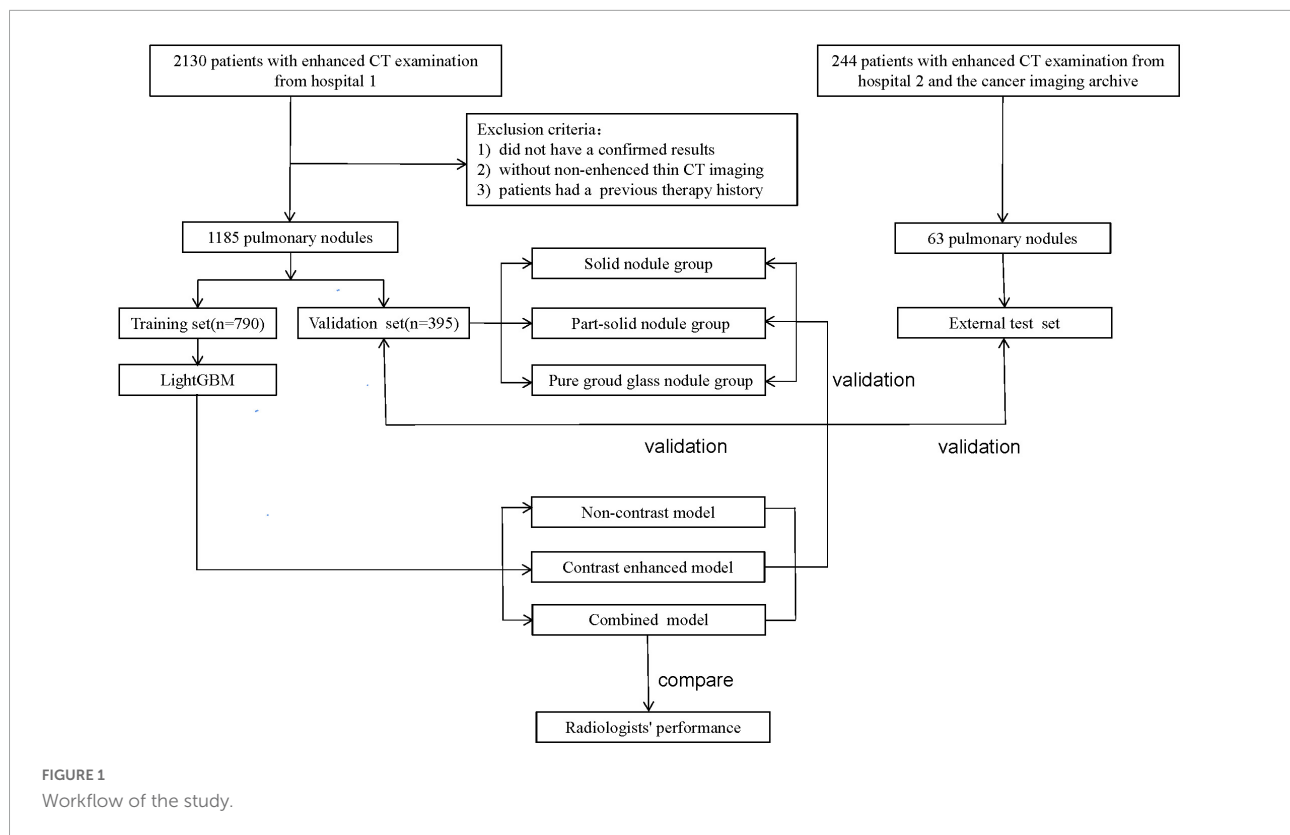
Materials and methods

Our institutional review board approved this retrospective study (No. 2019K134) and waived the requirement of obtaining informed consent from patients.

Study population

A total of 2,130 patients who underwent CECT examinations for pulmonary nodules between January 2014 and January 2019 were selected. Their medical records were reviewed for clinical characteristics, histopathological results, and serial chest CT scans. The inclusion criteria were as follows: (1) the presence of a pulmonary nodule; (2) histopathologically confirmed benign nodules, AAH, AIS, MIA, or IAC, or confirmed follow-up for inflammatory lesions; (3) NCCT and CECT scans were available and acquired sequentially in one examination; and (4) CT slice thickness \leq 1.25 mm. The exclusion criteria were: (1) prior treatment before surgery; (2). poor quality CT images, and (3). lesions that were difficult to delineate clearly.

Another 63 lung nodules met the inclusion and exclusion criteria were collected as external testing set to validate the stability and generalization of the models. Among the 63 nodules, 22 were selected from the cancer imaging archive (15) and 41 were collected from the Second Xiangya Hospital



of Central South University. The workflow is described in [Figure 1](#).

Computed tomography scanning

Chest CT scanning was performed using one of following the four CT systems: GE Discovery CT750 HD, 64-slice LightSpeed VCT (both from GE Medical Systems), Somatom Definition Flash, and Somatom Sensation-16 (both from Siemens Medical Solutions). The detailed scan and reconstruction parameters are listed in [Table 1](#). All patients received a bolus of 80–100 mL of intravenous contrast medium (Optiray; Mallinckrodt Imaging, MO, USA; 350 mg iodine per mL) at a rate of 3–4 mL/s using a power injector *via* an 18- or 20-gauge cannula into the antecubital vein. Enhanced CT scanning commenced 50–60 s after the administration of the contrast medium.

Pathological analysis

All resected specimens were formalin-fixed and stained with hematoxylin–eosin in accordance with the routine regulations of the hospital. A pathologist (with 10 years of experience in the pathological diagnosis of lung cancer) reviewed the specimens and recorded the pathological subtype of each nodule.

Nodule labeling and segmentation

One radiologist with 5 years of experience in chest CT interpretation used a medical image processing and navigation software, 3D Slicer (version 4.8; National Institutes of Health)¹, to manually delineate the volumes of interest of the 1,185 nodules at the voxel level in separate NCCT and CECT images. The volume of interest was confirmed by another radiologist with 12 years of experience in chest CT interpretation. DICOM images were imported into the software for delineation, and the label information was extracted with the nearly raw raster data format for further analysis. Each segmented nodule was given a specific label, non-IAC (inflammatory nodule, benign tumor, AAH, AIS, or MIA) or IAC. To assess the segmentation variability, a third radiologist with 3 years of experience in chest CT interpretation independently segmented a random set of 60 nodules to calculate the intra-class correlation coefficient (ICC) for each radiomic feature.

Observer study

Two radiologists (a junior and a senior radiologist with more than 3 and 10 years of experience, respectively), who

¹ <https://www.slicer.org>

TABLE 1 Detailed scan and reconstruction parameters.

Setting	Tube voltage (kV)	Tube current (mA)	Pitch	Slice thickness of reconstruction (mm)	Slice interval of reconstruction (mm)	Reconstruction algorithm
GE Discovery CT750 HD	120	200	0.984:1	1.25	1.25	STND
Lightspeed VCT	120	200	0.984:1	1.25	1.25	STND
Somatom definition flash	120	110	1	1	1	Medium sharp
Somatom sensation-16	120	110	0.8	1	1	Medium sharp

were blinded to the histopathological results and clinical data, independently classified and diagnosed all nodules in the validation set and external testing set. First, the two radiologists categorized the nodules as IAC or non-IAC based on the NCCT images, they then accessed to the folder containing the CECT images and diagnosed the nodules again using both the NCCT and CECT images.

Extraction of radiomic features

Radiomic features were extracted using PyRadiomics 2.2.0² (16), an open-source Python package for the extraction of radiomics. The process of extracting radiomic features is described in [Supplementary material](#). To minimize the effect of image heterogeneity, we normalized the image spatial resolution and voxels before radiomic features extraction. A total of 1,218 features were extracted, including shape class, first-order class, gray level co-occurrence matrix (GLCM) class, gray level dependence matrix (GLDM) class, gray level size zone matrix (GLSM) class, and gray level run length matrix (GLRLM) class. We also used Min-Max scaling to normalize features before model construction. For feature variability analysis, the ICC for each radiomic feature was calculated using a two-way random-effects model under an absolute agreement condition. The reproducibility of the radiomic features was considered to be either high ($ICC \geq 0.8$), intermediate ($0.5 \leq ICC < 0.8$), or poor ($ICC < 0.5$). The radiomic features with high reproducibility were used as the input variables for building the diagnostic models.

Building and validation of the diagnostic models

All patients were randomly assigned to a training set ($n = 790$) or a validation set ($n = 395$) at a ratio of 2:1 using the “scikit-learn” software packages for Python (17). The validation set was further divided into three subgroups, 91 solid nodules (Group 1), 239 part-solid nodules (Group 2), and 65 pure ground-glass nodules (Group 3). The distribution of

different nodules properties (non-IAC vs. IAC, solid nodules vs. part-solid nodules vs. pure ground glass nodules) was kept uniform in both the training set and the validation set. After assessing the reproducibility based on the re-segmentation data, the open-source framework LightGBM was used for feature selection and model building in the training set (15). LightGBM is a fast, distributed, and efficient gradient boosting framework based on decision tree algorithms. Finally, the NCCT, CECT, and combined models differentiating between non-IAC and IAC were established. The performances of these models were then tested in the validation set (also in three subgroups) and external testing set.

Statistical analysis

Differences in variables between the two patient groups were assessed using the independent-sample *t*-test or Mann–Whitney *U*-test for continuous variables and Fisher’s exact test or the chi-squared test for categorical variables. To assess the predictive performance of the study variables, receiver-operating characteristic (ROC) curves were plotted for the study variables to assess their predictive performance and compared using the DeLong test and the area under the curve (AUC) of the ROC curve was calculated. A two-sided *p*-value < 0.05 was considered statistically significant. Statistical analysis was performed using Python (Version 3.7.1) software and SPSS (Version 22.0, IBM).

Results

Patient profiles

A total of 1,185 nodules from 1,185 patients in our hospital were enrolled. Among the 1,185 patients, 690 were women (58.2%) and 495 were men (41.8%). The mean age of the patients was 58.95 ± 11.45 years (range: 20–81 years); the maximum diameter of the pulmonary nodules was 18.79 ± 11.32 mm (range: 5–82 mm). There were 478 (40.3%) nodules were diagnosed as non-IAC (123 inflammation or benign tumor; 11 AAH; 84 AIS; 260 MIA), and 707 (59.7%) IAC. Among the 1,185 nodules, 273 (23.0%) were solid nodules, 717 (60.5%) were part-solid nodules, and 195 (16.5%) were pure ground glass nodules.

² <https://pyradiomics.readthedocs.io/en/latest/index.html>

The patient information of the training set, validation set and external set are shown in [Table 2](#). The patient information of the three subgroups are shown in [Supplementary Table 1](#). Of the 63 pulmonary nodules in the external testing set, 22 were non-IAC and 41 were IAC. There were 28 (44.4%) solid nodules, 26 (41.3%) part-solid nodules and 9 (14.3) pure ground-glass nodules.

Model building and diagnostic validation

After reproducibility analysis, 534 features on NCCT and 559 features on CECT remained separate ($ICCs \geq 0.8$), and the details are shown in [Supplementary Tables 2, 3](#). The selected features were inputted into the LightGBM framework to construct the NCCT, CECT and combined models. LightGBM ranked the importance of features based on the number of times they were used in the decision tree.

In the validation set, the AUCs of the NCCT, CECT, and combined models were 0.91, 0.90 and 0.91 respectively, to distinguish IAC and non-IAC cases ([Figure 2A](#)). The DeLong test found no statistically significant difference among the three models (NCCT model vs. CECT model, $P = 0.247$; NCCT model vs. combined model, $P = 0.320$; CECT model

vs. combined model, $P = 0.277$). In the external testing set, the AUCs of the NCCT, CECT, and combined models were 0.89, 0.91, and 0.89, respectively ([Figure 2B](#)). Again, no statistically significant differences among the three models were identified by the DeLong test (NCCT model vs. CECT model, $P = 0.218$; NCCT model vs. combined model, $P = 0.436$; and CECT model vs. combined model, $P = 0.148$). The accuracies of the radiomics models were close to those of the senior radiologist and better than those of the junior radiologist for both the validation set and external testing set ([Table 3](#)).

Performance of the models in the subgroups

In Group 1, the AUCs of the NCCT, CECT, and combined models were 0.82, 0.79, and 0.81, respectively, without significant difference in the DeLong test (NCCT model vs. CECT model, $P = 0.247$; NCCT model vs. combined model, $P = 0.320$; and CECT model vs. combined model, $P = 0.277$) ([Figure 3A](#)). The accuracies of the radiomics models were slightly better than that of the junior radiologist but significantly lower than that of the senior radiologist ([Table 3](#)). In Group 2, the AUCs of the NCCT, CECT, and combined model were

TABLE 2 Patient information of the training set, validation set and external set.

Demographic and clinical characteristic	Training set ($n = 790$)	Validation set ($n = 395$)	p	External validation set ($n = 63$)
Age (years)	58.89 \pm 11.23	59.08 \pm 11.83	0.789	60.05 \pm 10.25
Size (mm)	18.56 \pm 10.75	18.91 \pm 10.53	0.598	22.8 \pm 10.92
Gender			0.708	
Female	463 (58.6)	227 (57.5)		31 (49.2)
Male	327 (41.4)	168 (42.5)		32 (50.8)
Pathology			0.933	
IAC	474 (60.0)	233 (59.0)		41 (65.1)
Non-IAC				
Benign lesions	83 (10.5)	40 (10.1)		5 (7.9)
AAH	9 (1.1)	2 (0.5)		0
AIS	49 (6.2)	35 (8.9)		4 (6.3)
MIA	175 (22.2)	85 (21.5)		13 (20.6)
Type			1.000	
Pure ground glass nodule	130 (16.5)	91 (23.0)		9 (14.3)
Part-solid nodule	478 (60.5)	239 (60.5)		26 (41.3)
Solid nodule	182 (23.0)	65 (16.5)		28 (44.4)
Location			0.585	
Right upper lobe	274 (34.1)	135 (34.2)		20 (31.7)
Right middle lobe	64 (8.1)	39 (9.9)		3 (4.8)
Right lower lobe	153 (19.4)	78 (19.7)		10 (15.9)
Left lower lobe	191 (24.2)	99 (25.1)		17 (27.0)
Left lower lobe	108 (13.7)	44 (11.5)		13 (20.6)

IAC, invasive adenocarcinoma; AAH, atypical adenocarcinoma hyperplasia; AIS, adenocarcinoma *in situ*; MIA, minimally invasive adenocarcinoma.

TABLE 3 Performance of the radiomics models and radiologists for lung IAC.

		Radiomics models			Junior radiologist		Senior radiologist	
		NCCT	CECT	NCCT + CECT	NCCT	NCCT + CECT	NCCT	NCCT + CECT
Validation set	Accuracy	82.74%	81.47%	83.50%	74.90%	76.90%	83.40%	83.90%
	F1	0.86	0.95	0.87				
	AUC	0.91	0.90	0.91				
Group 1	Accuracy	74.73%	68.13%	74.73%	66.70%	68.90%	80.00%	84.40%
	F1	0.82	0.79	0.82				
	AUC	0.82	0.79	0.81				
Group 2	Accuracy	85.71%	86.55%	86.55%	76.30%	77.50%	83.50%	82.60%
	F1	0.90	0.90	0.90				
	AUC	0.93	0.92	0.93				
Group 3	Accuracy	83.08%	81.54%	84.62%	81.50%	82.20%	87.70%	87.70%
	F1	0.86	0.84	0.88				
	AUC	0.90	0.90	0.89				
External testing set	Accuracy	84.13%	84.13%	84.13%	75.34%	76.12%	84.45%	85.21%
	F1	0.88	0.88	0.88				
	AUC	0.89	0.91	0.89				

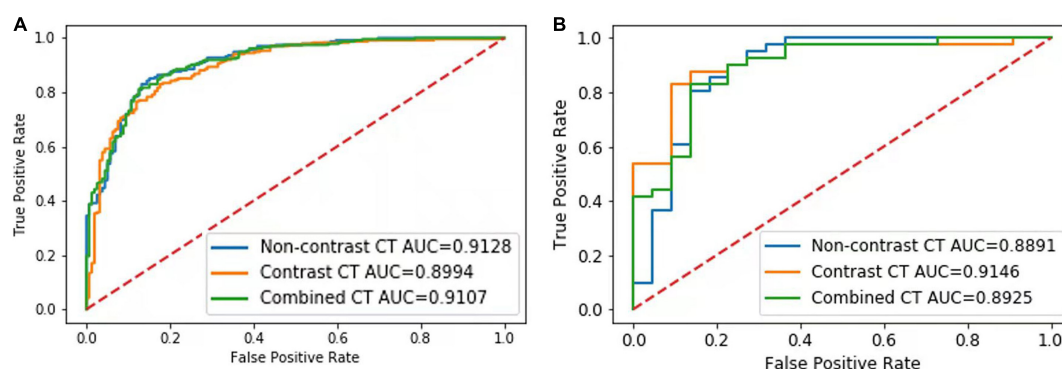


FIGURE 2

Results of the receiver-operating characteristic (ROC) curve analysis. The ROC curves of the NCCT, CECT, and combined models for identification of invasive adenocarcinoma (IAC) in the validation set (A) and external testing set (B) are shown.

0.93, 0.92, and 0.93, respectively (Figure 3B). The results of the DeLong test showed no statistically significant differences among the three models (NCCT vs. CECT model, $P = 0.159$; NCCT vs. combined model, $P = 0.402$; and CECT vs. combined model, $P = 0.160$). In this group, the accuracies of the radiomics models were better than those of the junior and senior radiologists (Table 3). In Group 3, the AUCs of the NCCT, CECT, and combined model in Group 3 were 0.90, 0.90, and 0.89, respectively (Figure 3C). The DeLong test showed no statistically significant differences among the three models (NCCT vs. CECT model, $P = 0.402$; NCCT vs. combined model, $P = 0.213$; and CECT vs. combined model, $P = 0.406$). The accuracies of the radiomics models were close to that of the junior radiologist but lower than that of the senior radiologist (Table 3).

Top 10 features of the non-contrast computed tomography, contrast-enhanced computed tomography, and combined models

The LightGBM framework ranked the importance of features according to the number of times they were used in the decision tree. The top 10 features of the models were listed in Figure 4. Most of the features were different, and only one feature (wavelet_gldm_DependenceEntropy) was same between the top 10 features of the NCCN model and CECT model. Seven of the combined model's top 10 features were from NCCT images and three features were from CECT images. Only six of the combined model's top 10 features appeared in the NCCT and

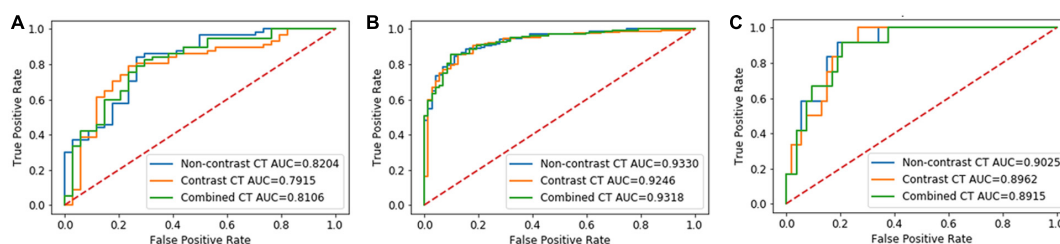


FIGURE 3

Results of the receiver-operating characteristic (ROC) curve analysis. The ROC curves of the NCCT, CECT, and combined models for identification of invasive adenocarcinoma (IAC) in solid nodule group (A), partly solid nodule group (B), and pure ground glass nodule group (C) are shown.

CECT models. Of all three models' top 10 features, thirteen were from GLSZM, seven were from GLCM, five were from GLDM, three were from first-order, one was from shape and one was from GLRLM separately. Of the thirteen features from GLSZM, four were in the NCCT model, four were in the NCCT model and five were in the combined model.

Discussion

This study investigated the value of radiomic features extracted from NCCT and CECT images in the diagnosis of IAC/non-IAC. The radiomics models showed good predictive performance in discriminating between IAC and non-IAC of the lung, especially those with in part-solid nodules. Generally, the accuracies of the radiomic models were close to that of the senior radiologist and better than that of the junior radiologist. However, the radiomic models based on CECT images provided no additional value compared to the NCCT models.

To date, several studies have documented that CT-based radiomics can identify lung IAC with AUCs of 0.77–0.90 (18). Our NCCT model also obtained good performance (AUC = 0.91 in validation set), verifying the ability of CT based radiomics for identifying IAC. However, previous radiomics studies were rarely based on CECT images. Recently, radiomics extracted from CECT images were investigated, however, the results were inconsistent. Chen et al. demonstrated that the radiomics model based on CECT could provide additional value in the prediction of invasiveness of subcentimeter ground glass nodules (AUC_CECT: 0.896 vs. AUC_NCCT: 0.851) (19). In the study of Fan et al., a radiomics model was constructed using NCCT images for IAC prediction showed similar performance in NCCT validation set and CECT validation set (7). This result suggests that contrast injection did not affect the two features included in their radiomics model (i.e., GLCM_correlation and GLCM_cluster_tendency). Other studies also constructed radiomics models separately based on NCCT and CECT images and compared their performance for predicting lung IAC. Gao et al. enrolled 34

IACs that appeared as ground glass nodules and constructed models using multivariate logistic regression analysis (14). Their results also suggested that CECT did not improve the performance of the radiomics model. For solid nodules, Yang et al. (18) constructed radiomics models for differentiating granulomatous nodules from lung adenocarcinoma; they came to the same conclusion. Our result showed that the NCCT, CECT, and combined model achieved similar performance for identifying lung IAC. In subgroups, the AUCs of the three models also showed no statistically significant difference. Our study enrolled pure ground-glass nodules, partly solid nodules, and solid nodules, built models that merged the three types of nodules, and validated them in three subgroups. To minimize interference factors caused by multiple scans (such as CT scanners and protocols), we excluded nodules whose NCCT and CECT images were not acquired in one examination. Our results suggested that CECT did not improve the radiomics performance for lung IAC prediction either in solid nodules or ground glass nodules. We considered the possible reasons were: (1) The existence of contrast agents within the tumor may reduce the biological heterogeneity that facilitates the differentiation between benign and malignant nodules. (2) Calibration before model building might reduce the image intensity.

In subgroup analysis, although there was no statistically difference between the AUCs of the radiomics models within the groups, there was a significant difference between groups. The performances of the models were significantly lower in solid nodule group than those in the part-solid nodule group and pure ground-glass nodule group. This result is with that of other studies, although they only included one type of nodules. While Wu et al. (20) showed that a radiomics model for the prediction of lung IAC (part solid nodules) obtained an AUC of 0.88., Yang et al. (18) reported that radiomics models for differentiating solitary granulomatous with solid IAC achieved low AUCs (AUC_NNCT = 0.78, AUC_CECT = 0.77, and AUC_combined = 0.80). In another study (21), a radiomics model achieved an AUC of 0.967 for differentiating solid lung adenocarcinoma from benign lesions; this is obviously better

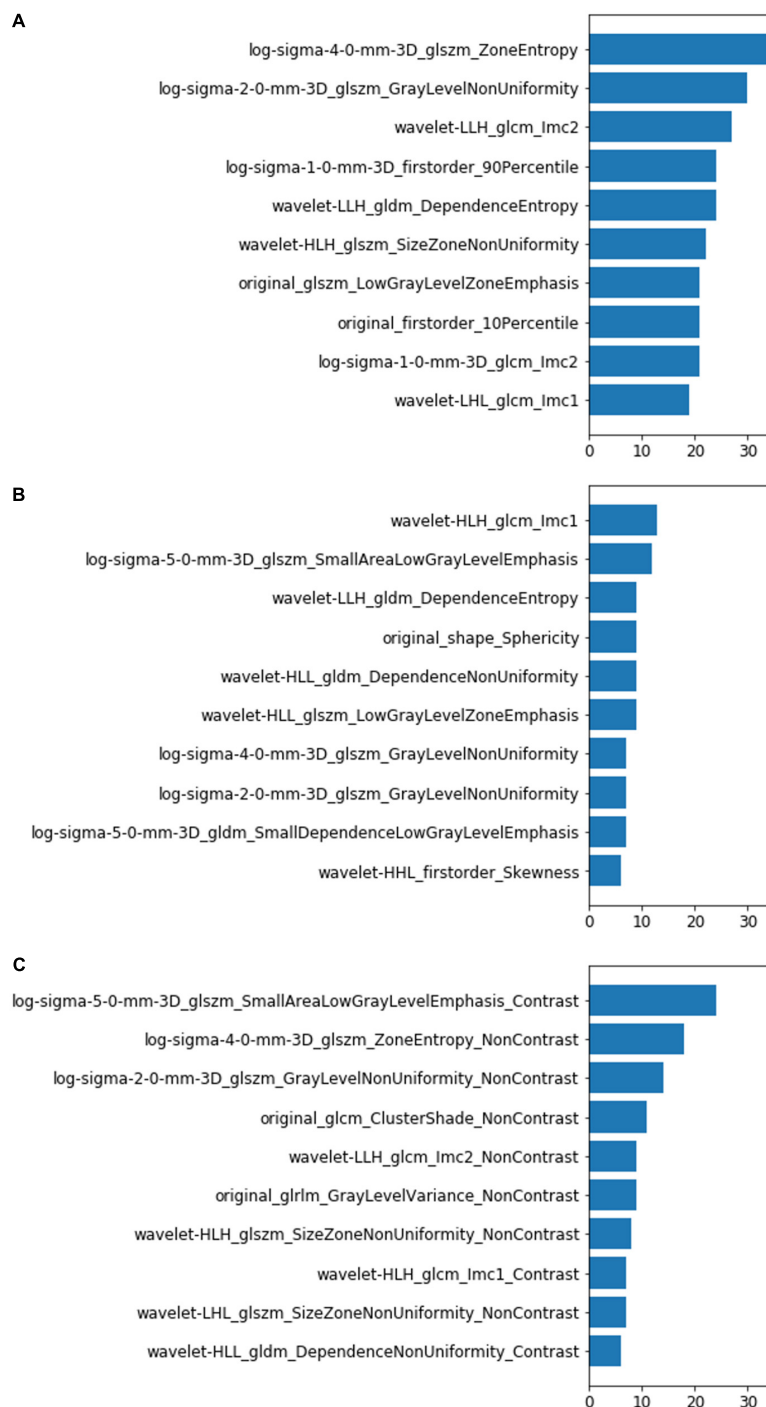


FIGURE 4

Top 10 most-used features of the NCCT model (A), CECT model (B), and combined model (C). The left vertical coordinates indicate the radiomic features; the horizontal coordinates indicate the number of times the features were used in the models.

than our result. A possible reason for this is that solid nodules only represented only a small percentage of our training set, and the model cannot generate diagnostic information. In addition, we found that the accuracies of the radiomics models were both superior to those of the junior radiologist and senior radiologist

for the part solid nodule group. The solid components in nodules, which are crucial for identifying IAC (GGN) with ground glass nodules, are diverse pathologically and include mucus, hemorrhage, mucus, granulation tissue, and alveolar collapse. It is rather difficult for radiologists to differentiate these

solid components in many cases, but some invisible radiomic feature may reflect their differences.

Although the performances of the NCCT and CECT models were similar, the top features they used differed greatly. This suggests that the contrast agent changed many radiomic features and affected their predictive power. In the combined model, more features were from NCCT (7/10) than from CECT (3/10); this phenomena may explain why CECT did not improve the model performance. Among the three model' top 10 features, 13/30 were from GLSZM class (4 in NCCT model, 4 in CECT model and 5 in combined model). GLSZM quantifies gray level zones in an image, which is defined as the number of connected voxels that share the same gray level intensity. This may indicate that GLSZM features are more stable and critical for lung IAC prediction.

This study has several limitations. First, it was limited by its retrospective nature. The heterogeneity of imaging protocols and image quality may have affected the result. Second, we did not validate the performance of models in subgroups of external set due to the limited data; therefore the subgroup results need to be confirmed. Third, the malignant group comprised only adenocarcinoma; thus, the results of this study cannot address the situation in other pulmonary malignant tumors.

In conclusion, the CT image based radiomics models showed good predictive performance in the diagnosis of lung invasive adenocarcinoma, especially those with part solid nodules; however, the radiomic model based on CECT images provided no additional value. In the diagnosis of pulmonary nodules, enhanced CT examinations should be selected cautiously, especially in young patients and patients with impaired renal function.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Our Institutional Review Board approved this Retrospective Study (no. 2019K134) and waived the requirement of obtaining informed consent from patients. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

YS and WZ: conceptualization, methodology, writing original draft preparation, and investigation. LJ, PG, YX, and JL: data curation and visualization. KK and SD: software and validation. ML and JL: writing—reviewing, supervision, and

editing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China 61976238 (ML), 82102157 (WZ), 81871405 (YX), the Science and Technology Planning Project of Shanghai Science and Technology Commission 22Y11910700 (ML), the Science and Technology Planning Project of Shanghai Science and Technology Commission 20Y11902900 (ML), the Shanghai “Rising Stars of Medical Talent” Youth Development Program “Outstanding Youth Medical Talents” SHWJRS [2021]-99 (ML), the Leading Talent of Huadong Hospital LIRC2202 (ML), and Academic Leaders in Health Sciences in Shanghai [2022] (ML). The Cancer Society of Shanghai SACA-CY21C12 (YS), the Clinical Research Center for Medical Imaging in Hunan Province 2020SK4001 (JL), and the Clinical Medical Technology Innovation Guidance Project in Hunan Province 2020SK53423 (WZ).

Acknowledgments

We are grateful to Jianying Li and Zegang Dong for English editing of the manuscript.

Conflict of interest

SD was employed by the company GE Healthcare.

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2022.939434/full#supplementary-material>

References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature.* (2018) 553:446–54. doi: 10.1038/nature25183
- Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol.* (2011) 6:244–85. doi: 10.1097/JTO.0b013e318206a221
- Yanagawa N, Shiono S, Abiko M, Ogata SY, Sato T, Tamura G. New Iaslc/Ats/Ers classification and invasive tumor size are predictive of disease recurrence in Stage I lung adenocarcinoma. *J Thorac Oncol.* (2013) 8:612–8. doi: 10.1097/JTO.0b013e318287c3eb
- Yoshiya T, Mimae T, Tsutani Y, Tsubokawa N, Sasada S, Miyata Y, et al. Prognostic role of subtype classification in small-sized pathologic N0 invasive lung adenocarcinoma. *Ann Thorac Surg.* (2016) 102:1668–73. doi: 10.1016/j.athoracsur.2016.04.087
- Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036
- Fan L, Fang M, Li Z, Tu W, Wang S, Chen W, et al. Radiomics signature: A biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground-glass nodule. *Eur Radiol.* (2019) 29:889–97. doi: 10.1007/s00330-018-5530-z
- Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, et al. Radiomics and Radiogenomics in lung cancer: A review for the clinician. *Lung Cancer.* (2018) 115:34–41. doi: 10.1016/j.lungcan.2017.10.015
- Beig N, Khorrami M, Alilou M, Prasanna P, Braman N, Orooji M, et al. Perinodular and intranodular Radiomic features on lung CT images distinguish adenocarcinomas from granulomas. *Radiology.* (2019) 290:783–92. doi: 10.1148/radiol.2018180910
- National Comprehensive Cancer Network. *The NCCN lung cancer screening (2020 [2019/04/14]). EB/OL* (version 1 2020). Fort Washington: NCCN (2020).
- Dennie C, Thornhill R, Sethi-Virmani V, Souza CA, Bayanati H, Gupta A, et al. Role of quantitative computed tomography texture analysis in the differentiation of primary lung cancer and granulomatous nodules. *Quant Imaging Med Surg.* (2016) 6:6–15. doi: 10.3978/j.issn.2223-4292.2016.02.01
- He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of Radiomics signature in solitary pulmonary nodule. *Sci Rep.* (2016) 6:34921. doi: 10.1038/srep34921
- Li M, Narayan V, Gill RR, Jagannathan JP, Barile MF, Gao F, et al. Computer-aided diagnosis of ground-glass opacity nodules using open-source software for quantifying tumor heterogeneity. *AJR Am J Roentgenol.* (2017) 209:1216–27. doi: 10.2214/AJR.17.17857
- Gao C, Xiang P, Ye J, Pang P, Wang S, Xu M. Can texture features improve the differentiation of infiltrative lung adenocarcinoma appearing as ground glass nodules in contrast-enhanced CT? *Eur J Radiol.* (2019) 117:126–31. doi: 10.1016/j.ejrad.2019.06.010
- Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging.* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7
- van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472.CAN-17-0339
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. *Scikit-learn: machine learning in python.* (2012). Available online at: <https://doi.org/10.48550/arXiv.1201.0490>
- Yang X, He J, Wang J, Li W, Liu C, Gao D, et al. CT-based Radiomics signature for differentiating solitary granulomatous nodules from solid lung adenocarcinoma. *Lung Cancer.* (2018) 125:109–14. doi: 10.1016/j.lungcan.2018.09.013
- Chen W, Li M, Mao D, Ge X, Wang J, Tan M, et al. Radiomics signature on CECT as a predictive factor for invasiveness of lung adenocarcinoma manifesting as subcentimeter ground glass nodules. *Sci Rep.* (2021) 11:20210211. doi: 10.1038/s41598-021-83167-3
- Wu G, Woodruff HC, Shen J, Refaee T, Sanduleanu S, Ibrahim A, et al. Diagnosis of invasive lung adenocarcinoma based on chest CT Radiomic features of part-solid pulmonary nodules: A multicenter study. *Radiology.* (2020) 297:451–8. doi: 10.1148/radiol.2020192431
- Liu J, Xu H, Qing H, Li Y, Yang X, He C, et al. Comparison of Radiomic models based on low-dose and standard-dose CT for prediction of adenocarcinomas and benign lesions in solid pulmonary nodules. *Front Oncol.* (2020) 10:634298. doi: 10.3389/fonc.2020.634298



OPEN ACCESS

EDITED BY

Igor Yakushev,
Technical University of
Munich, Germany

REVIEWED BY

Clovis Tauber,
INSERM U1253 Imagerie et Cerveau
(iBrain), France
Florence Muller,
Ghent University, Belgium

*CORRESPONDENCE

Anthime Flaus
anthime.flaus@univ-lyon1.fr

†These authors have contributed
equally to this work and share last
authorship

SPECIALTY SECTION

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 12 September 2022

ACCEPTED 21 October 2022

PUBLISHED 16 November 2022

CITATION

Flaus A, Deddah T, Reilhac A, Leiris ND,
Janier M, Merida I, Grenier T,
McGinnity CJ, Hammers A, Lartizien C
and Costes N (2022) PET image
enhancement using artificial
intelligence for better characterization
of epilepsy lesions.
Front. Med. 9:1042706.
doi: 10.3389/fmed.2022.1042706

COPYRIGHT

© 2022 Flaus, Deddah, Reilhac, Leiris,
Janier, Merida, Grenier, McGinnity,
Hammers, Lartizien and Costes. This is
an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction
in other forums is permitted, provided
the original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

PET image enhancement using artificial intelligence for better characterization of epilepsy lesions

Anthime Flaus^{1,2,3,4,5,6*}, Tahya Deddah⁶, Anthonin Reilhac⁷,
Nicolas De Leiris^{8,9}, Marc Janier^{1,2}, Ines Merida⁶,
Thomas Grenier⁴, Colm J. McGinnity⁴, Alexander Hammers^{3†},
Carole Lartizien^{4†} and Nicolas Costes^{5,6†}

¹Department of Nuclear Medicine, Hospices Civils de Lyon, Lyon, France, ²Faculté de Médecine Lyon Est, Université Claude Bernard Lyon 1, Lyon, France, ³King's College London and Guy's and St Thomas' PET Centre, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, ⁴Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, INSERM, CREATIS UMR 5220, Lyon, France, ⁵Lyon Neuroscience Research Center, INSERM U1028/CNRS UMR5292, Lyon, France, ⁶CERMEP-Life Imaging, Lyon, France, ⁷Brain Health Imaging Centre, Center for Addiction and Mental Health (CAHMS), Toronto, ON, Canada, ⁸Département of Nuclear Medicine, CHU Grenoble Alpes, University Grenoble Alpes, Grenoble, France, ⁹Laboratoire Radiopharmaceutiques Biocliniques, University Grenoble Alpes, INSERM, CHU Grenoble Alpes, Grenoble, France

Introduction: [¹⁸F]fluorodeoxyglucose ([¹⁸F]FDG) brain PET is used clinically to detect small areas of decreased uptake associated with epileptogenic lesions, e.g., Focal Cortical Dysplasias (FCD) but its performance is limited due to spatial resolution and low contrast. We aimed to develop a deep learning-based PET image enhancement method using simulated PET to improve lesion visualization.

Methods: We created 210 numerical brain phantoms (MRI segmented into 9 regions) and assigned 10 different plausible activity values (e.g., GM/WM ratios) resulting in 2100 ground truth high quality (GT-HQ) PET phantoms. With a validated Monte-Carlo PET simulator, we then created 2100 simulated standard quality (S-SQ) [¹⁸F]FDG scans. We trained a ResNet on 80% of this dataset (10% used for validation) to learn the mapping between S-SQ and GT-HQ PET, outputting a predicted HQ (P-HQ) PET. For the remaining 10%, we assessed Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Root Mean Squared Error (RMSE) against GT-HQ PET. For GM and WM, we computed recovery coefficients (RC) and coefficient of variation (COV). We also created lesioned GT-HQ phantoms, S-SQ PET and P-HQ PET with simulated small hypometabolic lesions characteristic of FCDs. We evaluated lesion detectability on S-SQ and P-HQ PET both visually and measuring the Relative Lesion Activity (RLA, measured activity in the reduced-activity ROI over the standard-activity ROI). Lastly, we applied our previously trained ResNet on 10 clinical epilepsy PETs to predict the corresponding HQ-PET and assessed image quality and confidence metrics.

Results: Compared to S-SQ PET, P-HQ PET improved PNSR, SSIM and RMSE; significantly improved GM RCs (from 0.29 ± 0.03 to 0.79 ± 0.04) and WM RCs (from 0.49 ± 0.03 to 1 ± 0.05); mean COVs were not statistically different. Visual lesion detection improved from 38 to 75%, with average RLA decreasing from 0.83 ± 0.08 to 0.67 ± 0.14 . Visual quality of P-HQ clinical PET improved as well as reader confidence.

Conclusion: P-HQ PET showed improved image quality compared to S-SQ PET across several objective quantitative metrics and increased detectability of simulated lesions. In addition, the model generalized to clinical data. Further evaluation is required to study generalization of our method and to assess clinical performance in larger cohorts.

KEYWORDS

Monte-Carlo simulation, residual network, brain, focal cortical dysplasia (FCD), clinical application, deep learning, deblurring, super resolution (SR)

Introduction

In the management of patients with epilepsy, approximately one third do not respond to medical therapy. For those with a focal onset, surgery could be their only potentially curative option (1). Identification of the epileptogenic zone (EZ), the zone where the seizure starts, is mandatory to allow planification of brain surgery. The EZ is the minimum brain tissue that needs to be resected to render the patient seizure-free, aiming at minimal functional impairment.

The presurgical evaluation workup includes history, semiology, EEG, video-EEG, and brain imaging (2). High-resolution brain magnetic resonance imaging (MRI) is the standard as it can identify structural lesions. However, in 35% of the cases, 3T MRI remains negative (3). In such cases, [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) positron emission tomography (PET) can be used to improve EZ detection (4–6). The EZ appears as glucose hypometabolism (decreased FDG uptake) on interictal FDG-PET, particularly relevant in focal cortical dysplasia type 2 (FCD2) (7–9).

However, several degrading factors, including a low signal to noise ratio (SNR) and an intrinsically limited spatial resolution of PET scanners compromise PET image quality. The low resolution of PET images results in the partial volume effect (10) which leads to the spill-over of estimated activity across different regions (11). These alterations could falsely normalize or attenuate the relative hypometabolism of the EZ, notably when it is small (such as for FCDs), limiting the detection performance of PET (12, 13). The most commonly used approaches to address the noise (denoising) and resolution (deblurring) challenges are: (1) within-reconstruction methods such as early iteration termination of the reconstruction algorithm (14) or point spread function modeling (15, 16) and (2) post-reconstruction methods, such as gaussian filtering, but as this decreases

the spatial resolution, many edge preserving alternatives were proposed (17–19). The most popular resolution recovery approaches in PET are partial volume correction (PVC) techniques but they rely on a segmented anatomical template based on MRI (20–23). Deconvolution methods that do not rely on structural information have also been proposed (24, 25). These methods partially correct the image but are still limited by the intrinsic resolution of PET physics and the statistical counting of the detection since they aim at converging to an explanatory distribution of the annihilation sites but not the emission sites of the positrons.

Artificial intelligence (AI)-based image enhancement is a very active field, but so far most of the publications focused on PET denoising rather than the deblurring problem (26). The deblurring problem involves the restoration of high-quality PET images (HQ) from lower-quality images [“standard quality (SQ)” PET images in our study] and not to restore a higher-count image from a low-count (low dose) PET image (denoising problem). Proof of concept of super-resolution PET has been validated with a 2D convolution neural network (CNN) in which the network was trained, using analytically simulated [^{18}F]FDG PET, to predict their corresponding ground truth for normal brains (27) and lung tumors (28). This network is neither a simple deconvolution algorithm nor a partial volume correction algorithm. The aim of this project was to develop a deep learning based deblurring method consisting in predicting the ground truth from the PET image to improve epilepsy lesion visualization. Originality of the method was that the training was performed from simulated data, for which the ground truth is known. In order to improve clinical translation of such methods, we created a new, realistic set of [^{18}F]FDG PET brain data using a validated Monte Carlo simulator (29–31) which were then reconstructed using Siemens e7 reconstruction tools. The 3D network trained to learn the mapping between the simulated

SQ PET (S-SQ PET) and the corresponding ground-truth HQ (GT-HQ) PET did not require anatomical input. We assessed the quality of the network-predicted HQ (P-HQ) PET. We repeated the process for simulated lesional brain PET data with cortical focal hypometabolism to simulate difficult-to-detect small EZ. Lastly, we used real PET data to illustrate the proof-of-concept that a model trained on Monte-Carlo simulated PET data is applicable on real data.

Materials and methods

Medical image data

We used an open, multi-vendor [General Electrics, Philips and Siemens 3T magnetic resonance imaging (MRI) scanners] brain MRI database, Calgary-Campinas (32), using 173 T1-weighted (T1w) 3D volumes (1 mm^3 voxels) from subjects with an average age of 53.4 ± 7.3 years (range 29–80, 50% women). Additionally, we used the publicly available database CERMEP-IDB-MRXFDG (33) which includes T1w MRI (Siemens 1.5T MRI) 3D volumes from 37 subjects (average age 38.11 ± 11.36 years; range 23–65, 54% women). It also includes 37 PET and computed tomography (CT) images from a Siemens Biograph mCT64, which we used to estimate a range of realistic FDG uptake values in brain PET as explained below. FDG PET data consisted in a static 10-min PET acquisition started 50 min after the injection of 122.3 ± 21.3 MBq of ^{18}F FDG. PET sinograms were reconstructed with Siemens' iterative ordered subset expectation maximization (OSEM) "High Definition" reconstruction, incorporating the spatially varying point spread function, with CT-based attenuation correction. To illustrate the capability of the developed AI deblurring method on clinical PET data, we used 10 datasets from epilepsy subjects with an average age of 23.3 ± 18.1 years (range 9–70, 50% women) acquired on the Siemens Biograph mMR at the King's College London and Guy's and St Thomas' PET Center, St Thomas' Hospital, London (Ethics Approval: 15/LO/0895). They consisted in a static 30-min PET acquisition started on average 120 ± 49 min after the injection of an average 120.6 ± 43.9 MBq of ^{18}F FDG.

PET simulation

Generation of numerical brain phantoms

Numerical brain phantoms are 3D labeled volume models built from segmented T1w 3D volumes. We performed MRI non-parametric non-uniformity intensity normalization, tissue class segmentation, and anatomical parcellation of the T1w 3D volumes with Freesurfer (34). To expand the segmentation to extracerebral tissues, we also used SPM12 (35). We were then able to create an anatomical brain model with nine labels: gray

matter (GM), white matter (WM) independently for the brain and the cerebellum (CEREB-WM, CEREB-GM), cerebrospinal fluid (CSF), basal ganglia (BG), bone, air, and soft tissue (SOFT).

Generation of ground truth high quality ^{18}F FDG PET

We created GT-HQ ^{18}F FDG PET by assigning activities to the nine labels of the numerical brain phantoms. Activities were derived from the distribution of normal ^{18}F FDG PET values from the CERMEP-IDB-MRXFDG database (33) after partial volume correction according to the Geometric Transfer Matrix (GTM) method (21).

We first simulated a series of normal brain SQ ^{18}F FDG PET scans. A total of 10 different brain activity distribution were generated for each anatomical brain model, resulting in 2100 (10×210) GT-HQ PETs. As a first step, WM activity was randomly chosen according to the observed distribution in (33). Activity ratios between cerebral GM and WM were then selected as 1.2, 1.8, 2.4, 3.0, 3.6, 4.2, 4.8, 5.4, 6.0, 6.6. Activities assigned to CSF, soft tissue and basal ganglia were randomly chosen according to the observed distribution in Mérida et al. (33). Cerebellum GM activity was set to 80% of the cerebrum.

Secondly, we created lesion GT-HQ PET phantoms with ROIs in the neocortex where we parametrically decreased assigned activity to simulate small metabolic lesions characteristic of FCDs. In 10 anatomical brain models with a GM/WM ratio of 3.6, we created one lesion each in the right frontal and in the left temporal region. The ROI for each lesion was manually defined as the largest component of the result of the multiplication of the GM mask and a sphere with volume of $1,008\text{ mm}^3$. In the same locations in the frontal and temporal lobes, we then repeated the process with two smaller spheres with volumes of 612 and 319 mm^3 . The resulting 60 lesion ROIs simulating small FCDs had volumes ranging from 17 to 570 mm^3 with a mean of $184 \pm 140\text{ mm}^3$: MRI volumetric values for FCDs ranged from 128 to $3,093\text{ mm}^3$ with a mean of $1,282 \pm 852\text{ mm}^3$ (36). Activity ratios between cerebral GM and the lesion were assigned values of 0.6 and 0.3. This resulted in 60 ($10\text{ models} \times 3\text{ sizes} \times 2\text{ activity ratios}$) lesion GT-HQ PET (120 lesions) with various morphologies and activities.

Monte-Carlo simulation of standard-quality PET

To generate realistic PET acquisitions, we used SORTEO, a Monte Carlo PET simulator developed by Reilhac et al. (31) and validated to provide realistic simulations for the Siemens Biograph mMR scanner (29, 31). The simulated 3D emission protocol consisted in the collection of data into a single timeframe for a 30-min period, as in our institution, starting 40 min post-injection, in accordance with international FDG PET guidelines (37). SORTEO generates the sinogram

(raw data), by simulating each disintegration occurring in labels where a constant activity was defined (GM, WM, CSF, CEREB-WM, CEREB-GM, GN and SOFT) including all physical phenomenon occurring from positron emission to detection. As for clinical scans, sinograms were normalized and corrected for randoms, scatter, attenuation, dead-time, and radioelement decay.

The simulations were performed at the IN2P3 (CNRS UAR6402) computing center. For each subject, simulation was divided into eight sub-processes to take advantage of multi-core processing and thus reducing the total simulation time.

Tomographic reconstruction

Corrected simulated sinograms were reconstructed with e7 reconstruction toolTM (Siemens Healthineers) using a 3D ordinary poisson-ordered subsets expectation maximization algorithm, incorporating the system point spread function, using 3 iterations and 21 subsets. Reconstructions were performed with a matrix size of $172 \times 172 \times 127$ and a zoom factor of 2, yielding a voxel size of $2.04 \times 2.04 \times 2.03 \text{ mm}^3$. The attenuation correction used a pseudo-CT synthesized with MaxProb multi-atlas attenuation correction method from the T1w MRI (38). Gaussian post-reconstruction 3D filtering (FWHM 4 mm isotropic) was applied to all PET images.

In the end, we have a database of 2100 pairs of GT-HQ and S-SQ PET images, with various anatomies and activity contrasts between brain structures. In addition, we simulated 120 small metabolic lesions characteristic of FCDs with various morphologies and activities.

Deep learning

Residual network architecture

Residual CNNs are commonly used algorithms for PET deblurring and are the main algorithms used for the generator in generative adversarial networks (26). The proof-of-concept of super-resolution PET was based on a very deep CNN (20 layers) (27) which was 2D because of computation limitation. As 3D images proved more successful in denoising tasks (39), we developed a 3D network for super resolution PET. Initially, we used a 3D U-Net, the main 3D network implemented for denoising PET (26). However, 3D U-Net did not achieve satisfactory results for our task and so we used a 3D sequential ResNet (40), similarly to recent papers by Spuhler et al. (41) and Sanaat et al. (42) with dilated kernels (model comparison shown in [Supplementary material](#)). They enlarge the field-of-view to incorporate multiscale context (43–45) and avoid the up-sampling layers of U-Net that degrade resolution, as spatial resolution of the input is maintained throughout the network (42). We implemented the model shown in [Figure 1](#). Each of the first 19 modules of the network exclusively uses convolutional

kernels of size $3 \times 3 \times 3$, along with batch normalization and Rectified Linear Unit (ReLU) activation function. In the first 7 modules, the network uses 16 kernels, the following 6 modules use 32 kernels, but with a dilation parameter of 2, and the next 6 modules use 64 kernels with dilation 4. The input of the deep learning model is the S-SQ PET.

Data preprocessing

Trilinear interpolation was used to resample all PET images to the same voxel size of $1 \times 1 \times 1 \text{ mm}$ with a $192 \times 256 \times 256$ grid size. The intensities in the input S-SQ PET images were standardized by dividing by the average of each individual image. Each GT-HQ PET was standardized by the average of the corresponding S-SQ PET image. The standardization factors were stored and subsequently applied to the network's predictions to rescale the resulting images, before performing any quantitative analysis [the PET unit was Becquerel (Bq) per: *centimetres cubed* (cm^3)].

Network implementation and optimization

The simulated images were split into training, validation, and testing datasets, with a ratio of 80/10/10%. Due to limitations of GPU memory during training, the network was trained with $32 \times 32 \times 32$ voxel patches. Twenty patches per volume were randomly chosen for the training and validation set. Mean absolute error was used as the loss function during training and the optimizer was AdamW (46). The learning rate was set to 10^{-4} and reduced by a factor of 0.1 when the validation loss stagnated for more than 10 epochs. The batch size was set to 50 and the maximum number of epochs to 200 using early stopping (validation loss not improving during more than 60 epochs).

We trained our model on a GPU server on 1 NVIDIA V100 GPU (32GB) running Python 3.9.10, Pytorch 1.10.0 (47), and TorchIO 0.18.71 (48).

For inference, patch of size $32 \times 32 \times 32$ voxels were used with $8 \times 8 \times 8$ overlapping tile stride. These patches were selected in sequence from the whole $192 \times 256 \times 256$ volume, then the P-HQ patches were put together to generate the entire P-HQ PET. Overlapping patches were combined using a weighted averaging strategy.

Evaluation

Evaluation of AI-enabled super-resolution PET was carried out on the P-HQ PET by comparing it to the S-SQ PET and the GT-HQ PET in brain masked images. We used the following quantitative evaluation metrics: (1) the Peak Signal-to-Noise Ratio (PSNR) (49), (2) the structural similarity index measure (SSIM) (50) which is a well-accepted measure of perceived

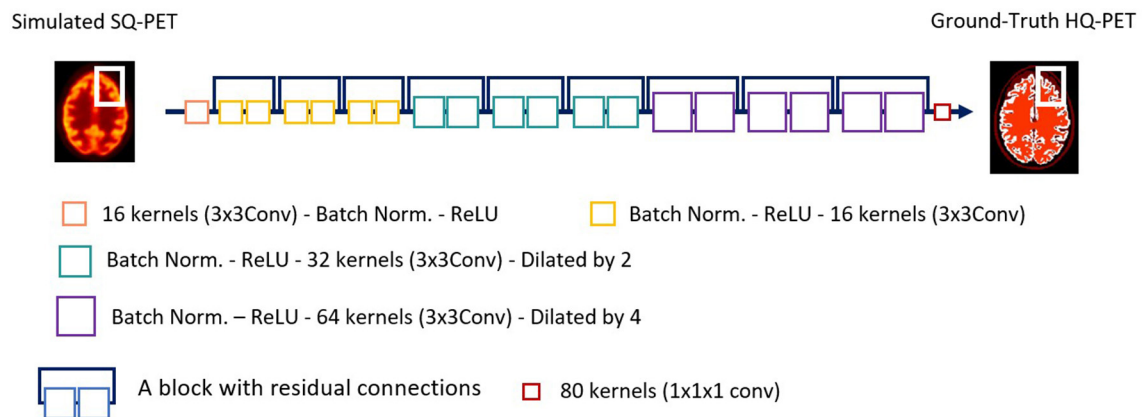


FIGURE 1

Architecture of the ResNet network used in this study. Conv, convolution; ReLU, Rectified Linear Unit; PET, Positron Emission Tomography; SQ, standard-quality; Norm, normalization; HQ, High-quality.

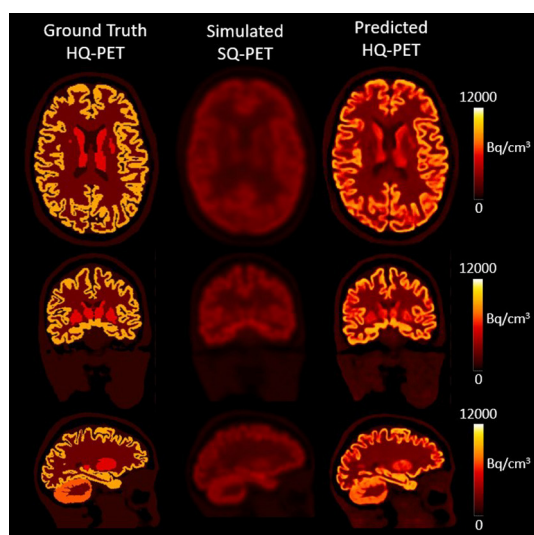


FIGURE 2

Results from one subject belonging to the test dataset. The first column depicts the Ground Truth High Quality (HQ) PET, the second column the corresponding simulated Standard Quality (SQ) PET and the third column the Predicted HQ PET, i.e., the output from the proposed network. For each set, from top to bottom, transverse, coronal, and sagittal slices are shown. Images are displayed using radiological conventions (subject's left on the right). Bq, Becquerel.

image quality s , and (3) the root mean squared error (RMSE) (Equations 1–3, respectively). An objective improvement in image quality is reflected by larger values in peak signal to noise ratio (PSNR) and structural similarity index metrics (SSIM) and smaller values for the root mean square error (RMSE).

$$\text{PSNR}(X, Y) = 20 \times \log_{10} \left(\frac{\text{Max}(X)}{\sqrt{\text{MSE}(X, Y)}} \right) \quad (1)$$

$$\text{SSIM}(X, Y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (2)$$

$$\text{RMSE}(X, Y) = \sqrt{\frac{\sum_{j=1}^L (X - Y)^2}{L}} \quad (3)$$

In Equation (1), given two images X and Y , $\text{Max}(X)$ indicates the maximum intensity value of X , whereas MSE is the mean squared error. In Equation (2), μ_x and μ_y denote the mean value of X and Y , respectively. σ_{xy} indicates the covariance of σ_x and σ_y , which in turn represent the variances of X and Y , respectively. The constant parameters c_1 and c_2 ($c_1 = 0.01$ and $c_2 = 0.03$) were used to avoid a division by very small numbers. In Equation (3), L is the total number of voxels in the head region, X and Y are the two compared images.

For the next evaluations, we used GM and WM ROIs, issued from the GM and WM probability maps resulting from T1w MRI segmentation using Freesurfer as described in 2.2.1. The WM ROI was obtained from the WM mask eroded by a radius of 6 voxels using ITK (51) to give a conservative WM ROI. The mean GM ROI volumes were $948,106 \pm 102,640 \text{ mm}^3$ and the mean eroded WM ROI volumes were $486,509 \pm 63,909 \text{ mm}^3$.

Recovery coefficients (RCs) defined as the ratio between the observed activity and the ground truth activity as shown in Equation (4), were calculated using μ the mean value in the GM ROI and the WM ROI for S-SQ PET and P-HQ PET compared to the GT-HQ PET.

$$\text{RC}_{\text{mean}} = \frac{\mu_{\text{measured}}}{\mu_{\text{ground truth}}} \quad (4)$$

We also computed the coefficient of variation (CoV) defined as the ratio between σ , the standard deviation, and μ , the mean value in the ROI, as shown in Equation (5). It is a metric for describing ensemble noise or statistical noise and it was computed in the GM ROI and the WM ROI for S-SQ PET and

P-HQ PET.

$$COV = \frac{\sigma_{measured}}{\mu_{measured}} \times 100 \quad (5)$$

TABLE 1 Mean and standard deviation of the root mean squared error (RMSE), peak signal to noise ratio (PSNR), and structural similarity index measure (SSIM) for simulated standard quality and predicted high quality (HQ) PET images in the test set.

	Root mean squared error	Peak signal-to-noise ratio (dB)	Structural similarity index measure
Simulated	2,393 \pm 1,496	16.6 \pm 1.1	0.876 \pm 0.013
Standard-quality PET			
Predicted	1,359 \pm 888	21.8 \pm 1.8	0.929 \pm 0.011
High-quality PET			

The comparator is the ground-truth HQ PET.

For lesion assessment, we performed first a visual assessment. The reader evaluated two sets of PET images: P-HQ PET images and S-SQ images in a random order. The reader determined whether a hypometabolic lesion was present (0 = none, 1 = visible lesion), and scored overall diagnostic confidence (ODC) in interpreting the images on a Likert scale of 1–5 (1 = none, 2 = poor, 3 = acceptable, 4 = good, 5 = excellent diagnostic confidence) (52) for each lesion. A second reader performed a visual assessment of a subset of lesioned S-SQ PET and G-HQ PET ($n = 84$ images) to assess inter-reader concordance. Secondly, to quantify lesion detectability, we computed a ratio between the measured activity in the ROI of the lesion over the same ROI in the P-HQ PET image without the lesion, termed relative lesion activity (RLA). We also computed the recovery coefficient in the lesion as in Equation (4).

For clinical data, we performed a visual assessment of the clinical PET and the P-HQ clinical PET computed with the trained network ($n = 20$ images) by two readers. The reader

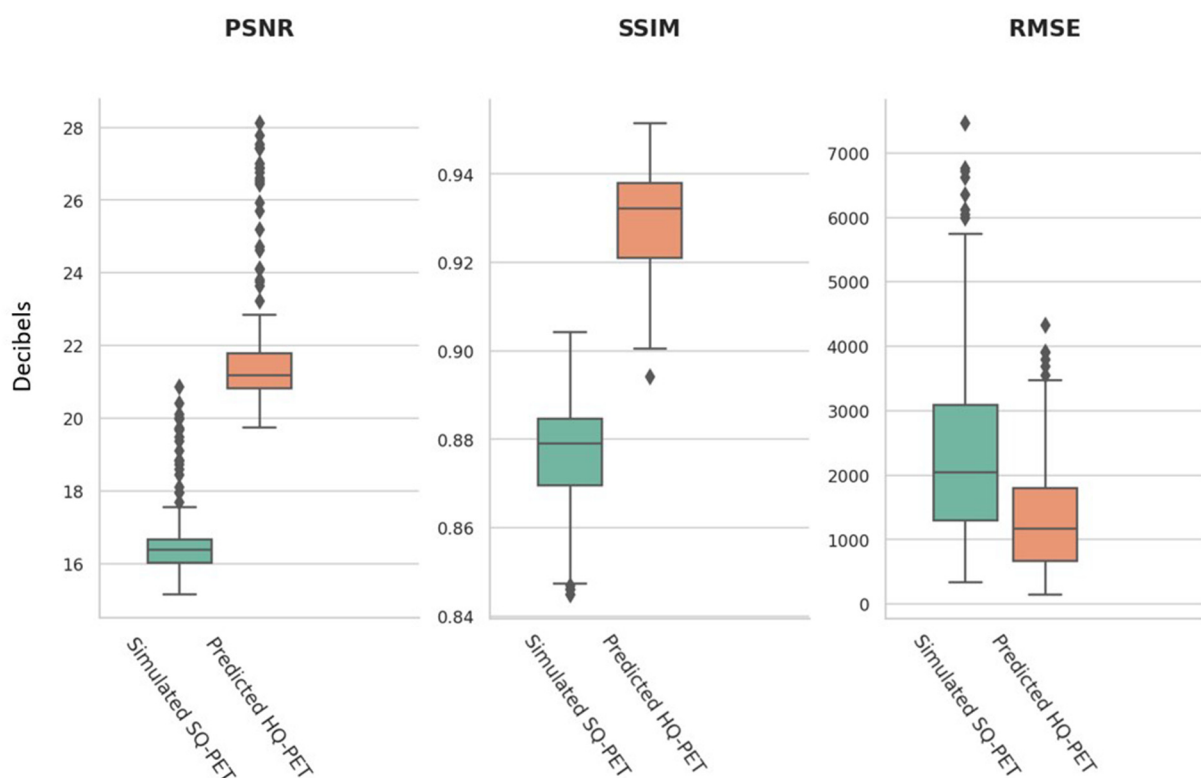


FIGURE 3

Image quality metrics from the simulated standard-quality (SQ) PET and the predicted high-quality (HQ) PET for the test set. An objective improvement in image quality is reflected by larger values in peak signal to noise ratio (PSNR) and structural similarity index metrics (SSIM) and smaller values for the root mean square error (RMSE).

TABLE 2 Mean, standard deviation (SD) and range of the recovery coefficient (RC) for the gray matter (GM) and the white matter (WM) for predicted high-quality (HQ) PET and simulated standard-quality (SQ) PET in the test set.

	Simulated SQ PET		Predicted HQ PET	
	GM RC	WM RC	GM RC	WM RC
Mean ± SD	0.29 ± 0.03	0.49 ± 0.03	0.79 ± 0.04	1 ± 0.05
Range	0.22–0.38	0.35–0.56	0.65–0.94	0.8–1.14

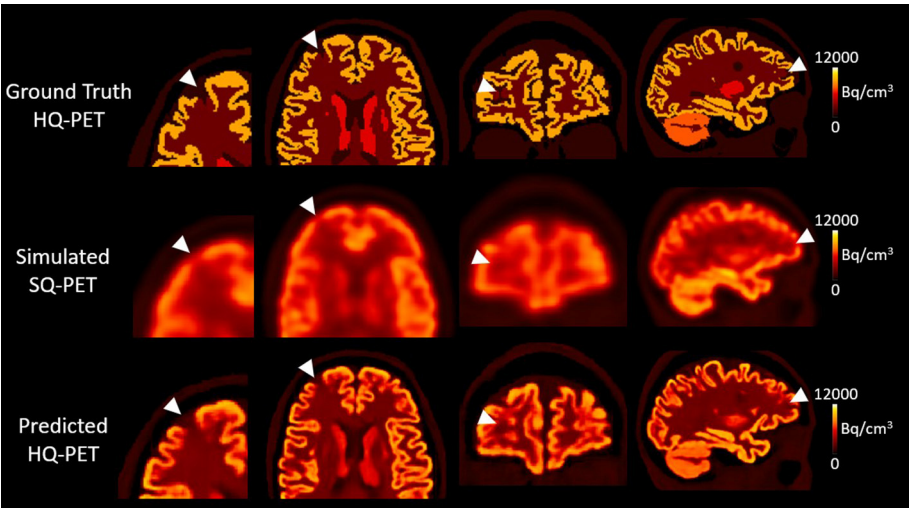


FIGURE 4 Results from one subject belonging to the test dataset with a simulated right frontal hypometabolic lesion with a volume of 0.327 cm³. First column, enlarged view of the lesion in transverse view; second column, transverse view, third column, coronal view, fourth column, sagittal view. The relative lesion activity was 0.3 in the ground-truth high-quality-PET, 0.75 in the simulated standard-quality PET, and decreased to 0.44 in the predicted HQ-PET. Arrowheads indicate the location of the simulated lesion. Images are displayed using radiological conventions (subject's left on the right). Bq, Becquerel cm³: centimetres cubed.

scored the diagnostic image quality on a 5-point Likert scale (1 = non-diagnostic, 2 = poor, 3 = standard, 4 = good, 5 = excellent image quality) (52) and as previously, indicated if a hypometabolic lesion was present and scored ODC.

We compared the quantitative results through the different metrics with pairwise *t*-tests or Wilcoxon rank sum test. Kappa coefficients were computed to assess inter-reader agreement. For all comparisons, the threshold of statistical significance was set at 5%.

Results

Non-lesioned simulated brains

The model was successfully trained to learn the mapping from the S-SQ PET to the GT-HQ PET after 105 epochs. Figure 2 showcases the result for one subject from the test dataset in transverse, coronal, and sagittal slices for the GT-HQ PET, its corresponding S-SQ PET, and the P-HQ PET.

The performance metrics computed on the test set for the P-HQ PET are shown in Table 1 and are plotted in Figure 3. The

values of those metrics on the S-SQ PET were also included for comparison. P-HQ PET showed improved image quality compared to the S-SQ PET (*p* < 0.0001 for all comparisons).

We computed the recovery coefficient of the GM and the WM in the test set for the S-SQ PET and the P-HQ PET. Recovery coefficients were significantly improved in the P-HQ PET for the WM and the GM compared to the S-SQ PET (*p* ≤ 0.0001). Mean, standard deviation (SD) and range of the recovery coefficient (RC) for the gray matter and the white matter for P-HQ-PET and S-SQ PET in the test set are shown in Table 2.

We further analyzed by GM/WM ratios (Boxplots shown in Supplementary Figures 2, 3). For all GM/WM ratios, GM recovery as well as WM recovery were significantly improved for the P-HQ compared to the S-SQ PET (*p* < 0.0001). In the S-SQ PET, across all GM/WM ratio, the mean GM RC ranged from 0.26 to 0.36 and standard deviation ranged from 0.008 to 0.220; the WM RC ranged from 0.45 to 0.52 and standard deviation range from 0.008 to 0.044. For the P HQ-PET, the mean GM RC ranged from 0.77 to 0.86 and standard deviation range from 0.016 to 0.042 and for the WM RC, it

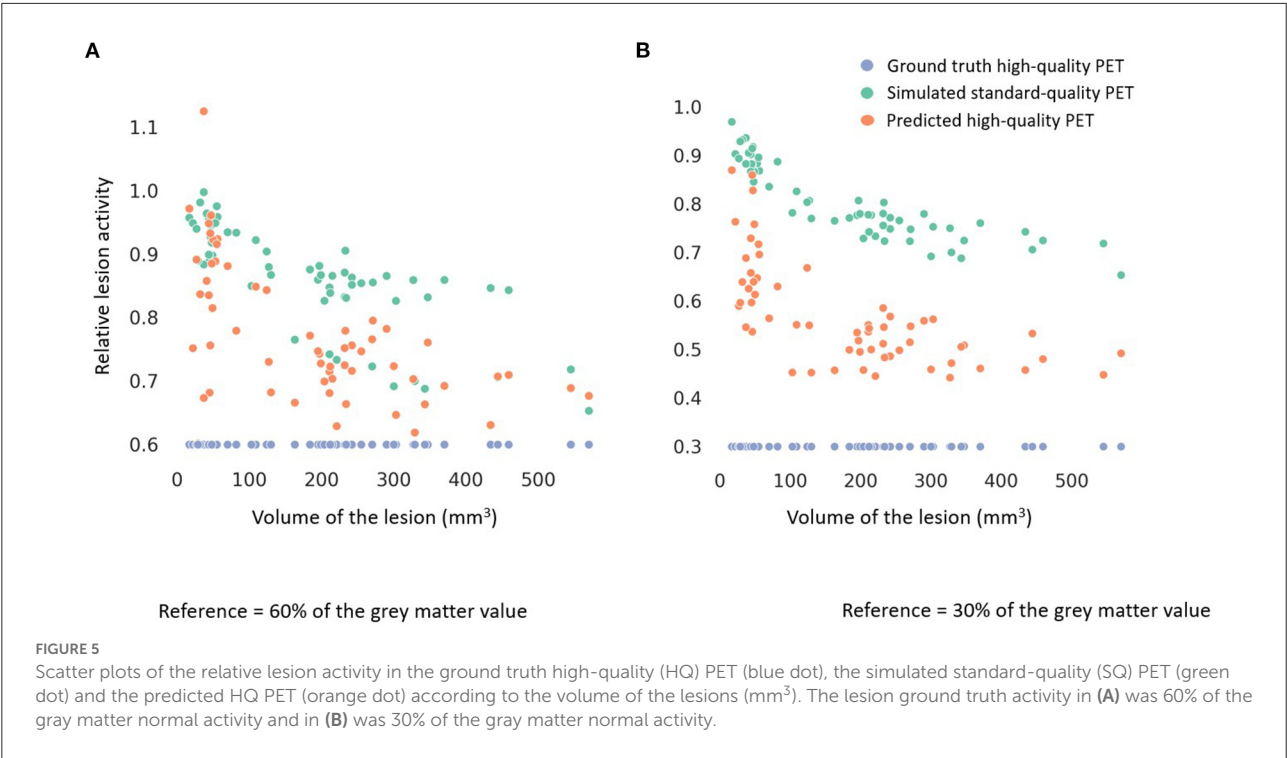


TABLE 3 Mean and standard deviation of the lesion relative lesion activity (RLA), the relative RLA error, and the lesion recovery coefficient (RC) for simulated standard quality, and predicted high quality (HQ) PET images in the test set.

	GT-HQ PET RLA lesion = 0.3			GT-HQ PET RLA lesion = 0.6		
	Lesion RLA (Target: 0.3)	Relative RLA error (Target: low)	Lesion RC (Target: 1)	Lesion RLA (Target: 0.6)	Relative RLA error (Target: low)	Lesion RC (Target: 1)
Simulated standard-quality PET	0.80 ± 0.08	1.68 ± 0.26	0.71 ± 0.10	0.86 ± 0.08	0.44 ± 0.17	0.39 ± 0.05
Predicted high-quality PET	0.57 ± 0.11	0.89 ± 0.35	1.44 ± 0.33	0.77 ± 0.11	0.28 ± 0.18	0.98 ± 0.17

The comparator is the ground-truth (GT) HQ PET. Left panel, lesion RLA was 0.3; right panel, lesion RLA was 0.6.

ranged from 0.99 to 1.04 and standard deviation ranged from 0.024 to 0.047. *Post-hoc* Anova analysis showed a significant difference for GM and WM RC, with a better RC for the lowest ratio (1.2).

The mean COV across all test datasets in the GM ROI was 38.9 ± 2.0 in the S-SQ PET and minimally higher at 39.3 ± 2.0 in the P-HQ PET (difference not significant, $p = 0.051$). The mean COV in the WM ROI was very similar at 4.90 ± 0.89 for S-SQ PET and 4.91 ± 0.89 for P-HQ PET ($p = 0.97$).

Lesioned simulated brain

At the group level, the visual detection rate was 38% in the S-SQ PET increasing to 75% in the P-HQ PET ($p < 0.05$) with a

similar overall diagnostic confidence score of 3.3 ± 1.6 vs. 3.5 ± 1.5 ($p > 0.05$). Kappa coefficients for inter-reader concordance were 0.77 for all images, 0.88 for P-HQ PET and 0.72 for S-SQ PET. Overall mean visual detection rates (44 vs. 42% in the S-SQ PET and 75 vs. 72% in the P-HQ PET) and diagnostic confidence scores (3.2 ± 1.7 vs. 3.1 ± 1.5 in the S-SQ PET and 3.4 ± 1.5 vs. 3.5 ± 1.3 in the P-HQ PET) were not statistically different between readers.

Figure 4 shows an example of one subject with a right frontal hypometabolic lesion of 327 mm³ from the test dataset for the GT-HQ PET, the S-SQ PET and the P-HQ PET. Through visual inspection, the hypometabolic lesion was easier to detect and with more confidence on the P-HQ PET. The RLA was 0.75 in the S-SQ PET, decreasing to 0.44 in the P-HQ PET, closer to the ground truth of 0.3.

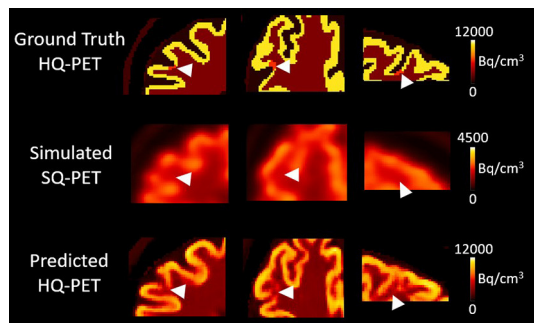


FIGURE 6

Results from one subject belonging to the test dataset with a simulated right frontal hypometabolic lesion with a volume of 22 mm^3 . First column, transverse view, second column, coronal view, third column, sagittal view centered on the lesion. The relative lesion activity was 0.6 in the ground-truth high-quality (HQ) PET, 0.95 in the simulated standard-quality (SQ) PET, and decreased to 0.75 in the predicted HQ PET. Arrowheads indicate the location of the simulated lesion. Images are displayed using radiological conventions (subject's left on the right). Bq, Becquerel cm^3 : centimetres cubed.

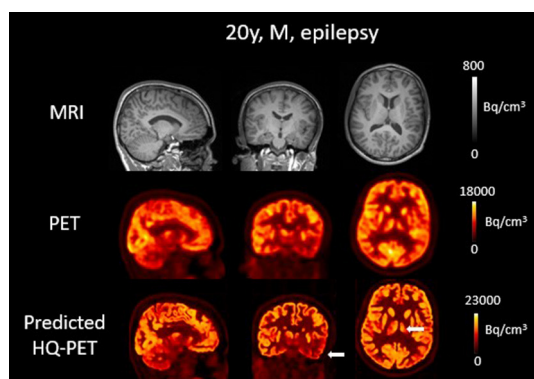


FIGURE 7

Brain T1w MRI and clinical $[^{18}\text{F}]\text{FDG}$ PET as well as predicted high-quality (HQ) PET (predicted by the network developed in this work) from one patient with drug-resistant epilepsy. Images are displayed using radiological conventions (subject's left on the right) and white arrows are used to highlight areas of hypometabolism. The first two rows show images from the scanner and the third row shows the AI-enhanced high-quality PET. There was no clear anomaly on the MR but a hypometabolism in the left temporal lobe as well as in the left thalamus on both PET images. Bq, Becquerel cm^3 : centimetres cubed.

Among all the lesions (GT-HQ PET RLA 0.3 or 0.6), RLA was substantially higher at 0.83 ± 0.08 (0.65–1) in the S-SQ PET but decreased toward the GT-HQ PET with 0.67 ± 0.14 (0.44–1.12) ($p < 0.0001$) in the P-HQ PET. RLA according to lesion volumes (mm^3) are plotted in Figure 5 for both the ground truths set at 0.3 or 0.6. There is a negative relation between the size of the lesion and the RLA value. For each subgroup whose

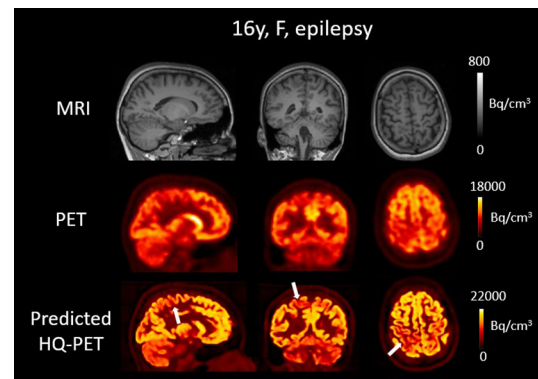


FIGURE 8

Brain T1w MRI and clinical $[^{18}\text{F}]\text{FDG}$ PET as well as predicted high-quality (HQ) PET (predicted by the network developed in this work) from one patient with drug-resistant epilepsy. Images are displayed using radiological conventions (subject's left on the right) and white arrows are used to highlight areas of hypometabolism. The first two rows show images from the scanner and the third row shows the AI-enhanced high-quality PET. MRI depicted a blurred white matter gray matter border in the right postcentral gyrus. The PET showed a correlated blurred and mild hypometabolism extending toward the precuneus. The predicted HQ PET showed a clearer hypometabolism very well correlated with the lesion that also extended to the precuneus. Bq, Becquerel cm^3 : centimetres cubed.

GT-HQ PET RLA was 0.3 or 0.6, mean RLA, relative RLA error and RC and their standard deviations are presented in Table 3. For the subgroup whose GT-HQ PET RLA was 0.3 (high contrast between lesion and surrounding GM), the mean RLA value for S-SQ PET was 0.80 ± 0.08 (0.65–0.97) and decreased to 0.57 ± 0.11 (0.44–0.87) in P-HQ PET. Values were significantly lower in the P-HQ PET ($p < 0.0001$) but remained significantly superior to the GT-HQ PET RLA of 0.3 ($p < 0.0001$). The mean relative RLA error in the S-SQ PET was 1.68 ± 0.26 (1.18–2.23) vs. 0.89 ± 0.35 (0.47–1.90) in P-HQ PET ($p < 0.0001$). The mean RC in the lesion ROI was 0.71 ± 0.10 (0.55–0.97) for the S-SQ PET vs. 1.44 ± 0.33 (1.05–2.5) for the P-HQ PET ($p < 0.0001$). For the subgroup whose GT-HQ PET RLA was 0.6 (low contrast between lesion and surrounding GM), the mean RLA value for the S-SQ PET was 0.86 ± 0.08 (0.65–1) and decreased to 0.77 ± 0.11 (0.60–1.12) in P-HQ-PET. Values were significantly lower in the P-HQ PET ($p < 0.0001$) but remained significantly superior to the GT-HQ PET RLA of 0.6 ($p < 0.0001$). The mean relative RLA error for the S-SQ PET was 0.44 ± 0.17 (0.09–0.66) vs. 0.28 ± 0.18 (0.00–0.87) in the P-HQ PET ($p < 0.0001$). Finally, the mean RC in the lesion ROI was 0.39 ± 0.05 (0.32–0.51) for the S-SQ PET vs. 0.98 ± 0.17 (0.69–1.45) for the P-HQ PET ($p < 0.0001$). Mean RC in P-HQ PET and GT-HQ PET were not different ($p = 0.32$). In Figure 6, we show a small lesion in the frontal lobe. The RLA was 0.6 in the GT-HQ PET, 0.95 in the S-SQ PET, and decreased to 0.75 in the P-HQ PET.

Epilepsy patients

The result of the trained model for clinical data is illustrated in Figures 7, 8 showing brain T1w MRI, [^{18}F]FDG PET and the P-HQ PET from two different patients with drug-resistant epilepsy. Across the cohort of epilepsy patients, the mean diagnostic image quality ratings for the clinical PETs were 2.9 ± 0.3 vs. 3.9 ± 0.5 for the predicted HQ PET ($p < 0.01$). Inter-reader mean quality scores were not significantly different. The mean diagnostic confidence ratings were 3.4 ± 1.1 for the clinical PET vs. 4.2 ± 0.8 for the predicted HQ ($p = 0.02$). Inter-reader mean confidence rating scores were not significantly different. Lesion detection rates were identical for both readers (7/10) for both the clinical PET and the predicted HQ PET.

Discussion

In this work, we trained a network to map Monte-Carlo S-SQ PET to their GT-HQ PET. In an independent test set, the P-HQ PET showed improved image quality compared to S-SQ PET across several objective quantitative metrics. In an independent dataset with small, simulated epilepsy lesions, the P-HQ PET significantly improved the relative lesion activity and visual detectability. Lastly, we have shown that the model was able to generalize to clinical data, illustrating the proof-of-concept that a model trained on Monte-Carlo simulated PET data is applicable on real data.

To train our model we had to overcome the limited availability of high quality training data, a common challenge for the deblurring problem (53) and so we chose to use simulation. We developed a pipeline based on a Monte-Carlo based PET simulator as it can accurately model the PET acquisition process including physical effects resulting in realistic sinograms (29) that have the same data distribution as the real PET. Compared to the few papers about PET deblurring with AI in image space, our simulated PET were more realistic: two studies used physically unrealistic degradation methods for their S-SQ PET adding Gaussian noise to an inverted T1w MR or down-sampling the standard PET image (54, 55). The latter approach also does not allow improvement beyond S-SQ PET. Two other studies used PET simulated analytically rather than with a Monte Carlo method (27, 28). While the main drawback of Monte-Carlo simulation is the computational burden, we were able to simulate PET acquisitions in a reasonable amount of time (about 3 h per scan) using PET SORTEO (29) which has been validated to provide realistic simulation of the Siemens Biograph mMR PET-MR (31), a system available across both our institutions. To be as close as possible to the clinical PET images, we reconstructed the generated sinogram using e7toolsTM (Siemens Healthineers), which is also used for clinical data. Next, we used the same pipeline to generate data with a simulated epileptogenic lesion. This pipeline now enables

creation of a whole range of realistic datasets for training if needed.

Our model has several particularities. We went beyond previous published PET deblurring methods with AI which used 2D models (27, 28, 54–56). We developed a 3D model, following results in the PET denoising field where 3D models tend to outperform 2D or 2.5D models (39, 57) because of additional features in 3D space. To prevent the impact of regional homogeneity on GT-HQ PET on the model parameters, we trained using small brain patches ($32 \times 32 \times 32 \text{ mm}^3$) from PET data simulated with of large number of GM/WM ratios. Thus, at the end of the training, the network weights were defined to respond to a wide range of voxel values (including hypometabolism) and patterns. The inference was also computed using the same size of patches which were then put together to obtain the predicted P-HQ PET. Compared to many deblurring methods (including some PVC methods and AI-based approaches) which rely on anatomical information (26), we provide a model that only relies on PET data which offers multiples advantages. Firstly, as the method works in the image space it can be applied on previously acquired PET even if raw data (sinograms) are not or no longer available, as will be the case in most clinical centers. Secondly, with the current development of dedicated standalone brain PET scanners (58), a PET-only method offers a unique opportunity to be combined with novel high-performance, high-resolution hardware to detect very small lesions. Thirdly, using a PET-only method prevents potential performance degradation that could stem from inter-modality alignment errors (59) which can occur even with simultaneous PET-MR if the MR sequence used for deblurring has been acquired at a different time to the emission data under study.

Our model achieved very good performance for relative lesion activity, which depicts lesion contrast, among all the lesions, despite different localization or shape. In the S-SQ PET, RLA was substantially higher at 0.83 ± 0.08 but decreased toward the GT-HQ PET ground truth (0.45) with 0.67 ± 0.14 in P-HQ PET. There was one outlier in the 0.6 RLA group with a P-HQ PET RLA value above one for a 37 mm^3 lesion in the lateral temporal lobe. This occurred because in the S-SQ PET, the lesion had a RLA value (0.997) so close to 1 that the information about presence of a lesion was lost during the simulation process. This is a principal limitation of our model which will only be overcome with higher resolution hardware. However, these results suggested that the model improved the RLA for most lesions even largely inferior to the nominal average 1D spatial resolution of 4.3 mm in full width at half maximum of the Siemens Biograph mMR (60), which defined a volumetric resolution near 80 mm^3 . The quantitative results correlated well to the visual analysis of the P-HQ PET images showing increased visibility of the simulated lesions as well as slight improvement in the confidence of the reader, suggesting the improvements

from P-HQ PET are relevant for future clinical application for epilepsy presurgical PET assessment.

Even if the SORTEO simulator is validated for the PET-MR, clinical PET images from the PET-MR will be slightly different requiring normalization, so the clinical P-HQ PET was expected to be different. Nevertheless, distributions of the simulated data and real data were close enough to enable use of both data types with the same network. We therefore consider that the clinical data application was successful in illustrating proof-of-concept that a model trained on Monte-Carlo simulated PET data is applicable on real data. Whereas, generalizability to out-of-distribution data is a common critical limiting factor for deep learning-based image processing (53), in our case this limitation could in principle be overcome by creating more simulations using the Monte-Carlo pipeline with settings tuned (29, 31) to simulate different scanners and reconstructions. Nevertheless, a study of generalization exceeds the scope of this manuscript. Such a study would need to be carefully planned to include reconstruction methods, scanner manufacturer, injection dose, uptake time and acquisition time to quantify the potential of such methods and their limits.

The realistic Monte-Carlo PET simulations and our training method allowed us to directly apply the trained network on clinical data. The P-HQ PET of the patients again showed an improved visual quality as well as an improved reader confidence. When we visually compared the GT-HQ PET and the P-HQ PET, it was apparent that the cortical structures were similar indicating that P-HQ PET from clinical data should not mislead physicians. Also, the clinical reading of epilepsy imaging does not rely on PET only. Indeed, physicians are trained to read both PET and MRI independently first, and jointly later, using the additional information to interpret the metabolism. In addition, and as in clinical practice (for example with non-attenuation corrected images), the non-enhanced standard quality image would always be made available to the reading physician to consult. One limitation of the clinical application was the small retrospective cohort of unselected epilepsy patients, but clinical evaluation of our model was not the main objective of this work. In addition, our ground truth was the visual assessment from the nuclear medicine physician using the standard PET which is an inherent limit to show the potential of the P-HQ PET. It would be interesting to evaluate our method in patients for which the standard quality PET was negative, but this is a very restrictive subpopulation where “ground truth” is often impossible to obtain as patients then neither undergo depth-electrode investigations nor surgery. Nevertheless, the patient with a small right post-central hypometabolism (Figure 8) underlines P-HQ PET’s potential for clinical application. This work can be put into perspective with the work of Baete and Goffin (12, 61) that used the anatomy-based maximum a-posteriori (A-MAP) reconstruction algorithm to improve detection of small areas of cortical hypometabolism. Their method showed promise to increase

detectability of hypometabolic areas on interictal [^{18}F]FDG PET in a cohort of 14 patients with FCD. FCDs are the most commonly resected epileptogenic lesions in children and the third most common lesions in adults (8). FCD type II is a malformation with disrupted cortical lamination and specific cytological abnormalities (62). Surgery remains the treatment of choice in drug resistant patients and relies on lesion localization (63). In Goffin et al. (12) improvement failed to reach significance due to the small sample size, but underlined the clinical potential of such methods. For epilepsy surgery, the outcome of seizures and long-term results, including discontinuation of antiepileptic drugs, is highly dependent on the discovery of an epileptogenic lesion in the surgical specimen; for example, for FCD the chance of being seizure-free increases to 67% for positive sample (64). Imaging has an important role to localize FCD (9) in particular [^{18}F]FDG PET (4, 65). In a study assessing the impact of imaging on FCD surgery outcome, there was no significant difference between FCDs detected on [^{18}F]FDG PET, whether MRI had been positive or negative (66).

We quantitatively and qualitatively validated our model on simulated data with and without epilepsy-typical lesions. We also illustrated its potential applicability to clinical data. The next step is to assess the performance of the P-HQ PET in a clinical study, ideally in a large cohort of patients with well-localized lesions (FCDs), such as seizure-free subjects after brain surgery. It will be also important to evaluate performance of nuclear medicine physicians with different levels of experience: P-HQ PET should be seen as a diagnostic support to improve reader detection and confidence, allowing non-expert readers to perform closer to expert reader performance. Another interesting perspective would be to assess the improvement of an AI based anomaly detection model (67) with the P-HQ PET compared to the standard PET.

Conclusion

In this work, we trained a deep learning model to map S-SQ PET to their GT-HQ PET using a new large realistic Monte-Carlo simulated database. In an independent test set, the P-HQ PET showed improved image quality compared to S-SQ PET across several quantitative objective metrics. Moreover, in the context of epilepsy simulated lesions, the P-HQ PET improved the relative lesion activity and their visual detection. Following this validation on simulated lesion data and the successful clinical application to illustrate the proof-of-concept that a model trained on Monte-Carlo simulated PET data is applicable on real data, next steps are to perform a generalization study and to assess the performance of the P-HQ PET in a cohort of epilepsy patients with well-characterized lesions and/or normal standard-quality PET.

Data availability statement

MRI data used are available at <https://sites.google.com/view/calgary-campinas-dataset/download?authuser=0>. The MRxFDG data used are available from the corresponding author at <https://doi.org/10.1186/s13550-021-00830-6>. A sample of standard and high quality pet data are available at <https://osf.io/4j6xu> and further access can be requested by contacting the corresponding author. Comparable deep learning model code are available at <https://github.com/Project-MONAI/MONAI>.

Ethics statement

The studies involving human participants were reviewed and approved by King's College London and Guy's and St Thomas' PET Center, St Thomas' Hospital, London (Ethics Approval: 15/LO/0895). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

AF designed the study, implemented the deep learning method, analyzed the data, and wrote the manuscript draft. NL performed a visual analysis of the PET images. TD, NC, AF, and IM performed the data simulations and image reconstructions. AR provided expertise for the simulation and network implementation. TG provided expertise with the network training. AH co-wrote the second manuscript draft. CM contributed to the acquisition and reconstruction of clinical PET data. AH, NC, and CL guided and supervised the project. All authors contributed to critically reviewing and approving the manuscript and read and approved the final manuscript.

Funding

This research was funded in whole, or in part, by the Wellcome Trust [WT 203148/Z/16/Z]. For the purpose of open access, the author has applied a CC BY public copyright

licence to any Author Accepted Manuscript version arising from this submission. AF received funding from the French branch of the International League Against Epilepsy (ILAE), Ligue Française contre l'Epilepsie (LFCE), the Labex Primes from Lyon University, Lyon and Hospices Civils de Lyon.

Acknowledgments

The School of Biomedical Engineering and Imaging Sciences is supported by the Wellcome EPSRC Centre for Medical Engineering at King's College London [WT 203148/Z/16/Z] and the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. This work was supported by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2022.1042706/full#supplementary-material>

References

1. Kwan P, Brodie MJ. Early identification of refractory epilepsy. *N Engl J Med.* (2000) 342:314–9. doi: 10.1056/NEJM200002033420503
2. Ryvlin P, Rheims S. Epilepsy surgery: eligibility criteria and presurgical evaluation. *Dialogues Clin Neurosci.* (2008) 10:91–103. doi: 10.31887/DCNS.2008.10.1/ryvlin

3. Hainc N, McAndrews MP, Valiante T, Andrade DM, Wennberg R, Krings T. Imaging in medically refractory epilepsy at 3 Tesla: a 13-year tertiary adult epilepsy center experience. *Insights Imaging*. (2022) 13:99. doi: 10.1186/s13244-022-01236-1
4. Chassoux F, Rodrigo S, Semah F, Beuvon F, Landre E, Devaux B, et al. FDG-PET improves surgical outcome in negative MRI Taylor-type focal cortical dysplasias. *Neurology*. (2010) 75:2168–75. doi: 10.1212/WNL.0b013e31820203a9
5. Gok B, Jallo G, Hayeri R, Wahl R, Aygun N. The evaluation of FDG-PET imaging for epileptogenic focus localization in patients with MRI positive and MRI negative temporal lobe epilepsy. *Neuroradiology*. (2013) 55:541–50. doi: 10.1007/s00234-012-1121-x
6. Flaus A, Mellerio C, Rodrigo S, Brulon V, Lebon V, Chassoux F. 18F-FDG PET/MR in focal epilepsy: a new step for improving the detection of epileptogenic lesions. *Epilepsy Res*. (2021) 178:106819. doi: 10.1016/j.eplepsyres.2021.106819
7. Taylor DC, Falconer MA, Bruton CJ, Corsellis JAN. Focal dysplasia of the cerebral cortex in epilepsy. *J Neurol Neurosurg Psychiatry*. (1971) 34:369–87. doi: 10.1136/jnnp.34.4.369
8. Blumcke I, Spreafico R, Haaker G, Coras R, Kobow K, Bien CG, et al. Histopathological findings in brain tissue obtained during epilepsy surgery. *N Engl J Med*. (2017) 377:1648–56. doi: 10.1056/NEJMoa1703784
9. Desarnaud S, Mellerio C, Semah F, Laurent A, Landre E, Devaux B, et al. 18F-FDG PET in drug-resistant epilepsy due to focal cortical dysplasia type 2: additional value of electroclinical data and coregistration with MRI. *Eur J Nucl Med Mol Imaging*. (2018) 45:1449–60. doi: 10.1007/s00259-018-3994-3
10. Hoffman EJ, Huang SC, Phelps ME. Quantitation in positron emission computed tomography: 1. Effect of object size. *J Comput Assist Tomogr*. (1979) 3:299–308. doi: 10.1097/00004728-197906000-00001
11. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. *J Nucl Med*. (2007) 48:932–45. doi: 10.2967/jnumed.106.035774
12. Goffin K, Van Paesschen W, Dupont P, Baete K, Palmieri A, Nuyts J, et al. Anatomy-based reconstruction of FDG-PET images with implicit partial volume correction improves detection of hypometabolic regions in patients with epilepsy due to focal cortical dysplasia diagnosed on MRI. *Eur J Nucl Med Mol Imaging*. (2010) 37:1148–55. doi: 10.1007/s00259-010-1405-5
13. Vaquero JJ, Kinahan P. Positron emission tomography: current challenges and opportunities for technological advances in clinical and preclinical imaging systems. *Annu Rev Biomed Eng*. (2015) 17:385–414. doi: 10.1146/annurev-bioeng-071114-040723
14. Tong S, Alessio AM, Kinahan P. Image reconstruction for PET/CT scanners: past achievements and future challenges. *Imaging Med*. (2010) 2:529–45. doi: 10.2217/iim.10.49
15. Labbé C, Froment JC, Kennedy A, Ashburner J, Cinotti L. Positron emission tomography metabolic data corrected for cortical atrophy using magnetic resonance imaging. *Alzheimer Dis Assoc Disord*. (1996) 10:141–70. doi: 10.1097/00002093-199601030-00005
16. Panin VY, Kehren F, Michel C, Casey M. Fully 3-D PET reconstruction with system matrix derived from point source measurements. *IEEE Trans Med Imaging*. (2006) 25:907–21. doi: 10.1109/TMI.2006.876171
17. Hofheinz F, Langner J, Beuthien-Baumann B, Oehme L, Steinbach J, Kotzerke J, et al. Suitability of bilateral filtering for edge-preserving noise reduction in PET. *EJNMMI Res*. (2011) 1:23. doi: 10.1186/2191-219X-1-23
18. Le Pogam A, Hanzouli H, Hatt M, Cheze Le Rest C, Visvikis D. Denoising of PET images by combining wavelets and curvelets for improved preservation of resolution and quantitation. *Med Image Anal*. (2013) 17:877–91. doi: 10.1016/j.media.2013.05.005
19. Chan C, Fulton R, Barnett R, Feng DD, Meikle S. Postreconstruction nonlocal means filtering of whole-body PET with an anatomical prior. *IEEE Trans Med Imaging*. (2014) 33:636–50. doi: 10.1109/TMI.2013.2292881
20. Müller-Gärtner HW, Links JM, Prince JL, Bryan RN, McVeigh E, Leal JP et al. Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. *J Cereb Blood Flow Metab*. (1992) 12:571–83. doi: 10.1038/jcbfm.1992.81
21. Roussel OG, Ma Y, Evans AC. Correction for partial volume effects in PET: principle and validation. *J Nucl Med*. (1998) 39:904–11.
22. Aston JAD, Cunningham VJ, Asselin M-C, Hammers A, Evans AC, Gunn RN. Positron emission tomography partial volume correction: estimation and algorithms. *J Cereb Blood Flow Metab*. (2002) 22:1019–34. doi: 10.1097/00004647-200208000-00014
23. Thomas BA, Erlandsson K, Modat M, Thurfjell L, Vandenberghe R, Ourselin S, et al. The importance of appropriate partial volume correction for PET quantification in Alzheimer's disease. *Eur J Nucl Med Mol Imaging*. (2011) 38:1104–19. doi: 10.1007/s00259-011-1745-9
24. Tohka J, Reilhac A. Deconvolution-based partial volume correction in Raclopride-PET and Monte Carlo comparison to MR-based method. *Neuroimage*. (2008) 39:1570–84. doi: 10.1016/j.neuroimage.2007.10.038
25. Golla SSV, Lubberink M, van Berckel BNM, Lammertsma AA, Boellaard R. Partial volume correction of brain PET studies using iterative deconvolution in combination with HYPR denoising. *EJNMMI Res*. (2017) 7:36. doi: 10.1186/s13550-017-0284-1
26. Liu J, Malekzadeh M, Mirian N, Song T-A, Liu C, Dutta J. Artificial intelligence-based image enhancement in PET imaging: noise reduction and resolution enhancement. *PET Clin*. (2021) 16:553–76. doi: 10.1016/j.cpet.2021.06.005
27. Song T-A, Chowdhury SR, Yang F, Dutta J. Super-resolution PET imaging using convolutional neural networks. *IEEE Trans Comput Imaging*. (2020) 6:518–28. doi: 10.1109/TCI.2020.2964229
28. Dal Toso L, Chalampalakis Z, Buvat I, Comtat C, Cook G, Goh V, et al. Improved 3D tumour definition and quantification of uptake in simulated lung tumours using deep learning. *Phys Med Biol*. (2022) 67:095013. doi: 10.1088/1361-6560/ac65d6
29. Reilhac A, Lartizen C, Costes N, Sans S, Comtat C, Gunn RN, et al. PET-SORTEO: a Monte Carlo-based simulator with high count rate capabilities. *IEEE Trans Nucl Sci*. (2004) 51:46–52. doi: 10.1109/TNS.2003.823011
30. Stute S, Tauber C, Leroy C, Bottlaender M, Brulon V, Comtat C. Analytical simulations of dynamic PET scans with realistic count rates properties. In: *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. San Diego, CA, USA: IEEE (2015). p. 1–3.
31. Reilhac A, Soderlund T, Thomas B, Irace Z, Merida I, Villien M, et al. Validation and application of PET-SORTEO for the geometry of the Siemens mMR scanner. In: *PSMR Conference*. Cologne (2016)
32. Souza R, Lucena O, Garrafa J, Gobbi D, Saluzzi M, Appenzeller S, et al. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *Neuroimage*. (2018) 170:482–94. doi: 10.1016/j.neuroimage.2017.08.021
33. Mérida I, Jung J, Bouvard S, Le Bars D, Lancelot S, Lavenne F, et al. CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F]FDG PET, T1 and FLAIR MRI, and CT images available for research. *EJNMMI Res*. (2021) 11:91. doi: 10.1186/s13550-021-00830-6
34. Fischl B. FreeSurfer. *Neuroimage*. (2012) 62:774–81. doi: 10.1016/j.neuroimage.2012.01.021
35. Ashburner J, Friston KJ. *Image Segmentation*. Human Brain Function. Academic Press (2003). Available online at: <https://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch5.pdf>
36. Besson P, Andermann F, Dubeau F, Bernasconi A. Small focal cortical dysplasia lesions are located at the bottom of a deep sulcus. *Brain*. (2008) 131:3246–55. doi: 10.1093/brain/awn224
37. Guedj E, Varrone A, Boellaard R, Albert NL, Barthel H, van Berckel B, et al. EANM procedure guidelines for brain PET imaging using [18F]FDG, version 3. *Eur J Nucl Med Mol Imaging*. (2022) 49:632–51. doi: 10.1007/s00259-021-05603-w
38. Mérida I, Reilhac A, Redouté J, Heckemann RA, Costes N, Hammers A. Multi-atlas attenuation correction supports full quantification of static and dynamic brain PET data in PET-MR. *Phys Med Biol*. (2017) 62:2834–58. doi: 10.1088/1361-6560/aa5f6c
39. Lu W, Onofrey JA, Lu Y, Shi L, Ma T, Liu Y, et al. An investigation of quantitative accuracy for deep learning based denoising in oncological PET. *Phys Med Biol*. (2019) 64:165019. doi: 10.1088/1361-6560/ab3242
40. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. *arXiv:1707.01992*. (2017) 10265:348–60. doi: 10.1007/978-3-319-59050-9_28
41. Spuhler K, Serrano-Sosa M, Cattell R, DeLorenzo C, Huang C. Full-count PET recovery from low-count image using a dilated convolutional neural network. *Med Phys*. (2020) 47:4928–38. doi: 10.1002/mp.14402
42. Sanaat A, Shooli H, Ferdowsi S, Shiri I, Arabi H, Zaidi H. DeepTOFSino: a deep learning model for synthesizing full-dose time-of-flight bin sinograms from their corresponding low-dose sinograms. *Neuroimage*. (2021) 245:118697. doi: 10.1016/j.neuroimage.2021.118697
43. Luo W, Li Y, Urtasun R, Zemel R. *Understanding the Effective Receptive Field in Deep Convolutional Neural Networks*. Barcelona: Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2016). p. 4898–906.
44. Chen L-C, Papandreou G, Schroff F, Adam H. *Rethinking Atrous Convolution for Semantic Image Segmentation*. (2017). Available online at: <http://arxiv.org/abs/1706.05587> (accessed August 21, 2022).

45. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell.* (2018) 40:834–48. doi: 10.1109/TPAMI.2017.2699184
46. Loshchilov I, Hutter F. *Decoupled Weight Decay Regularization.* (2019). Available online at: <http://arxiv.org/abs/1711.05101> (accessed May 24, 2022).
47. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. *PyTorch: An imperative style, high-performance deep learning library.* In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC (2019). p. 12.
48. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed.* (2021) 208:106236. doi: 10.1016/j.cmpb.2021.106236
49. Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: *2010 20th International Conference on Pattern Recognition.* Istanbul, Turkey: IEEE (2010). p. 2366–9.
50. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans on Image Process.* (2004) 13:600–12. doi: 10.1109/TIP.2003.819861
51. McCormick M, Liu X, Jomier J, Marion C, Ibanez L. ITK: enabling reproducible research and open science. *Front Neuroinform.* (2014) 8:e00013. doi: 10.3389/fninf.2014.00013
52. Chaudhari AS, Mittra E, Davidzon GA, Gulaka P, Gandhi H, Brown A, et al. Low-count whole-body PET with deep learning in a multicenter and externally validated study. *NPJ Digit Med.* (2021) 4:127. doi: 10.1038/s41746-021-00497-2
53. Pain CD, Egan GF, Chen Z. Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement. *Eur J Nucl Med Mol Imaging.* (2022) 49:3098–118. doi: 10.1007/s00259-022-05746-4
54. Garehdaghi F, Meshgini S, Afrouzian R, Farzamnian A. PET image super resolution using convolutional neural networks. In: *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPI)* (Shahrood). (2019). p. 1–5.
55. Chen W, McMillan A. Single subject deep learning-based partial volume correction for PET using simulated data and cycle consistent networks. *J Nucl Med.* (2020) 61:520. Available online at: http://jnm.snmjournals.org/content/61/supplement_1/520.abstract
56. Song T-A, Chowdhury SR, Yang F, Dutta J. PET. image super-resolution using generative adversarial networks. *Neural Networks.* (2020) 125:83–91. doi: 10.1016/j.neunet.2020.01.029
57. Wang Y, Yu B, Wang L, Zu C, Lalush DS, Lin W, et al. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *Neuroimage.* (2018) 174:550–62. doi: 10.1016/j.neuroimage.2018.03.045
58. Catana C. Development of dedicated brain PET imaging devices: recent advances and future perspectives. *J Nucl Med.* (2019) 60:1044–52. doi: 10.2967/jnumed.118.217901
59. Liu C-C, Qi J. Higher SNR PET image prediction using A deep learning model and MRI image. *Phys Med Biol.* (2019) 64:115004. doi: 10.1088/1361-6560/ab0dc0
60. Delso G, Fürst S, Jakoby B, Ladebeck R, Ganter C, Nekolla SG, et al. Performance measurements of the Siemens mMR integrated whole-body PET/MR scanner. *J Nucl Med.* (2011) 52:1914–22. doi: 10.2967/jnumed.111.092726
61. Baete K, Nuyts J, Van Paesschen W, Suetens P, Dupont P. Anatomical-based FDG-PET reconstruction for the detection of hypo-metabolic regions in epilepsy. *IEEE Trans Med Imaging.* (2004) 23:510–9. doi: 10.1109/TMI.2004.825623
62. Palmini A, Najm I, Avanzini G, Babb T, Guerrini R, Foldvary-Schaefer N, et al. Terminology and classification of the cortical dysplasias. *Neurology.* (2004) 62:S2–8. doi: 10.1212/01.WNL.0000114507.30388.7E
63. Guerrini R, Duchowny M, Jayakar P, Krsek P, Kahane P, Tassi L, et al. Diagnostic methods and treatment options for focal cortical dysplasia. *Epilepsia.* (2015) 56:1669–86. doi: 10.1111/epi.13200
64. Lamberink HJ, Otte WM, Blümcke I, Braun KPJ, Aichholzer M, Amorim I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. *Lancet Neurol.* (2020) 19:748–57. doi: 10.1016/S1474-4422(20)30220-9
65. Salamon N, Kung J, Shaw SJ, Koo J, Koh S, Wu JY, et al. FDG-PET/MRI coregistration improves detection of cortical dysplasia in patients with epilepsy. *Neurology.* (2008) 71:1594–601. doi: 10.1212/01.wnl.0000334752.41807.2f
66. Chassoux F, Landré E, Mellerio C, Turak B, Mann MW, Daumas-Duport C, et al. focal cortical dysplasia: electroclinical phenotype and surgical outcome related to imaging: phenotype and Imaging in TTFC. *Epilepsia.* (2012) 53:349–58. doi: 10.1111/j.1528-1167.2011.03363.x
67. Smith RL, Chandler H, Alsayed E, Bartley L, Fielding P, Marshall C. Deep learning PET epilepsy detection with a novel symmetric loss convolutional autoencoder. In: *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC).* Boston, MA, USA: IEEE (2020). p. 1–3.



OPEN ACCESS

EDITED BY

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC),
Switzerland

REVIEWED BY

Paola Mapelli,
Vita-Salute San Raffaele University, Italy
Laura Evangelista,
University of Padua, Italy

*CORRESPONDENCE

Nadja Rolim Gonçalves de Alencar
nadja.rolim@gmail.com

SPECIALTY SECTION

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 16 September 2022

ACCEPTED 10 November 2022

PUBLISHED 02 December 2022

CITATION

Alencar NRG, Machado MAD,
Mourato FA, Oliveira ML, Moraes TF,
Mattos Junior LAR, Chang TC,
Azevedo CRAS and Brandão SCS
(2022) Exploratory analysis of
radiomic as prognostic biomarkers in
 ^{18}F -FDG PET/CT scan in uterine
cervical cancer.
Front. Med. 9:1046551.
doi: 10.3389/fmed.2022.1046551

COPYRIGHT

© 2022 Alencar, Machado, Mourato,
Oliveira, Moraes, Mattos Junior,
Chang, Azevedo and Brandão. This is
an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided
the original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Exploratory analysis of radiomic as prognostic biomarkers in ^{18}F -FDG PET/CT scan in uterine cervical cancer

Nadja Rolim Gonçalves de Alencar^{1,2*},
Marcos Antônio Dórea Machado³, Felipe Alves Mourato^{1,2},
Mércia Liane de Oliveira⁴, Thauan Fernandes Moraes⁴,
Luiz Alberto Reis Mattos Junior⁵, Tien-Man Cabral Chang⁶,
Carla Rameri Alexandre Silva de Azevedo⁷ and
Simone Cristina Soares Brandão^{1,2,5}

¹Master of Science Surgery Post-Graduation Program, Federal University of Pernambuco, Recife, Pernambuco, Brazil, ²Department of Radiology and Nuclear Medicine, Hospital das Clínicas, Federal University of Pernambuco, Recife, Pernambuco, Brazil, ³Department of Radiology, Complexo Hospitalar Universitário Professor Edgard Santos/Universidade Federal da Bahia (UFBA), Salvador, Bahia, Brazil, ⁴Northeast Center for Strategic Technologies, Recife, Pernambuco, Brazil, ⁵Clinical Medicine, Center for Medical Sciences, Federal University of Pernambuco, Recife, Pernambuco, Brazil, ⁶Nuclear Medicine Service, Instituto de Medicina Integrada Fernandes Figueira, Recife, Pernambuco, Brazil, ⁷Department of Oncology, Instituto de Medicina Integrada Fernandes Figueira, Recife, Pernambuco, Brazil

Objective: To evaluate the performance of ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET/CT) radiomic features to predict overall survival (OS) in patients with locally advanced uterine cervical carcinoma.

Methods: Longitudinal and retrospective study that evaluated 50 patients with cervical epidermoid carcinoma (clinical stage IB2 to IVA according to FIGO). Segmentation of the ^{18}F -FDG PET/CT tumors was performed using the LIFEx software, generating the radiomic features. We used the Mann–Whitney test to select radiomic features associated with the clinical outcome (death), excluding the features highly correlated with each other with Spearman correlation. Subsequently, ROC curves and a Kaplan–Meier analysis were performed. A p -value < 0.05 were considered significant.

Results: The median follow-up was 23.5 months and longer than 24 months in all surviving patients. Independent predictors for OS were found—SUVpeak with an AUC of 0.74, sensitivity of 77.8%, and specificity of 72.7% ($p = 0.006$); and the textural feature gray-level run-length matrix GLRLM_LRLGE, with AUC of 0.74, sensitivity of 72.2%, and specificity of 81.8% ($p = 0.005$). When we used the derived cut-off points from these ROC curves (12.76 for SUVpeak and 0.001 for GLRLM_LRLGE) in a Kaplan–Meier analysis, we can see two different groups (one with an overall survival probability of approximately 90% and the other with 30%). These biomarkers are independent of FIGO staging.

Conclusion: By radiomic ^{18}F -FDG PET/CT data analysis, SUVpeak and GLRLM_LRLGE textural feature presented the best performance to predict OS in patients with cervical cancer undergoing chemo-radiotherapy and brachytherapy.

KEYWORDS

positron emission tomography, prognosis, uterine cervical neoplasms, ^{18}F -fluorodeoxyglucose, radiomics

Introduction

Cervical uterine cancer is an important cause of death in women, especially in regions of low socioeconomic development (1–3). In more advanced stages, fluorine-18-labeled fluorodeoxyglucose positron emission tomography associated with computed tomography (^{18}F -FDG PET/CT) is recommended for the adequate evaluation of lymph nodes and distant metastases (4–6).

The standardized uptake value (SUV) of ^{18}F -FDG is the most used semi-quantitative variable in ^{18}F -FDG PET/CT (7). This value translates the lesion glycolytic metabolism and the higher the value, the more aggressive the tumor (8). Other quantitative metrics extracted from the ^{18}F -FDG PET/CT scan are the metabolic tumor volume (MTV), which translates the measure of the tumor volume with a higher metabolism, and the total lesion glycolysis rate (TLG), which is the product of the mean SUV by the lesion MTV (9). These three variables reflect the tumor metabolic load and could help to predict the patient's prognosis (10, 11).

Radiomic is the extraction of mineable data from medical imaging that has emerged recently (12). It analyzes the lesion phenotype using mathematical formulas that dissect the image, quantifying and characterizing several tumoral features (13–15). Among the numerous variables of the radiomic analysis of ^{18}F -FDG PET/CT images, the textural features present a greater correlation with the heterogeneous biological behavior of the tumor. They may serve as predictive markers of overall survival (OS) and therapeutic response (16–18).

Therefore, this paper aims to find radiomic features and metabolic parameters predictive of OS from ^{18}F -FDG PET/CT scans of uterine cervical cancer.

Materials and methods

Patients and methods

The present study included 50 consecutive patients with histologically confirmed diagnoses of uterine cervical squamous cell carcinoma between 2013 and 2015 (Table 1).

The inclusion criteria were: women over 18 years that were undergone pretreatment ^{18}F -FDG PET/CT. All patients received standardized chemotherapy treatment with cisplatin and gemcitabine, with two cycles of neoadjuvant chemotherapy, with subsequent radiotherapy and brachytherapy according to the institutional protocol. The patients were followed up for at least 24 months. The exclusion criteria included ^{18}F -FDG PET/CT scans in disagreement with the acquisition, processing, or reconstruction parameters, according to the Image Biomarker Standardization Initiative (IBSI) (19). The selected patients were divided into two groups according to their progression after 24 months of follow-up: group 1, with overall survival of at least 24 months and group 2, deceased due to cancer in the follow-up period. The institutional research ethics committee approved this study. The demographic data and clinical information were obtained from the medical records and included: age, origin, education, smoking status, number of children, and number of sexual partners, in addition to the clinical and imaging staging data (FIGO) (4), and information regarding the treatment.

Protocol of the ^{18}F -FDG PET/CT scan

The scans were performed at the nuclear medicine and molecular imaging facility of the *Instituto de Medicina Integral Professor Fernando Figueira* using PET/CT scanner (Siemens Biography 16 channels, Germany), according to the guidelines of the European Society of Nuclear Medicine (20). Patients fasting for at least 4 h and with glycemic levels ≤ 150 mg/dL received 0.14 mCi/kg of ^{18}F -FDG intravenously. Approximately 60 min after the administration of ^{18}F -FDG, images were obtained from the skull to the thigh root. All the patients received 20 mg of furosemide after the first imaging; additionally, 120 min after the radiopharmaceutical injection, they returned to the scanner for late imaging of the pelvis. The acquisition parameters of the initial images were analyzed, with a reconstruction diameter of 500 mm, tube voltage of 130 kV, current of 75–310 mAs, and thickness of 3 mm. The images were reconstructed with 3D OSEM mode (four iterations and eight subsets) in a $4.07 \times 4.07 \times 5.00$ mm³ matrix.

TABLE 1 Clinical and demographic characteristics of the study patients.

Variable	n (%)	%
N = 47		
Age (mean ± SD)	47 ± 23 years	
Origin		
Metropolitan area	26	55.4
Inland cities	21	44.6
Education		
Illiterate	17	33.1
0 to 8 years	24	51.0
8 to 12 years	06	12.7
Smoking		
Non-smoker	23	48.9
<20 pack-year	04	8.5
>20 pack-year	13	27.6
Ex-smoker for > 5 years	08	14.8
Number of children		
1 child	6	12.7
2 children	7	14.8
3 or more children	34	72.3
Number of sexual partners		
Up to two partners	12	25.6
Three or two partners	35	74.4
Tumor size (cm) (mean ± SD)	5.43 cm (SD 1.49)	
FIGO staging		
IB2	02	4.20
II	04	8.0
III	21	44.6
IV	20	42.5

FIGO, international federation of gynecology and obstetrics.

Radiomic analysis

Segmentation

We use the free access multiplatform Local Image Features Extraction (LIFEx) software (V6.30—Inserm, Orsey, France) (21), as can be seen in **Supplementary Figure 1**. Initially, a semi-automatic segmentation of the uterine cervical lesion was performed (whole-body image only), identified by the ¹⁸F-FDG uptake on the CT fusion image, and manually outlined with a 3D design tool. Subsequently, the software selected the area of highest uptake, considering a fixed threshold of 40% of the standard uptake value (SUV) of the ROI volume (VOI), a method validated for cervical uterine neoplasms (22, 23). Notably, the details regarding the computation parameters and formulas are described at www.lifexsoft.org (21). A radiologist specialized in the female pelvis and supervised by a nuclear medicine specialist, both with 20 years of experience, did the segmentations for all patients.

Extraction

For each selected volume, a massive extraction of numerical data was performed by LIFEx, using 4 × 4 × 4 resizing,

0.25 fixed number width (FBW) intensity discretization method and histogram redefinition, obtaining 50 tumor features. These features were divided into categories, including: first-order statistics derived from the voxel intensity histogram (shape, volume, and histogram), and conventional indices (SUVpeak, SUVmean, SUVmax, MTV, and TLG); second-order statistics, including features based on the gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), neighboring gray-level dependence matrix (NGLDM), and gray-level zone length matrix (GLZLM) (12, 21, 24).

Selection of radiomic features

Initially, searching for clinically significant markers associated with OS, we performed an independent sample test with the Mann–Whitney to assess the distribution for each feature in the two groups, including those with a *p*-value < 0.05, to subsequent analysis.

After that, the data were submitted to dimension reduction through rank correlation with Spearman's coefficient, evaluating each pair of features. Later, we found which markers correlated with each other, excluding redundant markers using a correlation matrix and selecting those with a pre-established hypothetical *rho* lower than 0.85.

Statistical analysis

The absolute and relative frequency described categorical variables in percentage. Continuous variables with a normal distribution were analyzed by the mean and standard deviation; while non-parametric variables were analyzed by the median, maximum and minimum values, and interquartile range (IQR). For comparison between variables, we used the Mann–Whitney *U* test. We determined the cut-off points for variables with a *p*-value < 0.05 and the distinction between groups by ROC curves (DeLong methodology).

For prognostic evaluation, we correlated the selected radiomic features with the OS. Kaplan–Meier survival curves were constructed, with cut-off points obtained by the ROC curve for each variable, using the MedCalc software (MedCalc Software Ltd, Ostend, Belgium; <https://www.medcalc.org>; 2022). *P*-values lower than 0.05 were considered statistically significant.

Results

Clinical and demographic characteristics

The sample was initially composed of 50 consecutive patients. Three patients were excluded: one whose pretreatment baseline scan was unavailable and two other scans with divergence in the acquisition parameters (disagreement with IBSI standards).

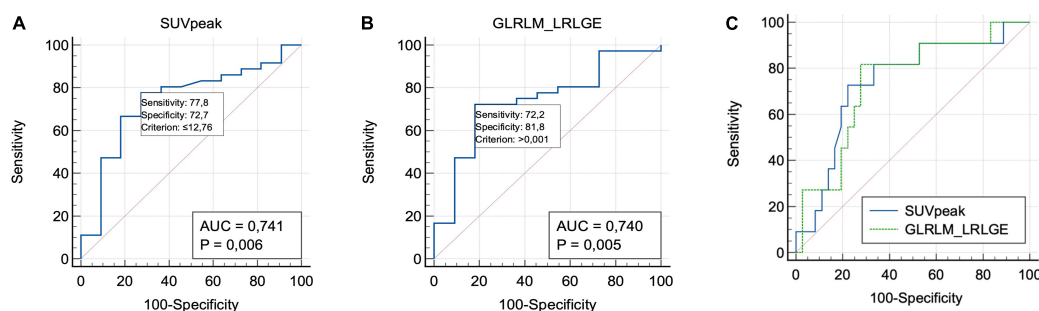


FIGURE 1

Receiver operating characteristic curve analyses of SUVpeak (A), GLRLM_LRLGE (B), and both (C) value for predicting overall survival.

Therefore, 47 patients were eligible for this study, with a mean age of 47 ± 23 years and ranging from 24 to 70 years. The majority of the patients had a low level of education, with reports of multiple sexual partners. Approximately 87% of patients presented with advanced stages of the disease (FIGO III and IV) (Table 1). Advanced stages of FIGO were correlated with lower overall survival (Supplementary Figure 2).

The median follow-up was 23.5 months (range: 3.73–39 months), with all surviving patients being followed up for at least 24 months. Of a total of 47 patients, 36 (77%) were alive at the end of 24 months (group 1) and 11 (23%) patients had died due to the disease (group 2).

Metabolic biomarkers and textural radiomic features

The data were extracted from 47 VOIs. We selected the features with discriminatory power for the selected outcome: three metabolic parameters—SUVmax ($p = 0.02$), SUVmean ($p = 0.02$), and SUVpeak ($p = 0.01$); and 13 textural markers—five markers from the GLZLM matrix (GLZLM_SZE, GLZLM_LGZE, GLZLM_HGZE, GLZLM_SZLGE, and GLZLM_SZHGE); six markers from the GLRLM matrix (GLRLM_LGRE, GLRLM_HGRE, GLRLM_SRLGE, GLRLM_SRHGE, GLRLM_LRLGE, and GLRLM_LRHGE); and two markers from the GLCM matrix (GLCM_Contrast variance and GLCM_Dissimilarity).

Among the metabolic parameters, the SUVpeak showed the best performance to differentiate the groups. The SUVpeak median in group 1 was 10.89 (IQR 7.60–12.69), while in group 2 was 13.87 (IQR 12.17–14.14), $p = 0.01$. The best cut-off point value (ROC curve analysis) was 12.76 with an AUC of 0.74, a sensitivity of 77.8%, and a specificity of 72.7%, $p = 0.006$.

The SUVmax median value in group 1 was 12.83 (IQR 9.09–14.90) vs. 15.98 in group 2 (IQR 13.52–19.09), $p = 0.02$. The best cut-off point was 14.32, AUC = 0.68, sensitivity = 72.3%, and specificity = 72.7% for the cut-off point of 14.32 ($p = 0.012$). The SUVmean median value in group 1 was 7.68 (IQR 9.09–14.90)

vs. 9.88 in group 2 (IQR 8.88–10.92), $p = 0.02$. It presented an AUC of 0.68, sensitivity of 72.3%, and specificity of 72.7% for a cut-off point of 8.8 ($p = 0.01$).

The other conventional metabolic metrics were not significant. The median MTV in group 1 was 31.9 (IQR: 18.5–51.0) vs. 37.8 (IQR: 24.6–72.4) in group 2 ($p = 0.49$). The median TLG in group 1 was 295.9 (IQR: 100.7–403.7) vs. 320.3 (IQR: 253.2–465.7) in group 2, $p = 0.33$.

Aiming for the redundancy feature reduction, we used the Spearman rank correlation for each of these 13 attributes. Three of them showed a rho value lower than 0.85: GLRLM_LGRE, GLRLM_SRLGE, and GLRLM_LRLGE. When we compared the AUC of these three indices, the GLRLM_LRLGE textural feature presented a little better performance. The GLRLM_LRLGE median in group 1 was 1.2×10^{-3} (IQR: 8×10^{-4} – 32×10^{-3}) vs. 7.7×10^{-3} in group 2 (IQR 6×10^{-4} – 9×10^{-3} , $p = 0.017$). The best cut-off point value was 1×10^{-3} (AUC: 0.74; sensitivity: 72.2%; specificity: 81.8%, $p = 0.005$).

For GLRLM_LGRE, the group 1 median value was 1.2×10^{-3} (IQR: 7×10^{-4} – 2.4×10^{-3}), and in group 2 was 7×10^{-4} (IQR: 5×10^{-4} to 8×10^{-4} , $p = 0.01$). It presented an AUC of 0.73, sensitivity of 81.8%, and specificity of 72.2% for a cut-off point of 9×10^{-4} ($p = 0.006$). For GLRLM_SRLGE, the group 1 median value was 1.2×10^{-3} (IQR: 7×10^{-4} – 2.3×10^{-3}), and in group 2 was 7×10^{-4} (IQR: 5×10^{-4} – 8×10^{-4} , $p = 0.01$). The AUC was 0.73, sensitivity of 81.8%, and specificity of 72.2% for a cut-off point of 8×10^{-4} ($p = 0.006$) (Figure 1). More information at Table 2.

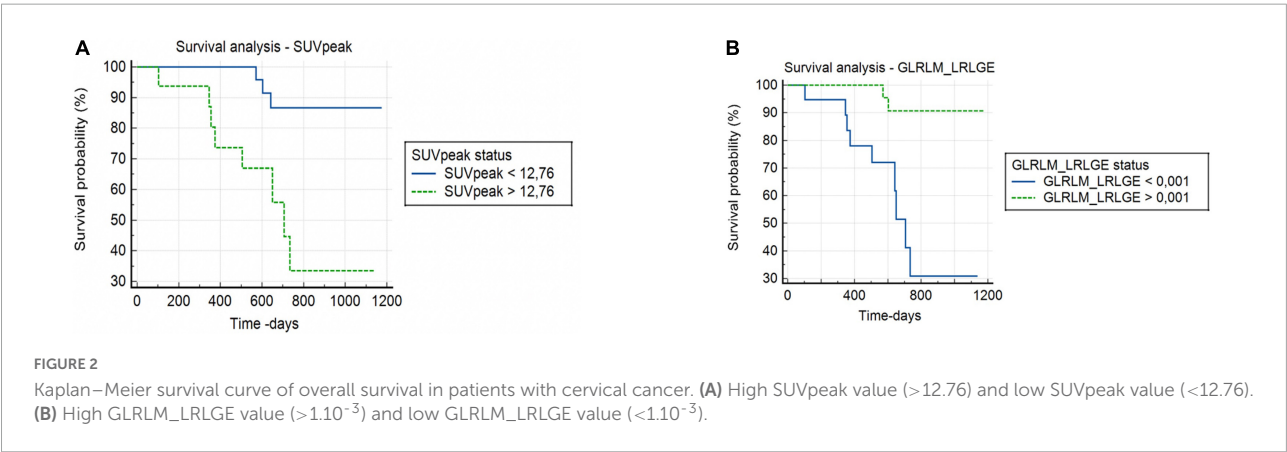
Correlations between conventional parameters ^{18}F -FDG PET/CT and textural features

The SUVpeak showed negative correlations with the GLRLM matrix. GLRLM_LRLGE ($r = -0.890$, $p < 0.01$). The SUVmax also showed a negative correlation with GLRLM_LRLGE ($r = 0.764$, $p < 0.01$).

TABLE 2 Results of independent-samples Mann–Whitney test analysis parameters 18F-FDG PET/CT for cervical cancer predicting overall survival.

	Group 1		Group 2		P
	Median	Range	Median	Range	
Image-based parameters					
SUV mean	7.68	9.09–14.90	9.88	8.88–10.92	0.02
SUV peak	10.89	7.60–12.69	13.87	12.17–14.14	0.01
SUV max	12.83	9.09–14.90	15.98	13.52–19.09	0.02
TLG (mL)	295.9	100.7–403.7	320.3	253.2–465.7	0.33
MTV (mL)	31.9	18.5–51.0	31.9	24.6–72.4	0.49
Texture parameters					
GLRLM_LGRE	1.2×10^{-3}	7×10^{-4} – 2.4×10^{-3}	7×10^{-4}	5×10^{-4} – 8×10^{-4}	0.01
GLRLM_SRLGE	1.2×10^{-3}	7×10^{-4} – 2.3×10^{-3}	7×10^{-4}	5×10^{-4} – 8×10^{-4}	0.01
GLRLM_LRLGE	1.2×10^{-3}	8×10^{-4} – 3.2×10^{-3}	7.7×10^{-3}	6×10^{-4} – 9×10^{-4}	0.01

¹⁸F-FDG PET/CT, ¹⁸fluorodeoxyglucose positron emission tomography; Group 1, survivors; Group 2, died; SUV, standardized uptake values; SUVmax, maximum standardized uptake value; SUVmean, mean standardized uptake value; SUVpeak, the peak of SUV in 1 mL; TLG, total lesion glycolysis; MTV, metabolic tumor volume; GLRLM, Gray level run length matrix; LGRE, low gray-level runs emphasis; SRLGE, short runs low gray-level emphasis; LRLGE, long runs gray-level emphasis.



Both SUVpeak and GLRLM_LRLGE were not correlated with FIGO staging (Supplementary Table 1).

Kaplan–Meier survival analysis

GLRLM_LRLGE showed a significant correlation with the OS ($p = 0.003$). Patients who died presented a GLRLM_LRLGE value lower than the cut-off point, with a shorter survival time: median of 708 days (CI: 505.0–734.0). A risk ratio of 10.8 (CI: 3.0–39.1) was observed.

SUVpeak showed a significant correlation with the OS ($p = 0.006$). Patients who died presented a higher SUVpeak value, with a shorter survival time: a median of 706 days (CI: 374.0–734.0). A risk ratio of 10.5 (CI: 2.7–40.3) was observed (Figure 2).

Discussion

This study demonstrated the prognostic association between radiomic biomarkers of primary uterine cervical cancer lesions

at ¹⁸F-FDG-PET/CT and overall survival. Among the evaluated metabolic parameters, SUVpeak showed the best discriminatory power; and among all the selected radiomic textural features, the GLRLM_LRLGE presented the best predictive performance. Moreover, SUVpeak and GLRLM_LRLGE demonstrated a greater correlation with OS compared with clinical and other more conventional ¹⁸F-FDG-PET/CT parameters, including MTV and TLG. These data reinforce the importance of metabolic radiomic evaluation in cervical uterine tumor staging.

Cervical cancer accounts for high morbidity and mortality in patients of productive and reproductive age worldwide, especially in vulnerable populations (25). The staging of this neoplasm is based on FIGO classification, which includes characteristics of the primary lesion, and lymph node or distant dissemination (4–6). However, FIGO classification presented a low accuracy in predicting therapy response and survival, especially among patients with advanced-stage cancer disease (4).

The search for non-invasive and robust prognostic biomarkers can improve the predictive power of therapy response. Radiomic is considered a promising analysis tool in

precision medicine (26, 27). Some studies have also reported the use of this technology in cervical tumor cases, based on several imaging methods, especially magnetic resonance imaging and ^{18}F -FDG-PET/CT (28). Usually, these studies aim to evaluate several aspects of the tumor, covering most frequently lymph node invasion (29–31), prognosis (28, 32), and therapeutic response (33); followed by histological grade (34–36), staging (37), and lymphovascular space invasion (29).

Standardized uptake value represents a semi-quantitative metric of ^{18}F -FDG-PET/CT with prognostic ability, including OS evaluation of patients with uterine cervix tumors (10, 33). All metrics of SUV are correlated with each other, providing information on the tumor metabolic activity. SUVpeak is reported as more robust and reproducible than SUVmax and SUVmean, although it is not widely disseminated in clinical practice (7). Studies report better performance of SUVpeak to demonstrate the aggressiveness of early-stage cervical tumors compared to SUVmax, maybe because SUVpeak measures several voxels in a more metabolically active spherical VOI of the lesion (7, 38).

The SUVpeak in our study presents a cut-off value similar to those described in other studies. Schernberg et al. (39) analyzed locally advanced disease treated with definitive chemoradiation and demonstrated that a high SUVpeak value was superior in predicting the OS and local recurrence, when compared with other ^{18}F -FDG-PET/CT parameters, like MTV and TLG. Other studies also evaluated early-stage cervical cancer, in which a low SUVpeak was significantly correlated with high progression-free survival (40).

A systematic review by Piñeiro-Fiel et al. evaluated the radiomics of ^{18}F -FDG-PET/CT in several neoplasms. Gynecological cancers were among the four most studied types, with 19 publications in a total of 741 studies. Of these 19 publications, cervical uterine cancer accounted for the largest number of publications (74%), followed by endometrial cancer (16%). As in our study, the textural features were correlated with the conventional metrics of ^{18}F -FDG-PET/CT, including SUV. In the analysis of gynecological cancers, the texture matrices that presented higher significance were GLCM, GLRLM, and GLZSM (41).

We showed that GLRLM_LRLGE could perform well in predicting OS in patients with advanced cervical cancer. The radiomic matrix GLRLM conceptually relates to the intensity of the gray level of pixels in an image, in a given direction, and LRLGE represents the distribution of long stretches with a high or low gray level, being an indicator of the uniformity of the homogeneous distribution of FDG uptake (42). GLRLM_LRLGE is a potential biomarker in other neoplasms too, as it can discriminate benign from malignant renal tumors (42), and can be used to assess recurrence in rectal cancer (43).

Additionally, some studies demonstrated a significant correlation between the GLRLM matrix ^{18}F -FDG-PET/CT

textural markers (LGRE, SRLGE, and LRLGE) and RNA-level immunological biomarkers of PD-L1 (programmed death ligand 1) in lung cancer (44). PD-L1 protein expression is also a predictive biomarker in uterine cervical cancer (45). Subsequently, we could find a possible intercorrelation between these textural markers (GLRLM_LRLGE) and PD-L1 expression, representing an important prognostic and selection factor for immunotherapy. This hypothesis may be evaluated in future prospective studies.

GLRLM_LRLGE possibly shows a relationship with tumor necrosis, as it assesses the homogeneity of ^{18}F -FDG uptake, and its highest value is documented in benign homogeneous lesions (42). On the other hand, several studies demonstrate a direct relationship between PD-L1 and tumor necrosis factor (TNF alpha) in oncologic diseases, including findings in which TNF alpha produced by adipocytes positively regulates PD-L1 (46). Based on these findings, we can assume that the textural factor GLRLM_LRLGE also correlates with TNF alpha.

However, this study has many limitations. It is a single-center study with a low number of patients and a retrospective analysis. However, the sample was derived from a clinical trial (47), with a relatively homogeneous and controlled group of patients with a good clinical follow-up. Additionally, the ^{18}F -FDG-PET/CT pretreatment images were reevaluated in order to collect new data regarding radiomic characteristics in the primary lesions. Moreover, we analyzed only the scans with a protocol following the parameters established by the IBSI. We also do not perform multiple correction tests in our data, mainly because of the low number of included patients.

In conclusion, in patients with advanced cervical tumors, this study investigated and identified two biomarkers with better prognostic performance (SUVpeak and GLRLM_LRLGE). These features denote metabolism and intratumoral textural homogeneity, respectively. In the future, the SUVpeak and GLRLM_LRLGE have the potential to be incorporated into clinical practice, helping to identify patients with a higher risk of death.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Comitê de Ética em Pesquisa do IMIP. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

NA, SB, and MM conceptualized and designed the study. T-MC and CA were acquired the data. NA and TM identified and marked the lesions. MM and FM inferred results, implemented the methods, and analyzed and interpreted the data. NA, MO, and SB drafted the manuscript and wrote the first draft of the manuscript. LM, FM, SB, and MO critically revised the manuscript and provided supervision, support, conceptualization, and guidance throughout the project. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2022.1046551/full#supplementary-material>

References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Arbyn M, Weiderpass E, Bruni L, Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob Health.* (2019) 8:e191–120. doi: 10.1016/S2214-109X(19)30482-6
- 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 Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660
- Bhatla N, Denny L. FIGO Cancer Report 2018. *Int J Gynaecol Obstet.* (2018) 143(Suppl. 2):2–3. doi: 10.1002/ijgo.12608
- Lee SI, Atri M. 2018 FIGO Staging system for uterine cervical cancer: enter cross-sectional imaging. *Radiology.* (2019) 292:15–24. doi: 10.1148/radiol.2019190088
- Salib MY, Russell JHB, Stewart VR, Sudderuddin SA, Barwick TD, Rockall AG, et al. 2018 FIGO staging classification for cervical cancer: added benefits of imaging. *Radiographics.* (2020) 40:1807–22. doi: 10.1148/rg.2020200013
- Sher A, Lacoëuille F, Fosse P, Vervueren L, Cahouet-Vannier A, Dabli D, et al. For avid glucose tumors, the SUV peak is the most reliable parameter for [(18)F]FDG-PET/CT quantification, regardless of acquisition time. *EJNMMI Res.* (2016) 6:21. doi: 10.1186/s13550-016-0177-8
- Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. European Association of Nuclear Medicine (EANM). FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging.* (2015) 42:328–54. doi: 10.1007/s00259-014-2961-x
- Im HJ, Bradshaw T, Solaiyappan M, Cho SY. Current methods to define metabolic tumor volume in positron emission tomography: which one is better? *Nucl Med Mol Imaging.* (2018) 52:5–15. doi: 10.1007/s13139-017-0493-6
- Herrera FG, Breuneval T, Prior JO, Bourhis J, Ohsahin M. 18F-FDG-PET/CT metabolic parameters as useful prognostic factors in cervical cancer patients treated with chemo-radiotherapy. *Rad Oncol.* (2016) 11:1–11. doi: 10.1186/s13014-016-0614-x
- Bollineni VR, Ytre-Hauge S, Gulati A, Halle MK, Woie K, Salvesen O. The prognostic value of preoperative FDG-PET/CT metabolic parameters in cervical cancer patients. *Eur J Hybrid Imaging.* (2018) 2:1–14. doi: 10.1186/s41824-018-0042-2
- Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* (2016) 278:563–77. doi: 10.1148/radiol.2015151169
- van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472
- Rizzo S, Botta F, Raimondi S, Origi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and challenges of image analysis. *Eur Radiol Exp.* (2018) 2:36. doi: 10.1186/s41747-018-0068-z
- Anbumani S, Jayaraman P, Anchaneay N, Bilimappa RS, Nambiraj A. Quantitative radiomic phenotyping of cervix cancer. *Int Clin Pathol J.* (2018) 6:26–8. doi: 10.15406/icpj.2018.06.00149
- Cook GJR, Siddique M, Taylor BP, Yip C, Chicklore S, Goh V. Radiomics in PET: principles and applications. *Clin Transl Imaging.* (2014) 2:269–76. doi: 10.1007/s40336-014-0064-0
- Ho KC, Fang YHD, Chung HW, Yen TC, Ho TY, Chou HH, et al. A preliminary investigation into textural features of intratumoral metabolic heterogeneity in (18)F-FDG PET for overall survival prognosis in patients with bulky cervical cancer treated with definitive concurrent chemo-radiotherapy. *Am J Nucl Med Mol Imaging.* (2016) 6:166–75.
- Mu W, Liang Y, Hall LO, Tan Y, Balagurunathan Y, Wenham R, et al. 18F-FDG PET/CT Habitat radiomics predicts outcome of patients with cervical cancer treated with chemo-radiotherapy. *Radiol Artif Intell.* (2020) 2:e190218. doi: 10.1148/ryai.2020190218
- Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv.* (2016) [Preprint]. doi: 10.48550/arXiv.1612.07003
- Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging.* (2010) 37:181–200. doi: 10.1007/s00259-009-1297-4
- Nioche C, Orhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res.* (2018) 78:4786–9. doi: 10.1158/0008-5472
- Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. *Eur J Nucl Med Mol Imaging.* (2013) 40:716–27. doi: 10.1007/s00259-012-2332-4

23. Tsujikawa T, Rahman T, Yamamoto M, Yamada S, Tsuyoshi H, Kiyono Y, et al. 18F-FDG PET radiomics approaches: comparing and clustering features in cervical cancer. *Ann Nucl Med.* (2017) 31:678–85. doi: 10.1007/s12149-017-1199-7
24. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, et al. Texture indexes and gray level size zone matrix application to cell nuclei classification. *Proceedings of the 10th International Conference on Pattern Recognition and Information Processing.* Minsk (2009).
25. Stelzle D, Tanaka LF, Lee KK, Khalil AI, Baussano I, Shah ASV, et al. Estimates of the global burden of cervical cancer associated with HIV. *Lancet Glob Health.* (2021) 9:e161–9. doi: 10.1016/S2214-109X(20)30459-9
26. Barbet J, Bernaudin M, Payoux P, Cicone F, Gaugler MH, Kraeber-Bodéré F. Editorial: nuclear medicine in the context of personalized medicine. *Front Med.* (2020) 7:252. doi: 10.3389/fmed.2020.00252
27. Ai Y, Zhu H, Xie C, Jin X. Radiomics in cervical cancer: current applications and future potential. *Crit Rev Oncol Hematol.* (2020) 152:102985. doi: 10.1016/j.critrevonc.2020.102985
28. Lucia F, Visvikis D, Vallières M, Desseroit MC, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemo-radiotherapy. *Eur J Nucl Med Mol Imaging.* (2019) 46:864–77. doi: 10.1007/s00259-018-4231-9
29. Li Z, Li H, Wang S, Dong D, Yin F, Chen A, et al. MR-based radiomics nomogram of cervical cancer in prediction of the lymph-vascular space invasion preoperatively. *J Magn Reson Imaging.* (2019) 49:1420–6. doi: 10.1002/jmri.26531
30. Chen X, Liu W, Thai TC, Castellano T, Gunderson CC, Moore K, et al. Developing a new radiomics-based CT image marker to detect lymph node metastasis among cervical cancer patients. *Comput Methods Programs Biomed.* (2020) 197:105759. doi: 10.1016/j.cmpb.2020.105759
31. Song J, Hu Q, Ma Z, Zhao M, Chen T, Shi H. Feasibility of TWI-MRI-based radiomics nomogram for predicting normal-sized pelvic lymph node metastasis in cervical cancer patients. *Eur Radiol.* (2021) 31:6938–48. doi: 10.1007/s00330-021-07735-x
32. Ferreira M, Lovinfosse P, Hermesse J, Decuyppere M, Rousseau C, Lucia F, et al. [F]FDG PET radiomics to predict disease-free survival in cervical cancer: a multi-scanner/center study with external validation. *Eur J Nucl Med Mol Imaging.* (2021) 48:3432–43. doi: 10.1007/s00259-021-05303-5
33. Reuzé S, Orlhac F, Chargari C, Nioche C. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. *Oncotarget.* (2017) 8:43169–79. doi: 10.18632/oncotarget.17856
34. Shen WC, Chen SW, Liang JA, Hsieh TC, Yen KY, Kao CH. [18F]Fluorodeoxyglucose positron emission tomography for the textural features of cervical cancer associated with lymph node metastasis and histological type. *Eur J Nucl Med Mol Imaging.* (2017) 44:1721–31. doi: 10.1007/s00259-017-3697-1
35. Liu Y, Zhang Y, Cheng R, Liu S, Qu F, Yin X, et al. Radiomics analysis of apparent diffusion coefficient in cervical cancer: a preliminary study on histological grade evaluation. *J Magn Reson Imaging.* (2019) 49:280–90. doi: 10.1002/jmri.26192
36. Su X, Chen N, Sun H, Liu Y, Yang X, Wang W, et al. Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain. *Neuro Oncol.* (2020) 22:393–401. doi: 10.1093/neuonc/noz184
37. Umutlu L, Nensa F, Demircioglu A, Antoch G, Herrmann K, Forsting M, et al. Radiomics Analysis of Multiparametric PET/MRI for N- and M-Staging in patients with primary cervical cancer. *Rofo.* (2020) 192:754–63. doi: 10.1055/a-1100-0127
38. Vanderhoek M, Perlman SB, Jerai R. Impact of the definition of peak standardized uptaken value on quantification of treatment response. *J Nucl Med.* (2012) 53:4–11. doi: 10.2967/jnumed.111.093443
39. Schernberg A, Reuze S, Orlhac F, Buvat I, Derclé L, Sun R, et al. A score combining baseline neutrophilia and primary tumor SUVpeak measured from FDG PET is associated with outcome in locally advanced cervical cancer. *Eur J Nucl Med Mol Imaging.* (2018) 45:187–95. doi: 10.1007/s00259-017-3824-z
40. Zhang L, Sun H, Du S, Xu W, Xin J, Guo Q. Evaluation of 18F-FDG PET/CT parameters for reflection of aggressiveness and prediction of prognosis in early-stage cervical cancer. *Nucl Med Commun.* (2018) 39:1045–52. doi: 10.1097/MNM.0000000000000909
41. Piñero-Fiel M, Moscoso A, Pubul V, Ruibal Á, Silva-Rodríguez J, Aguiar P. A systematic review of PET textural analysis and radiomics in cancer. *Diagnostics.* (2021) 11:380. doi: 10.3390/diagnostics11020380
42. Matsumoto S, Arita Y, Yoshida S, Fukushima H, Kimura K, Yamada I, et al. Utility of radiomics features of diffusion-weighted magnetic resonance imaging for differentiation of fat-poor angiomyolipoma from clear cell renal cell carcinoma: model development and external validation. *Abdom Radiol.* (2022) 47:2178–86. doi: 10.1007/s00261-022-03486-5
43. Park H, Kim KA, Jung JH, Rhie J, Choi SY. MRI features and texture analysis for the early prediction of therapeutic response to neoadjuvant chemo-radiotherapy and tumor recurrence of locally advanced rectal cancer. *Eur Radiol.* (2020) 30:4201–11. doi: 10.1007/s00330-020-06835-4
44. Kim BS, Kang J, Jun S, Kim H, Pak K, Kim GH, et al. Association between immunotherapy biomarkers and glucose metabolism from F-18 FDG PET. *Eur Rev Med Pharmacol Sci.* (2020) 24:8288–95. doi: 10.26355/eurrev_202008_22625
45. Rotman J, den Otter LAS, Bleeker MCG, Samuels SS, Heeren AM, Roemer MGM, et al. PD-L1 and PD-L2 expression in cervical cancer: regulation and biomarker potential. *Front Immunol.* (2020) 17:596825. doi: 10.3389/fimmu.2020.596825
46. Li Z, Zhang C, Du JX, Zhao J, Shi MT, Jin MW. Adipocytes promote tumor progression and induce PD-L1 expression via TNF- α /IL-6 signaling. *Cancer Cell Int.* (2020) 20:179. doi: 10.1186/s12935-020-01269-w
47. Azevedo CRAS, Thuler LCS, Mello MJG, Lima JTO, Fonte ALF, Fontão DFS, et al. Phase II trial of neoadjuvant chemotherapy followed by chemoradiation in locally advanced cervical cancer. *Gynecol Oncol.* (2017) 146:560–5. doi: 10.1016/j.ygyno.2017.07.006



OPEN ACCESS

EDITED BY

Salvatore Annunziata,
Fondazione Policlinico Universitario A. Gemelli
IRCCS, Italy

REVIEWED BY

Paola Mapelli,
Vita-Salute San Raffaele University, Italy
Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

*CORRESPONDENCE

M. C. Cannatà
✉ chiacanna@gmail.com

SPECIALTY SECTION

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 01 October 2022

ACCEPTED 03 January 2023

PUBLISHED 19 January 2023

CITATION

Chiesa S, Russo R, Beghella Bartoli F, Palumbo I, Sabatino G, Cannatà MC, Gigli R, Longo S, Tran HE, Boldrini L, Dinapoli N, Votta C, Cusumano D, Pignotti F, Lupattelli M, Camilli F, Della Pepa GM, D'Alessandris GQ, Olivi A, Balducci M, Colosimo C, Gambacorta MA, Valentini V, Aristei C and Gaudino S (2023) MRI-derived radiomics to guide post-operative management of glioblastoma: Implication for personalized radiation treatment volume delineation. *Front. Med.* 10:1059712. doi: 10.3389/fmed.2023.1059712

COPYRIGHT

© 2023 Chiesa, Russo, Beghella Bartoli, Palumbo, Sabatino, Cannatà, Gigli, Longo, Tran, Boldrini, Dinapoli, Votta, Cusumano, Pignotti, Lupattelli, Camilli, Della Pepa, D'Alessandris, Olivi, Balducci, Colosimo, Gambacorta, Valentini, Aristei and Gaudino. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

MRI-derived radiomics to guide post-operative management of glioblastoma: Implication for personalized radiation treatment volume delineation

S. Chiesa¹, R. Russo², F. Beghella Bartoli¹, I. Palumbo^{3,4}, G. Sabatino^{5,6}, M. C. Cannatà^{1*}, R. Gigli⁷, S. Longo¹, H. E. Tran¹, L. Boldrini¹, N. Dinapoli¹, C. Votta¹, D. Cusumano⁷, F. Pignotti^{5,6}, M. Lupattelli⁴, F. Camilli³, G. M. Della Pepa⁶, G. Q. D'Alessandris⁶, A. Olivi⁶, M. Balducci¹, C. Colosimo², M. A. Gambacorta¹, V. Valentini¹, C. Aristei^{3,4} and S. Gaudino²

¹Department of Radiology, Radiation Oncology and Hematology, Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Rome, Italy, ²Department of Diagnostic Imaging, Oncological Radiotherapy and Hematology, Institute of Radiology, Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Rome, Italy, ³Radiation Oncology Section, University of Perugia, Perugia, Italy, ⁴Perugia General Hospital, Perugia, Italy, ⁵Department of Neurosurgery, Mater Olbia Hospital, Olbia, Italy, ⁶Department of Neurosurgery, Agostino Gemelli University Polyclinic (IRCCS), Rome, Italy, ⁷Medical Physics, Mater Olbia Hospital, Olbia, Italy

Background: The glioblastoma's bad prognosis is primarily due to intra-tumor heterogeneity, demonstrated from several studies that collected molecular biology, cytogenetic data and more recently radiomic features for a better prognostic stratification. The GLIFA project (GLIoblastoma Feature Analysis) is a multicentric project planned to investigate the role of radiomic analysis in GB management, to verify if radiomic features in the tissue around the resection cavity may guide the radiation target volume delineation.

Materials and methods: We retrospectively analyze from three centers radiomic features extracted from 90 patients with total or near total resection, who completed the standard adjuvant treatment and for whom we had post-operative images available for features extraction. The Manual segmentation was performed on post gadolinium T1w MRI sequence by 2 radiation oncologists and reviewed by a neuroradiologist, both with at least 10 years of experience. The Regions of interest (ROI) considered for the analysis were: the surgical cavity \pm post-surgical residual mass (CTV_cavity); the CTV a margin of 1.5 cm added to CTV_cavity and the volume resulting from subtracting the CTV_cavity from the CTV was defined as CTV_Ring. Radiomic analysis and modeling were conducted in RStudio. Z-score normalization was applied to each radiomic feature. A radiomic model was generated using features extracted from the Ring to perform a binary classification and predict the PFS at 6 months. A 3-fold cross-validation repeated five times was implemented for internal validation of the model.

Results: Two-hundred and seventy ROIs were contoured. The proposed radiomic model was given by the best fitting logistic regression model, and included the following 3 features: F_cm_merged.contrast, F_cm_merged.info.corr.2, F_rlm_merged.rlnu. A good agreement between model predicted probabilities and observed outcome probabilities was obtained (p -value of 0.49 by Hosmer and Lemeshow statistical test). The ROC curve of the model reported an AUC of 0.78 (95% CI: 0.68–0.88).

Conclusion: This is the first hypothesis-generating study which applies a radiomic analysis focusing on healthy tissue ring around the surgical cavity on post-operative MRI. This study provides a preliminary model for a decision support tool for a customization of the radiation target volume in GB patients in order to achieve a margin reduction strategy.

KEYWORDS

radiomic, glioblastoma, target volume definition, heterogeneity, precision medicine

1. Introduction

Glioblastoma (GB) continues to be the most common and threatening primary brain tumors in adults and despite a multimodal treatment (maximum safe surgical resection followed by adjuvant radio-chemotherapy with Temozolomide) the prognosis remains poor, with a median overall survival (OS) of 14.6 months and a median progression free survival (PFS) of 6.9 months (1). In spite of decades of research, our knowledge of this neoplasm is still limited. This bad prognosis is primarily due to intra-tumor heterogeneity, demonstrated also from several studies that collected molecular biology and cytogenetic data for a better prognostic stratification of glioblastoma.

The implementation of these markers, however, depends in routine clinical practice on surgical tissue (2). On the contrary, the use of medical imaging, as a non-invasive tool to derive prognostic factors that can predict outcome such as survival, PFS, and response to therapy, is becoming increasingly popular. The images can be described not only qualitatively in order to highlight the presence of necrotic, edemigenous, malignant, suspected or metabolically active areas, but also quantitatively in order to generate numbers that become real measurable data (3–5).

Radiomics (6) is the process that involves the high-throughput extraction of quantitative features by computing local macro and micro-scale morphologic changes in texture patterns (e.g., roughness, image homogeneity, regularity, edges) with the intent of creating mineable databases from radiographic images.

Some experiences with glioblastoma are reported *via* radiomics approaches to predict tumor's histological features (7), progression (8), grade, treatment response (9), or even overall survival (10–13).

Magnetic resonance imaging (MRI) is the imaging modality for characterizing GB in these studies and generally has an integral role in diagnosis, response assessment, surveillance and radiation treatment, especially for defining the volume of irradiation (14).

Defining the optimal target volume for GB is still a challenge and represents a balance between minimizing treatment related toxicity, while ensuring efficacy in terms of tumor control and allowing a re-irradiation approach. The recent ESTRO-ACROP guidelines in macroscopically resected GB recommend to add an isotropic margin

of 2 cm, adjusted to anatomical border, to resection cavity plus any residual enhancing tumor on contrast-enhanced T1 weighted MRI, without considering the peri-tumoral oedema.

This size of safety margin had traditionally been defined around 2–3 cm based on early anatomic and clinical research. In fact, the recurrences reported in several studies are mainly central, in field or marginal (80–90%) with 10–20% of lesions outside the irradiated field (15, 16).

Several studies have been conducted to identify look for strategies of margin reduction, such as peritumoral zone investigation, the analysis of pattern of recurrence (15) or integration between different imaging methods (17), but no clear indication of reducing margin is yet available (18–22).

In light of all these considerations, the GLIFA project (GLIoblastoma Feature Analysis) is a multicentric project planned to investigate the role of radiomic features in GB management. In particular, in this study we aim to verify whether there are any radiomic features in the tissue around the resection cavity which may guide the target volume delineation allowing a margin reduction strategy toward a personalized medicine approach (23).

TABLE 1 Eligibility criteria for GLI.F.A. Project.

Inclusion criteria	Exclusion criteria
<ul style="list-style-type: none"> – Histological diagnosis of GB > 18 yrs; – ECOG performance status <4; – Total or near-total resection; – Platelet counting > 100 × 10⁹/L; – Hb > 11 g/L; – GB > 4000/mm³; – Neutrophils > 1900/mm³; – Total bilirubin and alkaline phosphatase at less than 1.25 normal concentration; – Informed consent that documents that the patient has been informed in a way that is clear and comprehensible to him and that fits all aspects of the study. 	<ul style="list-style-type: none"> – Biopsy – Degenerative neurological diseases or other neuropsychiatric disorders; – Pregnancy status; – Respiratory failure; – Immunodepression status; – Chronic renal failure.

ECOG, Eastern Cooperative Oncology Group.

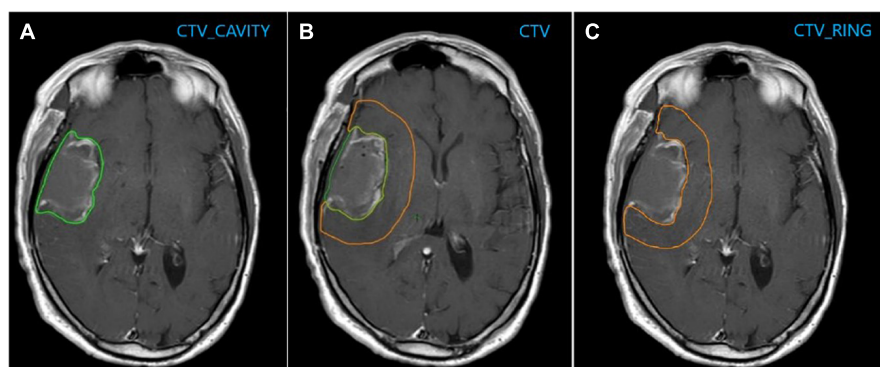


FIGURE 1

(A) CTV_cavity: Surgical cavity \pm post-surgical residual mass; (B) CTV: CTV_cavity + 1.5 cm; (C) CTV_Ring: CTV–CTV_Cavity.

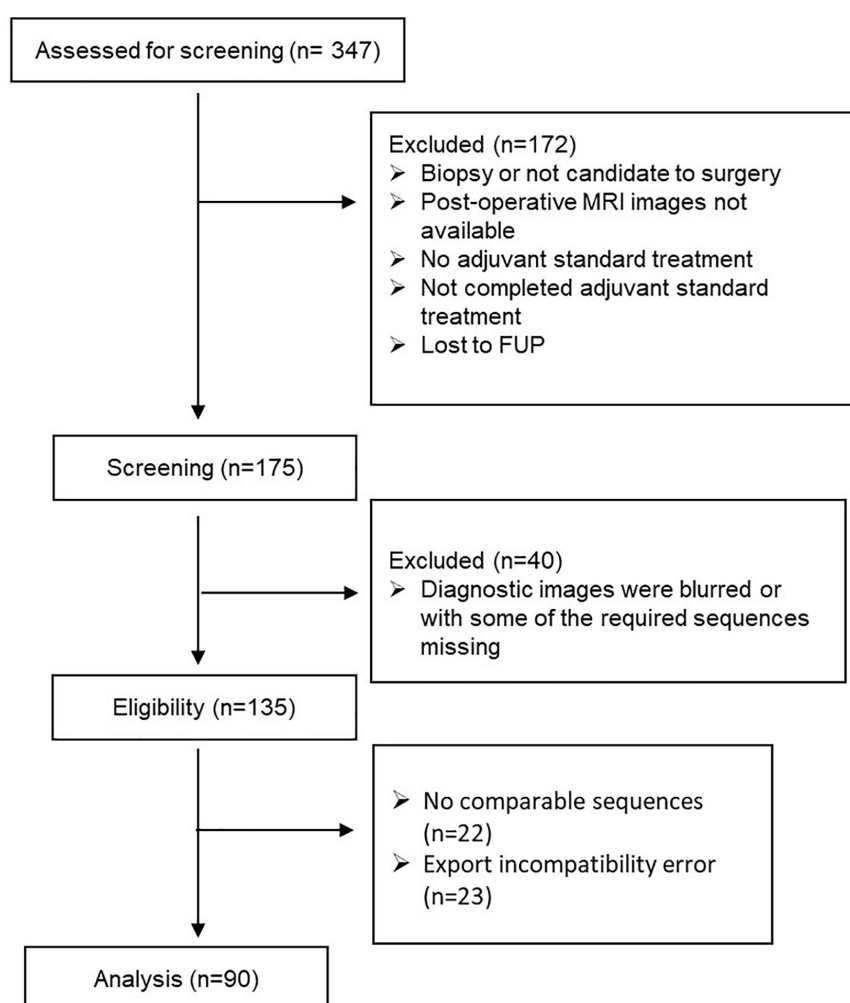


FIGURE 2

Patients' selection.

1.1. Patients selection

This is a multicentric retrospective study approved by the ethics committees of Institutions involved. All procedures performed were in accordance with the ethical standards of the institutional and/or national research committee and with the

1964 Helsinki declaration and its later amendments or comparable ethical standards.

All adult patients, with histologically proven glioblastoma Isocitrate dehydrogenase (IDH) wild-type underwent total or near-total resection of the enhancing tumor, followed by standard radio-chemotherapy and adjuvant chemotherapy (1), who have performed

MRIs according to a timeline protocol of image acquisition shared among the project participants, were considered eligible in this study (Table 1). All MRIs must contain at least the post-contrast T1-weighted sequences and T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) and relative images must be available in the required imaging protocol descriptions from Digital Imaging and Communication in Medicine, or DICOM format (24).

Patients, clinical data and MRI data of GB were obtained from three centers (Università degli studi di Perugia e Azienda ospedaliera di Perugia; Fondazione Policlinico Universitario Policlinico Agostino Gemelli, IRCCS; Mater Olbia Hospital).

Data were collected from patients treated from 2016 to October 2020, with total or near total resection, who completed the standard adjuvant treatment, with at least 9 months of follow-up and for whom we had post-operative images available for features extraction.

The MRIs of these patients were examined and patients whose diagnostic images were blurred or with some of the required sequences missing were excluded from the contouring phase.

1.2. Image acquisition and segmentation

Imaging was performed on 1.5 T MRI unit from different manufactures (Philips Medical Systems, SIEMENS, GE Medical Systems).

One sequence was included in the current study: gadolinium (Gd) enhanced T1-weighted FSPGR (T1c). The images were acquired with the following imaging parameters: slice thickness 4–5 mm, pixel spacing 0.35–0.90 mm.

The images were loaded in a radiation therapy delineation console (Eclipse, Varian Medical Systems, Palo Alto, CA, USA) and in the open-source software 3D Slicer for the definition of regions of interest (ROI).

Manual segmentation was performed on post gadolinium T1w MRI sequence by cooperation of 2 radiation oncologists expert in the management of brain cancer, with at least 10 years of experience (SC, FB), and all cases were individually reviewed by a neuroradiologist with at least 10 years of experience (SG, RR).

The ROI considered for the analysis were the following: the surgical cavity \pm post-surgical residual mass clinical target volume_cavity (CTV_cavity); a margin of 1.5 cm was added to CTV_cavity to obtain the CTV and the volume resulting from subtracting the CTV_cavity from the CTV was defined as CTV_Ring (Figure 1).

1.3. Radiomic feature extraction

Radiomic features were extracted from the CTV_Ring using MODDICOM, an open-source R library developed for radiomic feature extraction (25). This software was validated and calibrated within the Image Biomarker Standardization Initiative, which aimed to standardize the definition and computation of radiomic features among different software implementations (26).

In total, 226 radiomic features belonging to different feature families were extracted for each CTV_Ring. 17 statistical features provided statistical measures of the gray-level histogram of the ROI; 14 morphological features provided morphological descriptors of the ROI; 195 textural features described properties of the local

TABLE 2 Clinical data characteristics of patients with glioblastoma (GB) ($n = 90$).

Characteristics	<i>n</i> (%)
Gender	
Male	62 (68,9%)
Female	28 (31,1%)
Age	
Median	61,7 yrs
Min	80 yrs
Max	39 yrs
<50 yrs	12 (13, 3%)
≥ 50 yrs	78 (86, 7%)
MGMT-gene metylation	
Not	37 (41, 1%)
Yes	48 (53,3%)
NA	5 (5, 6%)
Type of surgery	
GTR	23 (25, 6%)
STR	67 (74, 4%)
IDH	
IDH wild-type	100 (100%)
PFS	
PFS ≤ 6 months	30 (33, 3%)
PFS > 6 months	60 (66, 7%)

GTR, gross total resection; STR, subtotal resection; IDH, Isocitrate dehydrogenase; MGMT, methylguanine-DNA methyl-transferase; yrs, years.

distribution of the gray levels within the ROI based on co-occurrence of gray levels, consecutive sequence of pixels or zones with the same gray level (27).

1.4. Radiomic feature selection and radiomics modeling

Radiomics analysis and modeling were conducted in RStudio (R version 3.6.3). Z-score normalization was applied to each radiomic feature before further analysis.

We generated a radiomic model using the features extracted from the CTV_Ring to perform a binary classification and predict the PFS at 6 months. Class 1 represented PFS below or equal to 6 months, while class 0 represented PFS above 6 months.

Feature selection was implemented to reduce the number of variables included in the model and prevent overfitting. A univariate analysis was performed using the Wilcoxon-Mann-Whitney statistical test, which tested the statistically significant difference between the two classes for each radiomic feature. A significance level of 0.05 was set for the univariate analysis. The collinearity of the statistically significant features was assessed by computing the Pearson cross-correlation coefficient. We set a threshold of 0.9 for the Pearson coefficient to exclude collinear (highly correlated) features.

Different logistic regression models were generated using the selected features. The best fitting model was determined with a

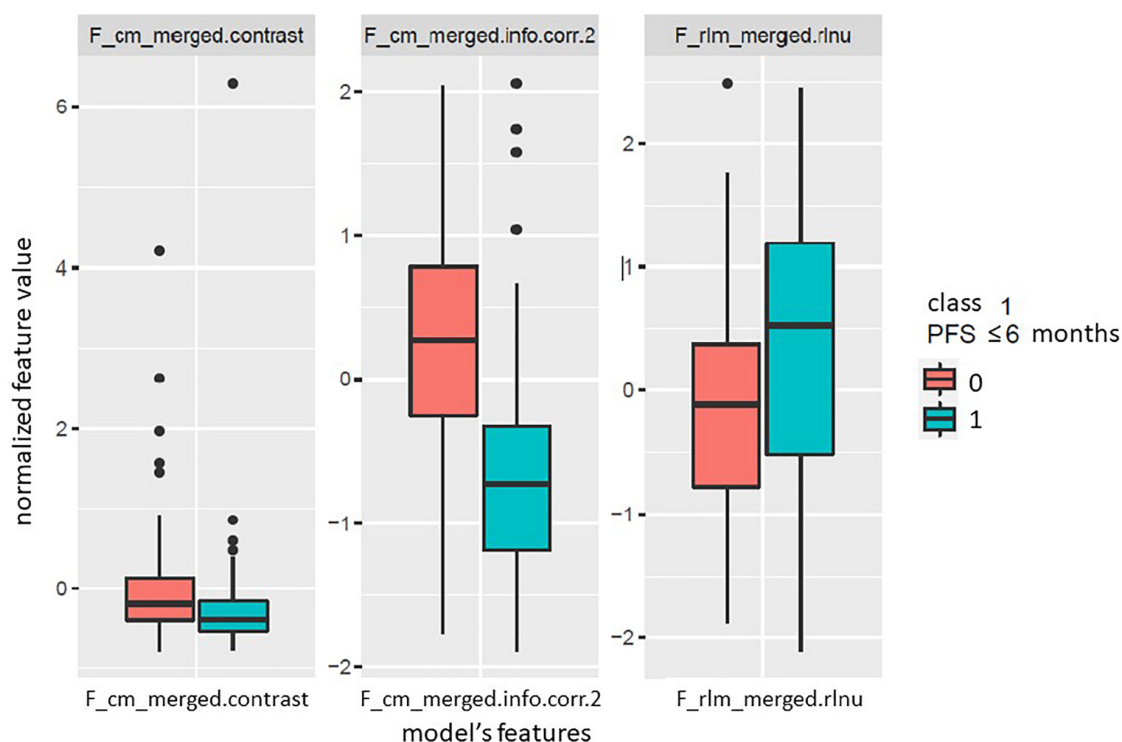


FIGURE 3

Boxplots of the radiomic features included in the developed logistic regression model for the two classes of the outcome. Class 1 (cyan) indicates progression free survival (PFS) below or equal to 6 months, while class 0 (red) indicates PFS above 6 months.

stepwise feature selection according to the Akaike Information Criteria (28), to compromise between model fitting goodness and model complexity.

1.5. Radiomic model performance and validation

The internal calibration of the proposed model was evaluated by producing the calibration plot, reporting model predicted probabilities against observed outcome probabilities, and by means of the Hosmer and Lemeshow goodness-of-fit statistic. A p -value > 0.05 indicated that there was no statistically significant difference between model predicted probabilities and observed outcome probabilities (29).

The discrimination performance of the proposed model was assessed by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and by computing the classification evaluation metrics.

The 95% confidence interval (CI) for the AUC was yielded by performing 2000 stratified bootstrap resampling. Sensitivity, specificity, positive and negative predictive values (PPV, NPV) were computed after defining the probability threshold as the best cut-off according to the Youden's index method. The 95% CI of these evaluation classification metrics was obtained by adopting the Jeffreys method for small sample sizes (30).

A 3-fold cross-validation repeated five times was implemented for internal validation of the model. Mean and standard deviations of the evaluation classification metrics were calculated over the five repetitions (31, 32).

2. Results

2.1. Patient population

From January 2016 to October 2020, we collected consecutive 347 newly pathologically confirmed patients with GB and screened these cases (Figure 2).

90 patients were considered to retrospectively analyze the pattern of radiomic features.

Patients' characteristics are reported in Table 2.

2.2. Development and validation of radiomic model

Based on the Wilcoxon–Mann–Whitney statistical test, 48 out of the extracted 226 radiomic features showed a statistically significant difference between the two classes. Following the correlation analysis with the Pearson coefficient, 12 out of the 48 remaining features

TABLE 3 Model coefficients and statistically significant p -values.

	Estimated model coefficient	Standard error	P -value
Intercept	−0.92	0.27	<0.001
F_cm_merged.contrast	0.89	0.36	0.013
F_cm_merged.info.corr.2	−1.10	0.34	0.0012
F_rlm_merged.rlnu	0.81	0.33	0.014

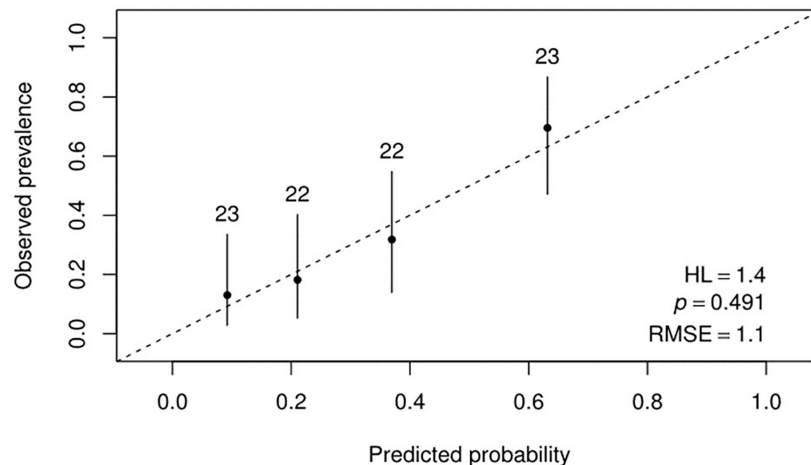


FIGURE 4

Calibration plot reporting the observed probabilities against the model predicted probabilities.

were retained for the model development phase. The proposed radiomic model was given by the best fitting logistic regression model, and included the following 3 features: F_cm_merged.contrast, F_cm_merged.info.corr.2, F_rlm_merged.rlnu.

The boxplots represented in Figure 3 show the distribution of the selected radiomic features used in the model for the two classes of outcome. Table 3 reports the estimated model coefficients and the statistically significant p -values (p -value < 0.05). The feature F_cm_merged.info.corr.2 which presented no overlap of the interquartile ranges of the two classes, as shown in Figure 3, was also associated to the most statistically significant p -value of the model coefficients.

A good agreement between model predicted probabilities and observed outcome probabilities was obtained, as showed in the calibration plot (Figure 4) and as indicated by the p -value of 0.49 resulting from the Hosmer and Lemeshow statistical test. Figure 5 represents the ROC curve of the model with an AUC of 0.78 (95% CI: 0.68–0.88). The discrimination performances of the model for the binary classification are reported in Table 4 for model fitting and internal validation. The cross-validation confirmed the performances

obtained during model fitting with a slight or no decrease of the metrics, suggesting that no overfitting had occurred. Specifically, the specificity decreased from 0.80 during model fitting to 0.75 for the cross-validation, while the NPV remained stable at 0.84.

3. Discussion

The emerging big challenge in the field of medical research is to identify multimodal predictive/prognostic factors (clinical, imaging and molecular data) and integrate them in a quantitative manner to provide prediction models that estimate patient outcomes as a function of the possible decisions toward an individualized or personalized medicine.

In the last years, the main effort of radiology research has been focused on quantifying imaging variations trying to understand their clinical and biological implications.

Radiomics uses high-throughput radiomic features and mathematical models to quantify tumor characteristics, allowing the non-invasive capture of microscale information hidden within medical imaging features undetectable by the human eye and add value to clinical visual perception by exposing underlying pathophysiology, including intra-tumoral heterogeneity (29, 33–35).

To date the application of radiomics in GM setting has shown considerable progress in demonstrating that it can be a tool capable of deriving much information, with implications in diagnostics, such as differentiating tumors based on texture analysis, differentiating treatment effects (radiation necrosis, pseudo-progression) and tumor recurrence, in prognosis such as survival stratification (1–4, 34–38) and applications in the choice of optimal therapy (39–41), e.g., stratification of response to anti-angiogenic treatment for recurrent glioblastoma.

Most radiomics studies have focused on analyzing features extrapolated from pre-operative MRI by studying the macroscopic site of the tumor, using ROIs such as tumor enhancement (ET), non-enhancement, tumor/necrosis (NET), and edema (ED).

Few studies (42, 43) have suggested that heterogeneity extends beyond the tumor margins into the peritumoral brain region (PBR), suggesting that the interaction of specific cells (i.e., glioma cells,

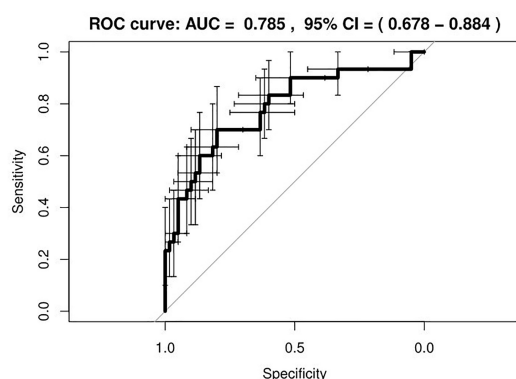


FIGURE 5

Receiver operating characteristic (ROC) curve of the developed radiomic model. The bars indicate the 95% confidence intervals (CI) for sensitivity and specificity.

TABLE 4 Model discrimination metrics for model fitting and internal validation with cross-validation.

	AUC	Sensitivity	Specificity	PPV	NPV
Model fitting	0.78 (0.68–0.88)	0.70 (0.52–0.84)	0.80 (0.68–0.89)	0.64 (0.47–0.78)	0.84 (0.73–0.92)
Cross-validation	0.79 (0.04)	0.70 (0.08)	0.75 (0.12)	0.60 (0.15)	0.84 (0.03)

Model fitting presents the 95% confidence interval (CI) of the metrics in brackets. Cross-validation presents mean and standard deviation values (in brackets). PPV, positive predictive value; NPV, negative predictive value.

vascular endothelial, neuroglial and microglial cells) (44, 45) and molecular events in the PBR contribute to tumor infiltration, blood-brain barrier impairment and micro-vascularization and ultimately affect overall survival in GB.

There has been also an increasing interest in understanding the role of the PBR in molecular pathogenesis, as the residual cells along the resection margin and in the surrounding region can represent resistant and rapidly proliferating clones (43), which can lead to disease recurrence (46).

On the other hand, as we know, the anatomy of the brain can be significantly altered after surgery and the characteristics of the tissue surrounding the surgical cavity can be affected by postoperative changes such as gliosis, ischemia, blood products and can be the site of resistant and rapidly proliferating clones. After all, in radiotherapy, postoperative MRI is the imaging of choice for volume definition: surgical cavity plus the margin because it may be the site of resistant and rapidly proliferating clones (43).

Few studies have focused on the radiomic analysis of features in postoperative MRI. Dasgupta et al. generated probabilistic maps by developing a radiomic signature using imaging data from low-grade glioma (LGG) (tumor marker) and brain metastasis (BM) PTR (edema marker) and applied on 10 cases of GB PTR. They found that a radiomic signature can demarcate areas of microscopic tumors from edema in the PTR of GB, which correlates with areas of future recurrence. The authors finally suggested the potential application of radiomic features in driving radiotherapy target volumes, as standard practice includes a wider margin empirically (46).

Our study aimed to develop a predictive model based on radiomic features analysis extracted from real data to guide the target volume delineation in radiotherapy, focusing on the open question of the margins to be given to the surgical cavity, in order to re-evaluate and to hypothesize a CTV contouring guided and personalized according to radiomic features.

Considering our homogeneous population of 90 GB IDH wild-type, the analysis focused on a healthy tissue ring around the surgical cavity resulting in a radiomic model able to discriminate between patients with low-risk and high-risk of relapse at 6 months with an AUC of 78.5%. We decided to considerate the clinical outcome of PFS at 6 months that could describe the local control after radio-chemotherapy, excluding the overall survival that could depend on other clinical and treatment variables. This predictive model with high NPV of 0.84 could allow us to select a population of patients with low-risk of relapse at 6 months, in whom it may be possible to reduce the total CTV by decreasing the margins to 1.5 cm, planning a dose strategy modulation in the surrounding tissue and potential reducing the toxicity of healthy tissue and critical structures.

The radiomic features included in the developed radiomic model were textural features computed from the gray-level co-occurrence matrix (F_{cm_merged.contrast}, F_{cm_merged.info.corr.2}), which is based on the combinations of the gray-levels of neighboring pixels, and from the gray level run length matrix (F_{rlm_merged.rlnu}),

which is based on the sequence of consecutive pixels with the same gray-level. Furthermore, the radiomic model presented a high NPV of 0.84 when compared to the null model, which was based on the prevalence of the majority class 0 (~67%). This result was confirmed in the internal validation, which was performed to assess the generalizability of the model. The limitations of this study include the lack of independent validation of the proposed radiomic model, the absence of images for all patients due to unsuitable imaging data, small sample size and the lack of correlation with other potential clinical prognostic factors of PFS or with recurrence pattern.

However, this is the first hypothesis-generating study that applies a radiomic analysis based on the irradiated target volume as region of interest (ROI) for GB, focusing on healthy tissue ring around the surgical cavity on post-operative MRI. Future steps will include performing an external validation of the model and verifying the applicability of the model in the clinical practice through clinical trials.

4. Conclusion

This study provides a preliminary model for a decision support tool employing radiomic features for a customization of the radiation target volume in GB IDH wild-type in order to achieve a margin reduction strategy.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Prof. Andrea Bacigalupo, Clinico Ematologo. Presidente Prof. Stefania Boccia, Biostatistico. Vicepresidente Dott. Paolo Angelo Bonini, Esperto di Bioetica Prof. Emilio Bria, Clinico-Oncologia Prof. Alessandro Caruso, Clinico-Ostetricia e Ginecologia Dott. Antonello Cocchieri, Rappresentante dell'area delle professioni sanitarie Dott. Alessio de luca, Farmacista SSR1 Dott. Francesco Filidoro, Farmacista esperto di Dispositivi Medici Avv. Danilo Gallitelli, Esperto in materia giuridica e assicurativa o Medico legale Prof. Fiorella Gurrieri, Esperto di Genetica Dott. Michele Lepore, Medico di Medicina Generale Prof. Giuseppina Loffredi, Rappresentante del Volontariato/Associazione Tutela Pazienti Prof. Camillo Marra, Clinico-Neurologia Dott.ssa Barbara Meini, Farmacista SSR1 Prof. Nadia Mores, Sostituto permanente del Direttore Sanitario Prof. Key Peris, Clinico-Dermatologo Prof. Giacomo Pozzoli, Farmacologo Prof. Riccardo

Riccardi, Pediatra Prof. Dario Sacchini, Esperto di Bioetica. Membri esterni: Avv. Filippo E. Leone, Responsabile Grant Office Prof. Antonio Gioacchino Spagnolo, Esperto di Bioetica. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SC, IP, SG, GS, LB, MG, GD, GD'A, ML, and MC contributed to conception and design of the study. MC, FC, and FP organized the database. MC, FB, SC, RR, and RG performed Manual segmentation. SG and RR reviewed all cases' Manual segmentation. HT, ND, CV, and DC performed the statistical analysis. MC and SC wrote the first draft of the manuscript. HT, MC, SL, and SC wrote sections of the manuscript. CA, AO, MB, CC, and VV participated in supervision. All authors contributed to manuscript revision, read, and approved the submitted version.

References

- Stupp R, Mason W, van den Bent M, Weller M, Fisher B, Taphoorn M, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. (2005) 352:987–96. doi: 10.1056/NEJMoa043330
- Tatli S, Gerbaudo V, Mamede M, Tuncali K, Shyn P, Silverman S. Abdominal masses sampled at PET/CT-guided percutaneous biopsy: initial experience with registration of prior PET/CT images. *Radiology*. (2010) 256:305–11. doi: 10.1148/radiol.10090931
- Seow P, Wong J, Ahmad-Annuar A, Mahajan A, Abdullah N, Ramli N. Quantitative magnetic resonance imaging and radiogenomic biomarkers for glioma characterisation: a systematic review. *Br J Radiol*. (2018) 91:20170930. doi: 10.1259/bjr.20170930
- Alksas A, Shehata M, Atef H, Sherif F, Alghamdi N, Ghazal M, et al. A novel system for precise grading of glioma. *Bioengineering*. (2022) 9:532. doi: 10.3390/bioengineering9100532
- Abdel Razek A, Alksas A, Shehata M, AbdelKhalek A, Abdel Baky K, El-Baz A, et al. Clinical applications of artificial intelligence and radiomics in neuro-oncology imaging. *Insights Imaging*. (2021) 12:152.
- Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout R, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036
- Chaddad A, Desrosiers C, Toews M. GBM heterogeneity characterization by radiomic analysis of phenotype anatomical planes. In: Martin A, Elsa D, editors. *Medical Imaging 2016: Image Processing*. Orlando, FL: International Society for Optics and Photonics (2016).
- Zacharakis E, Wang S, Chawla S, Yoo D, Wolf R, Melhem E, et al. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn Reson Med*. (2009) 62:1609–18. doi: 10.1002/mrm.22147
- Kickingereder P, Götz M, Muschelli J, Wick A, Neuberger U, Shinohara R, et al. Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. *Clin Cancer Res*. (2016) 22:5765–71. doi: 10.1158/1078-0432.CCR-16-0702
- Chaddad A, Daniel P, Desrosiers C, Toews M, Abdulkarim B. Novel radiomic features based on joint intensity matrices for predicting glioblastoma patient survival time. *IEEE J Biomed Health Inform*. (2019) 23:795–804. doi: 10.1109/JBHI.2018.2825027
- Chaddad A, Tanougast C. Extracted magnetic resonance texture features discriminate between phenotypes and are associated with overall survival in glioblastoma multiforme patients. *Med Biol Eng Comput*. (2016) 54:1707–18. doi: 10.1007/s11517-016-1461-5
- Chaddad A, Desrosiers C, Hassan L, Tanougast C. A quantitative study of shape descriptors from glioblastoma multiforme phenotypes for predicting survival outcome. *Br J Radiol*. (2016) 89:20160575. doi: 10.1259/bjr.2016.0575
- Cui Y, Ren S, Tha K, Wu J, Shirato H, Li R. Volume of high-risk intratumoral subregions at multi-parametric MR imaging predicts overall survival and complements molecular analysis of glioblastoma. *Eur Radiol*. (2017) 27:3583–92. doi: 10.1007/s00330-017-4751-x
- Villanueva-Meyer J, Mabray M, Cha S. Current clinical brain tumor imaging. *Clin Neurosurg*. (2017) 81:397–415.
- Minniti G, Amelio D, Amichetti M, Salvati M, Muni R, Bozzao A, et al. Patterns of failure and comparison of different target volume delineations in patients with glioblastoma treated with conformal radiotherapy plus concomitant and adjuvant temozolomide. *Radiother Oncol*. (2010) 97:377–81. doi: 10.1016/j.radonc.2010.08.020
- Brandes A, Tosoni A, Franceschi E, Sotti G, Frezza G, Amistà P, et al. Recurrence pattern after temozolomide concomitant with and adjuvant to radiotherapy in newly diagnosed patients with glioblastoma: correlation with MGMT promoter methylation status. *J Clin Oncol*. (2009) 27:1275–9. doi: 10.1200/JCO.2008.19.4969
- Wallner K, Galicich J, Krol G, Arbit E, Malkin M. Patterns of failure following treatment for glioblastoma multiforme and anaplastic astrocytoma. *Int J Radiat Oncol Biol Phys*. (1989) 16:1405–9.
- Hochberg F, Pruitt A. Assumptions in the radiotherapy of glioblastoma. *Neurology*. (1980) 30:907–11.
- Gaspar L, Fisher B, Macdonald D, Leber DV, Halperin E, Schold S, et al. Supratentorial malignant glioma: patterns of recurrence and implications for external beam local treatment. *Int J Radiat Oncol Biol Phys*. (1992) 24:55–7. doi: 10.1016/0360-3016(92)91021-e
- Chang E, Akyurek S, Avalos T, Rebuena N, Spicer C, Garcia J, et al. Evaluation of peritumoral edema in the delineation of radiotherapy clinical target volumes for glioblastoma. *Int J Radiat Oncol Biol Phys*. (2007) 68:144–50. doi: 10.1016/j.ijrobp.2006.12.009
- Aydin H, Sillenberger I, von Lieven H. Patterns of failure following CT-based 3-D irradiation for malignant glioma. *Strahlenther Onkol*. (2001) 177:424–31. doi: 10.1007/pl00002424
- Oppitz U, Maessen D, Zunterer H, Richter S, Flentje M. 3D-recurrence-patterns of glioblastomas after CT-planned postoperative irradiation. *Radiother Oncol*. (1999) 53:53–7. doi: 10.1016/s0167-8140(99)00117-6
- Buchlak Q, Esmaili N, Leveque J, Bennett C, Farrokhi F, Piccardi M. Machine learning applications to neuroimaging for glioma detection and classification: an artificial intelligence augmented systematic review. *J Clin Neurosci*. (2021) 89:177–98. doi: 10.1016/j.jocn.2021.04.043
- Bidgood W, Horii S, Prior F, van Syckle D. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J Am Med Inform Assoc*. (1997) 4:199–212. doi: 10.1136/jamia.1997.0040199
- Dinapoli N, Alitto A, Vallati M, Gatta R, Autorino R, Boldrini L, et al. Moddicom: a complete and easily accessible library for prognostic evaluations relying on image features. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Milan: EMBS (2015). doi: 10.1109/EMBC.2015.7318476
- Zwanenburg A, Vallières M, Abdalah M, Aerts H, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. (2020) 295:328–38. doi: 10.1148/radiol.2020191145
- Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E. A deep look into radiomics. *Radiol Med*. (2021) 126:1296–311. doi: 10.1007/s11547-021-01389-x
- Zhang Z. Variable selection with stepwise and best subset approaches. *Ann Transl Med*. (2016) 4:136. doi: 10.21037/atm.2016.03.35

Conflict of interest

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

The handling editor declared a shared parent affiliation with the authors SC, RR, FB, GS, MC, RG, SL, HT, LB, ND, CV, FP, GD, GD'A, AO, MB, CC, MG, VV, and SG at the time of review.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

29. Moons K, Altman D, Reitsma J, Ioannidis J, Macaskill P, Steyerberg E, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* (2015) 162:W1–73. doi: 10.7326/M14-0698
30. Brown L, Cai T, das Gupta A. Interval estimation for a binomial proportion. *Stat Sci.* (2001) 16:101–33.
31. Kim J. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal.* (2009) 53:3735–45.
32. Krstajic D, Buturovic L, Leahy D, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform.* (2014) 6:10.
33. Pavlou M, Ambler G, Seaman S, de Iorio M, Omar R. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med.* (2016) 35:1159–77. doi: 10.1002/sim.6782
34. Dinapoli N, Valentini A, Pesce A, Gatta R, Masciocchi C, Ninivaggi V, et al. OC-0317: MR radiomics and fractal dimension in cervical cancer predicting pathological complete response. *Radiother Oncol.* (2017) 123:S164–5.
35. Yip S, Aerts H. Applications and limitations of radiomics. *Phys Med Biol.* (2016) 61:R150–66.
36. Beig N, Patel J, Prasanna P, Hill V, Gupta A, Correa R, et al. Radiogenomic analysis of hypoxia pathway is predictive of overall survival in Glioblastoma. *Sci Rep.* (2018) 8:7. doi: 10.1038/s41598-017-18310-0
37. Rapisarda A, Melillo G. Overcoming disappointing results with antiangiogenic therapy by targeting hypoxia. *Nat Rev Clin Oncol.* (2012) 9:378–90. doi: 10.1038/nrclinonc.2012.64
38. Monteiro A, Hill R, Pilkington G, Madureira P. The role of hypoxia in glioblastoma invasion. *Cells.* (2017) 6:45. doi: 10.3390/cells6040045
39. Mallick S, Benson R, Hakim A, Rath G. Management of glioblastoma after recurrence: a changing paradigm. *J Egypt Natl Cancer Inst.* (2016) 28:199–210.
40. Bahrami N, Piccioni D, Karunamuni R, Chang Y, White N, Delfanti R, et al. Edge contrast of the FLAIR hyperintense region predicts survival in patients with high-grade gliomas following treatment with bevacizumab. *Am J Neuroradiol.* (2018) 39:1017–24.
41. Kickingeder P, Bonekamp D, Nowosielski M, Kratz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology.* (2016) 281:907–18. doi: 10.1148/radiol.2016161382
42. Engelhorn T, Savaskan N, Schwarz M, Kreutzer J, Meyer E, Hahnen E, et al. Cellular characterization of the peritumoral edema zone in malignant brain tumors. *Cancer Sci.* (2009) 100:1856–62. doi: 10.1111/j.1349-7006.2009.01259.x
43. Lemée J, Clavreul A, Menei P. Intratumoral heterogeneity in glioblastoma: don't forget the peritumoral brain zone. *Neuro Oncol.* (2015) 17:1322–32. doi: 10.1093/neuonc/nov119
44. Davies D. Blood-brain barrier breakdown in septic encephalopathy and brain tumours. *J Anat.* (2002) 200:639–46. doi: 10.1046/j.1469-7580.2002.00065.x
45. Badie B, Schartner J, Hagar A, Prabakaran S, Peebles T, Bartley B, et al. Microglia cyclooxygenase-2 activity in experimental gliomas: possible role in cerebral edema formations. *Clin Cancer Res.* (2003) 9:872–7.
46. Tseng C, Stewart J, Whitfield G, Verhoeff J, Bovi J, Soliman H, et al. Glioma consensus contouring recommendations from a MR-Linac international consortium research group and evaluation of a CT-MRI and MRI-only workflow. *J Neurooncol.* (2020) 149:305–14. doi: 10.1007/s11060-020-03605-6



OPEN ACCESS

EDITED BY

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

REVIEWED BY

Daniele Antonio Pizzuto,
Agostino Gemelli University Polyclinic (IRCCS),
Italy
Salvatore Annunziata,
Fondazione Policlinico Universitario Agostino
Gemelli IRCCS, Italy

*CORRESPONDENCE

Maria Picchio
✉ picchio.maria@hsr.it

†These authors share first authorship

SPECIALTY SECTION

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 28 December 2022

ACCEPTED 07 February 2023

PUBLISHED 23 February 2023

CITATION

Ghezzi S, Mongardi S, Bezzi C, Samanes
Gajate AM, Preza E, Gotuzzo I, Baldassi F,
Jonghi-Lavarini L, Neri I, Russo T, Brembilla G,
De Cobelli F, Scifo P, Mapelli P and Picchio M
(2023) External validation of a convolutional
neural network for the automatic
segmentation of intraprostatic tumor lesions
on ^{68}Ga -PSMA PET images.
Front. Med. 10:1133269.
doi: 10.3389/fmed.2023.1133269

COPYRIGHT

© 2023 Ghezzi, Mongardi, Bezzi, Samanes
Gajate, Preza, Gotuzzo, Baldassi,
Jonghi-Lavarini, Neri, Russo, Brembilla,
De Cobelli, Scifo, Mapelli and Picchio. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

External validation of a convolutional neural network for the automatic segmentation of intraprostatic tumor lesions on ^{68}Ga -PSMA PET images

Samuele Ghezzi^{1,2†}, Sofia Mongardi^{1†}, Carolina Bezzi^{1,2},
Ana Maria Samanes Gajate², Erik Preza², Irene Gotuzzo³,
Francesco Baldassi³, Lorenzo Jonghi-Lavarini³, Ilaria Neri^{1,2},
Tommaso Russo^{1,4}, Giorgio Brembilla^{1,4}, Francesco De Cobelli^{1,4},
Paola Scifo², Paola Mapelli^{1,2} and Maria Picchio^{1,2*}

¹Department of Medicine and Surgery, Vita-Salute San Raffaele University, Milan, Italy, ²Department of Nuclear Medicine, IRCCS San Raffaele Scientific Institute, Milan, Italy, ³School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy, ⁴Department of Radiology, IRCCS San Raffaele Scientific Institute, Milan, Italy

Introduction: State of the art artificial intelligence (AI) models have the potential to become a “one-stop shop” to improve diagnosis and prognosis in several oncological settings. The external validation of AI models on independent cohorts is essential to evaluate their generalization ability, hence their potential utility in clinical practice. In this study we tested on a large, separate cohort a recently proposed state-of-the-art convolutional neural network for the automatic segmentation of intraprostatic cancer lesions on PSMA PET images.

Methods: Eighty-five biopsy proven prostate cancer patients who underwent ^{68}Ga PSMA PET for staging purposes were enrolled in this study. Images were acquired with either fully hybrid PET/MRI ($N = 46$) or PET/CT ($N = 39$); all participants showed at least one intraprostatic pathological finding on PET images that was independently segmented by two Nuclear Medicine physicians. The trained model was available at <https://gitlab.com/dejankostyszyn/prostate-gtv-segmentation> and data processing has been done in agreement with the reference work.

Results: When compared to the manual contouring, the AI model yielded a median dice score = 0.74, therefore showing a moderately good performance. Results were robust to the modality used to acquire images (PET/CT or PET/MRI) and to the ground truth labels (no significant difference between the model's performance when compared to reader 1 or reader 2 manual contouring).

Discussion: In conclusion, this AI model could be used to automatically segment intraprostatic cancer lesions for research purposes, as instance to define the volume of interest for radiomics or deep learning analysis. However, more

robust performance is needed for the generation of AI-based decision support technologies to be proposed in clinical practice.

KEYWORDS

PSMA, convolutional neural network, segmentation, prostate cancer, external validation

1. Introduction

Prostate cancer (PCa) is the second most common cancer in men, with 1,414,259 new cases in 2020, accounting for 15.1% of all cancer diagnoses within the male population (1). Although histopathological examination of prostate biopsy cores is required for the diagnosis of PCa, imaging is pivotal to characterize the disease (2). Multiparametric (mp)-MRI has been used for years in clinical practice to guide biopsy and to drive the clinical management of PCa patients (2).

PSMA PET has been recently added to the EAU-ESTRO-SIOG guidelines for staging high-risk PCa (2) in view of its higher sensitivity compared to mp-MRI (3, 4). Therefore, a possible next step will be to use PSMA PET to diagnose clinically significant PCa (5–8) and to perform quantitative analysis that might allow for a better and more objective characterization of the disease (9–11).

Accurate contouring of intraprostatic gross tumor volume (GTV) is mandatory for an accurate assessment of PCa in several clinical settings, including both biopsy guidance and radiomic features extraction. However, this procedure is time consuming and largely affected by the experience of the contouring physicians, often resulting in non-reproducible segmentations (12).

Recently, there has been a surge in the development of artificial intelligence (AI) models in the medical field, with the first tools being already available for use (13, 14). Convolutional neural networks (CNN) have been shown to accurately segment medical images (15–17) and hold the potential to improve intraprostatic tumor delineation (18–21). The use of CNN in this setting could improve GTV definition by reducing the inter-reader variability while saving time by automating this task.

Kostyszyn and colleagues were the first to develop a CNN for the automatic segmentation of intraprostatic cancer lesions on PSMA (using both ^{68}Ga - and ^{18}F -PSMA) PET images (18). They used 152 patients examined at two centers (Germany and China) to train their model and a cohort composed by 57 patients to test it. However, only 20 patients in the testing cohort were studied at an external institution (center 3, Germany) not used for training, making it difficult to draw conclusions regarding the model's generalizability.

External validation of AI models on independent cohorts is necessary to assess with certainty their robustness and reproducibility, hence their possible application in clinical practice (22). Therefore, this study aims to evaluate the performance of the CNN for the automatic segmentation of intraprostatic cancer lesions on ^{68}Ga -PSMA PET images that was previously presented in (18) and that is publicly available at <https://gitlab.com/dejankostyszyn/prostate-gtv-segmentation>.

2. Materials and methods

2.1. Patients

All patients with biopsy proven PCa who underwent ^{68}Ga -PSMA PET at IRCCS San Raffaele Scientific Institute from June 2020 to January 2022 for staging purposes were considered for inclusion. A total of 124 patients was identified. Eligibility criteria were: (1) age greater than 18 years at the time of the PET examination (0 patients excluded), (2) presence of at least one intraprostatic pathological finding at ^{68}Ga -PSMA PET (30 patients excluded), (3) absence of neoadjuvant treatments prior to imaging (9 patients excluded). Eighty-five patients met the inclusion criteria and were included for analysis. See **Figure 1** for a flowchart showing the patients' selection process. Prostate specific antigen (PSA) level and the International Society of Urological Pathology (ISUP) grade were collected. This retrospective study was approved by the Institutional Ethics Committee of IRCCS San Raffaele Scientific Institute, and informed consent was waived due to the retrospective nature of the study.

2.2. PET imaging

PET scans were acquired using either Signa PET/MRI 3 Tesla system, GE Healthcare, Waukesha, WI, USA ($N = 46$) or PET/CT, Discovery-690, GE Healthcare ($N = 39$).

Fasting condition was requested on the day of ^{68}Ga -PSMA PET/MRI and PET/CT scan.

PET scans were acquired from the skull base to mid-thigh (5–6 FOVs, 4 min/FOV), and started approximately 60 min (mean \pm SD, 63 ± 6 min) after injection of 111–273 MBq (Mean \pm SD, 168 ± 33 MBq) of ^{68}Ga -PSMA. PET images, acquired with either PET/MRI or PET/CT scanner, were reconstructed using fully 3D ordered subset expectation-maximization (OSEM) algorithm, time-of-flight (TOF) and point-spread-function (PSF).

^{68}Ga PSMA PET image read-out was performed by two Nuclear Medicine physicians on an Advantage Workstation (AW, General Electric Healthcare, Waukesha, WI, USA) and the presence of ^{68}Ga -PSMA intraprostatic increased uptake was considered positive for malignancy.

2.3. Image segmentation

Two Nuclear Medicine physicians manually contoured the GTV on every slice of ^{68}Ga -PSMA PET images using 3D Slicer (Slicer; version 4.11.2) being aware of all the available patients' clinical and imaging information. The first reader (Exp 1)

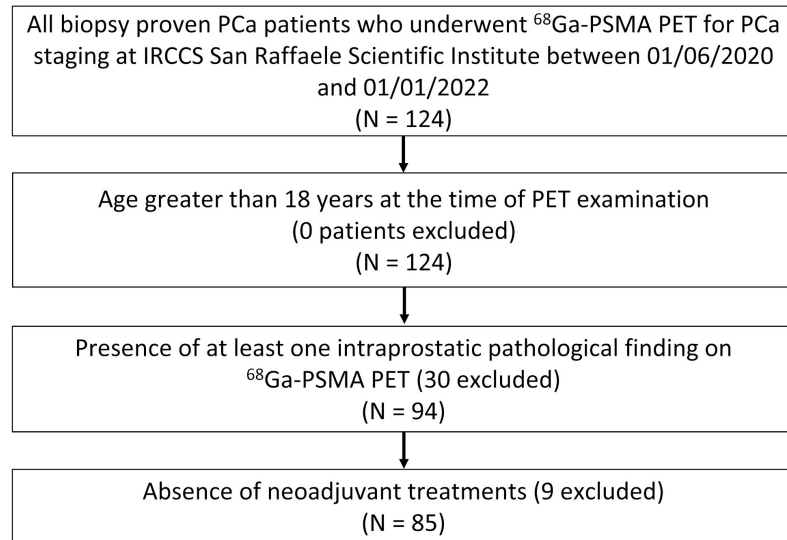


FIGURE 1
Flowchart illustrating the patients' selection process.

delineated the GTV using an inverted gray scale for display, windowed with SUVmin-max: 0–5, as previously described in Kostyszyn et al. (18). To ensure that the segmentation approach used in the reference work was not introducing any bias, a second reader (Exp 2), instead, contoured images independently without any fixed thresholding of voxel values, blind to any instruction on how images were evaluated in the reference work of Kostyszyn et al.

Additionally, two radiologists performed a manual contouring of the prostatic gland on CT and MRI scans by using 3D Slicer (Slicer; version 4.11.2). Since it is not always feasible to discriminate between prostatic tissue and bladder signal in ⁶⁸Ga-PSMA PET images, only contouring within the delineated prostatic gland were used for analyses, as described in Kostyszyn et al.

2.4. Resampling

To ensure that the CNN's performance in this study was not affected by discrepancies in the methods used as compared to the reference work, resampling and preprocessing of the images was performed exactly as described by Kostyszyn et al. (18).

Specifically, all PET images (nearly raw raster data format, nrrd) were resampled to standardize the voxel spacing to 2.0 mm × 2.0 mm × 2.0 mm using SimpleITK (version 1.2.4) since the PET images collected with PET/MRI scanner had original voxel size = 3.125 mm × 3.125 mm × 2.780 mm, while the original voxel size of images acquired with PET/CT scanner was 2.734 mm × 2.734 mm × 3.270 mm. Prostate and GTV segmentations were also resampled to a voxel size of 2.0 mm × 2.0 mm × 2.0 mm. PET volumes were resampled using both tri-linear interpolation and B-spline interpolation, whereas Nearest Neighbour interpolation was used to resample segmentation contours. All data were cropped using the manual contouring of the prostate gland as guidance to a size of 64 × 64 × 64 voxels, and then normalized with $x_i' = \frac{x_i - \bar{x}}{\sigma}$ where x_i is the PET data for patient i , and \bar{x} and σ are the arithmetic mean

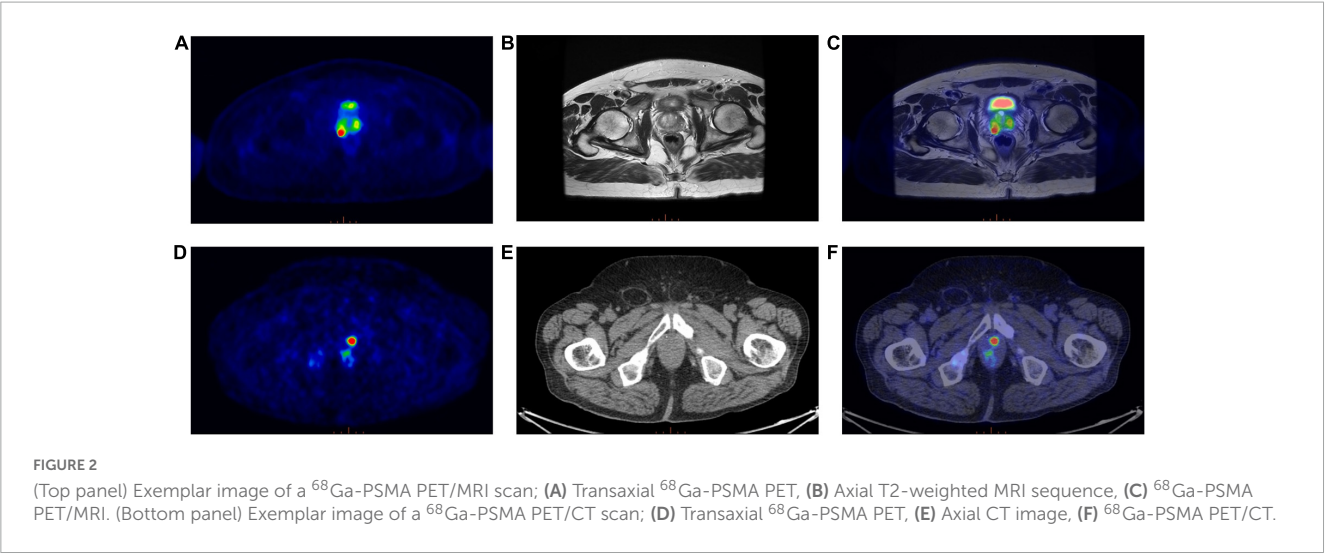
and the standard deviation calculated over the entire cropped PET training dataset.

2.5. Convolutional neural network

The model consists of 3 down sampling steps performed by $2 \times 2 \times 2$ max-pooling along the contracting path, and 3 up-sampling steps performed by $2 \times 2 \times 2$ transpose convolutions with padding of 1 and stride of 2 along the expanding paths. Skip connections from the contracting path are concatenated with their corresponding up-sampled feature maps. There are 14 $3 \times 3 \times 3$ convolutional layers in total, having stride and padding of 1. Each convolution is followed by batch normalization and ReLU activation function. The last layer in the model performs a $1 \times 1 \times 1$

TABLE 1 Patients' characteristics.

Statistics	
No. of patients	85
Median age, years	68 (range: 45–85)
Median PSA, ng/ml	7.82 (range: 1.72–1263)
ISUP grade, no. (%)	
1	3 (3.6%)
2	9 (10.6%)
3	17 (20.0%)
4	20 (23.5%)
5	29 (34.1%)
Unknown	7 (8.2%)
Scanner	
PET/MRI	46
PET/CT	39



convolution with no padding, followed by batch normalization and sigmoid activation function. The whole script of the trained CNN can be freely downloaded at <https://gitlab.com/dejankostyszyn/prostate-gtv-segmentation>.

2.6. Statistical analysis

Statistical analyses were performed with R statistical software (23). Dice score coefficient (DSC) was computed to estimate the performance of the trained CNN (GTV-CNN) presented in Kostyszyn et al. (18). Moreover, DSC was also used to quantitatively assess the agreement between the GTVs manually segmented by the different experts (GTV-Exp 1, GTV-Exp 2). As PET volumes in the dataset have been acquired using two different modalities, PET/MRI and PET/CT, Student’s *t*-test was carried out to determine whether the image modality of acquisition possibly affected the model performance. Student’s *t*-test was also employed to determine whether there was a statistically significant difference in CNN performance across the different GTV-Exp segmentations and to study whether the volume predicted by the CNN was different in size as compared to those manually delineated by experts. Ground truth PCa lesion volumes (GTV-Exp) were correlated with DSC scores using Pearson correlation. Finally, to investigate the impact of different interpolation algorithms, analyses were first conducted on PET images resampled using tri-linear interpolation and then on PET volumes resampled with B-spline interpolation. The obtained DSC were compared by means of Student’s *t*-test. *P* values lower than 0.05 were considered statistically significant.

3. Results

3.1. Patients

Eighty-five patients with biopsy proven PCa were enrolled in this study. The median age was 68 years (range: 45–85 years), whereas the median PSA level was 7.82 ng/ml. Patients’

characteristics are reported in Table 1. Forty-six out of 85 patients were examined on a PET/MRI scanner (see an example; Figure 2, top panel) and 39/85 on a PET/CT scanner (see an example; Figure 2, bottom panel).

3.2. CNN performance

Analyses were performed on PET volumes resampled with tri-linear interpolation and then repeated on images resampled using B-spline interpolation. The results based on tri-linear interpolation are reported here, while Supplementary Table 1 contains the results using B-spline interpolation for voxel resampling. The trained CNN, when validated on the lesion volumes manually contoured by the first reader (GTV-Exp 1), reached a median DSC = 0.74 (range: 0.07–0.93). When the ground truth label was drawn without fixed thresholding of voxel values by the second reader (GTV-Exp 2), the CNN obtained a median DSC = 0.69 (range: 0.07–0.96). However, this difference was not statistically significant (*P* value > 0.05). Using tri-linear or B-spline interpolation did not affect model’s performance (*P* value > 0.05). See Table 2 for a detailed description of CNN model performance, and Figure 3 for a representative image. To better show the performance of the

TABLE 2 External validation of the CNN performance.

	Mean DSC ± SD		Median DSC (range)	
	GTV-Exp 1 vs. GTV-CNN	GTV-Exp 2 vs. GTV-CNN	GTV-Exp 1 vs. GTV-CNN	GTV-Exp 2 vs. GTV-CNN
All	0.70 ± 0.18	0.67 ± 0.20	0.74 (0.07 – 0.93)	0.69 (0.07 – 0.96)
PET/MRI	0.69 ± 0.18	0.64 ± 0.21	0.72 (0.07 – 0.93)	0.68 (0.07 – 0.96)
PET/CT	0.71 ± 0.19	0.70 ± 0.19	0.77 (0.10 – 0.90)	0.75 (0.10 – 0.91)

Mean and median performance of the CNN for the automatic segmentation of intraprostatic cancer lesions considering the contouring made by reader 1 (Exp 1) and reader 2 (Exp 2) as ground truth.

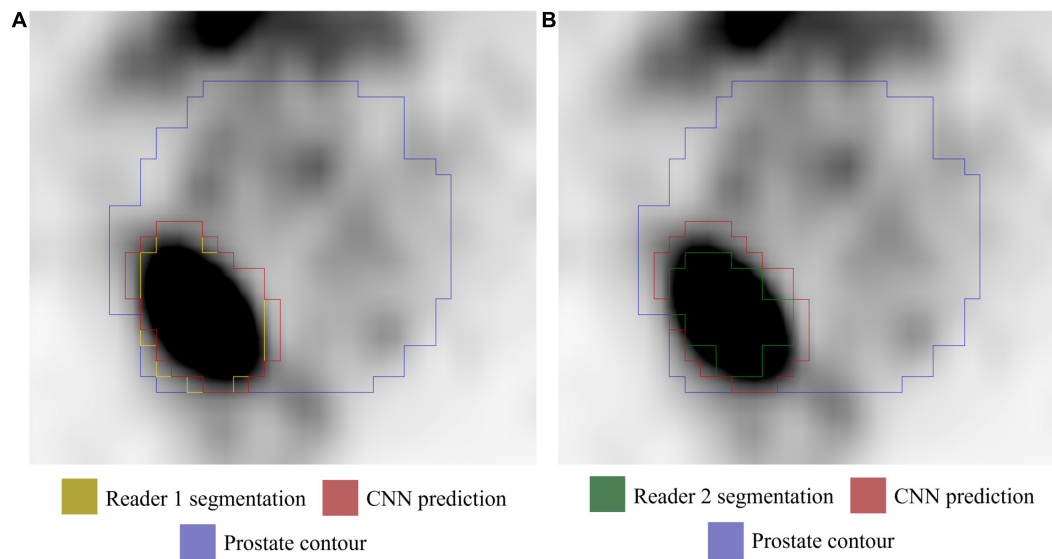


FIGURE 3

Axial ^{68}Ga -PSMA PET image (image windowing SUVmin-max: 0–5). (A) GTV-Exp 1 lesion contour (yellow). (B) GTV-Exp 2 lesion contour (green). The GTV-CNN contour is shown in red and the prostate contour in purple.

CNN, additional segmentation results for sequential ^{68}Ga -PSMA-PET slices are shown in Figure 4. Moreover, no statistically significant differences were identified in the volumes of the intraprostatic tumor lesions defined by the expert Nuclear Medicine physicians and those predicted by the CNN (P value > 0.05 , Table 3).

The DSC obtained by comparing the PCa lesion contouring manually defined by the two expert Nuclear Medicine physicians was 0.73 (range: 0.25–0.92).

No statistically significant differences in CNN performance between PET/MRI and PET/CT images, regardless of the method used to visualize and contour PET images (P value > 0.05 for both GTV-Exp 1 and GTV-Exp 2) were observed. Conversely, a positive correlation was found between DSC and GTV-Exp ($r = 0.43$, P value < 0.001 and $r = 0.44$, P value < 0.001 for GTV-Exp 1 and GTV-Exp 2, respectively), meaning that the CNN produced more accurate segmentations for bigger lesions.

4. Discussion

In the present work, an external validation of a CNN for the automatic segmentation of intraprostatic cancer lesions on ^{68}Ga -PSMA PET images previously presented by Kostyszyn and colleagues (18) has been performed. In our cohort, the trained CNN model reached a median DSC = 0.74 and its performance was independent from the imaging technique, PET/MRI or PET/CT, used to acquire PET images.

^{68}Ga -PSMA PET is widely used for the characterization of PCa in different settings and has been recently included into the EAU-ESTRO-SIOG guidelines for high-risk PCa staging (2). Several studies have been reported showing the potential utility of quantitative features extracted from ^{68}Ga -PSMA PET images

for the characterization of the disease (9–11). Considering the role of PSMA PET, a possible forthcoming application might be its use in the diagnosis of clinically significant PCa, including biopsy guidance in patients with equivocal mp-MRI findings (6, 24).

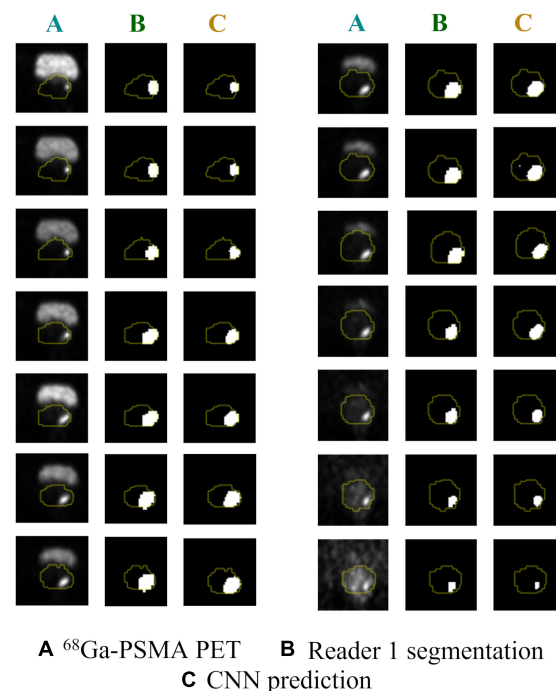


FIGURE 4

Predicted vs. actual lesion contours in sequential ^{68}Ga -PSMA PET slices. (A) Original ^{68}Ga -PSMA PET images; (B) ground truth GTV-Exp 1 contours; (C) CNN predicted contours. Prostate contours are shown in yellow.

TABLE 3 Gross tumor volume.

	GTV-Exp 1	GTV-Exp 2	GTV-CNN
All	12.23 ± 15.7 ml	12.15 ± 16.2 ml	16.55 ± 18.6 ml
PET/MRI	13.75 ± 18.3 ml	12.45 ± 18.3 ml	17.70 ± 20.1 ml
PET/CT	10.45 ± 12.1 ml	11.80 ± 13.5 ml	15.20 ± 17.0 ml

Mean volume, and standard deviation, of the intraprostatic cancer lesion (GTV) defined by Exp 1, Exp 2 and by the CNN.

Accurate contouring of intraprostatic GTV is required as the starting point both for biopsy guidance and for radiomic analysis. However, this procedure is extremely time consuming and affected by inter-reader heterogeneity, often resulting in non-replicable segmentations (12). Several CNNs have already been proposed for GTV segmentation in other oncological settings (19–21), bearing the potential to become a “one-stop shop” for improving the diagnostics and prognostics of various tumors, including PCa (25).

Kostyszyn and colleagues were the first to generate a CNN for the automatic segmentation of intraprostatic cancer lesions on PSMA PET images (18). This study was a joint effort of 3 different Institutions, 2 in Germany and 1 in China. The generated model was trained on 152 patients, employing images acquired with different tomographs in different centers (1 in Germany and 1 in China). However, only 20 patients in the testing cohort were studied at an external institution (center 3, Germany) not used for training, limiting conclusions regarding the model's generalizability.

Validation of AI models in external, independent cohorts is crucial to assess their robustness and, consequently, their potential utility. In our study, we tested the model generated by Kostyszyn and colleagues on a cohort of 85 patients examined with ⁶⁸Ga-PSMA PET at our Institution. Considering that image pre-processing can affect the model performance, as previously described in Kostyszyn et al. (18), all pre-processing steps were performed in agreement with the reference work. However, in the present study, images were independently reviewed by two Nuclear Medicine physicians. The first one (Exp 1) followed the instruction given in Kostyszyn et al. (18), while the second (Exp 2) was not informed on how images were viewed in the reference work, thus avoiding the introduction of any bias relative to the adopted segmentation method.

The trained CNN model achieved a moderately good performance on our cohort, reaching at best a median DSC = 0.74. Interestingly, results were independent of the modality used to acquire the images, despite the model being originally trained only on PET/CT images, as well as of the windowing of voxel values used when defining the ground truth labels. These results suggest that using images acquired with several different PET/CT scanners for training contributed to increasing model robustness. Moreover, it has been shown that the thresholding of voxel values SUV_{min-max}: 0–5 yields relatively stable contouring, as also reported in a previous work of the same group. (12). However, the CNN performance was affected by the volume of the ground truth labels (GTV-Exp 1 and GTV-Exp 2), resulting in more accurate segmentations for bigger lesions.

The main limitation of this study is its monocentric nature, as PET images were acquired in a single Institution. However, as our center was not included in the reference work of Kostyszyn

et al., our population represents a large independent and external testing cohort. Moreover, we included patients examined both with PET/CT (*N* = 39) or PET/MRI (*N* = 46), this could have potentially affected the results, but also allowed the comparison of model performance on images acquired with different modalities. Post-hoc analyses showed that no statistically significant differences in CNN performance was observed on images acquired with either PET/MRI or PET/CT. Nineteen patients studied with ¹⁸F-PSMA were included in the paper presented by Kostyszyn et al. All patients considered in this work underwent ⁶⁸Ga-PSMA PET, therefore, future studies are needed to assess the model's generalizability to ¹⁸F-PSMA PET findings.

In conclusion, the trained and publicly available CNN model presented by Kostyszyn et al. (18) yields fairly accurate contouring of intraprostatic cancer lesions on ⁶⁸Ga-PSMA PET images that could be used as a starting point for quantitative analysis using radiomics or deep learning approaches. Nonetheless, more robust performance is needed for the generation of AI-based decision support technologies that can be used and exploited in daily clinical practice.

Data availability statement

Data supporting the conclusions of this article will be made available by the corresponding author upon reasonable request.

Ethics statement

The studies involving human participants were reviewed and approved by Ethic Committee of IRCCS San Raffaele Scientific Institute. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

SG and MP: conceptualization and study design. SG, SM, and CB: formal analysis. PM, AS, FB, LJ-L, EP, TR, and GB: images acquisition and interpretation. SM, SG, and IN: data curation. SG and SM: writing—original draft preparation. PM, PS, FD, and MP: writing—review and editing. PS, FD, and MP: supervision. All authors read and approved the submitted version of the manuscript.

Funding

This research was funded by the Italian Association for Cancer Research (grant IG 2017 Id.20571) and Italian Ministry of Health (PE-2016-02361273); EUDRACT number: 2018-001034-18. Signa PET/MRI system (GEMS, Wakesha, WI, United States) used in the present work was purchased with funding from the Italian Ministry of Health.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1133269/full#supplementary-material>

References

- World Cancer, Research Fund International [WCRF]. *Cancer statistics*. (2020). Available online at: <https://www.wcrf.org/cancer-trends/prostate-cancer-statistics/> (accessed November 22, 2022).
- Mottet, N, Cornford P, van den Bergh R, Briers E, de Santis M, Gillessen S, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer. *Eur Assoc Urol. Proceedings of the EAU annual congress Amsterdam 2022*. ISBN 978-94-92671-16-5 (accessed October 24, 2022), Amsterdam (2022).
- Donato P, Morton A, Yaxley J, Ranasinghe S, Teloken P, Kyle S, et al. ^{68}Ga -PSMA PET/CT better characterises localised prostate cancer after MRI and transperineal prostate biopsy: is ^{68}Ga -PSMA PET/CT guided biopsy the future? *Eur J Nucl Med Mol Imaging*. (2020) 47:1843–51.
- Rhee H, Thomas P, Shepherd B, Gustafson S, Vela I, Russell P, et al. Prostate specific membrane antigen positron emission tomography may improve the diagnostic accuracy of multiparametric magnetic resonance imaging in localized prostate cancer. *J Urol*. (2016) 196:1261–7.
- Ferraro D, Becker A, Kranzbühler B, Mebert I, Baltensperger A, Zeimpekis K, et al. Diagnostic performance of ^{68}Ga -PSMA-11 PET/MRI-guided biopsy in patients with suspected prostate cancer: a prospective single-center study. *Eur J Nucl Med Mol Imaging*. (2021) 48:3315–24. doi: 10.1007/s00259-021-05261-y
- Kawada T, Yanagisawa T, Rajwa P, Sari Motlagh R, Mostafaei H, Quhal F, et al. Diagnostic performance of prostate-specific membrane antigen positron emission tomography-targeted biopsy for detection of clinically significant prostate cancer: a systematic review and meta-analysis. *Eur Urol Oncol*. (2022) 5:390–400. doi: 10.1016/j.euo.2022.04.006
- Emmett L, Buteau J, Papa N, Moon D, Thompson J, Roberts M, et al. The additive diagnostic value of prostate-specific membrane antigen positron emission tomography computed tomography to multiparametric magnetic resonance imaging triage in the diagnosis of prostate cancer (PRIMARY): a prospective multicentre study. *Eur Urol*. (2021) 80:682–9. doi: 10.1016/j.eururo.2021.08.002
- Liu C, Liu T, Zhang Z, Zhang N, Du P, Yang Y, et al. ^{68}Ga -PSMA PET/CT combined with PET/ultrasound-guided prostate biopsy can diagnose clinically significant prostate cancer in men with previous negative biopsy results. *J Nucl Med*. (2020) 61:1314–9.
- Ghezzi S, Bezzi C, Presotto L, Mapelli P, Bettinardi V, Savi A, et al. State of the art of radiomic analysis in the clinical management of prostate cancer: a systematic review. *Crit Rev Oncol Hematol*. (2022) 169:103544.
- Solari E, Gafita A, Schachoff S, Bogdanovici B, Villagrán Asiares A, Amiel T, et al. The added value of PSMA PET/MR radiomics for prostate cancer staging. *Eur J Nucl Med Mol Imaging*. (2022) 49:527–38. doi: 10.1007/s00259-021-05430-z
- Papp L, Spielvogel C, Grubmüller B, Grahovac M, Krajnc D, Ecsedi B, et al. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with ^{68}Ga -PSMA-11 PET/MRI. *Eur J Nucl Med Mol Imaging*. (2021) 48:1795–805. doi: 10.1007/s00259-020-05140-y
- Zamboglou C, Fassbender T, Steffan L, Schiller F, Fechter T, Carles M, et al. Validation of different PSMA-PET/CT-based contouring techniques for intraprostatic tumor definition using histopathology as standard of reference. *Radiother Oncol*. (2019) 141:208–13. doi: 10.1016/j.radonc.2019.07.002
- SIEMENS. *Auto ID*. (2022). Available online at: <https://www.siemens-healthineers.com/molecular-imaging/news/auto-id-for-pet-ct> (accessed November 24, 2022).
- MIM. *Contour protege AI*. (2022). Available online at: https://www.mimsoftware.com/radiation-oncology/contour-protegeai?utm_source=google_ads&utm_medium=ppc&utm_term=&utm_campaign=MIM+Maestro+Europe&hsa_src=g&hsa_acc=2475176161&hsa_ver=3&hsa_ad=538389917704&hsa_cam=1806236075&hsa_grp=127002588378&hsa_net=adwo (accessed November 24, 2022).
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A editors. *Medical image computing and computer-assisted intervention – MICCAI 2015*. Berlin: Springer (2015). p. 234–41.
- Lin G, Milan A, Shen C, Reid I. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway, NJ: IEEE (2017). p. 5168–77. doi: 10.1109/TPAMI.2019.2893630
- Liu C, Gardner S, Wen N, Elshaikh M, Siddiqui F, Movsas B, et al. Automatic segmentation of the prostate on CT images using deep neural networks (DNN). *Int J Radiat Oncol*. (2019) 104:924–32. doi: 10.1016/j.ijrobp.2019.03.017
- Kostyszyn D, Fechter T, Bartl N, Grosu A, Gratzke C, Sigle A, et al. Intraprostatic tumor segmentation on PSMA PET images in patients with primary prostate cancer with a convolutional neural network. *J Nucl Med*. (2021) 62:823–8. doi: 10.2967/jnumed.120.254623
- Wang J, Lu J, Qin G, Shen L, Sun Y, Ying H, et al. Technical note: a deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys*. (2018) 45:2560–4. doi: 10.1002/mp.12918
- Lin L, Dou Q, Jin Y, Zhou G, Tang Y, Chen W, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. (2019) 291:677–86. doi: 10.1148/radiol.2019182012
- Huang B, Chen Z, Wu P, Ye Y, Feng S, Wong C, et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. *Contrast Media Mol Imaging*. (2018) 2018:8923028. doi: 10.1155/2018/8923028
- Ramspek C, Jager K, Dekker F, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. (2021) 14:49–58. doi: 10.1093/ckj/sfaa188
- R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing (2022).
- Margel D, Bernstine H, Groshar D, Ber Y, Nezir O, Segal N, et al. Diagnostic performance of ^{68}Ga Prostate-specific membrane antigen PET/MRI compared with multiparametric MRI for detecting clinically significant prostate cancer. *Radiology*. (2021) 301:379–86.
- Shimizu H, Nakayama K. Artificial intelligence in oncology. *Cancer Sci*. (2020) 111:1452–60.



OPEN ACCESS

EDITED BY

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

REVIEWED BY

Francesco Dondi,
Università degli Studi di Brescia, Italy
Salvatore Annunziata,
Fondazione Policlinico Universitario A. Gemelli
IRCCS, Italy

*CORRESPONDENCE

Charline Lasnon
✉ c.lasnon@baclesse.unicancer.fr

SPECIALTY SECTION

This article was submitted to
Nuclear Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 04 January 2023

ACCEPTED 30 January 2023

PUBLISHED 13 March 2023

CITATION

Quak E, Weyts K, Jaudet C, Prigent A,
Foucras G and Lasnon C (2023) Artificial
intelligence-based ^{68}Ga -DOTATOC PET
denoising for optimizing $^{68}\text{Ge}/^{68}\text{Ga}$ generator
use throughout its lifetime.
Front. Med. 10:1137514.
doi: 10.3389/fmed.2023.1137514

COPYRIGHT

© 2023 Quak, Weyts, Jaudet, Prigent, Foucras
and Lasnon. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Artificial intelligence-based ^{68}Ga -DOTATOC PET denoising for optimizing $^{68}\text{Ge}/^{68}\text{Ga}$ generator use throughout its lifetime

Elske Quak¹, Kathleen Weyts¹, Cyril Jaudet^{1,2}, Anaïs Prigent^{1,3},
Gauthier Foucras^{1,3} and Charline Lasnon ^{1,4*}

¹Nuclear Medicine Department, Comprehensive Cancer Centre François Baclesse, UNICANCER, Caen, France, ²Radiophysics Department, Comprehensive Cancer Centre François Baclesse, UNICANCER, Caen, France, ³Radiopharmacy Department, Comprehensive Cancer Centre François Baclesse, UNICANCER, Caen, France, ⁴UNICAEN, INSERM 1086 ANTICIPE, Normandy University, Caen, France

Introduction: The yield per elution of a $^{68}\text{Ge}/^{68}\text{Ga}$ generator decreases during its lifespan. This affects the number of patients injected per elution or the injected dose per patient, thereby negatively affecting the cost of examinations and the quality of PET images due to increased image noise. We aimed to investigate whether AI-based PET denoising can offset this decrease in image quality parameters.

Methods: All patients addressed to our PET unit for a ^{68}Ga -DOTATOC PET/CT from April 2020 to February 2021 were enrolled. Forty-four patients underwent their PET scans according to Protocol_FixedDose (150 MBq) and 32 according to Protocol_WeightDose (1.5 MBq/kg). Protocol_WeightDose examinations were processed using the Subtle PET software (Protocol_WeightDose^{AI}). Liver and vascular SUV mean were recorded as well as SUVmax, SUVmean and metabolic tumour volume (MTV) of the most intense tumoural lesion and its background SUVmean. Liver and vascular coefficients of variation (CV), tumour-to-background and tumour-to-liver ratios were calculated.

Results: The mean injected dose of 2.1 (0.4) MBq/kg per patient was significantly higher in the Protocol_FixedDose group as compared to 1.5 (0.1) MBq/kg for the Protocol_WeightDose group. Protocol_WeightDose led to noisier images than Protocol_FixedDose with higher CVs for liver ($15.57\% \pm 4.32$ vs. $13.04\% \pm 3.51$, $p = 0.018$) and blood-pool ($28.67\% \pm 8.65$ vs. $22.25\% \pm 10.37$, $p = 0.0003$). Protocol_WeightDose^{AI} led to less noisy images than Protocol_WeightDose with lower liver CVs ($11.42\% \pm 3.05$ vs. $15.57\% \pm 4.32$, $p < 0.0001$) and vascular CVs ($16.62\% \pm 6.40$ vs. $28.67\% \pm 8.65$, $p < 0.0001$). Tumour-to-background and tumour-to-liver ratios were lower for protocol_WeightDose^{AI}: 6.78 ± 3.49 vs. 7.57 ± 4.73 ($p = 0.01$) and 5.96 ± 5.43 vs. 6.77 ± 6.19 ($p < 0.0001$), respectively. MTVs were higher after denoising whereas tumour SUVmax were lower: the mean% differences in MTV and SUVmax were $+11.14\%$ (95% CI = 4.84–17.43) and -3.92% (95% CI = -6.25 to -1.59).

Conclusion: The degradation of PET image quality due to a reduction in injected dose at the end of the $^{68}\text{Ge}/^{68}\text{Ga}$ generator lifespan can be effectively counterbalanced by using AI-based PET denoising.

KEYWORDS

PET, gallium-68, artificial intelligence, denoising, deep learning

Background

The half-life of the ^{68}Ga isotope is short (68 min) requiring on-site synthesis of ^{68}Ga -labeled tracers. The advent of commercially available $^{68}\text{Ge}/^{68}\text{Ga}$ generators and labeling kits has facilitated the synthesis of ^{68}Ga -labeled PET tracers in the hospital's radiopharmacy and contributed to its increased use. Frequently used ^{68}Ga -labeled PET tracers target somatostatin receptors in neuroendocrine tumours (NETs) (1) and prostate-specific membrane antigen (PSMA) in prostate cancer (2). The clinical benefits of ^{68}Ga -labeled PET tracers for imaging and diagnosis of NETs include improved sensitivity and specificity compared to other imaging modalities, as well as the ability to detect small and functional tumours. It is recommended as the first choice for PET/CT imaging of most NETs by international guidelines (3–6). Since the half-life of the parent ^{68}Ge isotope is 271 days, the generator lifespan is about 1 year. At the start of the lifespan, one generator elution allows the labeling of approximately four doses based on an injected dose of 3 MBq/kg. However, as the ^{68}Ge parent of the generator decays over time, the number of doses of tracer obtained per elution decreases. This means that during the lifespan of the generator, the number of examinations per elution and/or the activity injected in the patient in MBq/kg decreases, thereby negatively affecting the cost of the procedure or the quality of PET images due to increased image noise. Moreover, due to the short half-life of ^{68}Ga , the increase in image noise can hardly be counterbalanced by an increase in PET acquisition time, particularly if several patients injected with the same elution need to be scanned.

To optimize the use of the $^{68}\text{Ge}/^{68}\text{Ga}$ generator while maintaining PET image quality, innovative approaches based on artificial intelligence (AI) are opening up new perspectives. By using AI, the acquisition time per exam and/or the injected activity can be reduced without compromising image quality. Notably, several AI-based post-reconstruction PET/CT image enhancements have been recently developed (7). A post-reconstruction PET denoising software (SubtlePETTM, Subtle Medical®, Stanford, USA provided by Incepto®, France) that was recently developed by using a deep convolutional neural network on a library of millions of paired images (native and low-dose images) to learn and tune the optimal parameters to compute an estimate of the native image. Currently, only a few clinical publications have evaluated its use in oncology, all of them dealing with ^{18}F -FDG PET images (8–12).

At present, SubtlePETTM is FDA (Food and Drug Administration)-approved for use with ^{18}F -FDG and ^{18}F -Amyloid tracers and is now CE (European Conformity)-marked for use with ^{18}F -FDG, ^{18}F -Amyloid, ^{18}F -Fluciclovine, ^{18}F -DOPA, ^{18}F -Choline, ^{18}F -DCFPyL, Ga-68 Dotatate, and Ga-68 PSMA PET images (13). However, no clinical study has demonstrated the value of this software to enhance the quality of low-dose ^{68}Ga PET images, even though nuclear medicine departments are concerned about this issue. Various other deep learning-based methods have been evaluated for low-dose imaging and resolution enhancement, but none of them are currently validated for clinical use (14). Denoising techniques for ^{68}Ga -labeled radiotracers in PET imaging have been explored using both reconstruction-based methods and deep-learning techniques. It has been shown that both strategies can significantly improve the image quality by decreasing the noise level in low-dose ^{68}Ga PET scans (15).

Therefore, the aim of this prospective study was to explore the performance of this software to enhance the quality of ^{68}Ga -DOTATOC PET images, and to compare it to a standard Gaussian post-filtering approach. We hypothesized that to optimize the use of a $^{68}\text{Ge}/^{68}\text{Ga}$ generator throughout its lifetime, AI-based PET denoising might be a solution to maintain correct image quality.

Materials and methods

Population

All patients were informed about the use of their clinical and PET data for research purposes. Patients had the right to refuse the transmission of data covered by medical confidentiality used and processed in the context of this research. The procedure was declared to the National Institute for Health Data with the registration no. F20210720123322. Patients over 18 years old addressed to our PET unit for a ^{68}Ga -DOTATOC PET from April 2020 to February 2021 were enrolled. Sex, age and body mass index (BMI) were extracted from electronic patient records.

Positron emission tomography acquisition and reconstruction

All patients underwent their examinations on a VEREOS PET/CT system (Phillips). All PET emission acquisitions were performed 60 min after injection, from the skull to mid-thighs with 1 min 30 per bed position. Images were reconstructed with four iterations four subsets with point spread function (PSF) and 2-mm voxel size. All images were acquired and reconstructed according to the European guidelines (16). In the event of treatment with

Abbreviations: PET, positron emission tomography; AI, artificial intelligence; SUV, standardized uptake value; MTV, metabolic tumour volume; NETs, NeuroEndocrine tumours; VOI, volume of interest; GPF, Gaussian post filter; CV, coefficient of variation; BMI, body mass index; PSMA, prostate-specific membrane antigen; FDA, food and drug administration; CE, European conformity; SD, standard deviation; FDG, fluorodeoxyglucose.

somatostatin analogs, the treatment was stopped at least 21 days before the PET scan.

Between April and November 2020, corresponding to the first months of the generator's lifespan, patients were injected intravenously with a fixed dose of 150 MBq of ^{68}Ga -DOTATOC. This protocol is subsequently referred to as *protocol_FixedDose*.

Between December 2020 and February 2021, i.e., the last months of the generator's lifespan, patients were injected intravenously with 1.5 MBq/kg of ^{68}Ga -DOTATOC. This protocol is subsequently referred to as *protocol_WeightDose*. These PET examinations were then processed using Subtle PETTM software and was subsequently referred to as *protocol_WeightDose^{AI}*.

In addition, NEMA-NU2 image quality phantom acquisitions were performed and analyzed to find a specific Gaussian post-filter (GPF). This GPF will allow the *protocol_WeightDose* to recover a noise in the image equivalent to the former *protocol_FixedDose* (17). Measurements were made with a sphere-to-background ratio set at six and two background ^{68}Ga solution concentrations: 2.1 MBq/mL and 1.5 MBq/mL, corresponding to the average injected activities for *protocol_FixedDose* and *protocol_WeightDose*, respectively. CVs were measured in a VOI larger than 100 ml for both acquisitions. The width of the fitted GPF was optimized by dichotomy. This GPF was then applied to all *protocol_WeightDose* acquisitions and the resulting images referred to as *protocol_WeightDose^{Gaussian}*.

Clinical PET data extraction

Positron emission tomography scans were equally and randomly assigned to two senior nuclear physicians. PET

images were reviewed on MIM (MIM Software, Cleveland, OH, USA, version 5.6.5).

The following features were recorded separately for each PET acquisition:

- Liver SUV_{mean} (mean standard uptake value) and standard deviation (SD) from a 3 cm diameter spherical volume of interest (VOI) placed on the right liver lobe.
- Vascular SUV_{mean} and SD from a 2 cm diameter spherical-VOI placed on the descending aorta.
- Muscular SUV_{mean} and SD from a 2 cm spherical-VOI placed on the left erector spinae muscle at the height of the adrenals.
- Tumour SUV_{max} , SUV_{mean} and metabolic tumour volume (MTV) from a 40% isocontour VOI placed on the most intense lesion, as well as its location.
- The tumour background SUV_{mean} from a doughnut-shaped VOI surrounding the most intense lesion VOI.

Physiological noises were evaluated by means of coefficients of variations (CV) calculated as follows: $\frac{\text{SD}}{\text{SUV}_{\text{mean}}} \times 100$ (%). Lesion-to-background ratios were computed as follows: $\frac{\text{tumour } \text{SUV}_{\text{mean}}}{\text{background } \text{SUV}_{\text{mean}}}$.

Statistical analysis

Data was presented as mean (SD) unless otherwise specified.

Unmatched data were compared using Mann–Whitney and Kruskal–Wallis tests for quantitative data as appropriate. Wilcoxon and Friedman tests, and Bland–Altman analyses were used to compare paired quantitative data as appropriate.

Statistical analysis and figure design were performed using XLSTAT software (XLSTAT 2019: Data Analysis and Statistical

TABLE 1 Patients and PET examination characteristics.

Variables	<i>Protocol_FixedDose</i> (n = 44)	<i>Protocol_WeightDose</i> (n = 32)	P-value*
Patient characteristics			
Sex, n (%)			
• Female	18 (40.9)	18 (54.5)	0.246
• Male	26 (59.1)	14 (44.5)	
Age (yrs.), mean (SD)	65 (10)	63 (12)	0.521
BMI (kg/m ²), mean (SD)	25.7 (4.6)	25.4 (7.4)	0.858
PET indications, n (%)			
• Staging	7 (15.9)	7 (21.9)	0.545
• Disease monitoring	21 (47.7)	19 (59.4)	
• Suspected recurrence	3 (6.8)	2 (6.2)	
• Before PRRT	3 (6.8)	1 (3.1)	
• Metabolic lesion characterization	10 (22.7)	3 (9.4)	
PET examination characteristics			
Injected dose per patient (MBq), mean (SD)	151.6 (13.0)	111.8 (27.3)	<0.0001
Injected dose per patient (MBq/kg), mean (SD)	2.1 (0.4)	1.5 (0.1)	<0.0001
Uptake delay (min), mean (SD)	59 (5)	58 (3)	0.288

*Non-parametric Mann–Whitney tests p-values, except for PET indications and sex for which Fisher exact tests were performed. BMI, body mass index; PRRT, peptide receptor radionuclide therapy.

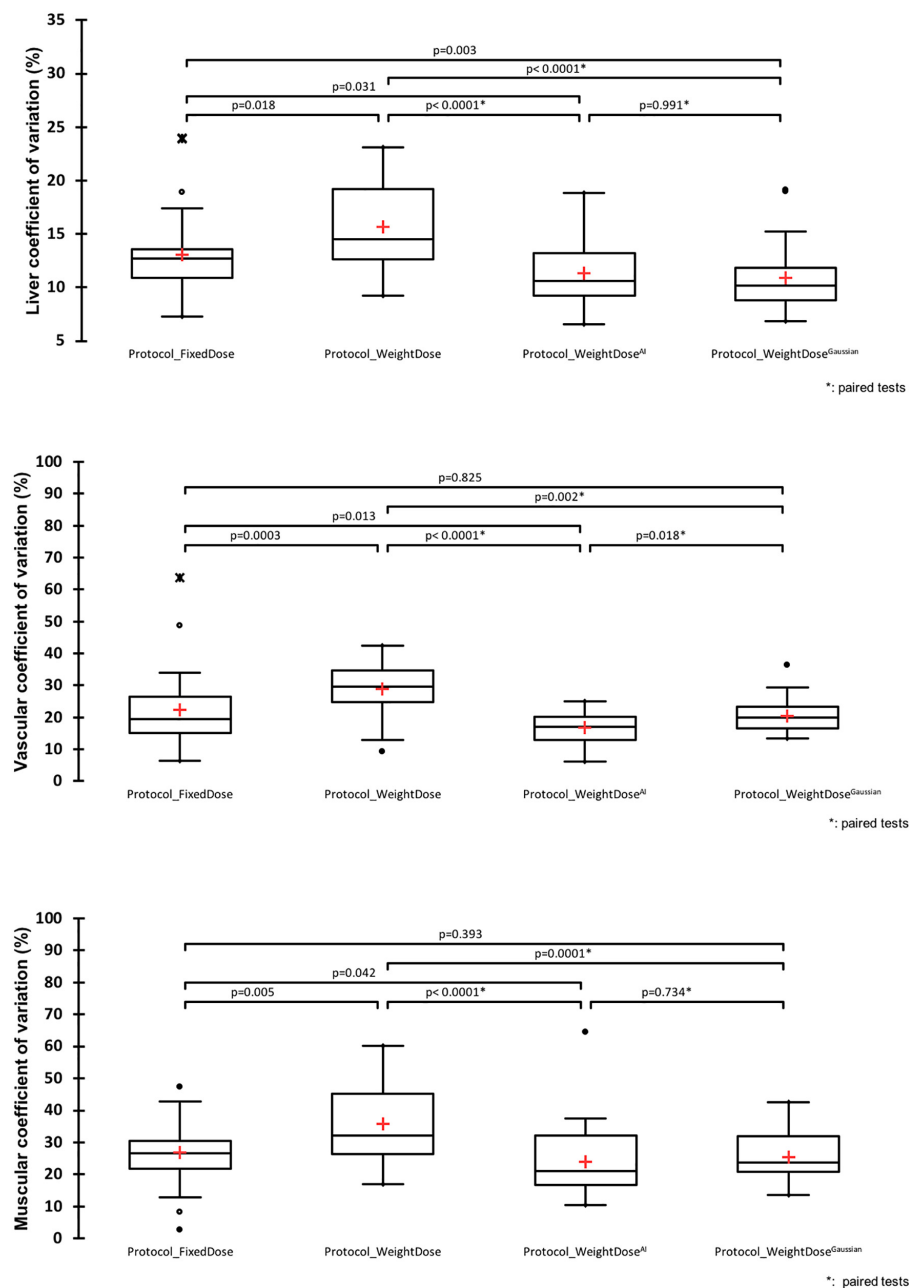


FIGURE 1

Image noise analysis. *Paired Wilcoxon tests were used to compare *protocol_WeightDose* and *protocol_WeightDose^{AI}* data; Otherwise Mann–Whitney tests were used.

Solution for Microsoft Excel, Addinsoft). *P*-values < 0.05 were considered statistically significant.

Results

Population characteristics

Sixty-seven patients were included. Forty-four patients underwent their PET scans according to *protocol_FixedDose* and 32 according to *protocol_WeightDose*. Of note, nine patients

underwent both protocols to monitor their disease over the inclusion period. Patients' characteristics can be found in [Table 1](#). Age, sex, BMI, PET indications and uptake delay were not different between *protocol_FixedDose* and *protocol_WeightDose* groups. The mean injected dose of 2.1 (0.4) MBq/kg per patient was significantly higher in the *protocol_FixedDose* group as compared to 1.5 (0.1) MBq/kg for the *protocol_WeightDose* group. Using the *protocol_FixedDose*, 93% of patients were injected with more than 1.5 MBq/kg, with an injected dose ranging from 1.4 MBq/kg in a severely obese patient (BMI = 41.2 kg/m²) to 3.0 MBq/kg injected in a normal weight patient (BMI = 19.1 kg/m²) ([Supplementary Figure 1](#)).

Comparison of *protocol_FixedDose* and *protocol_WeightDose*

Two patients in the *protocol_FixedDose* group had diffuse liver metastatic involvement that did not allow their hepatic CV to be calculated. Overall, *protocol_WeightDose* led to noisier images with higher liver, vascular and muscular CVs (Figure 1). The mean liver CVs were equal to $15.57\% \pm 4.32$ vs. $13.04\% \pm 3.51$ for *protocol_WeightDose* and *protocol_FixedDose*, respectively ($p = 0.018$). Mean vascular CVs were $28.67\% \pm 8.65$ vs. $22.25\% \pm 10.37$ for *protocol_WeightDose* and *Protocol_FixedDose*, respectively ($p = 0.0003$). Mean muscular CVs were $35.87\% \pm 12.46$ vs. $26.86\% \pm 8.63$ for *protocol_WeightDose* and *Protocol_FixedDose*, respectively ($p = 0.0005$).

Gaussian filter width determination for the *protocol_WeightDose*^{Gaussian}

The GPF width to be applied to the *protocol_WeightDose* acquisitions was determined from the NEMA-NU2 phantom acquisitions to ensure equivalent noise as compared to the *protocol_FixedDose*. A 2.6 mm GPF width was highlighted by dichotomization, applied and used thereafter. NEMA-NU2 CVs were equal to 23.15, 27.63 and 23.30% for *protocol_FixedDose*, *protocol_WeightDose*, and *Protocol_WeightDose*^{Gaussian}, respectively.

Performances of *protocol_WeightDose*^{AI} and *protocol_WeightDose*^{Gaussian}

Image quality: Noise and contrast

On paired comparison, *protocol_WeightDose*^{AI} led to less noisy images than *protocol_WeightDose* with lower liver, vascular and muscular CVs (Figure 1). Mean liver, vascular and muscular CVs were $11.42\% \pm 3.05$ vs. $15.57\% \pm 4.32$ ($p < 0.0001$), $16.62\% \pm 6.40$ vs. $28.67\% \pm 8.65$ ($p < 0.0001$) and $23.88\% \pm 10.58$ vs. $35.87\% \pm 12.46$ ($p < 0.0001$), respectively. Moreover, mean liver, vascular and muscular CVs using *protocol_WeightDose*^{AI} were slightly lower from those of *protocol_FixedDose* (Figure 1).

On paired comparison, *protocol_WeightDose*^{Gaussian} also led to less noisy images than *protocol_WeightDose* with lower liver, vascular and muscular CVs (Figure 1). *Protocol_WeightDose*^{Gaussian} mean liver, vascular and muscular CVs were $10.92\% \pm 3.00$ ($p < 0.0001$), $20.50\% \pm 5.12$ ($p = 0.002$) and $25.49\% \pm 7.14$ ($p = 0.0001$), respectively. The mean liver CV obtained with the *protocol_WeightDose*^{Gaussian} protocol was also lower than with the *protocol_FixedDose*. However, mean vascular and muscular CVs were not different (Figure 1). There were no significant differences between mean liver and muscular CVs of the *protocol_WeightDose*^{AI} and the *protocol_WeightDose*^{Gaussian}. In contrast, the mean vascular CV of the *protocol_WeightDose*^{Gaussian} was higher than that of the *protocol_WeightDose*^{AI}, $p = 0.018$ (Figure 1).

On paired comparison, tumour-to-background ratios and tumour-to-liver ratios were lower when using

protocol_WeightDose^{AI} with a mean tumour-to-background ratio of 6.78 ± 3.49 vs. 7.57 ± 4.73 for the *protocol_WeightDose* ($p = 0.04$) and a mean tumour-to-liver ratio of 5.96 ± 5.43 vs. 6.77 ± 6.19 ($p = 0.0001$). Using the *protocol_WeightDose*^{Gaussian} both these ratios were also lower than those obtained with the *protocol_WeightDose*, and even lower than those obtained with the *protocol_WeightDose*^{AI}. The mean tumour-to-background ratio was equal to 5.60 ± 2.95 ($p < 0.0001$ as compared to *protocol_WeightDose* and $p = 0.013$ as compared to *protocol_WeightDose*^{AI}) and the mean tumour-to-liver ratio was equal to 5.22 ± 4.93 ($p < 0.0001$ as compared to *protocol_WeightDose* and $p = 0.02$ as compared to *protocol_WeightDose*^{AI}).

Lesions quantitative values

Metabolic tumour volumes, SUV_{max} and SUV_{mean} of the hottest lesion were different between *protocol_WeightDose* and *protocol_WeightDose*^{AI} on paired comparison. Similar findings were observed between *protocol_WeightDose* and *protocol_WeightDose*^{Gaussian} (Figure 2).

Metabolic tumour volumes were significantly higher when using *protocol_WeightDose*^{AI} with a mean MTV of 9.11 ± 20.26 vs. 8.46 ± 18.87 for the *protocol_WeightDose* ($p = 0.044$). *Protocol_WeightDose*^{Gaussian} led to even higher MTV values (10.41 ± 21.44) with a p -value < 0.0001 as compared to *protocol_WeightDose* and equal to 0.001 as compared to *protocol_WeightDose*^{AI}.

SUV_{max} and SUV_{mean} were lower for the *protocol_WeightDose*^{AI} with a mean SUV_{max} of 66.65 ± 71.97 vs. 69.76 ± 77.29 for the *protocol_WeightDose* ($p = 0.09$) and a mean SUV_{mean} equal to 39.67 ± 42.95 vs. 41.72 ± 46.42 for the *protocol_WeightDose* ($p = 0.044$) (Figure 2). *Protocol_WeightDose*^{Gaussian} led to even lower SUV values than *protocol_WeightDose*^{AI}: 54.06 ± 59.11 for SUV_{max} ($p = 0.002$) and 32.32 ± 35.76 for SUV_{mean} ($p = 0.001$).

The mean % differences in MTV, SUV_{max} and SUV_{mean} before and after denoising by application of the *protocol_WeightDose*^{AI} were low, equal to $+11.14\%$ (95% CI = 4.84 – 17.43), -3.92% (95% CI = -6.25 to -1.59) and -4.32% (95% CI = -6.98 to -1.66), respectively (Figure 2). These mean % differences were higher by using the *Protocol_WeightDose*^{Gaussian}: $+42.69\%$ (95% CI = 25.23 – 60.15) for MTV, -24.66% (95% CI = -33.02 to -16.29) for SUV_{max} and -25.08 (95% CI = -30.00 to -20.15%) for SUV_{mean} .

Side-by-side representative images of a patient who underwent all four protocols during the inclusion period are displayed in Figure 3. Complete data for the nine patients who had all protocols are reported in Supplementary Table 1.

Discussion

This study shows that the degradation of PET image quality due to a reduction in injected dose at the end of the $^{68}\text{Ge}/^{68}\text{Ga}$ generator lifetime can be counterbalanced effectively by using AI-based PET denoising.

The EANM guidelines recommend an administered activity ranging from 100 to 200 Mbq, meaning that both fixed dose and ponderal dose strategies can be considered (16). To date, these two

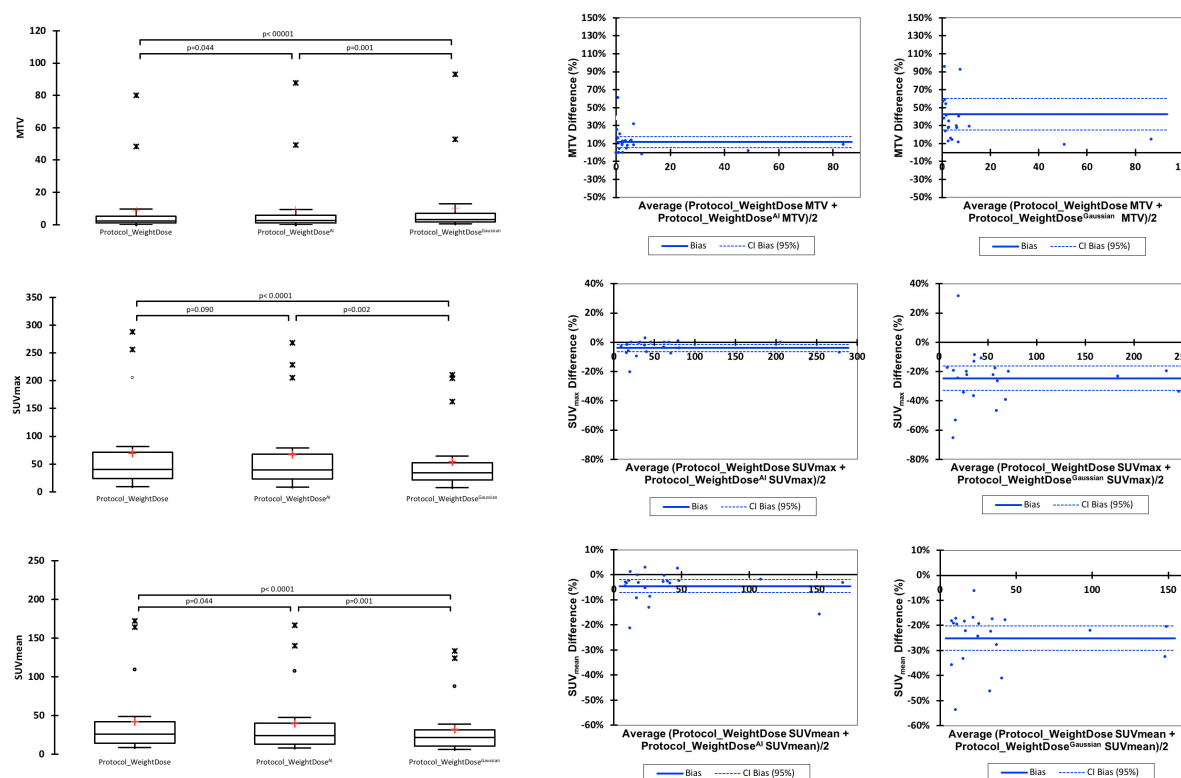


FIGURE 2

Paired comparison of *protocol_WeightDose*, *protocol_WeightDose^{AI}*, and *protocol_WeightDose^{Gaussian}* quantitative values.

strategies have not been compared and the use of either one is at the discretion of each nuclear medicine department.

In our center, at the start of the generator lifetime using the *protocol_FixedDose*, almost all patients were injected with more than 1.5 MBq/kg of ⁶⁸Ga-DOTATOC. This explains the better image quality parameters observed with *Protocol_FixedDose* than with *Protocol_WeightDose*. The use of *Protocol_WeightDose^{AI}* or *Protocol_WeightDose^{Gaussian}* led to an increase in image quality comparable to that of our former *protocol_FixedDose* with regard to image noise. To achieve comparable noise image quality performances at the end of the generator lifetime as per *Protocol_FixedDose* taken as reference in the present study, there are four possible solutions: (i) increasing the injected dose to 2.0 MBq/kg, which corresponds to the mean injected dose when using *Protocol_FixedDose*; (ii) increasing the PET acquisition time to compensate for the lower injected dose; (iii) adapting the reconstruction parameters, i.e., applying a Gaussian Filter; or (iv) exploring external solutions such as AI-based post-reconstruction PET denoising software.

Increasing the injected dose does not seem feasible as the eluted dose will inevitably decrease over time. Furthermore, it is always preferable for the patient's sake to decrease rather than increase the injected dose (8, 9). Increasing the acquisition time seems illusory in busy PET units, especially considering the short and therefore restrictive half-life of ⁶⁸Ga. The use of a Gaussian filter during reconstruction can certainly solve the problem of image noise but is detrimental to the quantitative values of the lesions. In the present study, the tumour volumes are overestimated on

average by more than 40% and the SUVs underestimated by more than 20%, which does not seem tolerable in clinical settings. This is consistent with previous results obtained with FDG-PET (12). Thus, applying PET denoising software to a *Protocol_WeightDose* ⁶⁸Ga-DOTATOC PET/CT images acquired rapidly and at “low-dose.” From an economic point of view, the costs of using an AI-based PET denoising solution should offset the costs related to the decreasing yield of the generator. As more and more ⁶⁸Ga-labeled tracers will probably be commercialized in the future, the value of AI will increase.

Previous work from our group on AI-based PET denoising in a large series of FDG PET scans showed the reassuringly high concordance rate in lesion detection between conventional and AI-processed PET images in the same patient (11). Therefore, the primary aim of PET imaging, which is lesion detection with high sensitivity, does not seem to be jeopardized by AI. Although FDG- and ⁶⁸Ga-labeled tracers target different diseases and show differences in biodistribution, we feel it is safe to extrapolate the detection rate obtained in AI-processed FDG PET scans to AI-processed ⁶⁸Ga PET scans, as the tumour contrast in the latter is often much higher than in the former. Also, the article by Liu et al. focusing on a cross-tracer and cross-protocol deep transfer learning method for noise reduction indicated that the network trained with FDG datasets can effectively reduce noise in low-dose PET images from less commonly used tracers (i.e., ⁶⁸Ga-DOTATATE) while preserving diagnostic information (18).

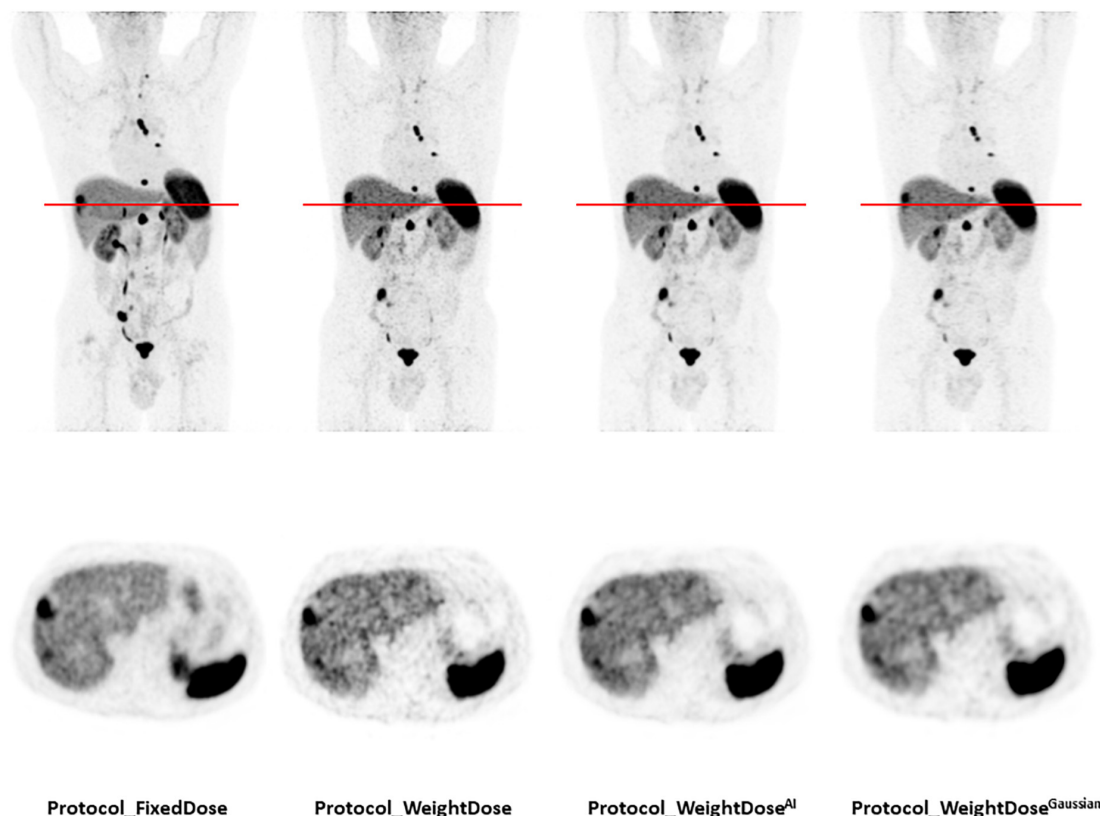


FIGURE 3

Representative images of a patient who underwent all four protocols during the inclusion period. A total of 77-year-old man of normal weight (BMI = 21.3 kg/m²) with a well-differentiated metastatic pancreatic neuroendocrine tumour (grade 1, Ki 67 < 1%). Injected doses were 158 MBq (2.4 MBq/kg) for *protocol_FixedDose* and 86 MBq (1.3 MBq/kg) for *protocol_WeightDose*, *protocol_WeightDose^{AI}* and *protocol_WeightDose^{Gaussian}*. All images are scaled to the same SUV_{max}.

We used two methods to evaluate tumour contrast: The tumour-to-background ratio using a doughnut-shaped VOI and the tumour-to-liver ratio. For the doughnut-shaped VOI, the choice of the tumour-contouring method was crucial to ensure the reliability of the resulting background noise measurements. We chose to use a thresholding value set in reference to SUV_{max}, which was previously demonstrated in the study by Reddy et al. (19) to be the most accurate measurement when compared to morphological volumes. Beyond tumour detectability, one must also take into account the risk of false positive results which increases with the noise in the image. In particular, an increase in liver background noise can easily lead to the overestimation of hepatic metastatic involvement by taking noise for small lesions, especially in patients followed for neuroendocrine tumours with high hepatic metastatic risk. Figure 3 illustrates this issue nicely.

We acknowledge our study has limitations. First, the use of semi-quantitative parameters for ⁶⁸Ga-peptide imaging has some limitations, although it is the most commonly used method in practice (20, 21). One of the main limitations is that it is subject to variations in PET device sensitivity, image acquisition parameters and patient-specific factors that can lead to inaccuracies in quantification (22). Another limitation is that it relies on the assumption that the tracer uptake is proportional to the density of the target receptor, which may not always be the case (23, 24).

Secondly, this is a single-center study on ⁶⁸Ga-DOTATOC PET images only. Although the cohort was small, it covers the lifetime of one generator, i.e., a period of approximately 1 year, during which all patients were included. The robustness of our findings need to be investigated in a multicenter study on different PET systems. Thirdly, only the *protocol_WeightDose* PET scans were AI-processed, leading to a limited number of pairwise comparisons. However, at the start of the generator lifetime, we did not feel the need to use AI processing in view of the good image quality of the *protocol_FixedDose* PET scans. The need to improve image quality became evident at the end of generator life. Finally, we could not properly evaluate the SUV_{peak} data because the small target lesions occurring in 57.7% of *protocol_FixedDose* patients (15/26) and 81.8% of *protocol_WeightDose* and *protocol_WeightDose^{AI}* patients (18/22) (25) were not sufficiently measurable. This was because most target lesions were small with a mean MTV around only 9cc for *protocol_WeightDose* and *protocol_WeightDose^{AI}*.

Conclusion

The degradation of PET image quality due to a reduction of injected dose at the end of the ⁶⁸Ge/⁶⁸Ga generator lifespan can be counterbalanced effectively by using an AI-based PET denoising solution.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of Centre François Baclesse. The patients/participants provided their written informed consent to participate in this study.

Author contributions

EQ and CL performed the image reading and wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript, read, and approved the final manuscript, contributed to the study's conception and design, material preparation, data collection, and analysis were performed.

Acknowledgments

Ray Cooke is thanked for copyediting the manuscript. We benefitted from a 1-month free trial period of SubtlePET™.

References

- Singh S, Poon R, Wong R, Metser U. 68Ga PET imaging in patients with neuroendocrine tumors: a systematic review and meta-analysis. *Clin Nucl Med*. (2018) 43:802–10. doi: 10.1097/RLU.0000000000002276
- Hofman M, Irvani A. Gallium-68 prostate-specific membrane antigen PET imaging. *PET Clin*. (2017) 12:219–34. doi: 10.1016/j.cpet.2016.12.004
- Hicks R, Kwekkeboom D, Krenning E, Bodei L, Grozinsky-Glasberg S, Arnold R, et al. ENETS consensus guidelines for the standards of care in neuroendocrine neoplasia: peptide receptor radionuclide therapy with radiolabeled somatostatin analogues. *Neuroendocrinology*. (2017) 105:295–309. doi: 10.1159/000475526
- Taieb D, Hicks R, Hindié E, Guillet B, Avram A, Ghedini P, et al. European association of nuclear medicine practice guideline/society of nuclear medicine and molecular imaging procedure standard 2019 for radionuclide imaging of pheochromocytoma and paraganglioma. *Eur J Nucl Med Mol Imaging*. (2019) 46:2112–37. doi: 10.1007/s00259-019-04398-1
- Howe J, Cardona K, Fraker D, Kebebew E, Untch B, Wang Y, et al. The surgical management of small bowel neuroendocrine tumors: consensus guidelines of the North American neuroendocrine tumor society. *Pancreas*. (2017) 46:715–31. doi: 10.1097/MPA.0000000000000846
- Ambrosini V, Kunikowska J, Baudin E, Bodei L, Bouvier C, Capdevila J, et al. Consensus on molecular imaging and theranostics in neuroendocrine neoplasms. *Eur J Cancer*. (2021) 146:56–73. doi: 10.1016/j.ejca.2021.01.008
- Liu J, Malekzadeh M, Mirian N, Song T, Liu C, Dutta J. Artificial intelligence-based image enhancement in PET imaging: noise reduction and resolution enhancement. *PET Clin*. (2021) 16:553–76. doi: 10.1016/j.cpet.2021.06.005
- Chaudhari A, Mittra E, Davidzon G, Gulaka P, Gandhi H, Brown A, et al. Low-count whole-body PET with deep learning in a multicenter and externally validated study. *NPJ Digit Med*. (2021) 4:127. doi: 10.1038/s41746-021-00497-2
- Katsari K, Penna D, Arena V, Polverari G, Ianniello A, Italiano D, et al. Artificial intelligence for reduced dose 18F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment. *EJNMMI Phys*. (2021) 8:25. doi: 10.1186/s40658-021-00374-7
- Bonardel G, Dupont A, Decazes P, Queneau M, Modzelewski R, Coulot J, et al. Clinical and phantom validation of a deep learning based denoising algorithm for F-18-FDG PET images from lower detection counting in comparison with the standard acquisition. *EJNMMI Phys*. (2022) 9:36. doi: 10.1186/s40658-022-00465-z
- Weyts K, Lasnon C, Ciappuccini R, Lequesne J, Corroyer-Dulmont A, Quak E, et al. Artificial intelligence-based PET denoising could allow a two-fold reduction in [(18)F]FDG PET acquisition time in digital PET/CT. *Eur J Nucl Med Mol Imaging*. (2022) 49:3750–60. doi: 10.1007/s00259-022-05800-1
- Jaudet C, Weyts K, Lechervy A, Batalla A, Bardet S, Corroyer-Dulmont A. The impact of artificial intelligence CNN based denoising on FDG PET radiomics. *Front Oncol*. (2021) 11:692973. doi: 10.3389/fonc.2021.692973
- Subtle Medical. *Subtlepet*. (2018). Available online at: <https://subtlemedical.com/usa/subtlepet/> (accessed February 15, 2023).
- Pain C, Egan G, Chen Z. Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement. *Eur J Nucl Med Mol Imaging*. (2022) 49:3098–118. doi: 10.1007/s00259-022-05746-4
- Gavrilidis P, Koole M, Annunziata S, Mottaghy F, Wierts R. Positron range corrections and denoising techniques for gallium-68 PET imaging: a literature review. *Diagnostics*. (2022) 12:2335. doi: 10.3390/diagnostics12102335
- Virgolini I, Ambrosini V, Bomanji J, Baum R, Fanti S, Gabriel M, et al. Procedure guidelines for PET/CT tumour imaging with 68Ga-DOTA-conjugated peptides: 68Ga-DOTA-TOC, 68Ga-DOTA-NOC, 68Ga-DOTA-TATE. *Eur J Nucl Med Mol Imaging*. (2010) 37:2004–10. doi: 10.1007/s00259-010-1512-3
- NEMA. Association. N.E.M., NEMA standards publication NU-2 2012 performance measurements of positron emission tomographs. Rosslyn: NEMA (2012).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1137514/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Injected dose scatter plot according to BMI for the 44 patients in Protocol_FixedDose group.

18. Liu H, Wu J, Lu W, Onofrey J, Liu Y, Liu C. Noise reduction with cross-tracer and cross-protocol deep transfer learning for low-dose PET. *Phys Med Biol.* (2020) 65:185006. doi: 10.1088/1361-6560/abae08
19. Reddy R, Schmidtlein C, Giancipoli R, Mauguen A, LaFontaine D, Schoder H, et al. The quest for an accurate functional tumor volume with (68)Ga-DOTATATE PET/CT. *J Nucl Med.* (2021) 121:262782. doi: 10.2967/jnumed.121.262782
20. Kratochwil C, Stefanova M, Mavriopoulou E, Holland-Letz T, Dimitrakopoulou-Strauss A, Afshar-Oromieh A, et al. SUV of [68Ga]DOTATOC-PET/CT predicts response probability of PRRT in neuroendocrine tumors. *Mol Imaging Biol.* (2015) 17:313–8. doi: 10.1007/s11307-014-0795-3
21. Sharma R, Wang W, Yusuf S, Evans J, Ramaswami R, Wernig F, et al. (68)Ga-DOTATATE PET/CT parameters predict response to peptide receptor radionuclide therapy in neuroendocrine tumours. *Radiother Oncol.* (2019) 141:108–15. doi: 10.1016/j.radonc.2019.09.003
22. Huizing D, Koopman D, van Dalen J, Gotthardt M, Boellaard R, Sera T, et al. Multicentre quantitative (68)Ga PET/CT performance harmonisation. *EJNMMI Phys.* (2019) 6:19. doi: 10.1186/s40658-019-0253-z
23. Velikyan I, Sundin A, Sörensen J, Lubberink M, Sandström M, Garske-Román U, et al. Quantitative and qualitative intrapatient comparison of 68Ga-DOTATOC and 68Ga-DOTATATE: net uptake rate for accurate quantification. *J Nucl Med.* (2014) 55:204–10. doi: 10.2967/jnumed.113.126177
24. Kaemmerer D, Peter L, Lupp A, Schulz S, Sängler J, Prasad V, et al. Molecular imaging with ⁶⁸Ga-SSTR PET/CT and correlation to immunohistochemistry of somatostatin receptors in neuroendocrine tumours. *Eur J Nucl Med Mol Imaging.* (2011) 38:1659–68. doi: 10.1007/s00259-011-1846-5
25. Sher A, Lacoeuille F, Fosse P, Vervueren L, Cahouet-Vannier A, Dabli D, et al. For avid glucose tumors, the SUV peak is the most reliable parameter for [(18)F]FDG-PET/CT quantification, regardless of acquisition time. *EJNMMI Res.* (2016) 6:21. doi: 10.1186/s13550-016-0177-8



OPEN ACCESS

EDITED BY

Giorgio Treglia,
Ente Ospedaliero Cantonale (EOC), Switzerland

REVIEWED BY

Chunhao Wang,
Duke University Medical Center, United States
Jing Wang,
Fudan University, China
Salvatore Annunziata,
Fondazione Policlinico Universitario A. Gemelli
IRCCS, Italy

*CORRESPONDENCE

Bart M. de Vries
✉ b.devries1@amsterdamumc.nl

RECEIVED 06 March 2023

ACCEPTED 17 April 2023

PUBLISHED 12 May 2023

CITATION

de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW and Boellaard R (2023) Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review.
Front. Med. 10:1180773.
doi: 10.3389/fmed.2023.1180773

COPYRIGHT

© 2023 de Vries, Zwezerijnen, Burchell, van Velden, Menke-van der Houven van Oordt and Boellaard. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review

Bart M. de Vries^{1*}, Gerben J. C. Zwezerijnen¹,
George L. Burchell², Floris H. P. van Velden³,
Catharina Willemien Menke-van der Houven van Oordt⁴ and
Ronald Boellaard¹

¹Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, ²Medical Library, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, ³Department of Radiology, Leiden University Medical Center, Leiden, Netherlands, ⁴Department of Oncology, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Rational: Deep learning (DL) has demonstrated a remarkable performance in diagnostic imaging for various diseases and modalities and therefore has a high potential to be used as a clinical tool. However, current practice shows low deployment of these algorithms in clinical practice, because DL algorithms lack transparency and trust due to their underlying black-box mechanism. For successful employment, explainable artificial intelligence (XAI) could be introduced to close the gap between the medical professionals and the DL algorithms. In this literature review, XAI methods available for magnetic resonance (MR), computed tomography (CT), and positron emission tomography (PET) imaging are discussed and future suggestions are made.

Methods: PubMed, [Embase.com](#) and Clarivate Analytics/Web of Science Core Collection were screened. Articles were considered eligible for inclusion if XAI was used (and well described) to describe the behavior of a DL model used in MR, CT and PET imaging.

Results: A total of 75 articles were included of which 54 and 17 articles described *post* and *ad hoc* XAI methods, respectively, and 4 articles described both XAI methods. Major variations in performance is seen between the methods. Overall, *post hoc* XAI lacks the ability to provide class-discriminative and target-specific explanation. *Ad hoc* XAI seems to tackle this because of its intrinsic ability to explain. However, quality control of the XAI methods is rarely applied and therefore systematic comparison between the methods is difficult.

Conclusion: There is currently no clear consensus on how XAI should be deployed in order to close the gap between medical professionals and DL algorithms for clinical implementation. We advocate for systematic technical and clinical quality assessment of XAI methods. Also, to ensure end-to-end unbiased and safe integration of XAI in clinical workflow, (anatomical) data minimization and quality control methods should be included.

KEYWORDS

deep learning, explainable artificial intelligence, magnetic resonance (MR) imaging, computed tomography (CT) imaging, positron emission tomography (PET) imaging

1. Introduction

Computer-aided diagnostics (CAD) using deep learning (DL) have been widely used in diagnostic imaging for various diseases and modalities (1–5). It shows almost similar or superior performance in comparison to medical professional aided diagnostics and therefore has great potential to be introduced in clinical workflow (6). However, despite the promising results, DL algorithms have not achieved high deployment in clinical practice yet. Unlike simpler machine learning (ML) approaches, DL algorithms do not require manual extraction of features depending on volumes of interest (VOIs) annotation. Instead, DL algorithms extract features in an unsupervised way, i.e., extract features without *a priori* defined assumptions and regulations. Ideally, efficient learning and explainability, i.e., understanding of the underlying DL model, should work together in synergy (Figure 1). Although DL algorithms have superior learning capabilities, they lack transparency due to this underlying black-box mechanism. Therefore, the DL algorithms are difficult to validate, i.e., which features trigger model decision, and lack trustworthiness which is one of the main causes of its low deployment (7–9).

To close this gap, transparency of these DL algorithms should be improved to provide the medical professional and other stakeholders with a pragmatic explanation of the model its decision (10). Explainable artificial intelligence (XAI) can mitigate this gap, because their attribution (i.e., feature importance) methods provide

the user with information on why a specific decision is made. This way the user can back propagate the models decision to target specific attributions present in the image. XAI may, therefore, have the potential to be used as a new imaging biomarker (IB) in routine management of patients. In other words, XAI may be able to function as an indicator of normal and/or pathogenic biological processes, which can complement medical professionals in medical decision-making. Also, XAI may provide new insight in disease characteristics, which alternatively can be used as an indicator of responses to an exposure or (therapeutic) intervention. However, XAI should also provide transparency about the quality/legibility of its decision, explanation, and (possible) associated errors. So, before XAI can be used as an useful and trustworthy IB for either testing research hypotheses, or clinical decision-making, it must cross “translational gaps,” through performing and reporting technical validation, clinical validation and assessment of cost-effectiveness (11, 12). Also, the new European Medical Device Regulation (EU MDR) endorses strict regulations regarding transparency that need to be met before such a tool can be implemented in clinical practice (13). XAI may be one of the keys to more transparent, ethical (unbiased) safe and trustworthy deployment of DL algorithms in clinical practice, but better understanding of current practice is required.

This literature review addresses the XAI methods related to DL algorithms in medical imaging. We limit the scope of this review to (functional) magnetic resonance (MR), computed tomography (CT),

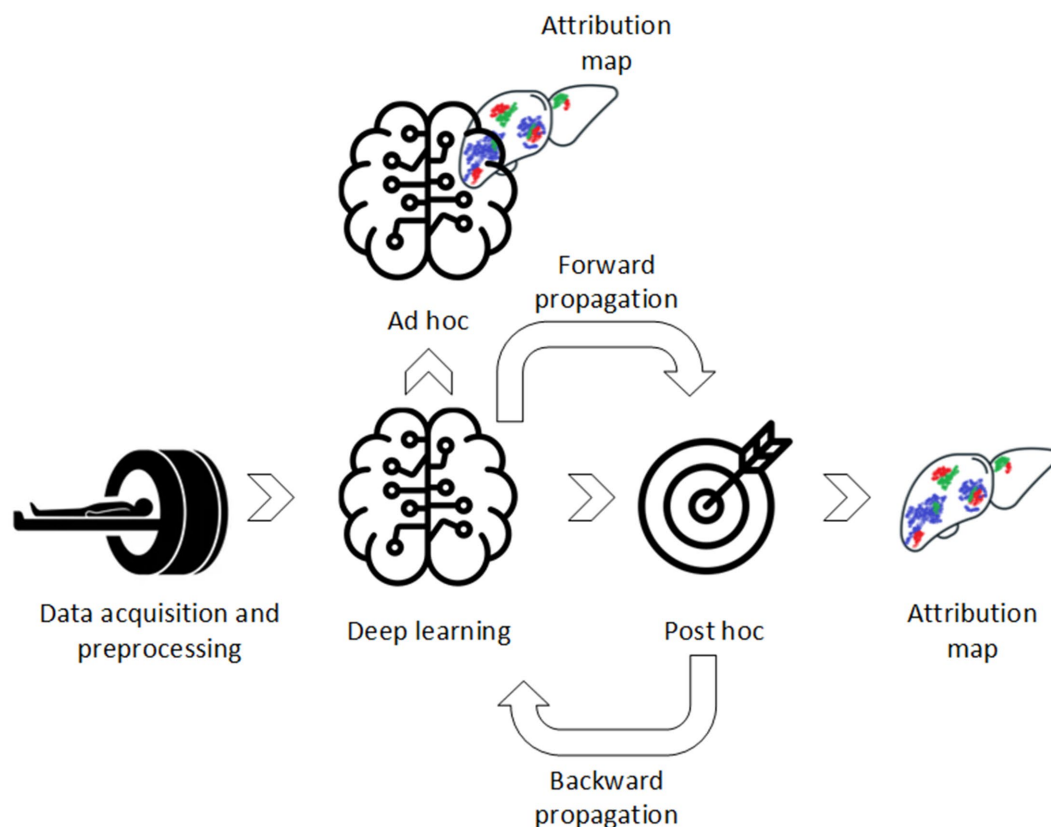


FIGURE 1
Conceptual difference between *post hoc* and *ad hoc* XAI methods.

and positron emission tomography (PET) imaging, which are three of the major cross-sectional imaging modalities. Also, we will try to establish a definition of what high quality explanation means at the end of this review.

2. Materials and methods

A systematic search was performed in the databases: PubMed, Embase.com and Clarivate Analytics/Web of Science Core Collection. The timeframe within the databases was from inception to 3rd October 2022 and conducted by GLB and BdV. The search included keywords and free text terms for (synonyms of) “explainable” or “interpretable” combined with (synonyms of) “artificial intelligence” combined with (synonyms of) “medical imaging.” A full overview of the search terms per database can be found in [Supplementary Tables 2–4](#). No limitations on date or language were applied in the search.

To be included in this literature review, studies had to meet the eligibility criteria presented in [Table 1](#). Included studies were classified based on *post* and/or *ad hoc* analysis ([Figure 1](#)):

- *Post hoc* methods: These refer to XAI methods that are used after DL model development;
- *Ad hoc* methods: These refer to XAI methods that are used during DL model development.

Additional background literature was included to provide (in-depth) information of the XAI methods. This was done through a specific search in PubMed.

In the result section, a general taxonomy of the attribution methods will be provided. Subsequently, per XAI method a (technical) conceptual explanation, its application, its advantages/disadvantages and a comparison with other XAI methods will be provided. Also, we will address the translation gaps present in the literature and a flowchart to *a priori* determine which XAI method to use in medical imaging will be provided. The structure of the flowchart is based on the taxonomy of the available XAI methods as present in the result section of this manuscript and based on from our perspective XAI important disease characteristics identified from the included literature. In addition, we will discuss metrics used in literature for technical and clinical quality assessment of these XAI models. Finally, the current and future direction in this field will be summarized. In [Supplementary material](#) a more extensive technical explanation is provided per XAI method.

3. Results

Searches of the literature databases resulted in the inclusion of a total of 117 studies ([Figure 2](#)). From the 117 studies, 10 did not have full-text available, 31 did not use or did not clearly describe the usage of XAI methods, eight did not use (medical) image data and three did not use DL, and therefore these were excluded from the review. Of the 75 studies included in the review, 54 studies reported data from *post hoc* analysis, 17 reported data from *ad hoc* analysis and four reported data from both *ad hoc* and *post hoc* analysis. A total of 24 additional

TABLE 1 Eligibility criteria for inclusion/exclusion.

Eligibility criteria	
Inclusion criteria	Exclusion criteria
XAI used and well describe in the method and result section	Either XAI is not used or is not well described in the method and/or result section
Medical image data available and used as input for DL model	Either no medical image data available or not used as input for DL model
- MRI;	- Either not MRI;
- CT;	- Or CT;
- PET	- Or PET
DL model used	No DL model used

studies were included to provide background information. [Supplementary Table 1](#) presents an overview of the 75 studies included in the review.

3.1. Taxonomy of XAI methods

The XAI methods in this study are classified based on the XAI taxonomy as shown in [Figure 3](#). *Post hoc* analysis provides model explanation after the classification is made, i.e., an AI model that is able to learn, but requires an additional model to provide an explanation. On the contrary, *ad hoc* explanation models are AI models, which are designed to be intrinsically explainable, i.e., a model that is both able to learn and to explain. Agnostic models are XAI methods that are able to explain multiple (technical) different AI models, while other XAI methods only work with one specific AI model such as a convolutional neural network (CNN). Global XAI methods are models, which are able to capture per-voxel attribution and inter-voxel dependencies, while local XAI methods are only able to provide per-voxel attribution. High-resolution XAI provides a per-voxel attribution value, while low resolution XAI provides a single attribution value for multiple voxels.

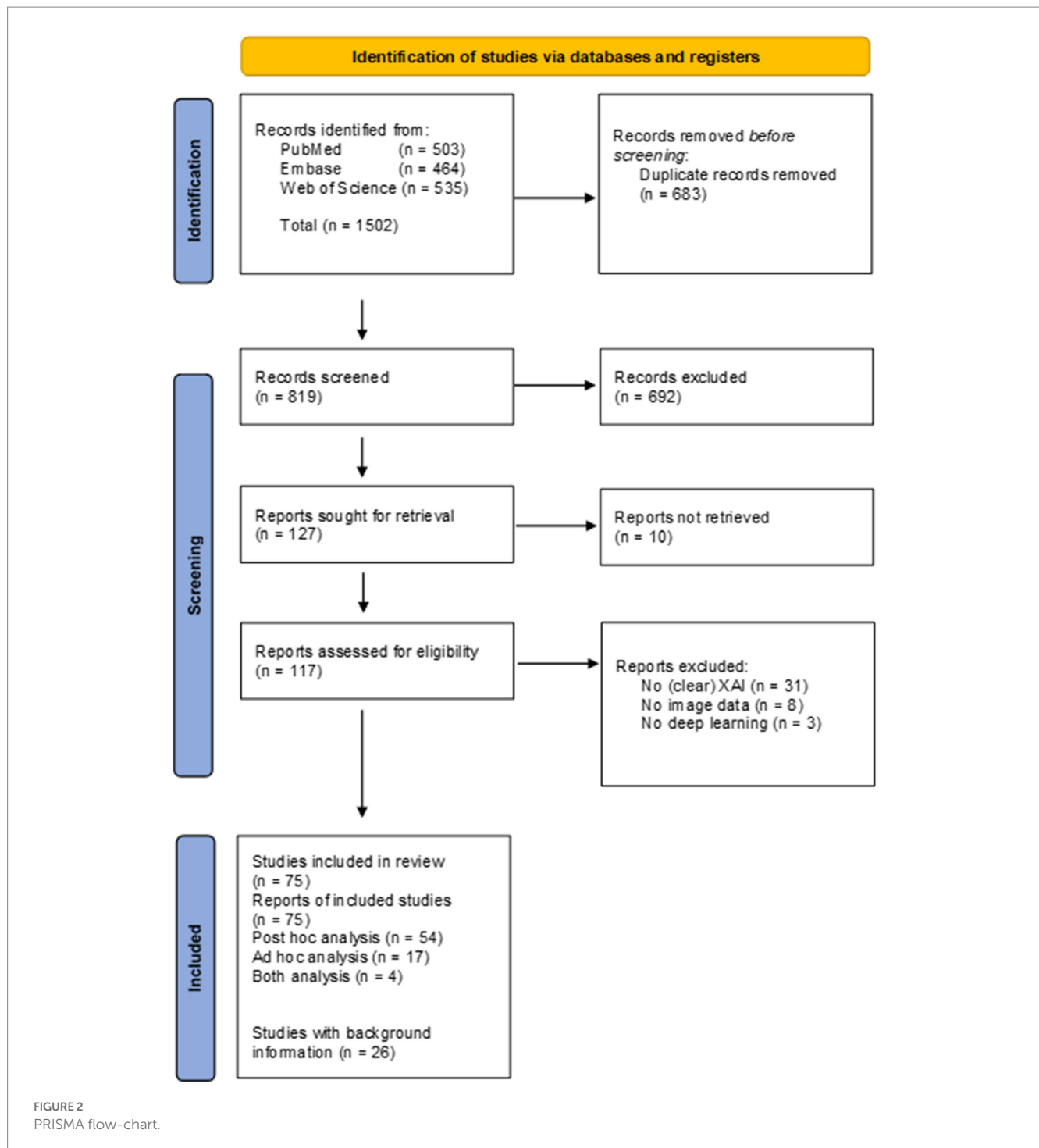
3.2. Post hoc XAI methods

The majority (~75%) of the DL algorithms in this literature study used *post hoc* XAI methods due to its wide availability and its plug-and-play deployment. In the following section, the *post hoc* methods will be divided into gradient-propagation methods, perturbation methods and briefly segmentation and radiomic methods will be discussed. An overview of the *post hoc* attribution methods are shown in [Table 2](#) (and a more extensive explanation in [Supplementary material](#): Appendix A).

3.2.1. Gradient-propagation approaches

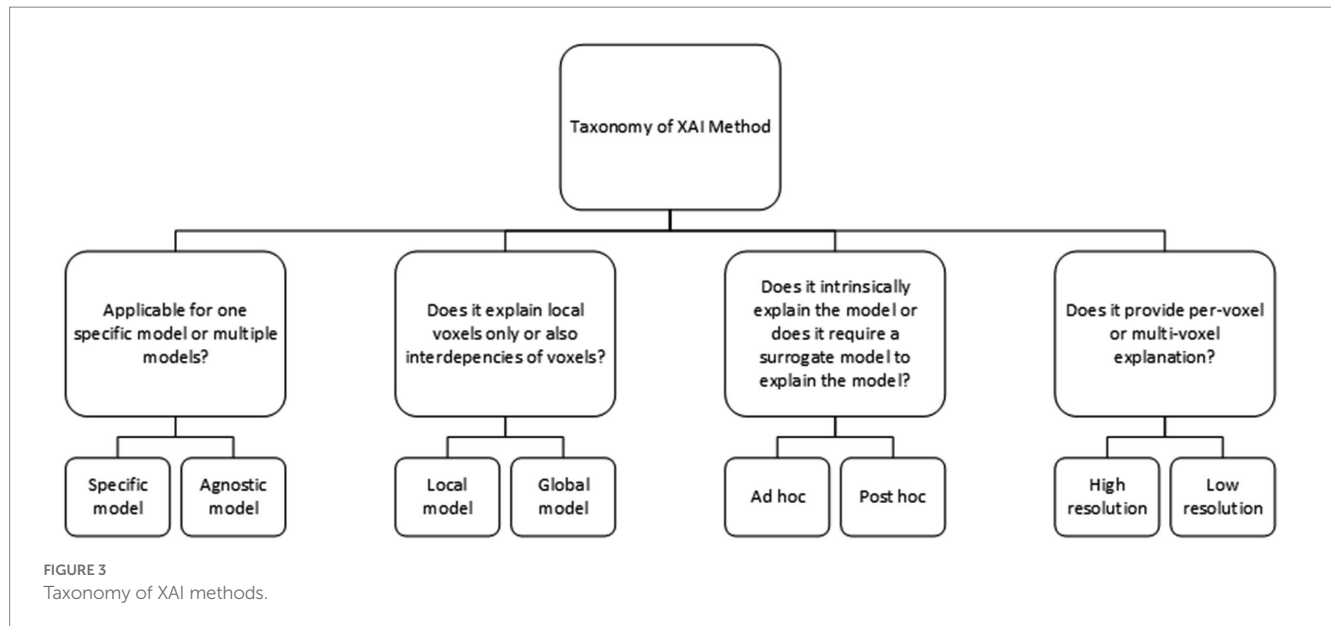
3.2.1.1. Vanilla gradient (VG)

VG is a XAI method that create an attribution map by calculating gradients over the layers using a single forward and backward propagation, i.e., the input image is fed into the AI



model and an output score is calculated (forward) and subsequently the dependence (gradient) between the neurons/convolution layers (subunit of the AI network that learns/extracts features from the input image) and the output is calculated (backward) to create an attribution map. Due to its simplicity, it is an intuitive attribution method and it requires low computational power. An attribution based framework called NeuroXAI compared VG and other attribution based visualization methods for MRI analysis of brain tumors (14). These methods were both utilized for classification and segmentation feature

visualization. In comparison to the other attribution methods, VG generated noisy attribution maps and suffers from gradient saturation, i.e., change in a neuron does not affect the output of the network and therefore cannot be measured. In a different study in which the contrast enhancement phase from CT images is predicted, similar results were seen using VG for feature visualization (15). In addition, VG lacks the ability to differentiate between classes (e.g., healthy vs. disease) (16). This illustrates that VG lacks ability to generate clear and class discriminative attribution maps.



3.2.1.2. DeconvNET

DeconvNET is effectively an equivalent of VG apart from the way it calculates the gradient over a Rectified Linear Unit (ReLU) function (17), i.e., a linear function that will output only positive input values and helps with improving model convergence during model training. TorchEsegeta, a framework for interpretable and explainable image-based DL algorithms, compared multiple attribution methods for interlayer CNN visualization in the segmentation of blood vessels in the human brain (18). VG and deconvNET provided more human-interpretable results than the other attribution methods (e.g., DeepLIFT and GradCAM++), since they mainly focused on the vessels, while other methods also showed non-vessel activation.

3.2.1.3. Guided back propagation (GBP)

GBP both incorporates the VG and the deconvNET (19). This results in fewer activated voxels and therefore in less noisy attribution maps than by using each method individually. In the NeuroXAI framework, GBP showed target specific attribution maps with indeed less noise in comparison to VG (14). In a study performed for predicting brain abnormalities using MRI, an additional smoothing function to the GBP was proposed to suppress the amount of noise and the effect of non-target specific attributions even more (20). The attribution maps showed low noise and accurate localization of a range of morphological distinct abnormalities. However, although GBP may show less noisy attribution maps, it may also result in overly sparse attribution maps, which are not useful for complete image characterization (21).

All three gradient based methods are very sensitive to understand how the neural network layers extract features, but are not class discriminative. Also, because of ReLU and pooling layers, local gradients may saturate. Therefore, important features may vanish over the layers in the network and that may result in incomplete model explanation or even focus on irrelevant features.

3.2.1.4. Layer-wise relevance propagation (LRP)

LRP is a XAI method that operates by propagating the class score backward over the neural layers to the input image using LRP specific rules (22). The concept of LRP is to conserve inter-neuron dependency,

i.e., what has been received by a neuron layer will be redistributed to the following lower layer in equal quantity. The decomposition is based on propagating relevance scores between the neurons instead of gradients and therefore, we tackle the difficulties of the saturation problem. In a study for screening of abdominal aortic aneurysm in CT images (23), LRP showed clear class difference based on activation difference in the lumen of the aorta. However, high activation for both classes was also seen in the vertebra, which indicates that either the DL model is biased, the DL model did not converge, the vertebra is a confounder, or that LRP also incorporates non-target specific features in its attribution map. A similar result was seen for COVID-19 classification, in which LRP was not able to visualize target-specific features (24). However, other studies showed class-discriminative regions and precise localization of lesions using LRP (25, 26). This difference may be explained by differences in DL model performance, biased data and LRP configuration, although there may not be one absolute reason.






3.2.1.5. DeepLIFT

DeepLIFT is a XAI method that uses a neutral reference activation (e.g., neuron activation of CT scan without pathology/disease) to solve the saturation problem (27). This reference activation is used to describe the change of a new neuron activation in comparison to the reference activation. From these differences, contribution scores are calculated for each neuron to compute an attribution map. DeepLIFT was compared with LRP and VG for identification of Multiple Sclerosis (MS) patients on MRI (26). This was done by perturbation of the three attribution maps for three VOIs. From quantitative assessment, it can be seen that DeepLIFT performs slightly better than LRP and much better than VG in extracting target-specific features. Both LRP and DeepLIFT are able to tackle gradient saturation, which may be the reason why it performs better than VG in this classification task.

3.2.1.6. Class activation map (CAM)

CAM is one of the most well-known model specific attribution methods (28, 29). It uses a Global Average Pooling (GAP) layer

TABLE 2 The different *post hoc* XAI methods scored [low/no (red), average (orange), and high/yes (green) performance] based on target specificity, spatial-resolution and local/global voxel dependency capability, model agnostic, and technical simplicity, respectively.

Post hoc	Characteristics					
						
VG	*					
DeconvNET	*					
GBP	*					
LRP	*					
DeepLIFT	*					
CAM	*					
Grad-CAM	*					
Occlusion	*	^				
LIME	*					
SHAP	*					

*Depends on DL model convergence. ^Depends on occlusion method.

instead of multiple dense layers, which introduces linearity after the last convolution layer and before the final dense layer. Since CAM only uses features from the last convolution layer, low-dimension attribution maps are generated. Therefore, the low-dimension CAM is able to visualize whether a model is able to roughly focus on specific targets, but due to its low specificity, it lacks discriminative power to accurately characterize class based features (30, 31). Perturbation analysis of multiple attribution methods also showed that gradient based methods have higher specificity than CAM (15). Yet, CAM can be discriminative in classification tasks in which the classes have clear visual differences, e.g., healthy brain vs. Alzheimer's brain (32) or by performing patch based (more focused) tumor analysis instead of whole image tumor analysis (33, 34).

3.2.1.7. Gradient-CAM (Grad-CAM)

Employment of XAI methods has showed tremendous growth due to COVID-19 detection (35). In general you can distinguish these methods based on using the whole CT image, or only using a segmentation of the lungs for COVID-19 detection. Especially, whole image based COVID-19 detection showed major performance difference in attribution mapping. Grad-CAM, an extension of CAM, was the most used attribution method and showed both very specific (36, 37) as also non-specific attributions (24, 38–41), but was overall able to roughly locate the potential COVID-19 lesions to make accurate predictions. To remove the influence of non-target specific features, *a priori* segmentations of the lungs was proposed (42–47). This way both the DL algorithms as the XAI methods can only extract features from the lungs. This anatomical based XAI method showed higher specificity than by using the whole CT image using Grad-CAM. This shows that DL and XAI methods benefit from medical based data minimization, in other words reducing the amount of trainable features and/or removing non-informative features from the input image.

Similar non-target specific attribution maps were also seen for the automated grading of enlarged perivascular spaces in acute stroke (48) and cerebral hemorrhage detection (49) using the whole image (without data minimization). Similar as for the COVID-19 studies to solve this specificity problem, *a priori* anatomical segmentation was used to classify and visualize mortality risks based on myocardial PET (50), Alzheimer's disease (51) and schizophrenia based on MRI (52). However, although data manipulation suppresses the presence of non-target specific features, Grad-CAM still suffers from low specificity due to its low-dimensional attribution maps (43, 53). In a study for classification of lung cancer histology based on CT images, the authors suggested that based on the Grad-CAM attribution maps, the activated features around the tumor correspond to regions harboring occult microscopic disease (2). However, this is more likely caused by this low-dimensionality characteristic of Grad-CAM, because CT does not have high enough spatial resolution to detect these microscopic diseases.

Similar to CAM, Grad-CAM can be class discriminative in case of classification tasks with clear radiological difference between the classes (5, 54–58). However, in case of tasks with less obvious radiological differences, e.g., predicting survival based on tumor characteristic, Grad-CAM lacks fine-grained details, complementary attribution methods should be used such as VG and GBP (15, 18). A study that combined GBP with Grad-CAM, a method called

guided Grad-CAM (gGrad-CAM), showed better localized attribution maps with higher resolution in MRI analysis of brain tumors (14). This advocates for combining the advantages of attribution methods for human-interpretable and precise model visualization.

There have been multiple other improved variation of Grad-CAM, such as Grad-CAM++. Grad-CAM++ has been introduced to provide better localization of target-specific features than Grad-CAM (59). Grad-CAM averages the gradients of the feature maps, which may suppress the difference in importance between the different gradients. Grad-CAM++ replaces this with a weighted average, which measures the importance of every unit of a feature map. It showed more target-specific attribution maps than Grad-CAM in the prediction of knee osteoarthritis using MRI (25).

The advantage of gGrad-CAM is clearly shown in a study where they compared different attribution methods for brain glioma classification (14). Grad-CAM provided the least noisy attribution maps and GBP provided attribution maps with high resolution but not class-discriminative. However, gGrad-CAM provided both class-discriminative as high resolution maps in which the edges of the tumor are highlighted instead of the whole tumor. Similar results were also seen for classification of frontotemporal dementia (60), although the skull was seen important for the classification as well.

However, non-target specific features in attribution maps do not only arise because of underperformance in DL algorithms and/or attribution methods. Artifacts can also play a major role in tricking DL algorithms and attribution methods (61). That is why it is important to have high quality data, perform (medical based) data minimization and have *a priori* (DL-based) quality control methods to detect bias present in the data (62, 63). In addition, it is also not always trivial what convolution layer should be used to compute the attribution map (64). Deeper layers may have higher hierarchical structures, but may suffer from low specificity and therefore using a shallower layer may contain more informative features.

3.2.2. Perturbation XAI methods

3.2.2.1. Occlusion mapping

Occlusion mapping is a simple to perform approach that reveals the feature importance of a model using systematic perturbation/conditioning over the image (e.g., replacing input pixels with zeros). In contrast to previous methods, occlusion maps do not take the feature maps into account, but only the different patches (grid- or atlas-wise combination of multiple pixels) of the input image. Therefore, it is a very intuitive method, which can easily be adapted to specific occlusion analysis. An example of this is a study that investigated the use of DL algorithms in predicting and visualizing Alzheimer's disease and Autism using MRI. But instead of rectangles, the Harvard-Oxford cortical and subcortical structural atlas was used for occlusion mapping (65). This provides a method that can easily be compared with more traditional atlas based analysis and therefore provides a medical based, transparent and intuitive visualization of the DL algorithm.

Randomized Input Sampling for Explanation (RISE) is an equivalent of occlusion mapping, but instead of systematic perturbation of the input image, it generates multiple random perturbation maps, which are pointwise multiplied with the input

image (66). Another occlusion method is square grid, where perturbation is performed using square grid divisions (62). These methods produce intuitive results, but are too rigid to follow anatomical/pathological structures present in the images, and require large computational power due to many forward and backward propagations.

3.2.2.2. Local interpretable model-agnostic explanations (LIME)

Instead of a predefined or random occlusion function, LIME perturbs super-pixels, which are a group of pixels that share common pixel/voxel characteristics. For COVID-19 detection using CT, super-pixels followed anatomical/pathological structures/characteristic of the image and therefore gave a better representation of the image than the previous occlusion methods (67–69). However, since LIME uses super-pixels as a whole, it provides occlusion maps with relatively low specificity. Also, from these COVID-19 studies it can be seen that non-target specific features (e.g., chest wall) show high activation. This suggests that also occlusion mapping suffers from non-target specific activation. In addition, LIME requires initialization parameters (kernel size, maximum distance, etc.) to compute super-pixels, which can be difficult to optimize.

3.2.2.3. SHapley additive exPlanations (SHAP)

SHAP is an advanced XAI algorithm that calculates SHAP values, which represent the attribution of each voxel to the change of the expected model prediction when conditioning on that voxel using reference samples (70). DeepSHAP is an extension of SHAP and works in an almost similar way as DeepLIFT. It can provide both local as global explanation based on individual pixels/voxels, but also whether a pixel/voxel is negatively associated or positively associated with the predictive class. Because of this, DeepSHAP may be difficult to interpret as is shown in a study to predict brain tumors using MRI (67). However, in a study in which the volumetric breast density on MRI was calculated using DeepSHAP, intuitive DeepSHAP maps were created (71). This difference may be the result of difference in data size and quality between the studies, but may also be impacted by the quality of the reference samples to create the attribution maps. Also, because of the required reference samples, DeepSHAP may not work optimal in classification tasks where there are substantial (non-)rigid anatomical/pathological variation present in the images. Feature explanation may therefore be negatively impacted by anatomical differences between the reference samples and the input image and therefore may show non-specific attributions.

3.2.3. Probability maps, deep feature maps, radiomics, and physics/clinical data

Previous described *post hoc* attribution methods predominantly focus on classification models, which are trained using weak labels, i.e., one label for the whole image. In contrast, segmentation DL algorithms use voxel-level annotations and compute voxel-level probability maps. Therefore, these probability maps are less complex to understand.

These probability maps were used to detect prostate lesion from multi-parametric MR sequences, which were easily interpretable and it allowed to perform prostate lesion analysis in new image data (72, 73). Similar probability maps were also created to detect lumbar spine MR intensity changes (74). However, further specific Modic type

categorization was performed using a non DL-based, but interpretable signal-intensity based nearest neighbor algorithm.

These segmentations can also be used to explore radiomic (e.g., intensity, morphology, and texture) based differences between classes. A joint detection and radiomic based classification algorithm was developed to explore the radiological difference between COVID-19 and community acquired pneumonia and showed clear difference between the two classes using understandable radiomic features (75). A similar approach was used for detection and classification of lung nodule malignancies (76, 77).

Although these methods (partly) tackle the problem of black-boxes, voxel-level annotation is very cumbersome and radiomic analysis depends on accurate VOI annotations, and *a priori* defined assumptions and regulations. This may suppress the full potential of DL algorithms and therefore have a possibility to underperform.

Another explainable method, is the use of deep feature maps (intermediate attribution maps) of the DL-based models (78). These deep feature maps provide the user with attributions maps of the intermediate model layers, which visualizes the underlying feature extraction mechanism used by the DL-based model. It therefore can give the user an understanding of what features are used, but more importantly how these features are processed throughout the model.

Physics-based AI models could also aid in higher transparency, as these models can explain feature extraction through well-defined mathematical formulas/assumptions, i.e., physics-aware AI. These models incorporate physics/mathematical knowledge prior to training. However, this approach is predominately used for image reconstruction and has low application/added-value for classification (yet) (79).

Also, clinical data (e.g., patient history) could aid in better performance and transparency of The AI algorithms. for detection of prostate cancer using MRI, clinical data improved The diagnostic performance significantly (73). In a different study, both clinical data and radiomics features showed a complementary role in the prediction of EGFR and PD-L1 status using CT images (30).







3.3. Ad hoc XAI models

Ad hoc XAI models are intrinsically able to learn and explain, which is different to the DL models that predominantly focusses on learning to achieve high performance (learning) and require a *post hoc* XAI algorithm to explain model behavior. An overview of the *ad hoc* attribution methods are shown in Table 3 (and a more extensive explanation in [Supplementary material: Appendix A](#)).

3.3.1. Explainable deep neural network (xDNN)

xDNN is a XAI method that uses a prototype identification layer in the network to identify new data samples based on similarity to predefined data samples (prototypes) (80). For this, representative prototypes need to be selected for each class, which can be a difficult task, especially in case of a cohort with a wide variety in disease morphology. Also, difference in class morphology is not always trivial and therefore obtaining representative prototypes can be difficult. However, xDNN can be very powerful in tasks where there is known difference between classes, as is the case for COVID-19 screening (81–84) and artifact detection (63). In these studies representative prototypes were used to assess new images based on their similarity.

TABLE 3 The different *ad hoc* XAI methods scored [low/no (red), average (orange), and high/yes (green) performance] based on target specificity, spatial-resolution and local/global voxel dependency capability, model agnostic, and technical simplicity, respectively.

Ad hoc	Characteristics					
						
xDNN	*					
Attention estimator	*					
Capsule network	*					

*Depends on DL model convergence.

This provides the user with transparent and intuitive model explanation, which in some way mimics the way we humans extract features based on previous experience.

3.3.2. Capsule networks

Capsule networks are described to be the new sensation in DL, since they are able to eliminate the pose and deformation challenges faced by CNNs, require less data and less computational power (85). A capsule tries to describe the presence and the instantiation parameters (orientation, thickness, skewed, position, etc.) of a particular object (e.g., tumor or lung) at a given location as a vector. Subsequently, the vectors from a lower capsule layer try to predict the output for the higher layer based on the instantiation parameters. Lower layer vectors with high agreement are routed to the following layer and the other vectors are suppressed, ideally resulting in only target specific attribution maps. A study proposed a novel capsule network-based mixture of expert (MIXCAPS) for detection and visualization of lung cancer nodules in CT images (86). MIXCAPS is an extension of the traditional capsule network, where instead of a single CNN, a mixture of (expert) CNNs specialized on a subset of the data and an ensemble of capsule networks is used. The authors compared MIXCAPS with a single capsule network, a single CNN and a mixture of CNNs and showed superior performance using MIXCAPS. However, its full potential has not been shown yet and requires further understanding before it will be used as the standard DL algorithm in this field.

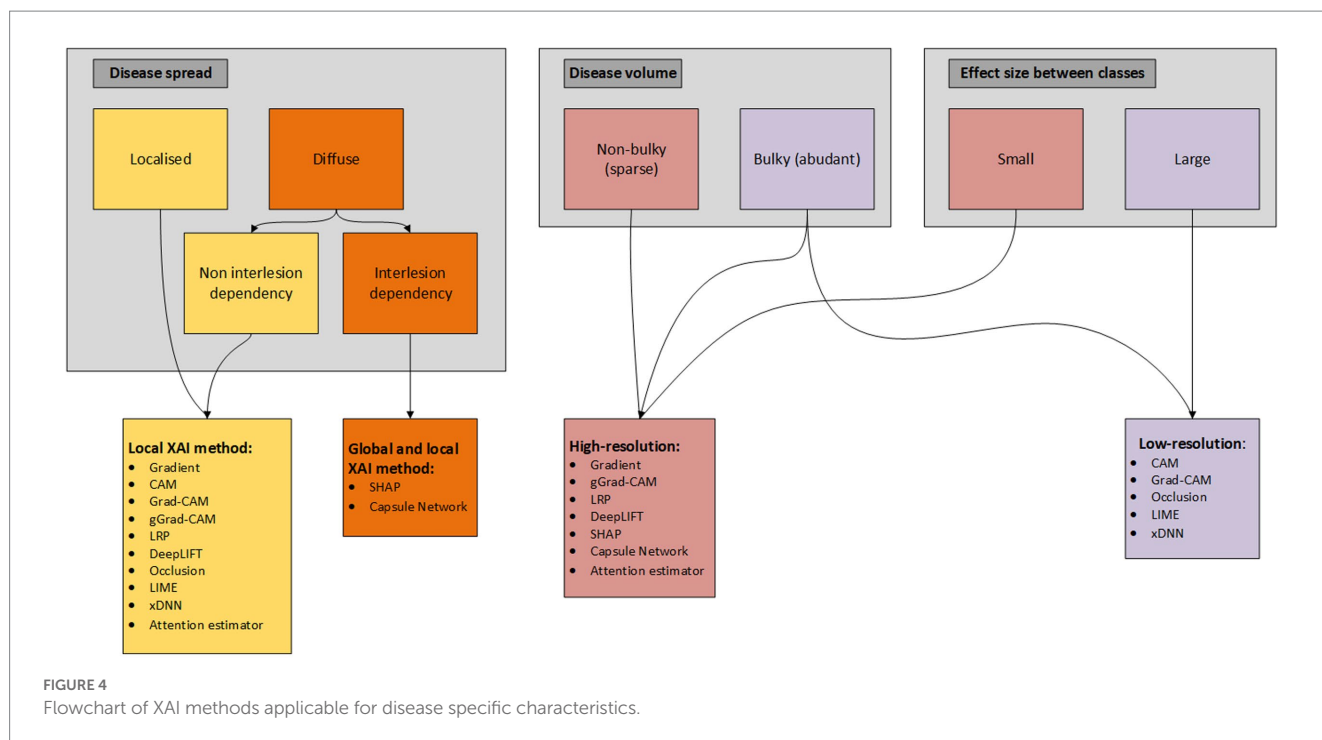
3.3.3. Attention mapping

A trainable spatial self-attention mechanism is in contrast to *post hoc* attention mechanisms, trained during model training to support (important) feature extraction (87) and replaces traditional non-learnable pooling operations (e.g., max-pooling). Spatial attention mapping uses attention estimators to compute attention

mask from a convolution layer as a goal to extract important local feature vectors. Attention mapping showed high correspondence between attention scores of specific regions and classification score in a study that assessed interpretable imaging biomarkers for Alzheimer's disease using MRI (88). In addition, attention mapping for COVID-19 detection showed better capabilities to extract more complex and scattered regions (24, 89). Attention mapping has also showed superior target-specific feature extraction in inverted papilloma and nasal polyp classification using CT (90), adenocarcinoma screening using CT (91) and segmentation of multiple organs from MRI (92).

Attention mapping has also been investigated in combination with Multi Instance Learning (MIL). MIL tries to tackle the downsides of weak labels and labor intensive per-voxel annotation. Instead MIL uses a set of labeled bags, each consisting of multiple instances (slices). In case of binary classification, a bag will be annotated negative if all the instances in the bag are negative (e.g., no presence of disease) and will be positive if there is at least one instance in the bag which is positive. Therefore, MIL intrinsically provides a more interpretable decision and in combination with attention mapping it gives insight into every voxel its contribution to the bag label. This combination have been used for the detection of COVID-19 using CT and showed more precise and complete detection of the infection areas of COVID-19 than weak labeled methods (93, 94). A similar method has been used to predict EGFR mutation status using CT and improved the interpretability of the model (95). This indicates that attention mechanisms (in combination with MIL) provide more spatial resilient CNNs, as it mimics the human behavior of focusing on more relevant features, while suppressing irrelevant features.

An alternative attention mechanism has been suggested for detection of COVID-19 from CT by feature encoding using a gated recurrent neural network in the horizontal and vertical direction using a feature block grid (96). In contrast to traditional CNNs, this mechanism allows to capture the horizontal and vertical dependencies



of the features present in the image. This attention mechanism helps to make the model interpretable. However, it lacks specificity due to its grid-wise attention mechanism.

3.4. Explainability quality of attribution methods

Performance assessment of DL algorithms is almost always expressed in terms of diagnostic performance (e.g., accuracy, sensitivity) or overlap (e.g., Dice coefficient) with the gold standard. Although CNNs are seen as the current state-of-the-art algorithms in this field, there is no clear consensus what XAI method has superior performance over the other methods. One of the problems with these XAI methods is that the performance of the attribution methods is often not expressed in measurable (quantitative) metrics. Most comparisons are performed solely on visual inspection, which is susceptible to human subjectivity, especially in case of non-trained readers. Current literature therefore lacks high-quality and objective technical and clinical assessment of the attribution methods, which makes objective comparison between the XAI methods difficult.

However, from the studies that used technical and clinical quality assessment, in general you can divide measurable metrics into human-based and computer-based derived metrics. Human-based metrics predominantly use correspondence scores to assess overlap between decision relevant VOIs and the gold standard VOIs. In a study where they assessed the correspondence of the attribution map with the aorta, the radiologist used a 5-point Likert scale to determine correspondence (23). An equivalent score, the mean alignment index (MAI) was used to evaluate the attribution map for COVID-19 detection (44). Another study measured the effect of diagnostic performance with and without attribution maps (spinal Modic maps) (74). First, they provided three radiologist with a MRI without the

attribution maps and after a 4-week washout period, the radiologist regraded the same dataset with the attribution maps. Although a 4-week washout period might not be sufficient, such methods are able to validate the effect of attribution methods in complementing medical professionals in medical decision making and therefore helps improving the trustworthiness of these algorithms in this field.

Computer-based metrics also use metrics to measure overlap between the attribution maps and a gold standard. A study calculated correspondence between the attribution maps with brain tumor segmentations using a localization hit and the intersection over Union metric (55). In other studies, correlation analysis was performed to compare pneumonia ratio between radiologists and thresholded attribution maps for COVID-19 detection (24) and between attribution scores of brain regions and classification accuracy in Alzheimer's disease (88). Another method proposed is the use of perturbation of the input image based on the attribution maps (26). The idea behind this is that important features from the attribution map should correspond with important features from the input image, which is expressed as the area over the perturbation curve (AOPC). So the more the prediction score decreases by perturbation, the better an attribution method is capable to identify relevant input features, resulting in a high AOPC.

3.5. Disease specific XAI

Utilization of disease-specific XAI is not unambiguous and therefore we propose a flowchart (based on taxonomy of the XAI methods) to determine what XAI methods present in the literature are from our perspective (most) applicable based on disease specific characteristics/patterns (Figure 4). In this flowchart we differentiate between local and global and low- and high-resolution XAI methods, what we think are two (important) taxonomies that can be determined

a priori for the development of XAI methodology. Differentiation of the XAI methods is based on disease spread, disease volume and effect size between the classes. Disease spread is divided into localized (e.g., only primary tumor) and diffuse (e.g., diffuse large B-cell non-Hodgkin lymphoma), where diffuse spread is again subdivided into non-interlesion (e.g., predicting non-Hodgkin vs. Hodgkin lymphoma) and interlesion (e.g., prediction of overall survival for Hodgkin lymphoma) dependency. Although the difference between the two seems small, a non-interlesion dependency can be described in terms of a regional (small ROI/VOI) linear relation with the output [e.g., (non-)presence of bone metastasis in Hodgkin vs. non-Hodgkin patients], while interlesion interaction requires an explanation/relation for all pixels/voxels (e.g., relationship between primary tumor, lymph node and distant metastases). Interlesion (voxel) interaction therefore requires both local as global XAI, while localized disease only requires a XAI method to extract local features. Disease volume is divided into non-bulky (e.g., stage I pancreatic cancer) and bulky (e.g., diffuse large B-cell non-Hodgkin lymphoma). The effect size, i.e., the magnitude of the difference between classes, may in some cases be more difficult to determine *a priori*. Yet, we divide the effect size in small (e.g., predicting progression free survival in stage III colon cancer) and large (e.g., predict presence of glioma in brain vs. healthy brain). This flowchart can be helpful for researchers to determine *a priori* what XAI methods currently present in literature can aid in explaining their DL model. However, in the end researchers should determine how the complexity of the AI task compares with the complexity of the XAI method and therefore the flowchart should only be seen as an additional tool for XAI application.

4. Discussion

There has been growing interest in the deployment of XAI to explain DL black-boxes in the field of MR, CT, and PET imaging. However, this review demonstrates that there is a variety of XAI methods available and that there is currently no clear consensus present in literature on how and what XAI should be deployed to realize utilization of DL algorithms in clinical practice. Although a variety of XAI methods are proposed in literature, technical and clinical quality assessment of these methods is rarely performed. Also, there is little evidence of the impact of attribution methods to complement medical professionals in medical decision making and what medical professionals expect and demand of XAI (74). This all illustrates that current XAI methods on their own may not be sufficient to realize deployment in clinical practice, but requires additional/tweaked (XAI) methods to improve transparency and trustworthiness. Therefore, we advocate for an end-to-end solution, which integrates *a priori* data-quality control, data pre-processing, (self-)explainable modules and technical and clinical (X)AI model quality control (26, 74). In addition, to the best of our knowledge we are the first study that provides a guide for current available XAI utilization based on disease/AI task specific characteristics (Figure 4). Also, we have provided a hands-on summary of the (dis-)advantages of each XAI method (Tables 2, 3). Both can be helpful for researchers to *a priori* determine which XAI method can be useful for their disease-specific AI task.

The majority of the studies utilized *post hoc* attribution methods to explain model behavior. For successful employment, these XAI

methods should be transparent, explainable and safe for all stakeholders. Current *post hoc* XAI methods are overall able to provide transparent and understandable attribution maps, but show low specificity, resulting in non-target specific attribution maps. Anatomical data minimization seems to suppress the effect of this, but due their intrinsic technical characteristics some still lack to provide class discriminative performance. In recent years, more advanced *post hoc* methods have been proposed, such as DeepSHAP. DeepSHAP uses multiple reference image samples from both classes and is therefore able to provide both positive as negative attributions. Therefore, DeepSHAP enables reasoning both for and against a models decision, which is important to consider for a complete image analysis and diagnosis. Although this provides high model transparency and greater insights, excessive information may result in lower understandability by the medical professional. Also, DeepSHAP may be negatively impacted by anatomical (non-)rigid variation in images and reference images and therefore may not work optimal in medical imaging.

From this perspective it is important to consider what medical professionals consider as complementing information for decision making. It is therefore critical to focus on addressing the epistemic and non-epistemic concerns of this group in specific contexts and occasions of these DL algorithms. These algorithms should be designed in the context of its user, which includes flawless integration in the user's clinical workflow, respect the autonomy of the user and provide transparent and effective outputs (97). One of the overall issues of XAI is the low specificity of the computed attribution maps, i.e., non-informative attributions make it overly difficult to interpret the attribution maps. This may be due to the wide non-medical application of these XAI methods, i.e., they are not optimized for medical imaging. Therefore, although these methods may be useful in more simple (non) medical AI tasks, more difficult medical AI tasks may require XAI methods specifically developed for medical imaging. In other words, these systems should be designed around stakeholders/imaging modalities to ensure both transparent and trustworthy outputs.

Although not extensively present in literature, *ad hoc* XAI models do provide intrinsic explanation of their decision and seem to be more target-specific than *post hoc* XAI methods. Self-attention mapping has showed great interest, because it is able to intrinsically explain, showed higher target specificity than *post hoc* algorithms and is also relatively simple to understand and integrate into current systems. However, self-attention mapping is not able to find global feature dependencies, which can be important in disease mapping. Yet, self-attention mapping in combination with DeepSHAP enables it to find global features, which supports to combine *ad hoc* and *post hoc* XAI methods in future research. Another promising XAI method is capsule networks, which are intrinsically able to handle spatial relationship between features and therefore have seen to be more resilient to spatial variance than CNNs. Also, agreement by routing provides an intuitive explanation of which feature belongs to which object. Therefore, capsule networks have been suggested as the new state-of-the-art DL model, but more research is required to explore its full potential.

To ensure unbiased and safe end-to-end integration of DL systems, also data quality control should be performed. Especially for systems with small data exposure, poor data quality can have high impact of the models its reliability. For example measuring the signal-to-noise-ratio for data quality harmonization, DL-based artifact detection model (62,

63) or simple visual inspection can be proposed to provide information of the quality of the data before utilization in the diagnostic DL systems. Also, quality control of the attribution maps should be performed to assess the use of XAI as potential IB. Unfortunately, only few studies (26, 55) implemented quality control systems to assess whether the attribution maps do present target-specific features. The absence of complete and transparent technical and clinical reporting limits the usability of finding in studies and in consequence, the acceptance of XAI as IB in clinical practice. In response to this, a new version of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement was published (12). The TRIPOD Statements provides a checklist for reporting of studies developing, validating, or updating a prediction model for diagnostic and/or prognostic purpose. In combination with standardized modality and/or disease specific implementation guidelines (11), higher transparency and effectivity of XAI as new IB can be utilized in future research.

5. Conclusion

High quality explanation is user and task subjective and therefore we require pragmatic explanations to address the concerns of DL algorithms for each stakeholder/imaging modality. *Ad hoc* XAI methods seem to provide state-of-the-art explanation algorithms, which advocates for shifting from *post hoc* to integrating self-explainable modules in the DL models. However, there is (still) no unambiguous (self-)explainable XAI method addressing all concerns, which advocates for combining XAI methods, perform anatomical data minimization and implement data quality systems to ensure end-to-end unbiased and safe system integration into the context of the stakeholder/imaging modality.

Although XAI shows a great potential to be used as IB in clinical practice, technical and clinical quality assessment is currently rarely reported. We recommend the utilization of developing and reporting

guidelines, accepted by the AI-community, to ensure a higher transparency and quality of future developed XAI algorithms.

Author contributions

BV, CM-v, FV, and RB: conceptualization. BV, CM-v, FV, GB, and RB: methodology. BV: investigation, writing—original draft preparation, and visualization. BV, GZ, CM-v, FV, GB, and RB: writing—review and editing. CM-v, FV, and RB: supervision. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1180773/full#supplementary-material>

References

- Alongi P, Stefano A, Comelli A, Spataro A, Formica G, Laudicella R, et al. Artificial intelligence applications on restaging [18F]FDG PET/CT in metastatic colorectal cancer: a preliminary report of morpho-functional radiomics classification for prediction of disease outcome. *Appl Sci*. (2022) 12:2941. doi: 10.3390/app12062941
- Chaunzwa TL, Hosny A, Xu Y, Shafer A, Dia N, Lanuti M, et al. Deep learning classification of lung cancer histology using CT images. *Sci Rep*. (2021) 11:5471. doi: 10.1038/s41598-021-84630-x
- de Vries BM, Golla SSV, Ebenau J, Verfaillie SCJ, Timmers T, Heeman F, et al. Classification of negative and positive 18F-florbetapir brain PET studies in subjective cognitive decline patients using a convolutional neural network. *Eur J Nucl Med Mol Imaging*. (2021) 48:721–8. doi: 10.1007/s00259-020-05006-3
- Prezioso E, Izzo S, Giampaolo F, Piccialli F, Dell'Aversana Orabona G, Cuocolo R, et al. Predictive medicine for salivary gland tumours identification through deep learning. *IEEE J Biomed Health Inform*. (2021) 26:4869–79. doi: 10.1109/JBHI.2021.3120178
- Gunasekar DD, Bielak L, Hägele L, Oerther B, Benndorf M, Grosu AL, et al. Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology. *Radiat Oncol*. (2022) 17:65. doi: 10.1186/s13014-022-02035-0
- Amisha P, Malik MP, Rathaur VK. Overview of artificial intelligence in medicine. *J Family Med Prim Care*. (2019) 8:2328–31. doi: 10.4103/jfmpc.jfmpc_440_19
- González-Gonzalo C, Thee EF, Klaver CCW, Lee AY, Schlingemann RO, Tufail A, et al. Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. *Prog Retin Eye Res*. (2022) 90:101034. doi: 10.1016/j.preteyeres.2021.101034
- Hasani N, Morris MA, Rahmim A, Summers RM, Jones E, Siegel E, et al. Trustworthy artificial intelligence in medical imaging. *PET Clin*. (2022) 17:1–12. doi: 10.1016/j.cpet.2021.09.007
- Ribeiro M. T., Singh S., Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
- He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0
- O'Connor JP, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. (2017) 14:169–86. doi: 10.1038/nrclinonc.2016.162
- Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. (2021) 11:e048008. doi: 10.1136/bmjopen-2020-048008
- Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. *Phys Med*. (2021) 83:1–8. doi: 10.1016/j.ejmp.2021.02.011
- Zeineldin RA, Karar ME, Elshaer Z, Coburger J, Wirtz CR, Burgert O, et al. Explainability of deep neural networks for MRI analysis of brain tumors. *Int J Comput Assist Radiol Surg*. (2022) 17:1673–83. doi: 10.1007/s11548-022-02619-x
- Philbrick KA, Yoshida K, Inoue D, Akkus Z, Kline TL, Weston AD, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *AJR Am J Roentgenol*. (2018) 211:1184–93. doi: 10.2214/ajr.18.20331

16. Martí-Juan G, Frías M, García-Vidal A, Vidal-Jordana A, Alberich M, Calderon W, et al. Detection of lesions in the optic nerve with magnetic resonance imaging using a 3D convolutional neural network. *Neuroimage Clin.* (2022) 36:103187. doi: 10.1016/j.nicl.2022.103187
17. Zeiler M., Fergus R. Visualizing and understanding convolutional networks. arXiv:1311.2901 (2013). doi: 10.48550/arXiv.1311.2901
18. Chatterjee S, das A, Mandal C, Mukhopadhyay B, Vipinraj M, Shukla A, et al. TorchEsegeta: framework for interpretability and explainability of image-based deep learning models. *Appl Sci.* (2022) 12:2022. doi: 10.3390/app12041834
19. Springenberg J, Dosovitskiy A., Brox T., Riedmiller M. Striving for simplicity: the all convolutional net. arXiv:1412.6806 (2014). doi: 10.48550/arXiv.1412.6806
20. Wood DA, Kafiabadi S, Busaidi AA, Guilhem E, Montvila A, Lynch J, et al. Deep learning models for triaging hospital head MRI examinations. *Med Image Anal.* (2022) 78:102391. doi: 10.1016/j.media.2022.102391
21. Saleem H, Shahid AR, Raza B. Visual interpretability in 3D brain tumor segmentation network. *Comput Biol Med.* (2021) 133:104410. doi: 10.1016/j.compbio.2021.104410
22. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview In: W Samek, G Montavon, A Vedaldi, L Hansen and KR Müller, editors. *Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science.* Cham: Springer (2019). 193–209.
23. Golla AK, Tönnies C, Russ T, Bauer DF, Froelich MF, Diehl SJ, et al. Automated screening for abdominal aortic aneurysm in CT scans under clinical conditions using deep learning (2021) *Diagnostics*, 11:2131. doi: 10.3390/diagnostics11112131
24. Shi W, Tong L, Zhu Y, Wang MD. COVID-19 automatic diagnosis with radiographic imaging: explainable attention transfer deep neural networks. *IEEE J Biomed Health Inform.* (2021) 25:2376–87. doi: 10.1109/jbhi.2021.3074893
25. Karim MR, Jiao J, Dohmen T, Cochez M, Beyan O, Rebholz-Schuhmann D, et al. DeepKneeExplainer: explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging. *IEEE Access.* (2021) 9:39757–80. doi: 10.1109/ACCESS.2021.3062493
26. Lopatina A, Ropele S, Sibgatulin R, Reichenbach JR, Güllmar D. Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. *Front Neurosci.* (2020) 14:609468. doi: 10.3389/fnins.2020.609468
27. Shrikumar A., Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv:1704.02685 (2017). doi: 10.48550/arXiv.1804.02391
28. Gulum MA, Trombley CM, Kantardzic M. A review of explainable deep learning cancer detection models in medical imaging. *Appl Sci.* (2021) 11:2021–5. doi: 10.3390/app11104573
29. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. *J Imaging.* (2020) 6:52. doi: 10.3390/jimaging6060052
30. Wang C, Ma J, Shao J, Zhang S, Liu Z, Yu Y, et al. Predicting EGFR and PD-L1 status in NSCLC patients using multitask AI system based on CT images. *Front Immunol.* (2022) 13:813072. doi: 10.3389/fimmu.2022.813072
31. Kumar A, Manikandan R, Kose U, Gupta D, Satapathy SC. Doctor's dilemma: evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Trans Multimedia Comput Commun Appl.* (2021) 17:1–26. doi: 10.1145/3457187
32. Uyulan C, Erguzel TT, Turk O, Farhad S, Metin B, Tarhan N. A class activation map-based interpretable transfer learning model for automated detection of ADHD from fMRI data. *Clin EEG Neurosci.* (2022):15500594221122699. doi: 10.1177/15500594221122699
33. Wang CJ, Hamm CA, Savic LJ, Ferrante M, Schober T, Schlachter T, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. *Eur Radiol.* (2019) 29:3348–57. doi: 10.1007/s00330-019-06214-8
34. Akatsuka J, Yamamoto Y, Sekine T, Numata Y, Morikawa H, Tsutsumi K, et al. Illuminating clues of cancer buried in prostate MR image: deep learning and expert approaches. *Biomolecules.* (2019) 9:673. doi: 10.3390/biom9110673
35. Fuhrman JD, Gorre N, Hu Q, Li H, El Naqa I, Giger ML. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med Phys.* (2022) 49:1–14. doi: 10.1002/mp.15359
36. Alshazly H, Linse C, Barth E, Martinetz T. Explainable COVID-19 detection using chest CT scans and deep learning. *Sensors.* (2021) 21:455. doi: 10.3390/s21020455
37. Hao J, Xie J, Liu R, Hao H, Ma Y, Yan K, et al. Automatic sequence-based network for lung diseases detection in chest CT. *Front Oncol.* (2021) 11:781798. doi: 10.3389/fonc.2021.781798
38. Lahsaini I, El Habib Daho M, Chikh MA. Deep transfer learning based classification model for covid-19 using chest CT-scans. *Pattern Recognit Lett.* (2021) 152:122–8. doi: 10.1016/j.patrec.2021.08.035
39. Garg A, Salehi S, Rocca M, Garner R, Duncan D. Efficient and visualizable convolutional neural networks for COVID-19 classification using chest CT. *Expert Syst Appl.* (2022) 195:116540. doi: 10.1016/j.eswa.2022.116540
40. Ullah F, Moon J, Naeem H, Jabbar S. Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model. *J Supercomput.* (2022) 78:19246–71. doi: 10.1007/s11227-022-04631-z
41. Lu SY, Zhang Z, Zhang YD, Wang SH. CGENet: a deep graph model for COVID-19 detection based on chest CT. *Biology.* (2022) 11:2022–1. doi: 10.3390/biology11010033
42. Jadhav S, Deng G, Zawin M, Kaufman AE. COVID-view: diagnosis of COVID-19 using chest CT. *IEEE Trans Vis Comput Graph.* (2022) 28:227–37. doi: 10.1109/tvcg.2021.3114851
43. Nagaoka T, Kozuka T, Yamada T, Habe H, Nemoto M, Tada M, et al. A deep learning system to diagnose COVID-19 pneumonia using masked lung CT images to avoid AI-generated COVID-19 diagnoses that include data outside the lungs. *Adv Biomed Eng.* (2022) 11:76–86. doi: 10.14326/abe.11.76
44. Suri JS, Agarwal S, Chabert GL, Carriero A, Paschè A, Danna PSC, et al. COVLIAS 20-cXAI: cloud-based explainable deep learning system for COVID-19 lesion localization in computed tomography scans. *Diagnostics.* (2022) 12:1482. doi: 10.3390/diagnostics12061482
45. Pennisi M, Kavasidis I, Spampinato C, Schinina V, Palazzo S, Salanitri FP, et al. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artif Intell Med.* (2021) 118:102114. doi: 10.1016/j.artmed.2021.102114
46. Draelos RL, Carin L. Explainable multiple abnormality classification of chest CT volumes. *Artif Intell Med.* (2022) 132:2022. doi: 10.1016/j.artmed.2022.102372
47. Li CF, Xu YD, Ding XH, Zhao JJ, du RQ, Wu LZ, et al. MultiR-net: a novel joint learning network for COVID-19 segmentation and classification. *Comput Biol Med.* (2022) 144:105340. doi: 10.1016/j.compbio.2022.105340
48. Williamson BJ, Khandwala V, Wang D, Maloney T, Sucharew H, Horn P, et al. Automated grading of enlarged perivascular spaces in clinical imaging data of an acute stroke cohort using an interpretable, 3D deep learning framework. *Sci Rep.* (2022) 12:788. doi: 10.1038/s41598-021-04287-4
49. Kim KH, Koo HW, Lee BJ, Yoon SW, Sohn MJ. Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning. *J Korean Phys Soc.* (2021) 79:321–7. doi: 10.1007/s40042-021-00202-2
50. Singh A, Kwiecinski J, Miller RJH, Otaki Y, Kavanagh PB, van Kriekinge S, et al. Deep learning for explainable estimation of mortality risk from myocardial positron emission tomography images. *Circ Cardiovasc Imaging.* (2022) 15:e014526. doi: 10.1161/circimaging.122.014526
51. Jain V, Nankar O, Jerrish DJ, Gite S, Patil S, Kotecha K. A novel AI-based system for detection and severity prediction of dementia using MRI. *IEEE Access.* (2021) 9:154324–46. doi: 10.1109/ACCESS.2021.3127394
52. Hu M, Qian X, Liu S, Koh AJ, Sim K, Jiang X, et al. Structural and diffusion MRI based schizophrenia classification using 2D pretrained and 3D naive convolutional neural networks. *Schizophr Res.* (2022) 243:330–41. doi: 10.1016/j.schres.2021.06.011
53. Islam MN, Hasan M, Hossain MK, Alam MGR, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Sci Rep.* (2022) 12:11440. doi: 10.1038/s41598-022-15634-4
54. Zhang X, Han L, Zhu W, Sun L, Zhang D. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE J Biomed Health Inform.* (2021) 26:5289–97. doi: 10.1109/jbhi.2021.3066832
55. Esmaeili M, Vettukattil R, Banitalebi H, Krogh NR, Geitung JT. Explainable artificial intelligence for human-machine interaction in brain tumor localization. *J Pers Med.* (2021) 11:1213. doi: 10.3390/jpm11111213
56. Wang SH, Govindaraj V, Gorriz JM, Zhang X, Zhang YD. Explainable diagnosis of secondary pulmonary tuberculosis by graph rank-based average pooling neural network. *J Ambient Intell Humaniz Comput.* (2021). doi: 10.1007/s12652-021-02998-0
57. Windisch P, Weber P, Fürweger C, Ehret F, Kufeld M, Zwahlen D, et al. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. *Neuroradiology.* (2020) 62:1515–8. doi: 10.1007/s00234-020-02465-1
58. Zhang F, Pan B, Shao P, Liu P, Alzheimer's Disease Neuroimaging Initiative/Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing et al. A single model deep learning approach for Alzheimer's disease diagnosis. *Neuroscience.* (2022) 491:200–14. doi: 10.1016/j.neuroscience.2022.03.026
59. Chattopadhyay A., Sarkar A., Howlader P, Balasubramanian V. N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018), 839–847.
60. Termine A, Fabrizio C, Caltagirone C, Petrosini L. On Behalf Of The Frontotemporal Lobar Degeneration Neuroimaging. A reproducible deep-learning-based computer-aided diagnosis tool for frontotemporal dementia using MONAI and clinica frameworks. *Life.* (2022) 12:947. doi: 10.3390/life12070947
61. Lin QH, Niu YW, Sui J, Zhao WD, Zhuo C, Calhoun VD. SSPNet: an interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data. *Med Image Anal.* (2022) 79:102430. doi: 10.1016/j.media.2022.102430

62. Palatnik de Sousa I, Vellasco M, Costa da Silva E. Explainable artificial intelligence for bias detection in COVID CT-scan classifiers. *Sensors*. (2021) 21:5657. doi: 10.3390/s21165657
63. Garcia M, Dosenbach N, Kelly C In: C Kelly, editor. *BrainQCNet: a deep learning attention-based model for multi-scale detection of artifacts in brain structural MRI scans*. Dublin: Trinity College Institute of Neuroscience (2022)
64. Dasanayaka S, Shantha V, Silva S, Meedeniya D, Ambegoda T. Interpretable machine learning for brain tumour analysis using MRI and whole slide images. *Softw Imp*. (2022) 13:100340–8. doi: 10.1016/j.simpa.2022.100340
65. Shahamat H, Saniee Abadeh M. Brain MRI analysis using a deep learning based evolutionary approach. *Neural Netw*. (2020) 126:218–34. doi: 10.1016/j.neunet.2020.03.017
66. Petsiuk V, Das A., Saenko K. Rise: randomized input sampling for explanation of black-box models. arXiv:180607421 (2018). doi: 10.48550/arXiv.1806.07421
67. Gaur L, Bhandari M, Razdan T, Mallik S, Zhao Z. Explanation-driven deep learning model for prediction of brain tumour status using MRI image data. *Front Genet*. (2022) 13:822666. doi: 10.3389/fgene.2022.822666
68. Ahsan MM, Gupta KD, Islam MM, Sen S, Rahman ML, Hossain MS. COVID-19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities. *Mach. Learn. Knowl. Extr*. (2020) 2:490–504. doi: 10.3390/make2040027
69. Ahsan MM, Nazim R, Siddique Z, Huebner P. Detection of COVID-19 patients from CT scan and chest X-ray data using modified *MobileNetV2* and *LIME*. *Healthcare*. (2021) 9:1099. doi: 10.3390/healthcare9091099
70. Lundberg S., Lee S.-I. A unified approach to interpreting model predictions. arXiv:1705.07874 (2017). doi: 10.48550/arXiv.1705.07874
71. van der Velden BH, Janse MH, Ragusi MA, Loo CE, Gilhuijs KG. Volumetric breast density estimation on MRI using explainable deep learning regression. *Sci Rep*. (2020) 10:18095–9. doi: 10.1038/s41598-020-75167-6
72. Sanyal J, Banerjee I, Hahn L, Rubin D. An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence. *AMIA Jt Summits Transl Sci Proc*. (2020) 2020:552–60.
73. Roest C, Kwee TC, Saha A, Fütterer JJ, Yakar D, Huisman H. AI-assisted biparametric MRI surveillance of prostate cancer: feasibility study. *Eur Radiol*. (2022) 33:89–96. doi: 10.1007/s00330-022-09032-7
74. Gao KT, Tibrewala R, Hess M, Bharadwaj UU, Inamdar G, Link TM, et al. Automatic detection and voxel-wise mapping of lumbar spine Modic changes with deep learning. *JOR Spine*. (2022) 5:e1204. doi: 10.1002/jsp2.1204
75. Wang X, Jiang L, Li L, Xu M, Deng X, Dai L, et al. Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis. *IEEE Trans Med Imaging*. (2021) 40:2463–76. doi: 10.1109/tmi.2021.3079709
76. Joshi A, Sivaswamy J, Joshi GD. Lung nodule malignancy classification with weakly supervised explanation generation. *J Med Imaging*. (2021) 8:2021. doi: 10.1117/1.JMI.8.4.044502
77. Wang WL, Charkborty G. Automatic prognosis of lung cancer using heterogeneous deep learning models for nodule detection and eliciting its morphological features. *Appl Intell*. (2021) 51:2471–84. doi: 10.1007/s10489-020-01990-z
78. Yang Z, Hu Z, Ji H, Lafata K, Vaio E, Floyd S, et al. A neural ordinary differential equation model for visualizing deep neural network behaviors in multi-parametric MRI-based glioma segmentation. *Med Phys*. (2023):1–14. doi: 10.1002/mp.16286
79. Decuyper M, Maebe J, Van Holen R, Vandenberghe S. Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Physics*. (2021) 8:81. doi: 10.1186/s40658-021-00426-y
80. Angelov P, Soares E. Towards explainable deep neural networks (xDNN). *Neural Netw*. (2020) 130:185–94. doi: 10.1016/j.neunet.2020.07.010
81. Teodoro AAM, Silva DH, Saadi M, Okey OD, Rosa RL, Otaibi SA, et al. An analysis of image features extracted by CNNs to design classification models for COVID-19 and non-COVID-19. *J Sign Process Syst*. (2023) 95:101–13. doi: 10.1007/s11265-021-01714-7
82. Singh G, Yow KC. Object or background: an interpretable deep learning model for COVID-19 detection from CT-scan images. *Diagnostics*. (2021) 11:1732. doi: 10.3390/diagnostics11091732
83. Qian X, Fu H, Shi W, Chen T, Fu Y, Shan F, et al. M (3)lung-sys: a deep learning system for multi-class lung pneumonia screening from CT imaging. *IEEE J Biomed Health Inform*. (2020) 24:3539–50. doi: 10.1109/jbhi.2020.3030853
84. Singh G. Think positive: an interpretable neural network for image recognition. *Neural Netw*. (2022) 151:178–89. doi: 10.1016/j.neunet.2022.03.034
85. Kwabena Patrick M, Felix Adekoya A, Abra Mighty A, Edward BY. Capsule networks – a survey. *J King Saud Univ Comput Inf Sci*. (2022) 34:1295–310. doi: 10.1016/j.jksuci.2019.09.014
86. Afshar P, Naderkhani F, Oikonomou A, Rafiee MJ, Mohammadi A, Plataniotis KN. MIXCAPS: a capsule network-based mixture of experts for lung nodule malignancy prediction. *Pattern Recogn*. (2021) 116:107942–8. doi: 10.1016/j.patcog.2021.107942
87. Jetley S., Lord N. A., Lee N., Torr P. H. Learn to pay attention. arXiv:180402391 (2018). doi: 10.48550/arXiv.1804.02391
88. Jin D, Zhou B, Han Y, Ren J, Han T, Liu B, et al. Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. *Adv Sci*. (2020) 7:2000675. doi: 10.1002/advs.202000675
89. Wang X, Yuan Y, Guo D, Huang X, Cui Y, Xia M, et al. SSA-net: spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Med Image Anal*. (2022) 79:102459. doi: 10.1016/j.media.2022.102459
90. Li X, Zhao H, Ren T, Tian Y, Yan A, Li W. Inverted papilloma and nasal polyp classification using a deep convolutional network integrated with an attention mechanism. *Comput Biol Med*. (2022) 149:105976. doi: 10.1016/j.combiomed.2022.105976
91. Wang J, Yuan C, Han C, Wen Y, Lu H, Liu C, et al. IMAL-net: interpretable multi-task attention learning network for invasive lung adenocarcinoma screening in CT images. *Med Phys*. (2021) 48:7913–29. doi: 10.1002/mp.15293
92. Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest J, et al. CA-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans Med Imaging*. (2021) 40:699–711. doi: 10.1109/tmi.2020.3035253
93. Han Z, Wei B, Hong Y, Li T, Cong J, Zhu X, et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. *IEEE Trans Med Imaging*. (2020) 39:2584–94. doi: 10.1109/tmi.2020.2996256
94. Li M, Li X, Jiang Y, Zhang J, Luo H, Yin S. Explainable multi-instance and multi-task learning for COVID-19 diagnosis and lesion segmentation in CT images. *Knowl Based Syst*. 252:(2022):109278. doi: 10.1016/j.knsys.2022.109278
95. Zhao W, Chen W, Li G, Lei D, Yang J, Chen Y, et al. GMILT: a novel transformer network that can noninvasively predict EGFR mutation status. *IEEE Trans Neural Netw Learn Syst*. (2022). doi: 10.1109/tnnls.2022.3190671
96. Zokaeinikoo M, Kazemian P, Mitra P, Kumara S. AIDCOV: an interpretable artificial intelligence model for detection of COVID-19 from chest radiography images. *ACM Trans Manag Inf Syst*. (2021) 12:1–20. doi: 10.1145/3466690
97. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. (2020) 46:205–11. doi: 10.1136/medethics-2019-105586

Frontiers in Medicine

Translating medical research and innovation into
improved patient care

A multidisciplinary journal which advances our
medical knowledge. It supports the translation
of scientific advances into new therapies and
diagnostic tools that will improve patient care.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Medicine

