






Article

An Automated Method for Artificial Intelligence Assisted Diagnosis of Active Aortitis Using Radiomic Analysis of FDG PET-CT Images

Lisa M. Duff ^{1,2,*} , Andrew F. Scarsbrook ^{1,3} , Nishant Ravikumar ^{1,4}, Russell Frod ^{1,3} , Gijs D. van Praagh ⁵, Sarah L. Mackie ^{1,6}, Marc A. Bailey ^{1,7}, Jason M. Tarkin ⁸, Justin C. Mason ^{9,†}, Kornelis S. M. van der Geest ¹⁰, Riemer H. J. A. Slart ^{5,11} , Ann W. Morgan ^{1,6} and Charalampos Tsoumpas ^{1,5} 

¹ School of Medicine, University of Leeds, Leeds LS2 9JT, UK

² Institute of Medical and Biological Engineering, University of Leeds, Leeds LS2 9JT, UK

³ Department of Radiology, St. James University Hospital, Leeds LS9 7TF, UK

⁴ Center for Computational Imaging and Simulation Technologies in Biomedicine, University of Leeds, Leeds LS2 9JT, UK

⁵ Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands

⁶ NIHR Leeds Biomedical Research Centre and NIHR Leeds MedTech and In Vitro Diagnostics Co-Operative, Leeds Teaching Hospitals NHS Trust, Leeds LS7 4SA, UK

⁷ The Leeds Vascular Institute, Leeds General Infirmary, Leeds LS2 9NS, UK

⁸ Division of Cardiovascular Medicine, University of Cambridge, Cambridge CB2 0QQ, UK

⁹ National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK

¹⁰ Department of Rheumatology and Clinical Immunology, University of Groningen, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands

¹¹ Department of Biomedical Photonic Imaging, Faculty of Science and Technology, University of Twente, 7522 NB Enschede, The Netherlands

* Correspondence: l.duff@beatson.gla.ac.uk

† The co-author has passed away.



Citation: Duff, L.M.; Scarsbrook, A.F.; Ravikumar, N.; Frod, R.; van Praagh, G.D.; Mackie, S.L.; Bailey, M.A.; Tarkin, J.M.; Mason, J.C.; van der Geest, K.S.M.; et al. An

Automated Method for Artificial Intelligence Assisted Diagnosis of Active Aortitis Using Radiomic Analysis of FDG PET-CT Images. *Biomolecules* **2023**, *13*, 343. <https://doi.org/10.3390/biom13020343>

Academic Editors: Jorge Joven, Fernández-Arroyo Salvador, Anna Hernández-Aguilera and Nuria Canela

Received: 23 November 2022

Revised: 30 January 2023

Accepted: 1 February 2023

Published: 9 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The aim of this study was to develop and validate an automated pipeline that could assist the diagnosis of active aortitis using radiomic imaging biomarkers derived from [18F]-Fluorodeoxyglucose Positron Emission Tomography-Computed Tomography (FDG PET-CT) images. The aorta was automatically segmented by convolutional neural network (CNN) on FDG PET-CT of aortitis and control patients. The FDG PET-CT dataset was split into training (43 aortitis:21 control), test (12 aortitis:5 control) and validation (24 aortitis:14 control) cohorts. Radiomic features (RF), including SUV metrics, were extracted from the segmented data and harmonized. Three radiomic fingerprints were constructed: A—RFs with high diagnostic utility removing highly correlated RFs; B used principal component analysis (PCA); C—Random Forest intrinsic feature selection. The diagnostic utility was evaluated with accuracy and area under the receiver operating characteristic curve (AUC). Several RFs and Fingerprints had high AUC values (AUC > 0.8), confirmed by balanced accuracy, across training, test and external validation datasets. Good diagnostic performance achieved across several multi-centre datasets suggests that a radiomic pipeline can be generalizable. These findings could be used to build an automated clinical decision tool to facilitate objective and standardized assessment regardless of observer experience.

Keywords: aortitis; radiomics; machine learning; convolutional neural network; positron emission tomography/computed tomography

1. Introduction

Aortitis refers to inflammatory conditions affecting the aortic wall that cannot be explained by atherosclerosis alone [1–3]. It can be an isolated disorder or observed in association with several diseases, including giant cell arteritis (GCA) and Takayasu arteritis

(TAK) [3,4]. However, diagnosis of active aortitis presents challenges as symptoms and blood tests can be nonspecific, and treatment can result in severe side effects meaning informed decisions are required [3,5,6].

[¹⁸F]-Fluorodeoxyglucose Positron Emission Tomography—Computed Tomography (FDG PET-CT) is frequently used to assess patients with suspected aortitis related to large vessel vasculitis (LVV) as FDG avidity identifies areas of high glycolytic activity in the inflamed vessel wall [3,7–10]. The imaging is often qualitatively assessed based on consensus imaging guidelines [11,12]. Although grading is conducted by imaging specialists, the visual assessment can be subjective and inconsistent [11,13–15]. Some semi-quantitative parameters have been utilised but can be vulnerable to several factors and have limited information [16].

Radiomics is a data mining technique involving extraction of quantitative information from medical images referred to as radiomic features (RF) which may help better understand and stratify disease [15,17–19]. Radiomics may be a useful technique for aiding in the diagnosis of active aortitis, but the process needs to be automated to deal with a large quantity of data efficiently and facilitate routine clinical use. In particular, vascular segmentation, if conducted manually, can be very time consuming, and is not always reproducible [19,20]. Fully automated segmentation methods using deep learning (DL) convolutional neural networks (CNN) have become increasingly popular as they are both fast and reproducible [21–23].

In previous work, a methodological framework for assisting the diagnosis of active aortitis using radiomic analysis of FDG PET-CT was established [20]. This study utilised a small single center dataset to develop a radiomic method and provide a proof of concept that radiomic analysis could add value to the diagnosis of active aortitis.

In this study, the aim was to continue this work by developing, testing and validating an automated radiomic analysis pipeline to assist the diagnosis of active aortitis. This study progresses from the original publication in two ways: firstly, the method was automated by replacing manual segmentation with a CNN and exploring the effect this change had, and secondly the initial findings were validated with data from multiple centres to determine the generalizability and transferability of the method. The pipeline combines automated segmentation, radiomic analysis and machine learning (ML) with the aim of producing a reproducible and standardized method which could be applied to a clinical decision support tool in the future.

2. Materials and Methods

There are four key stages to a radiomic diagnostic model (Figure 1: image acquisition, image processing and segmentation, feature extraction, and classification). Each step has multiple sources of variation that can influence final results making the diagnostic model vulnerable to poor reproducibility [24,25]. To mitigate this, TRIPOD guidelines were used [26]. The protocols followed for each stage are described in the following sections.

2.1. Image Acquisition

2.1.1. Patient Selection

Figure 2 demonstrates the distribution of the imaging cohorts—training, test and validation. The data acquired from Leeds Teaching Hospitals NHS Trust were split into training and test (80:20) datasets. The training dataset was used to train ML models including optimization of hyper-parameters, and the test dataset was used to confirm initial findings were generalizable. The validation dataset acquired from external centres and used to determine if model performance was transferable to imaging acquired elsewhere.

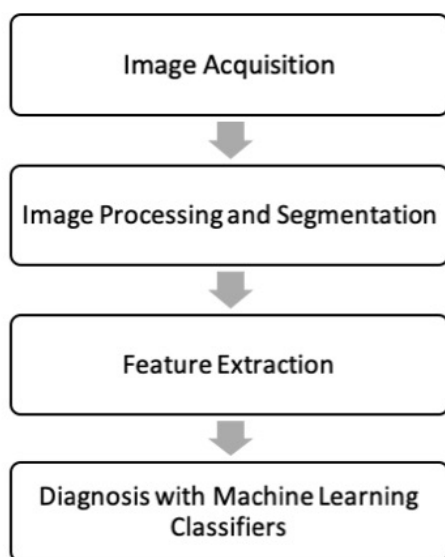


Figure 1. The steps of a radiomic diagnostic model.

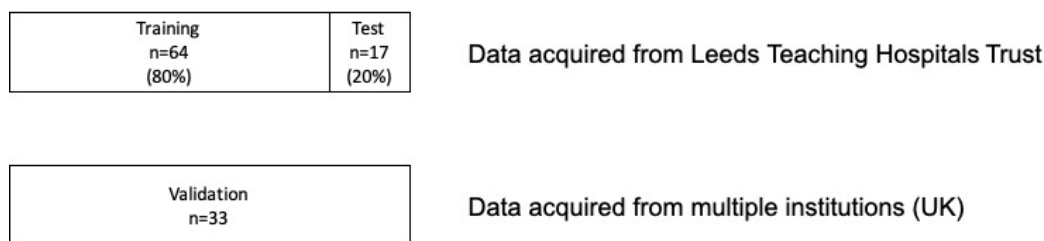


Figure 2. The distribution of datasets into training, test and validation cohorts.

Training and Testing Dataset

The training and test dataset was procured from Leeds Teaching Hospitals NHS Trust from imaging taken between January 2011 and December 2019. The collated data were then split into the training and test dataset (80:20). The inclusion and exclusion criteria for the training and test datasets are described as follows. Patients undergoing FDG PET-CT with a systemic inflammatory response (pyrexia of unknown origin, high acute phase response, weight loss) or suspected active aortitis were identified retrospectively. The ground truth diagnoses for all patients and controls were confirmed by a consultant rheumatologist with 17 years of experience of vasculitis (co-author AWM) based on clinical assessment, blood tests, biopsies and qualitative assessment of FDG PET-CT scans by a dual certified radiologist and nuclear medicine physician (co-author AFS) with more than 15 years of experience of reporting FDG PET-CT. Exclusion criteria included synchronous metabolically active conditions obscuring or interfering with the aorta, such as malignancy. Patients with known LVV were excluded if they did not have imaging evidence of active aortitis. Control patients were excluded if they had activity in the aorta related to atherosclerosis. For LVV patients who had undergone multiple FDG-PET scans, only the first scan that showed aortitis was selected. This study included a combination of newly-diagnosed patients and patients with relapse. The imaging data for the selected aortitis patients and controls were extracted from the institutional PACS (Picture Archiving and Communication System) and pseudoanonymised.

Validation Dataset

To evaluate multi-centre transferability, a validation dataset was formed using data from external institutions. The same inclusion and exclusion criteria were followed but was conducted at the centre of origin. Data from patients recruited to the UK GCA consortium (REC Ref. 05/Q1108/28) [27] with suspected aortitis, and who had FDG PET-CT scans performed as part of routine clinical care at Alliance Medical Ltd (AML) centres

in England, were extracted from the organizational PACS (IntelPACS Version 4, Intelrad Medical Systems). The AML centres included Addenbrookes Hospital, Freeman Hospital, Norfolk and Norwich PET CT Centre, Musgrove Park PET-CT Centre, Derriford Hospital, Bradford Royal Infirmary, Guildford Diagnostic Imaging, Sheffield PET-CT Centre, Poole Hospital and The Royal Liverpool University Hospital. The validation cohort was further supplemented by data from the PITA (PET Imaging of Giant Cell and Takayasu Arteritis) (REC approval: 19/EE/0043 Clinical trials registration: NCT04071691, PMID: 36697134) study in the University of Cambridge and Imperial College London.

2.1.2. Imaging Protocol

FDG PET-CT scans were acquired using a standard protocol [11,28,29]. Images were acquired from the upper thighs to the skull vertex in the supine position. Patients fasted for 6 hours before FDG injection, and scanning was conducted 1 hour after injection. Where possible, patients were not currently being treated with glucocorticoids (GC). Nine scanners from three different manufacturers were used (Table 1). Appendix A describes the acquisition parameters in further detail.

Table 1. Distribution of participants across scanners.

Scanner	Training		Test		Validation		Harmonization Batch
	Aortitis	Control	Aortitis	Control	Aortitis	Control	
Discovery 710	14	7	4	4	3	3	1
Gemini TF64	14	11	3	0	0	0	2
Discovery 690	15	3	5	1	9	2	3
Biograph 6 and Biograph 6 True Point	0	0	0	0	5	2	4
Biograph 64 mCT	0	0	0	0	1	2	5
Discovery MI DR	0	0	0	0	6	3	6
Discovery ST and STE	0	0	0	0	0	2	7

Discovery scanners from GE Healthcare—Chicago, IL, USA. Gemini Scanner from Philips Healthcare—Best, Netherlands. Biograph scanners from Siemens Healthineers—Erlangen, Germany.

The retrospectively gathered FDG PET-CT imaging was converted from DICOM to Nifti file format including converting the PET component to SUV using Simple ITK and PET DICOM (3D slicer extension from the University of Iowa (www.slicer.org/wiki/Documentation/Nightly/Modules/SUVFactorCalculator) (accessed on 1 September 2021)).

2.2. Image Processing and Segmentation

The segmentation method built into the overall pipeline was a CNN. A subset of the training and test patient dataset (aortitis $n = 50$, control $n = 25$) was manually segmented and given as input to the CNN in order to provide ground truth data to learn. Each FDG PET-CT scan of these patients was segmented manually using 3D slicer and the entire aorta was delineated (Version 4.10.2 (<https://www.slicer.org/>) (accessed on 1 April 2020)) [30,31]. The CT component was used as the main reference as it provides more anatomical information, but the result was checked against the PET scan. The Dice similarity coefficient (DSC) (Equation (1)) [32] was used to evaluate segmentation quality:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

The PET and CT components, and segmented masks were then resampled to a 4 mm isotropic voxel size to ensure uniform sampling across the entire cohort. Linear interpolation in Simple ITK was used for downsampling. This voxel size was selected as it was the lowest resolution of the three scanners in the training and test dataset meaning downsam-

pling alone was applied. A lower resolution was present in multi-centre data collected later (5.47 mm), but the 4 mm voxel size was maintained to ensure a valid comparison, and to keep an integer voxel size preventing rounding errors. The images and masks were also cropped to the same window size (144 × 144 pixels) as the CNN required the same slice sizes. Data were manually checked to ensure that the aorta was central and unaffected by the crop.

A CNN with U-Net architecture was built for automated segmentation (Tensorflow Version 2.4.1). The full architecture is shown in Figure 3. Training was undertaken on ARC4, and part of the high-performance computing facilities at the University of Leeds, UK. On ARC4, a single NVIDIA V100 GPU (graphics processing unit) was used. In total, training and then segmentation of all data took 11:51:20 (HH:MM:SS). The average segmentation time per patient was 1 min 12 s compared to an average of 30 min per patient for manual segmentation.

The manually segmented dataset was split into training and testing cohorts for the development of the CNN(70:30), and each CT image was read in slice by slice with its corresponding labelled slice as the input layer. The performance of the CNN was measured using the DSC. The batch size was set to 32 slices. The number of epochs was set to 100 with early stopping if the loss function (DSC loss) did not improve, which led to training stopping at 41 epochs. The activation function was a leaky rectified linear unit (ReLU). Convolution stride was 1, and pooling stride was 2. Kernel size was 3 × 3 for convolution and 2 × 2 for pooling. Once trained, the entire patient dataset was provided as input, and the predicted segmentations were output. Small 'islands' were found in the predicted segmentations. These were clusters of pixels in the background of the scan that were several orders of magnitude smaller than the aorta. These were removed by creating new segmentations that only retained the largest cluster of pixels in the slice using Python packages Numpy (Version 1.18.1) and Simple ITK (Version 2.01). The segmented slices were then reassembled into 3D volumes for use in feature extraction (Section 2.3.2).

2.3. Feature Extraction

2.3.1. Qualitative Grading of Vessel Wall FDG Activity

All scans were evaluated based on EANM/SNMMI guidelines [11] and assigned a vascular uptake score by an experienced radiologist (supervisor AFS):

- 0: no uptake (less than mediastinum)
- 1: low-grade uptake (less than liver)
- 2: intermediate-grade uptake (equal to liver), (possible aortitis)
- 3: high-grade uptake (greater than liver), (positive active aortitis)

2.3.2. Feature Extraction

Radiomic features encompass a large number of quantitative parameters. These features range from simple, such as SUV (standardised uptake value) metrics (Equation (2)), to more complex descriptors of the shape and spatial relationships between individual voxels of imaging data [33,34].

SUV metrics are part of the larger group of radiomic features but will be referred to independently when studied separately. As they are more established in clinical use, their diagnostic utility was explored alone and as part of the larger group. The SUV of a region of interest (ROI) can be averaged (SUV_{mean}) or the maximum determined (SUV_{max}). However, SUV measurements can be influenced by several factors such as the size of the volume of interest, image noise, concentration of glucose in plasma and body habitus [16]. SUV metrics were used instead of target-to-blood pool ratio as liver is a more common reference point as discussed in Section 2.3.1:

$$SUV = \frac{\text{radioactivity concentration}}{\text{injection dose (MBq) / patient's weight (kg)}} \quad (2)$$

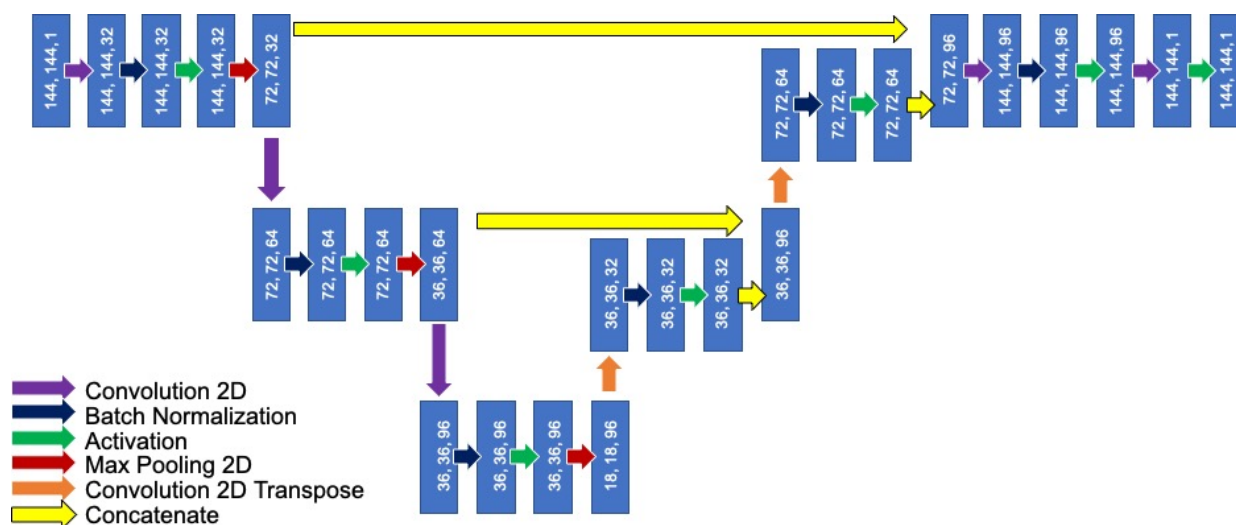


Figure 3. Architecture of convolutional neural network (CNN) used to segment the aorta.

Radiomic features ($n = 102$) were extracted with Pyradiomics (Version 3.0.1, radiomics.io/pyradiomics). A further five SUV metrics (SUV_x) were calculated separately using Numpy (Version 1.18.1) and Simple ITK (Version 2.01) and added to the radiomic features dataset. Each SUV metric was calculated as follows:

- SUV 90th Percentile—90% of the voxel's SUV value fall below this number;
- SUV mean—the mean SUV value in the region of interest;
- SUV maximum—the maximum SUV value in the region of interest;
- SUV_x ($x = 50, 60, 70, 80, 90$)—mean of the voxels that are equal or greater than $x\%$ of SUV maximum.

In both cases, the radiomic features were extracted from the entire segmented 3D volume of the aorta in the PET image [35]. In most cases, Pyradiomics is broadly compliant with the IBSI standards but deviates in some cases as described in their documentation (<https://pyradiomics.readthedocs.io/en/latest/faq.html> (accessed on 1 November 2022)). This will affect some of the extracted features where they rely on gray value discretization. Features were calculated with a SUV bin width of 0.075. This bin width was determined by dividing the maximum SUV value in the segmented areas across the whole dataset by 64—a commonly used bin number in radiomics. No filters were applied through Pyradiomics, and all other parameters were left as default.

A complete list of all radiomic features and SUV features ($n = 107$) extracted is provided in Table A2.

2.4. Diagnosis with Machine Learning Classifiers

2.4.1. Diagnostic Utility of Individual SUV Metrics and Radiomic Features

The diagnostic utility, also referred to as diagnostic performance, of the following methods was measured with AUC primarily, along with balanced accuracy as confirmation. Balanced accuracy was used as it adjusts for imbalanced datasets and allowed for comparison between our training, test and validation datasets. The AUC of the validation dataset was prioritised as it demonstrated both generalizability to other datasets and transferability to other institutions which is vital for clinical use [23]. As the benchmark AUCs for qualitative assessment of PET-CT in suspected aortitis quoted in the literature are 0.81–0.98 [11], any AUC value greater than 0.8 was considered a good performance. Where possible, methods with any balanced accuracy across the three cohorts $\leq 50\%$ were discounted. Cases where AUC was high but accuracy was low occur due to a bias towards the positive diagnosis.

The diagnostic utility of all radiomic features and SUV metrics were first evaluated individually using logistic regression classifiers (Sci-kit Learn Version 0.23.2). While SUV

metrics can be included as radiomic features (Section 2.3.2), they were separated and compared to all remaining radiomic features at this stage to determine if the newer radiomic features added value. To train the logistic regression classifiers, the hyper-parameters for each feature were tuned using the Sci-kit Optimise function BayesSearchCV using the training cohort with stratified 5-fold cross validation meaning the ratio of patients to controls in each fold was equal to the ratio in the total cohort. The hyperparameter optimization method was changed to BayesSearchCV from GridSearchCV from the previous study as it more thoroughly searches the parameter options [20]. The final diagnostic model for each individual feature was then trained with the best hyper-parameters on the training cohort with stratified 5-fold cross validation. The trained model was then applied to the test and validation dataset.

2.4.2. Forming Radiomic Fingerprints

Individually radiomic features can be used as metrics, but, when used collectively, they can provide complimentary information to improve diagnostic performance [36]. Using all or most extracted radiomic features can introduce a significant amount of redundant information and creates noise in the diagnostic model [37]. Therefore, radiomic fingerprints were created with the extracted radiomic features. Three radiomic fingerprints were built using the methods described below.

Fingerprint A was produced by selecting features with high individual diagnostic utility based on their training dataset performance in Section 2.4.1 : $AUC \geq 0.5$, balanced accuracy ≥ 0.5 . Features were filtered using Python package Pandas (Version 1.1.4). Highly correlated features were then removed. For every combination of feature pairs, if the correlation coefficient was >0.9 , the feature with the lower AUC was removed.

Fingerprint B was formed using principal component analysis (PCA). PCA represents a large set of variables as a smaller set of principal components by finding relationships between features and combining them to reduce redundancy and minimize loss of information. PCA was applied using Sci-kit Learn (Version 0.23.2). The fingerprint was formed with principal components needed to account for at least 90% of variance in the radiomic data.

Fingerprint C used the Sci-kit Learn (Version 0.23.2) random forest ML classifier. The classifier has intrinsic feature selection so all 107 extracted features were provided as input, and the classifier will select the features that produce the best performance.

2.4.3. Diagnostic Utility of Fingerprints

The diagnostic utility of radiomic fingerprints A and B was evaluated using the same methodology described in Section 2.4.1, but additional ML classifiers were tested alongside logistic regression [38–40]. Ten different ML classifiers were built, trained and tested (Sci-kit Learn Version 0.23.2): support vector machine, random forest, passive aggressive, logistic regression, k nearest neighbours, perceptron, multi-layered perceptron, decision tree, stochastic gradient descent and gaussian process classification. Stochastic gradient descent refers to the SGDClassifier within Sci-kit learn that uses stochastic gradient descent optimization on several different linear classifiers. The specific linear classifier is determined as part of hyper-parameter tuning.

Fingerprint C was evaluated as in Section 2.4.1 using only the Random Forest Classifier as it uses the embedded feature selection in this ML classifier.

2.4.4. Statistical Analysis

AUC was used as the key metric for determining diagnostic utility of the tested diagnostic models and ranking accordingly. Balanced accuracy used a confirming metric. The confidence intervals were determined using Delong's test [41]. The final comparison of models was conducted using the *p*-value derived from Delong's Test.

2.5. The Influence of Variation in Method

2.5.1. Harmonization

The effect of the ComBat Harmonization method (neuroCombat, Version 0.2.7) was explored. It is a technique used to reduce the effect of different imaging protocols on radiomic features [42–44]. These factors cannot be standardized retrospectively without reducing the size of the dataset, so harmonization is recommended to minimize the effect [44]. The overall dataset (training, test and validation combined) was grouped in batches as shown in Table 1 based on similar imaging protocol parameters. The effect of harmonization was explored in the previous study [20]. No significant improvement was achieved but as there is more variation in data in this study the comparison was repeated. The key results with and without harmonization were compared.

2.5.2. Segmentation

A CNN was utilised as the segmentation method in this study to automate the radiomic workflow. A subset of patients (50 aortitis and 25 controls) had both manual and automatic segmentations. The above methods were repeated on these patients using both segmentations separately in order to compare the effect the segmentation method had on performance. There were insufficient numbers to have a test and validation cohort so only training values were compared.

2.5.3. Imaging Sources

This study used multi-centre data from the UK as a validation cohort. This cohort was varied but generally followed the same diagnosis pathways set out by the National Health Service (NHS), standard of care imaging protocol, and were imaged in the same time period. Later in the study timeline, another dataset was collated from UMCG, Groningen, The Netherlands consisting of 40 GCA patients and 20 controls. Different imaging protocols were followed, and the images were acquired on scanners with higher image resolution—a range of Siemens Biograph scanners using the reconstruction methods PSF + TOF 3i21s most often, and PSF + TOF 4i5s and OSEM3D 3i24s occasionally. These variations are known to highly influence radiomic results [45].

The robustness of our method to highly varied data was tested with and without harmonization. The images were preprocessed and segmented in the same way as the previous cohorts before being included in the validation cohort and the diagnostic utility retested.

3. Results

3.1. Image Acquisition—Patient Characteristics

Overall, 114 participants were included in the training, test and validation datasets collectively (Table 2). The age of the patients and female predominance reflects the typical demographic of patients with LVV, the most common cause of which is GCA. The sensitivity of FDG PET-CT is significantly reduced within a few days of starting GC treatment, so GC (prednisolone) doses were zero at the time of scanning unless stated otherwise [46]. CRP and ESR are laboratory markers of inflammation used in clinical care.

Less clinical data were available for the validation dataset, but, as shown in Table 2, the gender distribution, LVV Type, prednisolone dose, CRP, ESR, blood glucose and median age of all datasets are similar.

Table 2. A description of patient demographics across the three datasets Key—Large vessel vasculitis (LVV), giant cell arteritis (GCA), Takayasu’s arteritis (TAK), Not applicable (n/a), CRP (C-reactive protein), ESR (erythrocyte sedimentation rate).

	Training		Test		Validation	
	Aortitis	Controls	Aortitis	Controls	Aortitis	Controls
Number of Participants	43	21	12	5	19	14
Age at time of scan, years -median (range)	67 (23–85)	67 (41–84)	70 (58–76)	60.5 (49–70)	67 (55–85)	68 (50–79)
Sex (male/female)	11/32	11/10	4/8	2/3	4/15	5/9
LVV type	40 GCA 3 TAK	n/a	12 GCA	n/a	17 GCA 2 TAK	n/a
Prednisolone dose at time of scan, mg -median (range)	0 (0–40)	0 (0–30)	0 (0–40)	0 (0–60)	0 (0–40)	3.5 (0–40)
CRP (mg/L) -median (range)	41 (5–165), not done (n = 8)	n/a	39 (11–149), not done (n = 3)	n/a	36 (10–112), not known (n = 15)	n/a
ESR (mm/Hr) -median (range)	71 (3–143), not done (n = 29)	n/a	37 (n = 1), not done (n = 11)	n/a	90 (12–120), not known (n = 15)	n/a
Blood Glucose (mmol/L) -median (range)	5.5 (4.2–9.9), not known (n = 11)	5.9 (4.6–12), not known (n = 13)	5.8 (5–7.3), not known (n = 3)	5.9 (5.1–7.4), not known (n = 2)	5.8 (4.4–7.5), not known (n = 7)	6.65 (5.4–9.5), not known (n = 2)

3.2. Segmentation

The manually segmented data had a mean DSC of 0.91 when a sample was compared to segmentations conducted by a second observer. The CNN achieved a mean DSC of 0.66 (median 0.72) before small ‘islands’ were removed and 0.71 (median = 0.80) after when compared to the original manual segmentations used for training. The time taken to segment the aorta automatically per patient was 1 min 12 s.

An example of a CNN segmentation is shown in Figure 4.



Figure 4. Segmentation from CNN with small ‘islands’ removed.

3.3. Qualitative Grading of Vessel Wall FDG Activity

Recent guidelines advocate qualitative grading of PET-CT scans based on FDG activity in the aortic wall relative to the liver [11]. Table 3 shows the grades assigned to the training, test and validation cohorts respectively by an experienced radiologist on retrospective review of the images.

Table 3. Grading of patient dataset based on EANM/SNMMI guidelines [11]

Grade	Training		Test		Validation	
	Aortitis	Control	Aortitis	Control	Aortitis	Control
0	0	21	0	5	0	11
1	1	0	0	0	0	3
2	2	0	0	0	2	0
3	40	0	12	0	17	0
Ground Truth	Grade 3 n = 43	Grade 0 n = 21	Grade 3 n = 12	Grade 0 n = 5	Grade 3 n = 19	Grade 0 n = 14

3.4. Diagnostic Utility of Individual SUV Metrics and Radiomic Features

Figure 5 demonstrates the performance of SUV metrics in a logistic regression classifier where higher accuracy and AUC indicate good diagnostic utility. In general, SUV metrics performed poorly when accuracy was considered. The SUV 90th percentile performed better consistently across all three cohorts with a validation AUC of 0.85 and a balanced accuracy of 71%.

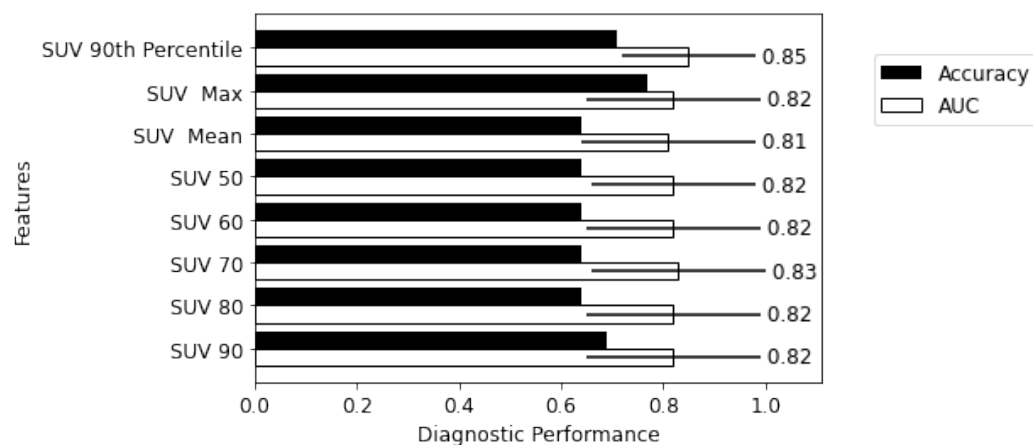


Figure 5. Diagnostic utility of individual SUV metrics.

The five-best performing radiomic features, when used individually in a logistic regression classifier, are shown in Figure 6. Performance was based on validation AUC but a minimum balanced accuracy of 50% had to be met across the training, testing and validation cohorts. In some cases, a radiomic feature would perform well in either testing or validation AUC but had poor accuracy. The radiomic features given in Figure 6 suggests that heterogeneity is an important characteristic in distinguishing aortitis from controls.

The performance of all individual radiomic features and SUV metrics in logistic regression classifiers, and in all three cohorts, are listed in Table A2.

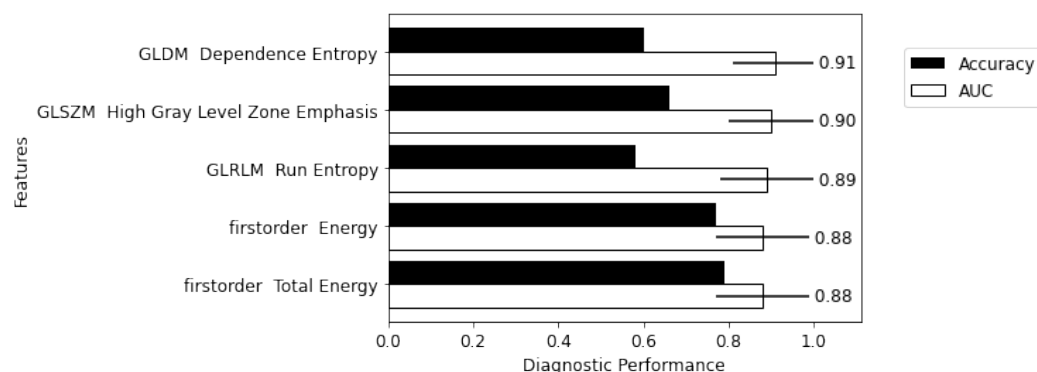


Figure 6. Diagnostic utility of the five highest performing individual radiomic features—performance ranked by validation AUC with a balanced accuracy above 50%.

3.5. Diagnostic Utility of Fingerprints

Fingerprint A was based on minimum thresholds of diagnostic performance for each feature and a maximum correlation to other features. Random Forest performed consistently across the training, testing and validation cohorts in both AUC and balanced accuracy (Figure 7), suggesting that this method may have multi-centre transferability. The performance of all explored ML classifiers is shown in Appendix Table A3.

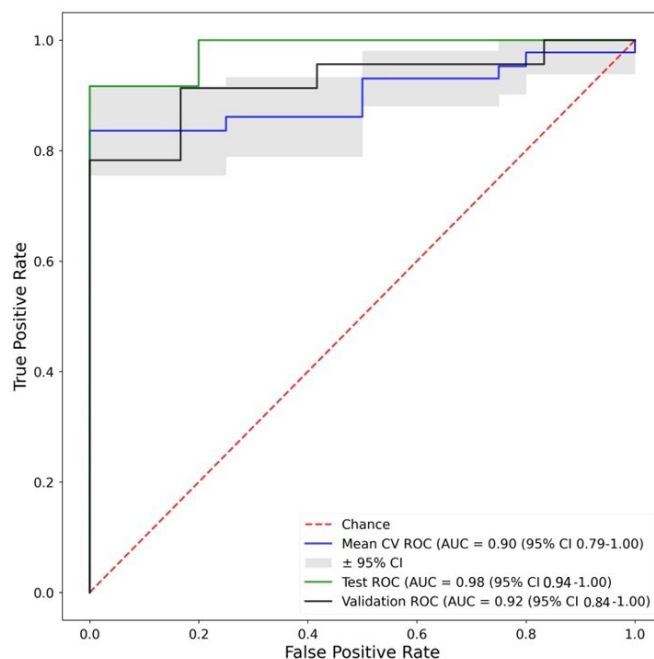


Figure 7. ROC curves of the best performing (by validation AUC and minimum accuracies) machine learning classifier trained on Fingerprint A—Random Forest. Corresponding Accuracies—Training: 0.74, Test: 0.8, Validation: 0.75. Key: Mean CV ROC—Mean cross validation ROC from training dataset, Test ROC—ROC from test dataset, Validation ROC—ROC from validation dataset.

Fingerprint B was based on PCA. For this fingerprint, the best validation AUC was achieved by a neural network classifier with a validation AUC = 0.90 (Figure 8). The performance of all explored ML classifiers is shown in Appendix Table A4. Appendix Table A4 also shows that, despite the neural network being the best performing classifier by validation AUC, the ranking metric used in this study, several other classifiers performed similarly and more consistently across datasets and with smaller confidence intervals.

Fingerprint C used the feature selection that is intrinsically part of Random Forest classification and did not include any other ML classifiers. This method produced good results

in both the testing and validation cohorts, demonstrating that Fingerprint C is a promising method for the diagnosis of aortitis. Figure 9 displays the ROC curves for Fingerprint C. As this method only used Random Forest, it was not tested in all ML classifiers.

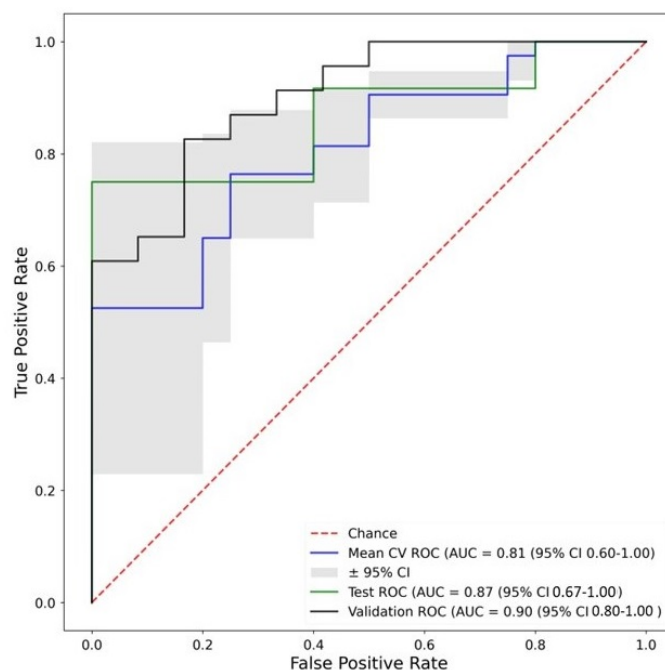


Figure 8. ROC curves of the best performing (by validation AUC) machine learning classifier trained on Fingerprint B—Neural Network. Corresponding Accuracies—Training: 0.78, Test: 0.68, Validation: 0.81. Key: Mean CV ROC—Mean cross validation ROC from training dataset, Test ROC—ROC from test dataset, Validation ROC—ROC from validation dataset.

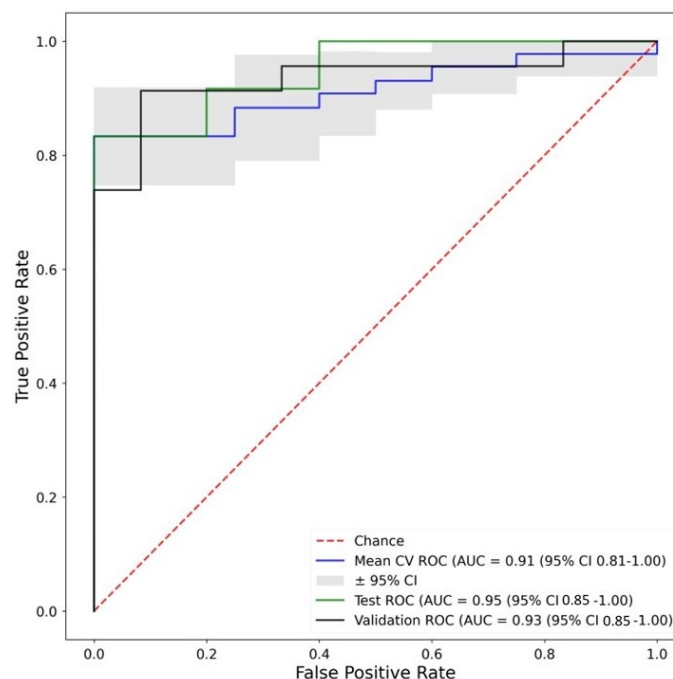


Figure 9. ROC curves of the Random Forest classifier in Fingerprint C. Corresponding Accuracies—Training: 0.79, Test: 0.8, Validation: 0.73. Key: Mean CV ROC—Mean cross validation ROC from training dataset, Test ROC—ROC from test dataset, Validation ROC—ROC from validation dataset.

3.6. Comparison of Selected Features

Table 4 shows the top 10 features (by Validation AUC and feature importance metric respectively) selected in Fingerprint A and Fingerprint C. As Fingerprint B used PCA and produces new components, it is not simple to directly compare them. Table 4 demonstrates that heterogeneity is important in distinguishing aortitis from controls. This is confirmed by earlier results in Section 3.4.

Table 4. Features selected for Fingerprint A and C. Key—SUV (standardized uptake value), GLDM (Gray-Level Dependence Matrix), GLCM (Gray-Level Co-Occurrence Matrix), GLRLM (Gray-Level Run Length Matrix), and GLSZM (Gray-Level Size Zone Matrix).

Top Ten Features Selected in:	
Fingerprint A	Fingerprint C
GLDM Small Dependence High Gray Level Emphasis	GLRLM Long Run Low Gray Level Emphasis
GLSZM Size Zone Non-Uniformity Normalized	GLSZM High Gray Level Zone Emphasis
GLRLM Gray Level Variance	GLDM Dependence Entropy
GLDM Large Dependence Low Gray Level Emphasis	GLDM Small Dependence High Gray Level Emphasis
GLRLM Long Run Emphasis	GLCM Autocorrelation
GLSZM Gray Level Variance	GLRLM Short Run Emphasis e
First Order Total Energy	GLDM Dependence Non-Uniformity Normalized
GLSZM Large Area Emphasis	First Order Entropy
GLSZM Size Zone Non-Uniformity	GLDM Gray Level Variance
First Order 10-Percentile	GLDM Large Dependence Emphasis

3.7. Summary of Key Results

Table 5 summarizes the best results from each of the explored methods for diagnosis of aortitis. The best result was determined as described in each of the previous sections, but, in all cases, validation AUC was used to initially rank the results and where possible results with a balanced accuracy $\leq 50\%$ were removed. While the displayed results ranked the best in each method by the given criteria, none were significantly better than each other with respect to AUC ($p > 0.05$, DeLong's Algorithm [41,47]).

Table 5. A summary of the diagnostic utility of each explored method.

Qualitative Assessment	Literature AUC 0.81–0.98 [11]					
	Training Accuracy	Training AUC (95% CI)	Testing Accuracy	Testing AUC (95% CI)	Validation Accuracy	Validation AUC (95% CI)
SUV Feature -SUV 90th Percentile	0.77	0.91 (0.73–1.00)	0.7	0.93 (0.79–1.00)	0.71	0.85 (0.72–0.99)
Radiomic Feature -GLDM Dependence Entropy	0.55	0.80 (0.61–1.00)	0.7	0.92 (0.74–1.00)	0.60	0.91 (0.82–1.00)
Fingerprint A -Random Forest	0.74	0.90 (0.79–1.00)	0.8	0.98 (0.94–1.00)	0.75	0.92 (0.84–1.00)
Fingerprint B -Neural Net	0.66	0.81 (0.60–1.00)	0.68	0.87 (0.67–1.00)	0.81	0.90 (0.80–1.00)
Fingerprint C -Random Forest	0.79	0.91 (0.81–1.00)	0.8	0.95 (0.85–1.00)	0.73	0.93 (0.85–1.00)

3.8. Influence of Variations in Method

The results given in Table 6 are from the same diagnostic models evaluated in Table 5 but with harmonized data as input instead. This demonstrates that feature harmonization has little influence on diagnostic utility in this scenario.

Table 6. A summary of the diagnostic utility of each explored method when feature harmonization is applied.

Qualitative Assessment	Literature AUC 0.81–0.98 [11]					
	Training Accuracy	Training AUC (95% CI)	Testing Accuracy	Testing AUC (95% CI)	Validation Accuracy	Validation AUC (95% CI)
SUV Feature -SUV 90th Percentile	0.69	0.86 (0.66–1.00)	0.7	0.93 (0.79–1.00)	0.67	0.83 (0.68–0.99)
Radiomic Feature—GLDM Small Dependence High Gray Level Emphasis	0.66	0.85 (0.73–0.97)	0.8	0.98 (0.94–1.00)	0.77	0.82 (0.67–0.96)
Fingerprint A—Logistic Regression	0.76	0.86 (0.69–1.00)	0.72	0.90 (0.75–1.00)	0.79	0.93 (0.86–1.00)
Fingerprint B—Neural Net	0.64	0.74 (0.57–0.91)	0.72	0.90 (0.75–1.00)	0.77	0.90 (0.80–1.00)
Fingerprint C—Random Forest	0.81	0.88 (0.72–1.00)	0.62	0.88 (0.71–1.00)	0.7	0.89 (0.79–1.00)

The results given in Table 7 compare the training AUC values for manual and CNN segmentation. These results show comparable performance for both segmentation methods.

Table 7. A comparison of key results from different segmentation methods.

	Manual Segmentation Training AUC Mean (95% CI)	Automated Segmentation Training AUC Mean (95% CI)
SUV Feature—SUV 90th Percentile	0.85 (0.77–0.93)	0.86 (0.81–0.91)
Radiomic Feature—GLSZM High Gray Level Zone Emphasis/GLCM Difference Variance	0.91 (0.84–0.98)	0.89 (0.87–0.91)
Fingerprint A—Random Forest	0.91 (0.80–1.0)	0.85 (0.81–0.89)
Fingerprint B—Random Forest/Support Vector Machine	0.88 (0.81–0.95)	0.91 (0.84–0.98)
Fingerprint C—Random Forest	0.86 (0.78–0.94)	0.81 (0.74–0.89)

The results given in Table 8 demonstrate which methods are affected by using a heterogeneous data set. The heterogeneity mostly involved larger variations in image acquisition protocol. Only validation data are shown as the new dataset was added to the validation cohort. CNN segmentations on this data achieved a median DSC of 0.71 when compared to manual segmentations. This demonstrates that standardization of imaging acquisition is required for most of the explored radiomic based diagnosis methods. Fingerprint A demonstrates the most robustness, albeit a small decrease in diagnostic utility.

Table 8. The diagnostic utility of the explored methods when data with an altered imaging acquisition protocol is added.

	Non-Harmonized		Harmonized	
	Validation Accuracy	Validation AUC (95% CI)	Validation Accuracy	Validation AUC (95% CI)
SUV Feature—SUV Mean	0.6	0.72 (0.62–0.83)	0.59	0.72 (0.61–0.82)
Radiomic Feature—First Order Energy/GLDM Dependence Entropy	0.63	0.72 (0.61–0.82)	0.58	0.83 (0.75–0.91)
Fingerprint A—Random Forest/K Nearest Neighbours	0.71	0.80 (0.71–0.89)	0.69	0.72 (0.61–0.82)
Fingerprint B—Perceptron	0.7	0.72 (0.61–0.82)	0.7	0.70 (0.59–0.81)
Fingerprint C—Random Forest	0.48	0.61 (0.50–0.72)	0.6	0.68 (0.57–0.78)

4. Discussion

This study presents an automated pipeline to assist diagnosis of active aortitis using CNN segmentation, radiomic analysis, and ML classifiers. Key results are summarised in Table 5. The different diagnostic models performed well and similarly to both each other and the standard of care qualitative assessment. Using radiomic fingerprints had the advantage of reducing the size of the confidence intervals. Fingerprint A demonstrated the most robustness.

4.1. Segmentation Automation

The main component in automation was aortic segmentation using a CNN which achieved a median DSC of 0.80 and allowed the diagnostic models to achieve a good performance. A good diagnostic performance or utility was defined as a Validation AUC ≥ 0.8 , and is therefore similar to the benchmark AUCs for qualitative assessment of PET-CT in suspected aortitis [11], and a (balanced) accuracy > 0.5 in all three cohorts. When compared to the performance using manual segmentations (Table 7), comparable results were achieved. Automating the method reduces the likelihood of inter and intra observation, increasing reproducibility [48]. It also makes routine clinical adoption a more realistic proposition.

4.2. Multi-Centre Transferability

Sollini et al. concluded in their systematic review that the lack of external validation was the key issue preventing radiomics translating into routine clinical practice [49]. Some multi-centre diagnostic performance was shown in the proposed methods. SUV metrics performed well in training cohorts but did not demonstrate good transferability to the testing or validation cohort from other institutions. SUV 90th Percentile demonstrated the most diagnostic utility from all the explored SUV metrics and did so in all three cohorts. In future work, it may be worth investigating the effect of adjusting for lean body mass rather than body weight, as is the case for SUV metrics, as the results from van Praagh et al. suggest that this could be more reliable [16]. Several individual radiomic features produced high AUC values and met the minimum accuracy values. In particular, features based on heterogeneity performed well across all three cohorts with the highest validation AUC coming from GLDM Dependence Entropy (AUC = 0.91) and first order Energy achieving the best compromise between AUC and accuracy. The features selected in Fingerprint A and C further demonstrate that heterogeneity is important in distinguishing aortitis (Table 4).

Fingerprint A was formed with high performing individual features with highly correlated features removed. It performed well with Random Forest in all three cohorts achieving AUC and accuracy values above the minimum thresholds stated earlier. This fingerprint also displayed the most robustness to using a heterogeneous imaging protocol in Table 8. This demonstrates good generalizability and transferability which are important for clinical use [49]. Fingerprint C (Random Forest—all features) performed similarly but was detrimentally affected by the changes in imaging protocol. This suggests it is beneficial to still pre-select features before using random forest.

4.3. Limitations

It would be preferable to only analyse the aortic wall, but a segmentation method which reliably distinguishes wall from the lumen in non-contrast enhanced CT has not been developed to the extent that it reliably worked on the patient cohort in this study. A segmentation method that achieves this with thresholding can be applied to PET scans with high aortic wall activity, but it would not identify non-inflamed aortic wall in the control dataset so this segmentation method would not be feasible for diagnostic purposes. Future studies that used only aortitis cases with radiomic analysis might be able to overcome this issue using thresholding: for example, prognostic/stratification studies. Location of inflammation in the aortic wall could also be considered for differentiating causes of aortitis as this varies. These were beyond the scope of the present work. Most other aortic wall segmentation methods require contrast enhancement. One method that does not use

contrast enhancement was developed by Piri et al. [50]. They utilised a CNN to produce whole aorta segmentations and then labelled the wall as a predefined thickness inside and outside of the edge, giving a 5 mm wall thickness in total. This gives a good approximation of the aortic wall but is not precise, does not account for inter and intra patient variation, and is not flexible to deal with anatomical alterations caused by aortitis. While a 5 mm thickness is achievable on CT, it would result in a one pixel thick wall on PET used in this study (voxel size $4 \times 4 \times 4$ mm) making any second or higher order features limited. Future work could explore whether the trade off between limiting potential features or including the lumen is worthwhile.

Another limitation is that patients with atherosclerosis were excluded from the study cohort, for both aortitis patients and controls, as atherosclerosis can also lead to FDG uptake in the vessel wall [51–53]. This was part of the exclusion criteria as the purpose of the study was to initially develop an artificial intelligence-based pipeline using unequivocal cases and controls. The next stage would be to test this pipeline on the whole spectrum of those presenting with suspected aortitis; atherosclerosis is common in this age group [12]. Another group has reported promising results using SUV metrics [54]. Similarly, it would be of interest to determine whether any of the radiomic features with high diagnostic utility can detect aortitis after treatment has started as this currently limits the use of PET imaging. FDG PET-CT is mostly used for baseline imaging of aortitis for diagnosis as GCs reduce its sensitivity [46,55,56]. While the diagnostic accuracy decreases significantly after 10 days, uptake is not eradicated completely. Van der Geest et al. determined that FDG PET still had some moderate diagnostic utility for monitoring treatment but that the individual results were highly variable and any conclusions drawn from imaging should only be interpreted in the context of clinical presentation [57]. This evaluation was based on visual assessment which is based on vessel activity compared to the liver and the distribution of uptake. Potentially, some of the radiomic features that demonstrated a high diagnostic performance, but are based on information that is not easily appreciated by eye, could help utilise FDG PET for monitoring aortitis.

Several components of radiomics can be replaced by DL methods adding many benefits such as automation and reproducibility. However, these methods require large datasets that are not always conceivable. They also need to be interpretable and understandable in a clinical context to encourage trust, avoid unnoticed bias in training data, and overcome privacy, legal and accountability issues [23,37,58]. These limitations do not eliminate the use of DL and are likely to be easier to overcome in coming years. It does leave room for other techniques like handcrafted radiomic features and simpler ML classifiers. While DL is popular, its application to steps in the radiomic workflow does not always produce better results than other well-established methods.

4.4. Harmonization and Standardization

There is still some debate as to the validity of harmonisation with ComBat [19,59,60]. Orlhac et al. stated that ComBat is only appropriate in situations described in their guide [61]. Papadimitroulas et al. described several other alternatives to ComBat but also concluded that ComBat performed well overall [23]. As this study uses data from several institutions and scanners, harmonization was explored and showed no significant difference in diagnostic utility. A key disadvantage to using ComBat for harmonization is that scans that do not belong to a predefined batch cannot be harmonized. In order to define a batch, a minimum of 20 samples are required, which was not achieved in some of our batches and is not feasible in many smaller centers.

A key point discussed in numerous radiomic studies and reviews is the need for standardized methodology. Standardization of imaging protocols was not feasible as this was a retrospective study with insufficient data to exclude patients based on imaging protocol. All steps after reconstruction and before feature extraction were kept consistent as this has been proven to have a significant effect [?]. Table 8 demonstrates the effect of using input data from centres following different imaging protocols. Fingerprint A

showed the most robustness, making it promising for transferability. TRIPOD reporting guidelines were adhered to in this project to ensure transparency of methodological details [26]. Feature extraction software (PyRadiomics) that mostly adheres to IBSI radiomic feature standardization was utilized. The IBSI definitions are discussed in their paper by Zwannenburg et al. [34]. Deviations from these definitions are discussed in the user documentation (<https://pyradiomics.readthedocs.io/en/latest/features> (accessed on 1 November 2022)) and accompanying publication [35]. While standardization is important, specific recommendations for steps are rarely made as optimal methods vary based on modality, condition and application. Some specific recommendations are published for PET imaging in LVV, but there are little or no studies reporting use of radiomics in this setting, meaning there is no specific guidance. The results of this project provide initial results but further optimization of each step could be explored to produce specific advice to this application. Meanwhile, thorough reporting of methods is sufficient to overcome most issues caused by a lack of standardization. As IBSI found, even with well-defined rules, there can be discrepancies in application, so following guidelines such as IBSI, STARD or TRIPOD when reporting can help convey the most important details [26,34]. Some decisions made in this project have been in areas still debated in literature. Examples include how to define the bin width or bin number when conducting gray level discretization, or whether to upsample or downsample when spatially resampling. While they are not limitations, they may be improved upon in future studies [22,34].

5. Conclusions

The purpose of this study was to develop and validate an automated pipeline that assists the diagnosis of active aortitis. The pipeline included an automated segmentation method with a CNN, radiomic analysis and ML.

The different diagnostic models performed well and similarly. Fingerprint A demonstrated the most robustness and was built using the best performing individual features whilst removing correlated features.

These findings demonstrate a radiomic pipeline can be generalizable and transferable. They could be used to build an automated clinical decision tool which would facilitate objective and standardized assessment regardless of observer experience.

Author Contributions: L.M.D. contributions include: a portion of the data collation, processing input data, developing and validating a radiomic methodology, formal analysis of all results, software development, writing the original draft manuscript and editing, and project management. A.F.S., S.L.M., M.A.B., A.W.M. and C.T. contributions include: supervision, conceptualization, project management, advice and discussions about the results, helping collate the required data and help with accessing the required software and hardware to conduct the experiments, paper reviewing and editing. R.F. conducted similar experiments and helped with trouble shooting when problems in the method and code arose. N.R. helped develop the initial version of the automated segmentation method used in this study. G.D.v.P. and R.H.J.A.S. helped validate our results by applying the method to their own data. J.M.T., J.C.M. and K.S.M.v.d.G. contributed data from their respective studies for use in validation. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Engineering and Physical Sciences Research Council Centre for Doctoral Training in Tissue Engineering and Regenerative Medicine; Innovation in Medical and Biological Engineering—Grant No. EP/L014823/1. Morgan is the principal investigator of the Medical Research Council TARGET (Treatment According to Response in Giant cELL arTeritis) Partnership grant (MR/N011775/1) and is also funded by the National Institute for Health Research (NIHR) Leeds Biomedical Research Centre and NIHR Medtech and In Vitro Diagnostics Co-operative. Bailey is funded by a British Heart Foundation Intermediate Clinical Research Fellowship (FS/18/12/33270) and Tsoumpas by a Royal Society Industry Fellowship (IF170011). Frood and Scarsbrook receive salary support from Innovate UK via the National Consortium for Intelligent Medical Imaging. Scarsbrook acknowledges academic salary support from Leeds Cares (Leeds Hospitals' Charity). Sarah Mackie is supported by the NIHR Leeds Biomedical Research Centre. Tarkin is supported by a Wellcome Trust Clinical Research Career Development Fellowship [211100/Z/18/Z]. Mason is

supported by The Imperial College NIHR Biomedical Research Centre. This publication presents independent research supported by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Institutional Review Board Statement: The study was approved by the institutional research data access committee and local governance team as a clinical service evaluation, which did not require external ethics committee review. Written consent was collected from all patients at the time of imaging for use of their anonymised data in research and service development projects. Patient meta-data were collated by clinical direct care teams and pseudo-anonymised for use within this study. The validation cohort was obtained from other institutions where all patients were part of clinical trials and provided written and informed consent for inclusion in other studies—UK GCA Consortium (REC Ref. 05/Q1108/28) [27] and PITA (PET Imaging of Giant Cell and Takayasu Arteritis) (REC approval: 19/EE/0043 Clinical trials registration: NCT04071691).

Informed Consent Statement: Written consent was collected from all patients at the time of imaging for use of their anonymised data in research and service development projects.

Data Availability Statement: Code and some data available with reasonable request.

Acknowledgments: This work was undertaken on ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. We acknowledge the clinical data collection work conducted by Louise Sorensen, the data shared by Pratik Adusumilli, and the infrastructure support from MRC TARGET, LICAMM and the University of Leeds. We also acknowledge Johann Alberts and Brad Miles from Alliance Medical who provided anonymized external imaging datasets, and Roie Manavaki for collating data from the PITA study. We acknowledge Alejandro Frangi for his advice in method development.

Conflicts of Interest: Lisa M Duff declares that she has no conflict of interest. Andrew F. Scarsbrook declares that he has no conflict of interest. Nishant Ravikumar declares he has no conflict of interest. Gijs D. van Praagh declares that he has no conflict of interest. Sarah L. Mackie reports Consultancy on behalf of her institution for Roche/Chugai, Sanofi, AbbVie, AstraZeneca; Investigator on clinical trials, Sanofi and GSK; speaking/lecturing on behalf of her institution for Roche/Chugai, Vifor and Pfizer; chief investigator on STERLING-PMR trial, funded by NIHR; patron of the charity PMRGCAuk. No personal remuneration was received for any of the above activities. Support from Roche/Chugai to attend EULAR2019 in person and from Pfizer to attend ACR Convergence 2021 virtually. Russell Frood declares that he has no conflict of interest. Marc Bailey declares that he has no conflict of interest. Jason M. Tarkin declares that he has no conflict of interest. Justin C. Mason declares that he has no conflict of interest. Riemer H.J.A. Slart declares that he has no conflict of interest. Kornelis S.M. van der Geest reports speaker fees from Roche and research support from AbbVie. Ann W. Morgan declares that she has no conflict of interest. Charalampos Tsoumpas declares that he has no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SUV	Standardized Uptake Value
PCA	Principal Component Analysis
DSC	Dice Similarity Coefficient
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic
GCA	Giant Cell Arteritis
LVV	Large Vessel Vasculitis
FDG PET-CT	[¹⁸ F]-Fluorodeoxyglucose Positron Emission Tomography—Computed Tomography
FDG	[¹⁸ F]-Fluorodeoxyglucose
PET	Positron Emission Tomography
CT	Computed Tomography
GLDM	Gray-Level Dependence Matrix
GLCM	Gray-Level Co-Occurrence Matrix

GLRLM	Gray-Level Run Length Matrix
GLSZM	Gray-Level Size Zone Matrix
DLYD	Delayed Event Subtraction
TAK	Takayasu’s arteritis
CRP	C-Reactive Protein
ESR	Erythrocyte Sedimentation Rate
ML	Machine Learning
DL	Deep Learning
DICOM	Digital Imaging and Communications in Medicine
GPU	Graphics Processing Unit
PITA	PET Imaging of Giant Cell and Takayasu Arteritis
TARGET	Treatment According to Response in Giant Cell arTeritis
CNN	Convolutional Neural Network
EULAR	European Alliance of Associations for Rheumatology
EANM	European Association of Nuclear Medicine
SNMMI	Society of Nuclear Medicine and Molecular Imaging
ReLU	Rectified Linear Unit
AI	Artificial Intelligence
ROI	Region of Interest
IBSI	International Biomarker Standardisation Initiative
PACS	Picture Archiving and Communication
VPFX	Vue Point FX (3D time of flight)
SS-SIMUL	Single-scatter Simulation
BLOB-OS-TF	Spherically symmetric basis function ordered subset algorithm

Appendix A. Imaging Protocol

Table A1. PET reconstruction parameters for each PET-CT system. Key—BLOB-OS-TF = spherically symmetric basis function ordered subset algorithm; SS-SIMUL = single-scatter simulation; VPFX = Vue Point FX (3D time of flight); DLYD = delayed event subtraction.

Scanner	Reconstruction	Scatter Correction	Randoms Correction	Matrix	Voxel Size
Gemini TF64	BLOB-OS-TF	SS-SIMUL	DLYD	144	4.00 × 4.00 × 4.00
Discovery 710	VPFX, QCFX, or VPHD	Model based	Singles	192	3.65 × 3.65 × 3.27
Discovery 690	VPFX or VPFX	Model-based	Singles	193	3.65 × 3.65 × 3.28
Discovery MI DR	VPFX, QCFX, or VPHD	Model-based	SING	256	2.73 × 2.73 × 3.27
Discovery ST	OSEM	Convolution subtraction	DLYD	128	4.69 × 4.69 × 3.27
Discovery STE	OSEM	Convolution subtraction	SING	128	5.47 × 5.47 × 3.27
Biograph 6 True Point	OSEM2D 4i8s	Model-based	DLYD	168	4.07 × 4.07 × 3.00
Biograph 6	OSEM2D 4i8s	Model-based	DLYD	168	4.07 × 4.07 × 3.00
Biograph 64 mCT	PSF + TOF 2i21s or OSEM3D 2i24s	Model-based	DLYD	200	4.07 × 4.07 × 3.00

Appendix B. ML Parameters and Diagnostic Performance of Individual SUV metrics and Radiomic Features

Table A2. ML Parameters and Diagnostic Performance of Individual SUV metrics and Radiomic Features : Key—ML (Machine Learning), ACC (Accuracy), CI (Confidence Interval), AUC (Area Under the Receiver Operating Characteristic Curve), Val (Validation), SUV (Standardized Uptake Value), GLDM (Gray-Level Dependence Matrix), GLCM (Gray-Level Co-Occurrence Matrix), GLRLM (Gray-Level Run Length Matrix), and GLSZM (Gray-Level Size Zone Matrix).

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC Test CI	Test Val	ACC Val	AUC Val	AUC Val CI
original shape Elongation	('C', 3.8056144605552977), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 6310), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.017923654340198922)	0.500	0.585	0.455–0.716	0.500	0.617	0.239–0.994	0.500	0.614	0.497–0.732	
original firstorder 10 Percentile	('C', 2.8817940478533313), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 3403), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.011518315256582621)	0.676	0.825	0.554–1.000	0.558	0.700	0.387–1.000	0.637	0.784	0.690–0.879	
original firstorder Energy	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10,000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02338512988384992)	0.689	0.852	0.769–0.934	0.800	0.967	0.888–1.000	0.625	0.717	0.611–0.822	
original firstorder Entropy	('C', 3.898563938816272), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1057), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09065383006063307)	0.710	0.901	0.793–1.000	0.800	0.917	0.780–1.000	0.470	0.555	0.440–0.670	
original firstorder Interquartile Range	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 8442), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.004402437057502587)	0.712	0.834	0.771–0.896	0.658	0.850	0.652–1.000	0.448	0.531	0.415–0.647	
original firstorder Kurtosis	('C', 3.634835218565156), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 6758), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0013191637682750674)	0.500	0.441	0.297–0.586	0.500	0.450	0.130–0.770	0.500	0.468	0.346–0.590	
original firstorder Maximum	('C', 3.5515515962056434), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 9296), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.008690899693926512)	0.606	0.764	0.456–1.000	0.800	0.967	0.888–1.000	0.537	0.508	0.384–0.632	
original firstorder Mean	('C', 3.9294940929254794), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 3769), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.034137309056045194)	0.767	0.890	0.668–1.000	0.800	0.883	0.647–1.000	0.598	0.724	0.620–0.828	
original firstorder Mean-AbsoluteDeviation	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10,000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.678	0.879	0.720–1.000	0.700	0.900	0.741–1.000	0.479	0.555	0.440–0.670	
original firstorder Median	('C', 3.9959470312560192), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 9514), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.003443979629449537)	0.756	0.853	0.574–1.000	0.600	0.833	0.591–1.000	0.568	0.690	0.581–0.799	
original firstorder Range	('C', 2.9625927896408317), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 5285), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04665607138780834)	0.592	0.756	0.495–1.000	0.700	0.967	0.888–1.000	0.465	0.484	0.360–0.608	

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original firstorder RobustMeanAbsoluteDeviation	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 6466), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.654	0.856	0.767–0.945	0.658	0.867	0.679–1.000	0.495	0.535	0.419–0.650
original firstorder Root-MeanSquared	('C', 3.9502570349836175), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 7402), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08258347919775297)	0.767	0.900	0.710–1.000	0.600	0.883	0.647–1.000	0.585	0.718	0.614–0.822
original firstorder Skewness	('C', 2.238056324193022), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1545), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08427086764521606)	0.489	0.568	−0.004–1.000	0.500	0.850	0.613–1.000	0.500	0.613	0.495–0.731
original firstorder Total Energy	('C', 1.292188045736975), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 9696), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00010802104590742912)	0.689	0.852	0.769–0.934	0.800	0.967	0.888–1.000	0.662	0.717	0.611–0.822
original firstorder Uniformity	('C', 3.39826558360898), ('dual', False), ('fit intercept', False), ('intercept scaling', 5), ('max iter', 7542), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07122003912935275)	0.500	0.138	0.000–0.277	0.500	0.100	−0.059–0.259	0.500	0.460	0.344–0.577
original firstorder Variance	('C', 2.8454630117539024), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 5096), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.007605136714617995)	0.723	0.899	0.756–1.000	0.800	0.967	0.890–1.000	0.453	0.551	0.437–0.665
original glcm Autocorrelation	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.798	0.894	0.786–1.000	0.700	0.950	0.839–1.000	0.469	0.604	0.493–0.716
original glcm Cluster-Prominence	('C', 1.2014550932380232), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1049), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02992266728475855)	0.718	0.850	0.710–0.990	0.700	1.000	nan–nan	0.424	0.558	0.444–0.673
original glcm Cluster-Shade	('C', 1.1562338815842363), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.500	0.573	0.069–1.000	0.500	0.533	0.255–0.812	0.500	0.703	0.596–0.809
original glcm Cluster-Tendency	('C', 3.3987720010803155), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 2281), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04646204360821551)	0.743	0.894	0.753–1.000	0.800	0.933	0.813–1.000	0.451	0.579	0.465–0.692
original glcm Contrast	('C', 2.794213383090132), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7568), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09298732537139162)	0.802	0.901	0.776–1.000	0.758	0.967	0.890–1.000	0.477	0.452	0.333–0.571

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original glcm Correlation	('C', 2.3265163816288474), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 7116), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0188936121493807)	0.500	0.423	0.244–0.603	0.500	0.367	0.068–0.665	0.500	0.768	0.658–0.877
original glcm Difference Average	('C', 3.765281617061703), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 3203), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04678697115468039)	0.762	0.896	0.769–1.000	0.800	0.900	0.742–1.000	0.502	0.436	0.316–0.556
original glcm Difference Entropy	('C', 2.5454905419682645), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7232), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.013588632723062564)	0.748	0.896	0.769–1.000	0.800	0.933	0.814–1.000	0.518	0.436	0.316–0.555
original glcm Difference Variance	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.774	0.901	0.771–1.000	0.800	0.967	0.890–1.000	0.487	0.462	0.343–0.580
original glcm Id	('C', 2.876700911377709), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 5860), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.021468961778094153)	0.628	0.862	0.741–0.982	0.600	0.900	0.742–1.000	0.475	0.401	0.282–0.521
original glcm Idm	('C', 3.5521886140988324), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.639	0.862	0.741–0.982	0.600	0.900	0.742–1.000	0.475	0.394	0.275–0.513
original glcm Idmn	('C', 2.5729756259638257), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 6648), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0722906499673705)	0.500	0.369	0.187–0.551	0.500	0.500	0.222–0.778	0.500	0.561	0.437–0.685
original glcm Idn	('C', 1.5612853647891844), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 8056), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.031534182504912224)	0.500	0.390	0.205–0.575	0.500	0.517	0.236–0.798	0.500	0.586	0.464–0.709
original glcm Imc1	('C', 3.9438796427354554), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 9320), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09874439787032405)	0.500	0.440	0.182–0.698	0.500	0.333	0.067–0.600	0.500	0.796	0.693–0.900
original glcm Imc2	('C', 2.543363070723238), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 2274), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00834386841277177)	0.500	0.500	nan–nan	0.500	0.517	0.208–0.825	0.500	0.783	0.684–0.882
original glcm Inverse Variance	('C', 3.2073129630396777), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 6658), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0074434857032721225)	0.628	0.862	0.741–0.982	0.700	0.900	0.742–1.000	0.490	0.398	0.278–0.518

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original firstorder RobustMeanAbsoluteDeviation	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 6466), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.654	0.856	0.767–0.945	0.658	0.867	0.679–1.000	0.495	0.535	0.419–0.650
original glcm JointAverage	('C', 2.884522523620075), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.018889093800706636)	0.714	0.894	0.802–0.987	0.700	0.900	0.742–1.000	0.501	0.591	0.478–0.703
original glcm JointEnergy	('C', 1.199944589432079), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 45), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09140428650296156)	0.500	0.851	0.767–0.934	0.500	0.933	0.814–1.000	0.500	0.478	0.361–0.594
original glcm JointAverage	('C', 2.884522523620075), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.018889093800706636)	0.714	0.894	0.802–0.987	0.700	0.900	0.742–1.000	0.501	0.591	0.478–0.703
original glcm JointEntropy	('C', 1.199944589432079), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 45), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09140428650296156)	0.500	0.851	0.767–0.934	0.500	0.933	0.814–1.000	0.500	0.478	0.361–0.594
original glcm JointEntropy	('C', 2.8485013507009964), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 1666), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0003388234219221265)	0.710	0.883	0.758–1.000	0.800	0.933	0.814–1.000	0.462	0.500	0.383–0.616
original glcm MCC	('C', 3.7544969295345636), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7337), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02150123778633736)	0.500	0.454	0.342–0.565	0.500	0.533	0.213–0.854	0.500	0.593	0.470–0.717
original glcm MaximumProbability	('C', 3.749171778156825), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 9343), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.039031034588965174)	0.500	0.156	0.076–0.235	0.500	0.100	−0.058–0.258	0.500	0.558	0.441–0.674
original glcm SumAverage	('C', 1.2810898302218976), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 3097), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07161938424684904)	0.714	0.894	0.802–0.987	0.700	0.900	0.742–1.000	0.501	0.591	0.478–0.703
original glcm SumEntropy	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 8498), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.705	0.901	0.793–1.000	0.800	0.900	0.741–1.000	0.452	0.586	0.473–0.699
original glcm SumSquares	('C', 1.305357350462431), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 339), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0006889623638318466)	0.755	0.895	0.754–1.000	0.800	0.967	0.890–1.000	0.452	0.554	0.440–0.668
original gldm DependenceEntropy	('C', 3.7717457451609153), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 14), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.009597244103991846)	0.539	0.807	0.605–1.000	0.500	0.917	0.743–1.000	0.500	0.812	0.729–0.896

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original gldm Dependence Non-Uniformity	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.598	0.696	0.359–1.000	0.800	0.933	0.814–1.000	0.468	0.502	0.381–0.623
original gldm Dependence Non-Uniformity Normalized	('C', 3.9397386998320325), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 5358), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.058125565483952035)	0.500	0.805	0.647–0.963	0.500	0.883	0.718–1.000	0.500	0.370	0.253–0.486
original gldm Dependence Variance	('C', 3.908072013043827), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 1551), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.026849446139380312)	0.639	0.795	0.620–0.970	0.658	0.850	0.660–1.000	0.443	0.303	0.191–0.415
original gldm Gray Level Non-Uniformity	('C', 3.1826657805845513), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 9909), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0008888365484206898)	0.676	0.779	0.447–1.000	0.558	0.600	0.288–0.912	0.474	0.468	0.344–0.593
original gldm Gray Level Variance	('C', 1.8552309464866807), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 4836), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0019131772861938168)	0.730	0.899	0.756–1.000	0.800	0.967	0.890–1.000	0.468	0.551	0.437–0.666
original gldm High Gray Level Emphasis	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04047765497803527)	0.823	0.906	0.812–0.999	0.700	0.967	0.888–1.000	0.484	0.602	0.490–0.713
original gldm Large Dependence Emphasis	('C', 1.0588224326821771), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 9519), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.047865873089018005)	0.673	0.828	0.648–1.000	0.700	0.883	0.718–1.000	0.459	0.351	0.234–0.468
original gldm Large Dependence High Gray Level Emphasis	('C', 3.1168548949137724), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 1217), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0838087387972799)	0.500	0.557	0.111–1.000	0.500	0.467	0.123–0.810	0.500	0.753	0.640–0.865
original gldm Low Gray Level Emphasis	('C', 2.8291349845222524), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 7077), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03740604869362484)	0.500	0.500	nan–nan	0.500	0.500	nan–nan	0.500	0.500	nan–nan
original gldm Small Dependence Emphasis	('C', 3.9838817245979876), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1542), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0008766419147272496)	0.710	0.867	0.699–1.000	0.800	0.900	0.742–1.000	0.534	0.407	0.287–0.527
original gldm Small Dependence High Gray Level Emphasis	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 2813), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05195113774499899)	0.773	0.929	0.832–1.000	0.800	0.983	0.937–1.000	0.491	0.525	0.409–0.641

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI	Test	ACC Val	AUC Val	AUC Val CI
original firstorder RobustMeanAbsoluteDeviation	(‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 6466), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 1×10^{-7})	0.654	0.856	0.767–0.945	0.658	0.867	0.679–1.000	0.495	0.535	0.419–0.650	
original gldm Small Dependence Low Level Emphasis	(‘C’, 2.030332175989515), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 5234), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.006906386148325776)	0.500	0.381	−0.008–0.770	0.500	0.333	0.034–0.633	0.500	0.282	0.176–0.387	
original glrlm Gray Level Non-Uniformity	(‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 10), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.1)	0.662	0.751	0.421–1.000	0.558	0.550	0.211–0.889	0.467	0.486	0.361–0.612	
original glrlm Gray Level Non-Uniformity Normalized	(‘C’, 3.8458368662008255), (‘dual’, False), (‘fit intercept’, False), (‘intercept scaling’, 3), (‘max iter’, 3365), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.014454416417432375)	0.500	0.133	0.011–0.255	0.500	0.117	−0.050–0.283	0.500	0.457	0.341–0.573	
original glrlm Gray Level Variance	(‘C’, 1.6722052556540499), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 10), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.1)	0.723	0.905	0.764–1.000	0.700	0.967	0.890–1.000	0.432	0.554	0.440–0.668	
original glrlm High Level RunEmphasis	(‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 10000), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.05118583169143218)	0.812	0.906	0.812–0.999	0.700	0.967	0.888–1.000	0.476	0.606	0.494–0.717	
original glrlm Long Run Emphasis	(‘C’, 3.833856019580739), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 1631), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.03501141624853724)	0.639	0.844	0.675–1.000	0.600	0.883	0.718–1.000	0.467	0.362	0.244–0.480	
original glrlm Long Run High Gray Level Emphasis	(‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 4806), (‘penalty’, ‘l1’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.018730143457553885)	0.753	0.884	0.740–1.000	0.700	0.967	0.890–1.000	0.500	0.654	0.545–0.762	
original glrlm LongRun-LowGray LevelEmphasis	(‘C’, 2.6289220027281743), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 1), (‘max iter’, 7056), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.09003136591880305)	0.500	0.861	0.738–0.984	0.500	0.867	0.680–1.000	0.500	0.657	0.547–0.767	
original glrlm Low Gray Level Run Emphasis	(‘C’, 2.2376337828274573), (‘dual’, False), (‘fit intercept’, False), (‘intercept scaling’, 4), (‘max iter’, 7747), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.07030672418295693)	0.500	0.157	−0.012–0.325	0.500	0.133	−0.054–0.320	0.500	0.310	0.204–0.417	
original glrlm RunEntropy	(‘C’, 3.3114257300096157), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 121), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.0194354228138649)	0.690	0.889	0.748–1.000	0.700	0.917	0.780–1.000	0.464	0.606	0.493–0.718	
original glrlm Run Length Non-Uniformity	(‘C’, 1.4366614662702664), (‘dual’, False), (‘fit intercept’, False), (‘intercept scaling’, 1), (‘max iter’, 2709), (‘penalty’, ‘l2’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.03928484749498973)	0.500	0.611	0.104–1.000	0.500	0.967	0.890–1.000	0.500	0.555	0.434–0.676	

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original glrlm Run Length Non-Uniformity Normalized	('C', 2.6720853872765495), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 3774), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07427424237615005)	0.500	0.839	0.684–0.993	0.500	0.900	0.742–1.000	0.500	0.378	0.259–0.496
original glrlm RunPercentage	('C', 2.6023123244300974), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 291), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.049781425459564835)	0.500	0.839	0.684–0.993	0.500	0.883	0.718–1.000	0.500	0.370	0.251–0.488
original glrlm RunVariance	('C', 3.61481821528191), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.564	0.834	0.651–1.000	0.600	0.867	0.689–1.000	0.467	0.347	0.231–0.464
original glrlm Short-RunEmphasis	('C', 2.732198485448542), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 974), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.01734471841788434)	0.500	0.839	0.684–0.993	0.500	0.900	0.742–1.000	0.500	0.377	0.258–0.495
original glrlm Short Run High Gray Level Emphasis	('C', 3.9351523410188607), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 2339), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06576459057456105)	0.787	0.906	0.812–0.999	0.800	0.967	0.888–1.000	0.476	0.594	0.482–0.706
original glrlm Short Run Low Gray Level Emphasis	('C', 3.4237606654811574), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 1707), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06073466346615761)	0.500	0.833	0.660–1.000	0.500	0.867	0.680–1.000	0.500	0.695	0.590–0.801
original glszm Gray Level Non-Uniformity	('C', 3.8458029875767066), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 2474), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04148253163117249)	0.500	0.635	0.355–0.915	0.500	0.850	0.661–1.000	0.500	0.428	0.306–0.549
original glszm Gray Level Non-Uniformity Normalized	('C', 3.600665976879018), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 1678), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05947499893500212)	0.500	0.894	0.770–1.000	0.500	0.983	0.937–1.000	0.500	0.573	0.459–0.687
original glszm Gray Level Variance	('C', 2.901196262126115), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03774426454693599)	0.749	0.860	0.721–0.999	0.800	1.000	nan–nan	0.453	0.575	0.461–0.688
original glszm High Gray Level Zone Emphasis	('C', 3.478204583582132), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 1616), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09703590654231291)	0.762	0.905	0.829–0.981	0.800	0.983	0.937–1.000	0.453	0.601	0.489–0.712
original glszm Large Area Emphasis	('C', 1.6046715254557193), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 6822), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07800900631220041)	0.648	0.851	0.652–1.000	0.700	0.867	0.689–1.000	0.490	0.378	0.260–0.496

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original Area Emphasis	glszm Large Gray Level (‘C’, 1.8487035352590049), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 1179), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.05701958294758005)	0.553	0.812	0.550–1.000	0.700	0.850	0.653–1.000	0.483	0.357	0.236–0.478
original Area Emphasis	glszm Large Gray Level (‘C’, 2.3369045710562144), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 7065), (‘penalty’, ‘12’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.08366919609286493)	0.673	0.853	0.674–1.000	0.558	0.867	0.690–1.000	0.467	0.425	0.307–0.543
original Level Zone Emphasis	glszm Low Gray Level (‘C’, 2.0825064626116987), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 469), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.05187162786375553)	0.500	0.500	nan–nan	0.500	0.500	nan–nan	0.500	0.500	nan–nan
original Size Zone Non-Uniformity	glszm Size Zone Non-Uniformity (‘C’, 3.4968658660509293), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 5), (‘max iter’, 5119), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.09985912291392018)	0.705	0.849	0.793–0.905	0.800	0.967	0.890–1.000	0.475	0.489	0.370–0.609
original Size Zone Normalized	glszm Size Zone Normalized (‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 1220), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.0010447413452487947)	0.653	0.907	0.773–1.000	0.700	0.967	0.888–1.000	0.496	0.429	0.309–0.549
original Area Emphasis	glszm Small Area (‘C’, 3.6221746139345705), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 813), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.004130953766069213)	0.614	0.896	0.738–1.000	0.600	0.967	0.888–1.000	0.520	0.427	0.308–0.547
original Area Emphasis	glszm Small Area High Gray Level (‘C’, 2.6240187371586012), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 3622), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.08556868775028845)	0.762	0.900	0.795–1.000	0.800	1.000	nan–nan	0.478	0.553	0.439–0.667
original Area Emphasis	glszm Small Area Low Gray Level (‘C’, 1.7064389360641852), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 1), (‘max iter’, 9885), (‘penalty’, ‘12’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.07794842268900995)	0.500	0.837	0.733–0.941	0.500	0.867	0.688–1.000	0.500	0.590	0.478–0.702
original Zone Entropy	glszm Zone Entropy (‘C’, 2.8550691194553095), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 3433), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.06520910207277585)	0.500	0.708	0.456–0.960	0.500	0.883	0.678–1.000	0.500	0.772	0.680–0.864
original Zone Percentage	glszm Zone Percentage (‘C’, 4.0), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 2), (‘max iter’, 956), (‘penalty’, ‘11’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 1×10^{-7})	0.710	0.874	0.698–1.000	0.800	0.900	0.742–1.000	0.534	0.409	0.290–0.529
original Zone Variance	glszm Zone Variance (‘C’, 2.2299373803328835), (‘dual’, False), (‘fit intercept’, True), (‘intercept scaling’, 3), (‘max iter’, 8739), (‘penalty’, ‘12’), (‘random state’, 1), (‘solver’, ‘liblinear’), (‘tol’, 0.05624148106047668)	0.623	0.850	0.654–1.000	0.700	0.867	0.689–1.000	0.490	0.375	0.258–0.493

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original shape Flatness	('C', 3.26351153620514), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 4285), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07375765313681576)	0.500	0.500	nan–nan	0.500	0.500	nan–nan	0.500	0.500	nan–nan
original shape Least Axis Length	('C', 2.226917940773356), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 9542), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05189302090817439)	0.500	0.587	0.222–0.952	0.500	0.800	0.566–1.000	0.500	0.650	0.529–0.772
original shape Major Axis Length	('C', 1.1143625492476652), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 6549), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00031345515541475905)	0.550	0.645	0.410–0.881	0.500	0.367	0.019–0.715	0.500	0.518	0.397–0.639
original shape Maximum2D DiameterColumn	('C', 1.0021818324304934), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 8894), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0929830142217613)	0.500	0.451	0.057–0.844	0.500	0.583	0.220–0.947	0.500	0.492	0.372–0.611
original shape Maximum2D DiameterRow	('C', 3.3024304072681905), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 4036), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09059059637632817)	0.500	0.415	0.083–0.748	0.500	0.650	0.331–0.969	0.500	0.491	0.370–0.612
original shape Maximum2D DiameterSlice	('C', 1.5124941175869826), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 8130), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0765168304164115)	0.500	0.497	0.287–0.706	0.500	0.833	0.632–1.000	0.500	0.691	0.581–0.802
original shape Maximum3D Diameter	('C', 3.1109766854613072), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 6978), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.017060855705831286)	0.500	0.429	0.070–0.789	0.500	0.617	0.251–0.983	0.500	0.490	0.369–0.610
original shape Mesh Volume	('C', 3.3281428672743534), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 7246), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06893000237976868)	0.500	0.527	0.048–1.000	0.500	0.850	0.655–1.000	0.500	0.588	0.467–0.708
original shape Minor Axis Length	('C', 1.5859938721531885), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 6160), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09283436723767415)	0.500	0.447	0.166–0.727	0.500	0.900	0.742–1.000	0.500	0.716	0.609–0.822
original shape Sphericity	('C', 2.003303146456421), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 7530), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07357167804542747)	0.500	0.494	0.201–0.787	0.500	0.383	0.035–0.732	0.500	0.522	0.398–0.646
original shape Surface Area	('C', 1.9839469572178487), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 2239), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0922515508141599)	0.500	0.516	0.019–1.000	0.500	0.850	0.660–1.000	0.500	0.553	0.432–0.674

Table A2. Cont.

Feature	Params	ACC Training	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC Val CI
original shape Surface Volume Ratio	('C', 3.499156350599231), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 6863), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08640779329555863)	0.500	0.500	nan–nan	0.500	0.500	nan–nan	0.500	0.500	nan–nan
original shape Voxel Volume	('C', 3.7936962680325133), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 4309), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.051928537070625225)	0.500	0.532	0.046–1.000	0.500	0.850	0.655–1.000	0.500	0.587	0.466–0.707
SUV 50	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10,000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.667	0.803	0.557–1.000	0.700	1.000	nan–nan	0.518	0.534	0.412–0.656
SUV 60	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 4189), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1×10^{-7})	0.581	0.763	0.403–1.000	0.700	1.000	nan–nan	0.525	0.518	0.393–0.642
SUV 70	('C', 3.159539333749517), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 2546), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04748372286529402)	0.556	0.737	0.372–1.000	0.700	0.967	0.888–1.000	0.517	0.511	0.385–0.637
SUV 80	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1110), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.023117444595346284)	0.581	0.730	0.398–1.000	0.800	0.950	0.839–1.000	0.522	0.495	0.369–0.622
SUV 90	('C', 3.1704526969890265), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.027440275486626378)	0.606	0.753	0.423–1.000	0.800	0.933	0.791–1.000	0.508	0.493	0.368–0.619

Appendix C. Diagnostic Performance of Fingerprint A in All Classifiers

Table A3. Diagnostic performance of Fingerprint A.

ML Type	ACC Training	ACC CI	AUC Training	AUC CI	ACC Test	AUC Test	AUC Test CI	ACC Val	AUC Val	AUC Val CI
rf	0.738	0.141	0.900	[0.789 1.]	0.8	0.983	[0.937 1.]	0.748	0.923	[0.835 1.]
lgr	0.736	0.149	0.905	[0.808 1.]	0.8	0.966	[0.890 1.]	0.769	0.880	[0.762 0.998]
dt	0.850	0.023	0.850	[0.798 0.902]	0.858	0.858	[0.646 1.]	0.748	0.748	[0.591 0.905]
gpc	0.5	0	0.5	[nan nan]	0.5	0.5	[nan nan]	0.5	0.5	[nan nan]
sgd	0.525	0.043	0.525	[0.427 0.622]	0.5	0.5	[nan nan]	0.5	0.5	[nan nan]
perc	0.642	0.153	0.895	[0.846 0.943]	0.5	0.966	[0.887 1.]	0.5	0.880	[0.766 0.994]
pasagr	0.552	0.086	0.891	[0.750 1.]	0.7	0.95	[0.854 1.]	0.519	0.891	[0.770 1.]
nnet	0.602	0.132	0.613	[0.276 0.951]	0.716	0.7	[0.435 0.964]	0.831	0.824	[0.680 0.967]
kneigh	0.718	0.129	0.816	[0.712 0.919]	0.758	0.833	[0.603 1.]	0.806	0.840	[0.699 0.981]

Appendix D. Diagnostic Performance of Fingerprint B in All Classifiers

Table A4. Diagnostic performance of Fingerprint B.

ML Type	ACC Training	ACC CI	AUC Training	AUC CI	ACC Test	AUC Test	AUC CI Test	ACC Val	AUC Val	AUC CI Val
rf	0.8	0.056	0.768	[0.51346925 1.]	0.958	0.958	[0.86887363 1.]	0.81	0.893	[0.78622024 1.]
lgr	0.819	0.05	0.864	[0.69097375 1.]	0.875	0.967	[0.89005795 1.]	0.786	0.895	[0.79316199 0.99669309]
svm	0.722	0.088	0.769	[0.59542127 0.94291207]	0.6	0.833	[0.63196618 1.]	0.667	0.859	[0.73004966 0.98734165]
dt	0.5	0.125	0.73	[0.40364462 1.]	0.8	0.617	[0.15542111 1.]	0.853	0.857	[0.72781114 0.98595698]
gpc	0.686	0.089	0.836	[0.70011143 0.97211079]	0.7	0.9	[0.74118659 1.]	0.685	0.884	[0.76955265 0.9985633]]
sgd	0.819	0.06	0.858	[0.70163664 1.]	0.875	0.967	[0.89005795 1.]	0.786	0.902	[0.80403173 1.]
perc	0.776	0.1	0.894	[0.73675941 1.]	0.717	0.883	[0.70838888 1.]	0.79	0.783	[0.60065059 0.9645668]]
pasagr	0.736	0.115	0.881	[0.72343443 1.]	0.775	0.883	[0.70825031 1.]	0.788	0.862	[0.74025575 0.98438193]
nnet	0.661	0.128	0.809	[0.59762327 1.]	0.675	0.867	[0.68965505 1.]	0.81	0.902	[0.80195173 1.]
kneigh	0.668	0.094	0.735	[0.5274969 0.9425031]	0.6	0.6	[0.26131101 0.93868899]	0.768	0.88	[0.76858261 0.99228696]

References

- Gornik, H.L.; Creager, M.A. Aortitis. *Circulation* **2008**, *117*, 3039–3051. [\[CrossRef\]](#)
- Stone, J.R.; Bruneval, P.; Angelini, A.; Bartoloni, G.; Basso, C.; Batoroeva, L.; Buja, L.M.; Butany, J.; d'Amati, G.; Fallon, J.T.; et al. Consensus statement on surgical pathology of the aorta from the Society for Cardiovascular Pathology and the Association for European Cardiovascular Pathology: I. Inflammatory diseases. *Cardiovasc. Pathol.* **2015**, *24*, 267–278. [\[CrossRef\]](#)
- Pugh, D.; Grayson, P.; Basu, N.; Dhaun, N. Aortitis: Recent advances, current concepts and future possibilities. *Heart* **2021**, *107*, 1620–1629. [\[CrossRef\]](#)
- Monti, S.; Águeda, A.F.; Luqmani, R.A.; Buttgerit, F.; Cid, M.; Dejaco, C.; Mahr, A.; Ponte, C.; Salvarani, C.; Schmidt, W.; et al. Systematic literature review informing the 2018 update of the EULAR recommendation for the management of large vessel vasculitis: Focus on giant cell arteritis. *RMD Open* **2019**, *5*, e001003. [\[CrossRef\]](#)

5. Parikh, M.; Miller, N.R.; Lee, A.G.; Savino, P.J.; Vacarezza, M.N.; Cornblath, W.; Eggenberger, E.; Antonio-Santos, A.; Golnik, K.; Kardon, R.; et al. Prevalence of a normal C-reactive protein with an elevated erythrocyte sedimentation rate in biopsy-proven giant cell arteritis. *Ophthalmology* **2006**, *113*, 1842–1845. [CrossRef]
6. Monach, P.A. Biomarkers in vasculitis. *Curr. Opin. Rheumatol.* **2014**, *26*, 24. [CrossRef] [PubMed]
7. Lee, S.W.; Kim, S.J.; Seo, Y.; Jeong, S.Y.; Ahn, B.C.; Lee, J. F-18 FDG PET for assessment of disease activity of large vessel vasculitis: A systematic review and meta-analysis. *J. Nucl. Cardiol.* **2019**, *26*, 59–67. [CrossRef] [PubMed]
8. Pelletier-Galarneau, M.; Ruddy, T.D. PET/CT for diagnosis and management of large-vessel vasculitis. *Curr. Cardiol. Rep.* **2019**, *21*, 34. [CrossRef] [PubMed]
9. Veeranna, V.; Fisher, A.; Nagpal, P.; Ghosh, N.; Fisher, E.; Steigner, M.; Creager, M.A.; Dorbala, S.; Di Carli, M.F. Utility of multimodality imaging in diagnosis and follow-up of aortitis. *J. Nucl. Cardiol.* **2016**, *23*, 590–595. [CrossRef]
10. DeJaco, C.; Ramiro, S.; Duftner, C.; Besson, F.L.; Bley, T.A.; Blockmans, D.; Brouwer, E.; Cimmino, M.A.; Clark, E.; Dasgupta, B.; et al. EULAR recommendations for the use of imaging in large vessel vasculitis in clinical practice. *Ann. Rheum. Dis.* **2018**, *77*, 636–643. [CrossRef]
11. Slart, R.H.; et al. FDG-PET/CT (A) imaging in large vessel vasculitis and polymyalgia rheumatica: Joint procedural recommendation of the EANM, SNMMI, and the PET Interest Group (PIG), and endorsed by the ASNC. *Eur. J. Nucl. Med. Mol. Imaging* **2018**, *45*, 1250–1269. [CrossRef]
12. Slart, R.H.; Glaudemans, A.W.; Gheysens, O.; Lubberink, M.; Kero, T.; Dweck, M.R.; Habib, G.; Gaemperli, O.; Saraste, A.; Gimelli, A.; et al. Procedural recommendations of cardiac PET/CT imaging: Standardization in inflammatory-, infective-, infiltrative-, and innervation (4Is)-related cardiovascular diseases: a joint collaboration of the EACVI and the EANM. *Eur. J. Nucl. Med. Mol. Imaging* **2020**, *48*, 1–24. [CrossRef]
13. Mackie, S.L.; DeJaco, C.; Appenzeller, S.; Camellino, D.; Duftner, C.; Gonzalez-Chiappe, S.; Mahr, A.; Mukhtyar, C.; Reynolds, G.; De Souza, A.W.S.; et al. British Society for Rheumatology guideline on diagnosis and treatment of giant cell arteritis. *Rheumatology* **2020**, *59*, e1–e23. [CrossRef] [PubMed]
14. Versari, A.; Pipitone, N.; Casali, M.; Jamar, F.; Pazzola, G. Use of imaging techniques in large vessel vasculitis and related conditions. *Q. J. Nucl. Med. Mol. Imaging: Off. Publ. Ital. Assoc. Nucl. Med. Int. Assoc. Radiopharmacol. Sect. Soc.* **2018**, *62*, 34–39. [CrossRef]
15. Grayson, P.C.; Alehashemi, S.; Bagheri, A.A.; Civelek, A.C.; Cupps, T.R.; Kaplan, M.J.; Malayeri, A.A.; Merkel, P.A.; Novakovich, E.; Bluemke, D.A.; et al. Positron emission tomography as an imaging biomarker in a prospective, longitudinal cohort of patients with large vessel vasculitis. *Arthritis Rheumatol.* **2018**, *70*, 439. [CrossRef]
16. van Praagh, G.D.; Nienhuis, P.H.; de Jong, D.M.; Reijrink, M.; van der Geest, K.S.M.; Brouwer, E.; Glaudemans, A.W.J.M.; Sinha, B.; Willemsen, A.T.M.; Slart, R.H.J.A. Toward Reliable Uptake Metrics in Large Vessel Vasculitis Studies. *Diagnostics* **2021**, *11*, 1986. [CrossRef]
17. Dellavedova, L.; Carletto, M.; Faggioli, P.; Sciascera, A.; Del Sole, A.; Mazzone, A.; Maffioli, L. The prognostic value of baseline 18 F-FDG PET/CT in steroid-naïve large-vessel vasculitis: Introduction of volume-based parameters. *Eur. J. Nucl. Med. Mol. Imaging* **2016**, *43*, 340–348. [CrossRef]
18. Motwani, M. Hiding beyond plain sight: Textural analysis of positron emission tomography to identify high-risk plaques in carotid atherosclerosis, 2019. [CrossRef] [PubMed]
19. Hatt, M.; Le Rest, C.C.; Antonorsi, N.; Tixier, F.; Tankyevych, O.; Jaouen, V.; Lucia, F.; Bourbonne, V.; Schick, U.; Badic, B.; et al. Radiomics in PET/CT: Current status and future AI-based evolutions. In *Seminars in Nuclear Medicine*; Elsevier: Amsterdam, The Netherlands, 2021; Volume 51, pp. 126–133.
20. Duff, L.; Scarsbrook, A.F.; Mackie, S.L.; Frood, R.; Bailey, M.; Morgan, A.W.; Tsoumpas, C. A methodological framework for AI-assisted diagnosis of active aortitis using Radiomic analysis of FDG PET–CT Images: Initial analysis. *J. Nucl. Cardiol.* **2022**, *29*, 3315–3331. [CrossRef] [PubMed]
21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
22. van Timmeren, J.E.; Cester, D.; Tanadini-Lang, S.; Alkadhi, H.; Baessler, B. Radiomics in medical imaging—“How-to” guide and critical reflection. *Insights Imaging* **2020**, *11*, 1–16. [CrossRef] [PubMed]
23. Papadimitroulas, P.; Brocki, L.; Chung, N.C.; Marchadour, W.; Vermet, F.; Gaubert, L.; Eleftheriadis, V.; Plachouris, D.; Visvikis, D.; Kagadis, G.C.; et al. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys. Med.* **2021**, *83*, 108–121. [CrossRef]
24. Lovinfosse, P.; Visvikis, D.; Hustinx, R.; Hatt, M. FDG PET radiomics: A review of the methodological aspects. *Clin. Transl. Imaging* **2018**, *6*, 379–391. [CrossRef]
25. Ferreira, M.; Lovinfosse, P.; Hermesse, J.; Decuypere, M.; Rousseau, C.; Lucia, F.; Schick, U.; Reinhold, C.; Robin, P.; Hatt, M.; et al. Comparison of radiomic pre-processing steps in the reproducible prediction of disease free survival across multi-scanners/centers. 2021, under review.
26. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *J. Br. Surg.* **2015**, *102*, 148–158. [CrossRef]
27. LIDA. Target. Online Resource. <https://lida.leeds.ac.uk/target-2/> (accessed on 1 November 2022).

28. Brown, P.; Zhong, J.; Froom, R.; Currie, S.; Gilbert, A.; Appelt, A.; Sebag-Montefiore, D.; Scarsbrook, A. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2790–2799. [[CrossRef](#)] [[PubMed](#)]
29. Boellaard, R.; Delgado-Bolton, R.; Oyen, W.J.; Giammarile, F.; Tatsch, K.; Eschner, W.; Verzijlbergen, F.J.; Barrington, S.F.; Pike, L.C.; Weber, W.A.; et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: Version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* **2015**, *42*, 328–354. [[CrossRef](#)]
30. Kikinis, R.; Pieper, S.D.; Vosburgh, K.G. 3D Slicer: A platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative Imaging and Image-Guided Therapy*; Springer: New York, NY, USA, 2014; pp. 277–289.
31. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341. [[CrossRef](#)]
32. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [[CrossRef](#)]
33. Dashora, H.R.; Rosenblum, J.S.; Quinn, K.A.; Alessi, H.; Novakovich, E.; Saboury, B.; Ahlman, M.A.; Grayson, P. Comparing semi-quantitative and qualitative methods of vascular FDG-PET activity measurement in large-vessel vasculitis. *J. Nucl. Med.* **2021**, *63*, 280–286. [[CrossRef](#)] [[PubMed](#)]
34. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **2020**, *295*, 328. [[CrossRef](#)] [[PubMed](#)]
35. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [[CrossRef](#)]
36. Xing, H.; Hao, Z.; Zhu, W.; Sun, D.; Ding, J.; Zhang, H.; Liu, Y.; Huo, L. Preoperative prediction of pathological grade in pancreatic ductal adenocarcinoma based on 18F-FDG PET/CT radiomics. *EJNMMI Res.* **2021**, *11*, 19. [[CrossRef](#)]
37. Visvikis, D.; Le Rest, C.C.; Jaouen, V.; Hatt, M. Artificial intelligence, machine (deep) learning and radio (geno) mics: Definitions and nuclear medicine imaging applications. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2630–2637. [[CrossRef](#)] [[PubMed](#)]
38. Langs, G.; Röhrich, S.; Hofmanninger, J.; Prayer, F.; Pan, J.; Herold, C.; Prosch, H. Machine learning: From radiomics to discovery and routine. *Der Radiol.* **2018**, *58*, 1–6. [[CrossRef](#)] [[PubMed](#)]
39. Nappi, C.; Cuocolo, A. The machine learning approach: Artificial intelligence is coming to support critical clinical thinking. *J. Nucl. Cardiol.* **2020**, *27*, 156–158. [[CrossRef](#)] [[PubMed](#)]
40. Shrestha, S.; Sengupta, P.P. Machine learning for nuclear cardiology: The way forward. *J. Nucl. Cardiol.* **2019**, *26*, 1755–1758. [[CrossRef](#)]
41. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [[CrossRef](#)] [[PubMed](#)]
42. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [[CrossRef](#)]
43. Fortin, J.P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **2018**, *167*, 104–120. [[CrossRef](#)]
44. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J. Nucl. Med.* **2018**, *59*, 1321–1328. [[CrossRef](#)]
45. Hatt, M.; Le Rest, C.C.; Tixier, F.; Badic, B.; Schick, U.; Visvikis, D. Radiomics: Data are also images. *J. Nucl. Med.* **2019**, *60*, 385–445. [[CrossRef](#)] [[PubMed](#)]
46. Fuchs, M.; Briel, M.; Daikeler, T.; Walker, U.A.; Rasch, H.; Berg, S.; Ng, Q.K.; Raatz, H.; Jayne, D.; Kötter, I.; et al. The impact of 18 F-FDG PET on the management of patients with suspected large vessel vasculitis. *Eur. J. Nucl. Med. Mol. Imaging* **2012**, *39*, 344–353. [[CrossRef](#)]
47. Sun, X.; Xu, W. Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **2014**, *21*, 1389–1393. [[CrossRef](#)]
48. Larue, R.T.; Defraene, G.; De Ruysscher, D.; Lambin, P.; Van Elmpt, W. Quantitative radiomics studies for tissue characterization: A review of technology and methodological procedures. *Br. J. Radiol.* **2017**, *90*, 20160665. [[CrossRef](#)] [[PubMed](#)]
49. Sollini, M.; Antunovic, L.; Chiti, A.; Kirienko, M. Towards clinical application of image mining: A systematic review on artificial intelligence and radiomics. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2656–2672. [[CrossRef](#)]
50. Piri, R.; Edenbrandt, L.; Larsson, M.; Enqvist, O.; Nøddeskou-Fink, A.H.; Gerke, O.; Høilund-Carlsen, P.F. Aortic wall segmentation in 18F-sodium fluoride PET/CT scans: Head-to-head comparison of artificial intelligence-based versus manual segmentation. *J. Nucl. Cardiol.* **2022**, *29*, 2001–2010. [[CrossRef](#)]
51. Zerizer, I.; Tan, K.; Khan, S.; Barwick, T.; Marzola, M.C.; Rubello, D.; Al-Nahhas, A. Role of FDG-PET and PET/CT in the diagnosis and management of vasculitis. *Eur. J. Radiol.* **2010**, *73*, 504–509. [[CrossRef](#)]
52. Soussan, M.; Nicolas, P.; Schramm, C.; Katsahian, S.; Pop, G.; Fain, O.; Mekinian, A. Management of large-vessel vasculitis with FDG-PET: A systematic literature review and meta-analysis. *Medicine* **2015**, *94*, e622. [[CrossRef](#)]
53. Tatsumi, M.; Cohade, C.; Nakamoto, Y.; Wahl, R.L. Fluorodeoxyglucose uptake in the aortic wall at PET/CT: Possible finding for active atherosclerosis. *Radiology* **2003**, *229*, 831–837. [[CrossRef](#)]

54. Espitia, O.; Schanus, J.; Agard, C.; Kraeber-Bodéré, F.; Hersant, J.; Serfaty, J.M.; Jamet, B. Specific features to differentiate Giant cell arteritis aortitis from aortic atheroma using FDG-PET/CT. *Sci. Rep.* **2021**, *11*, 17389. [[CrossRef](#)]
55. Nielsen, B.D.; Gormsen, L.C.; Hansen, I.T.; Keller, K.K.; Therkildsen, P.; Hauge, E.M. Three days of high-dose glucocorticoid treatment attenuates large-vessel 18F-FDG uptake in large-vessel giant cell arteritis but with a limited impact on diagnostic accuracy. *Eur. J. Nucl. Med. Mol. Imaging* **2018**, *45*, 1119–1128. [[CrossRef](#)]
56. Stellingwerff, M.D.; Brouwer, E.; Lensen, K.J.D.; Rutgers, A.; Arends, S.; Van Der Geest, K.S.; Glaudemans, A.W.; Slart, R.H. Different scoring methods of FDG PET/CT in giant cell arteritis: Need for standardization. *Medicine* **2015**, *94*, e1542. [[CrossRef](#)]
57. Van der Geest, K.; Treglia, G.; Glaudemans, A.; Brouwer, E.; Sandovici, M.; Jamar, F.; Gheysens, O.; Slart, R. Diagnostic value of [18F] FDG-PET/CT for treatment monitoring in large vessel vasculitis: A systematic review and meta-analysis. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 3886–3902. [[CrossRef](#)] [[PubMed](#)]
58. Ford, R.A.; Price, W.; Nicholson, I. Privacy and accountability in black-box medicine. *Mich. Telecommun. Technol. Law Rev.* **2016**, *23*, 1.
59. Ibrahim, A.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; Lambin, P.; et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers* **2021**, *13*, 1848. [[CrossRef](#)] [[PubMed](#)]
60. Orhac, F.; Buvat, I. Comment on Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848. *Cancers* **2021**, *13*, 3037. [[CrossRef](#)] [[PubMed](#)]
61. Orhac, F.; Eertink, J.J.; Cottreau, A.S.; Zijlstra, J.M.; Thieblemont, C.; Meignan, M.A.; Boellaard, R.; Buvat, I. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J. Nucl. Med.* **2021**, *63*, 172–179. [[CrossRef](#)]
62. Bettinelli, A.; Marturano, F.; Avanzo, M.; Loi, E.; Menghi, E.; Mezzenga, E.; Pirrone, G.; Sarnelli, A.; Strigari, L.; Strolin, S.; et al. A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. *Radiology* **2022**, *303*, 211604. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.