Intelligent feature analysis of FDG PET - CT images for more accurate diagnosis in large vessel vasculitis

Lisa Mairi Duff

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

UNIVERSITY OF LEEDS

University of Leeds
School of Mechanical Engineering
Institute of Medical and Biological Engineering

November 2022

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The candidate, Lisa M Duff, conducted the majority of the work presented in this thesis. This includes but is not limited to a portion of the data collation, processing input data, developing and validating a radiomic workflow, and analysing all results. This is the case for all publications mentioned below. Contributions from others in these publications are listed.

The chapter titled 'Background' is partially based on work from our publication:

Book Chapter – "Automated Diagnosis and Prediction in Cardiovascular Diseases Using Tomographic Imaging" in Big Data in Multimodal Medical Imaging (2019), Taylor and Francis Group

Authors - Lisa Duff, Charalampos Tsoumpas

This work was conducted by the candidate with advice and proofreading from Charalampos Tsoumpas.

The chapter titled 'Experiment Set 1 - Method Development' is based on work from our publication:

Journal Article – "A methodological framework for AI-assisted diagnosis of active aortitis using radiomic analysis of FDG PET–CT images: Initial analysis"

Authors - Lisa Duff, Andrew F Scarsbrook, Sarah L Mackie, Russell Frood, Marc Bailey, Ann W Morgan and Charalampos Tsoumpas Journal of Nuclear Cardiology (2022), pp. 1–17.

https://doi.org/10.1007/s12350-022-02927-4

The chapter titled 'Experiment Set 2 - Method Automation and Validation' is based on work from our publication:

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

Authors - Lisa M. Duff, Andrew F. Scarsbrook, Nishant Ravikumar, Russell Frood, Gijs D. van Praagh, Sarah L. Mackie, Marc A. Bailey, Jason M. Tarkin, Justin C. Mason, Kornelis S. M. van der Geest, Riemer H. J. A. Slart, Ann W. Morgan and Charalampos Tsoumpas Biomolecules 13(2), 343 (2023)
https://doi.org/10.3390/biom13020343

Andrew F Scarsbrook, Sarah L Mackie, Marc Bailey, Ann W Morgan and Charalampos Tsoumpas are the candidate's PhD supervisors. Their contributions included advice and discussions about the results, helping collate the required data and help accessing the required software and hardware to conduct the experiments. They also contributed to the paper writing process with proof reading and advice about the best paper form.

Russell Frood conducted similar experiments and helped with trouble shooting when problems in the method and code arose. Louise Sorensen conducted clinical data collection and Pratik Adusumilli shared data for segmentation validation.

In the chapter titled 'Experiment Set 2 - Method Automation and Validation' the above contributions apply along with the following.

Nishant Ravikumar helped develop the initial version of the automated segmentation method used in this chapter.

Alliance Medical, Jason M Tarkin and Justin C Mason contributed data from their respective studies for use in validation. This process was supported by Johann Alberts, Brad Miles and Roie Manavaki.

Acknowledgements

Doing a PhD has been a transformational experience. Firstly, I would like to thank my PhD supervisors Andrew F Scarsbrook, Sarah L Mackie, Marc Bailey, Ann W Morgan, and Charalampos Tsoumpas whose guidance and feedback made this project possible. Their support in the ups but especially the downs of the last 4 years has been appreciated greatly. Additionally, I would like to thank my Centre for Doctoral Training directors Claire Brockett and Anthony Herbert for making my time in Leeds so well rounded, and enjoyable, and for always going the extra mile when I needed help.

The help provided by Nishant Ravikumar, Russell Frood, and Harry Tunnicliffe was invaluable for overcoming problems with deep learning, radiomics and python coding respectively. I kindly acknowledge the clinical data collection work conducted by Louise Sorensen, the data shared by Pratik Adusumilli, and the infrastructure support from MRC TARGET, LICAMM, St. James's University Hospital and the University of Leeds. This work was undertaken on ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. I also acknowledge Johann Alberts and Brad Miles from Alliance Medical who provided anonymized external imaging datasets, Jason M Tarkin and Roie Manavaki for providing data from the PITA study, and similarly Justin C Mason for contributing validation data. I would like to highlight Justin Mason's contributions specifically as he sadly passed away earlier this year. I only had a few interactions with him while he was helping with my project but his kindness left a lasting impression.

I would like to give a sincere thank you to the patients who gave consent for their information to be used in this project. Without them none of this would have been possible.

Thank you to everyone in the Department of Nuclear Medicine and Molecular Imaging, University of Groningen for welcoming me into your department. In particular Gijs van Praagh for his help in validating my results, and Philipp Mohr, Laura Providência, Mostafa Roya, and Samaneh Mostafapour for their friendship when I was away from home.

My time at the University of Leeds provided me with too many life long friends to name but I hope I have let them know who they are. Special mentions go to my PhD cohort for being there from day one, my undergraduate friends - Jess, Kathleen and Emily

- for listening to all my stressed out ramblings, and Richa Gandhi - the only other student to be in the same PhD programme and research group - for being a fantastic friend and example to follow.

Last but not least, I have to thank my family. I cannot put into words the love and appreciation I have for my siblings and parents. Eilidh and Calum, thank you for being my best friends since the day you were born. Together we make a great team and I could not have finished this PhD without your encouragement. Mum and Dad, it will probably be a while before I fully understand everything you have ever done for me but I do not take you for granted. Thank you for everything, from practical help such as having me back home when a pandemic appears 18 months into my PhD, to more abstract help like supporting all of my dreams and passions. I love you all.

Abstract

Aortitis refers to inflammatory conditions affecting the aortic wall that cannot be explained by atherosclerosis alone. Large Vessel Vasculitis (LVV) is the main type of non-infectious aortitis and can affect any of the large arteries. Aortitis and LVV are difficult to diagnose and treat due to a variety of reasons such as non-specific symptoms and diagnostic tests, a large number of potential causes, and risks involved in providing incorrect or delayed treatment. [18F]-Fluorodeoxyglucose Positron Emission Tomography-Computed Tomography (FDG PET-CT) imaging plays a key role in diagnosis of LVV due to its ability to detect inflammation early and non-invasively, but is mostly assessed qualitatively making its interpretation vulnerable to bias and inter-observer variation. Therefore, there is a need for more reliable imaging biomarkers which can be achieved with radiomic analysis.

The aim of this thesis is to explore the diagnostic ability of radiomic features in FDG PET-CT imaging of aortitis. First, the feasibility is determined and a methodological pipeline established. Next, the findings are validated with data from multiple centres and the overall method automated. The first step of the method is the aortic segmentation using an artificial intelligence which produces similar radiomic features as the manual segmentation. When used as input in the diagnostic models, several individual radiomic features and groups of radiomic features demonstrated high diagnostic performance across the training, test and validation cohorts. In particular features based on heterogeneity perform well. The method displayed good generalizability and transferability which is important prerequisites for clinical use. These findings could be used to build an automated clinical decision tool which would facilitate objective and standardized assessment regardless of observer experience.

Contents

List of Tables

List of Equations

List of Figures

List of Abbreviations

6P - 6-Phosphate

AI - Artificial Intelligence

AML - Alliance Medical

AUC - Area Under the ROC Curve

BLOB-OS-TF - Spherically symmetric basis function ordered subset algorithm

CNN - Convolutional Neural Network

CRP - C-Reactive Protein

CT - Computed Tomography

CTA - Computed Tomography Angiography

DICOM - Digital Imaging and Communications in Medicine

DL - Deep Learning

DLYD - Delayed Event Subtraction

DSC - Dice Similarity Coefficient

EANM - European Association of Nuclear Medicine ESR - Erythrocyte Sedimentation Rate

EULAR - European Alliance of Associations for Rheumatology

FDG [^{18}F]-Fluorodeoxyglucose

FDG - [^{18}F]-Fluorodeoxyglucose

GCA - Giant Cell Arteritis

G6Pase - Glucose 6-Phosphatase

GPU - Graphics Processing Unit

GLCM - Gray-Level Co-Occurrence Matrix

GLDM - Gray-Level Dependence Matrix

GLRLM - Gray-Level Run Length Matrix

GLSZM - Gray-Level Size Zone Matrix

IBSI - International Biomarker Standardisation Initiative

IgG4 - IgG4 related disease

LVV - Large Vessel Vasculitis

ML - Machine Learning

MLEM - Maximum-Likelihood Expectation Maximization MRA - Magnetic Resonance Angiography
MRI - Magnetic Resonance Imaging
OSEM - Ordered Subsets Expectation Maximization
PACS - Picture Archiving and Communication
PCA - Principal Component Analysis
PET-CT - Positron Emission Tomography – Computed Tomography
PET - Positron Emission Tomography
PETVAS - PET Vascular Activity Score
PITA - PET Imaging of Giant Cell and Takayasu Arteritis
PMR - Polymyalgia Rheumatica
PSF - Point Spread Function
PMT - Photomultiplier Tubes
ReLU - Rectified Linear Unit
ROC - Receiver Operating Characteristic
ROI - Region of Interest
RPF - Retroperitoneal Fibrosis
SNMMI - Society of Nuclear Medicine and Molecular Imaging
SS-SIMUL - Single-scatter Simulation
SSS - Single-scatter Simulation
SUV - Standardized Uptake Value
TARGET - Treatment According to Response in Giant Cell arTeritis
TAK - Takayasu's arteritis
TCA - Tricarboxylic Acid
TLG - Total Lesion Glycolysis
TOF - Time-of-Flight US - Ultrasound
VPFX - Vue Point FX (3D time of flight)

Chapter 1

Preface

Overview of Thesis

Aortitis and large vessel vasculitis (LVV) are notoriously difficult to diagnose and treat. This is due to a variety of reasons such as non-specific symptoms and diagnostic tests, a large number of potential causes, and risks involved in providing incorrect or delayed treatment. Symptoms can include fever, headache, weight loss and lethargy which may not immediately be picked up as potential aortitis or LVV [1]. Similarly diagnostic tests include a combination of laboratory tests, imaging and review of clinical signs and symptoms which are non specific [2, 3]. Aortitis can be caused by a range of infectious or non-infectious conditions. More information about the underlying cause can prevent the use of immunosuppression in infectious cases and to prevent delays in providing the correct treatment that could reduce the risk of complications. Imaging plays an important role in the diagnosis of aortitis but is mostly assessed qualitatively making it vulnerable to bias and inter-observer variation.

Applying radiomic analysis and machine learning to [18F]-Fluorodeoxyglucose Positron Emission Tomography-Computed Tomography (FDG PET-CT) scans of aortitis patients produces a large quantity of previously unexplored information contained within the images. These imaging biomarkers can be used to address some of the difficulties in diagnosis and treatment by constructing diagnostic models that can aid in clinical decision-

making while also standardizing assessment.

This thesis is comprised of four central chapters which explore the development of a FDG PET radiomic pipeline to assist in the diagnosis of aortitis.

An overview of the chapters is presented below.

Background: This chapter describes in detail the fundamental concepts behind LVV / aortitis, PET imaging , and radiomic analysis. It explores the questions surrounding the diagnosis of aortitis and the theory behind the methods used to address these questions.

Experiment Set 1 - Method Development: The first goal of this project was to explore the diagnostic utility of radiomic features in aortitis. This chapter serves as both a proof of concept and explores developing a method using our initial single centre datasets.

Experiment Set 2 - Method Automation and Validation: The second set of experiments look to validate the findings of the previous chapter in a multi centre study. Another goal of these experiments was to automate the process using a convolutional neural network to segment the aorta as this was a bottle neck in earlier versions of this pipeline.

Synopsis: This chapter summarises the findings of the project starting with the proof of concept, development of the method and then validation. Limitations both in this study and in the field of radiomics more generally are discussed and a comprehensive analysis of where this work may progress is then given.

Key Contributions

The key contributions of this PhD are described below.

1. Several radiomic features and combinations of radiomic features have shown to have diagnostic utility in aortitis similar to that of the current standard of care, qualitative assessment.

2. An automated aortic segmentation method was developed and validated.

3. An open access automated methodology has been established which could standardize diagnosis of aortitis regardless of experience on observer variability. (https://github.com/LisaDuff/ClassificationRadiomicsModelBuilder)

4. Our findings and methodology open up the possibility of radiomic analysis to provide information about treatment response, outcome prediction and subtype classification within aortitis and large vessel vasculitis.

Outputs

Publications

Book Chapter – "Automated Diagnosis and Prediction in Cardiovascular Diseases Using Tomographic Imaging" in Big Data in Multimodal Medical Imaging (2019), Taylor and Francis Group. Authors - Lisa Duff, Charalampos Tsoumpas.

Journal Article – "A methodological framework for AI-assisted diagnosis of active aortitis using radiomic analysis of FDG PET–CT images: Initial analysis". Authors - Lisa Duff, Andrew F Scarsbrook, Sarah L Mackie, Russell Frood, Marc Bailey, Ann W Morgan and Charalampos Tsoumpas. Journal of Nuclear Cardiology (2022), pp. 1–17. https://doi.org/10.1007/s12350-022-02927-4

Journal Article – "Exploring the utility of radiomic feature extraction to improve the diagnostic accuracy of cardiac sarcoidosis using FDG PET". Authors - Nouf A Mushari, Georgios Soultanidis, Lisa Duff, Maria G Trivieri, Zahi A Fayad, Philip Robson, and Charalampos Tsoumpas. Frontiers in medicine 9 (2022). https://doi.org/10.3389/fmed.2022.840261

Journal Article – "An automated method for AI assisted diagnosis of active aortitis using radiomic analysis of FDG PET-CT images". Authors - Lisa M. Duff, Andrew F. Scarsbrook, Nishant Ravikumar, Russell Frood, Gijs D. van Praagh, Sarah L. Mackie, Marc A. Bailey, Jason M. Tarkin, Justin C. Mason, Kornelis S. M. van der Geest, Riemer H. J. A. Slart, Ann W. Morgan and Charalampos Tsoumpas. Biomolecules 13(2), 343 (2023). https://doi.org/10.3390/biom13020343

Journal Article (Submitted) – "Exploring the utility of cardiovascular magnetic resonance radiomic feature extraction of cardiac sarcoidosis". Authors - Nouf A. Mushari *, Georgios Soultanidis, Lisa Duff, Maria G. Trivieri, Zahi A. Fayad, Philip Robson, Charalampos Tsoumpas.

Journal Article (To be submitted) – "Fully automated multilabel segmentation, quantification, and visualization of the diseased aorta on hybrid PET/CT". Authors - G.D. van Praagh, P.H. Nienhuis, M. Reijrink, M. Davidse, L.M. Duff, B.S. Spottiswoode, D.J. Mulder, N.H.J. Prakken, C. Tsoumpas, J.M. Wolterink, K.B. Mouridsen, R.J.H. Borra, B. Sinha, R.H.J.A. Slart

Conferences

IEEE Nuclear Science Symposium and Medical Imaging Conference

2019 (Workshop Presentation) – ‘Automated Diagnosis and Prediction in Cardiovascular Diseases Using Tomographic Imaging’

European Molecular Imaging Meeting

2020 (Poster) – ‘Automated Diagnosis in Large Vessel Vasculitis using FDG PET-CT’

2021 (Poster) – ‘A multicentre study of AI-assissted diagnosis of active aortitis using radiomic analysis of FDG PET-CT images’

European Association of Nuclear Medicine Congress

2021 (Poster and Poster Presentation) – ‘Methodological Framework for AI-assisted Diagnosis of Active Aortitis using Radiomic Analysis of FDG PET-CT’

2022 (Oral Presentation) – ‘Automated AI-assisted diagnosis of active aortitis using radiomic analysis of FDG PET-CT imaging’

EANM’22 Congress – Short listed for Young Authors Award

Declarations

This PhD 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 number EP/L014823/1. Prof. Morgan is 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. Dr. Bailey is funded by a British Heart Foundation intermediate clinical research fellowship (FS/18/12/33270) and Prof. Tsoumpas by a Royal Society Industry fellowship (IF170011). Dr. Frood and Prof. Scarsbrook receive salary support from Innovate UK via the National Consortium for Intelligent Medical Imaging. Prof Scarsbrook acknowledges academic salary support from Leeds Hospitals Charity. Dr Sarah Mackie is supported by the NIHR Leeds Biomedical Research Centre. Dr Tarkin is supported by a Wellcome Trust Clinical Research Career Development Fellowship [211100/Z/18/Z]. Prof. Mason was 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 (National Health Service), the NIHR or the Department of Health and Social Care. Conflicts of interest/Competing interests - None to declare. The institutional research data access committee confirmed that formal ethics committee approval was not required for this study which was considered to represent evaluation of an established clinical service. Routinely collected patient meta-data was extracted by the clinical direct care team and rendered pseudo-anonymous for the purposes of analysis within this study and the institutional clinical governance team confirmed that this was also exempt from formal research ethics committee approval. Prospective written consent was obtained from all patients at the time of imaging for use of their anonymised FDG PET-CT imaging data in research and service development projects. All patient data were prospectively entered into a departmental database used for retrospective identification and audit. Availability of data and material Available on reasonable request. Code availability Available on reasonable request.

Chapter 2

Background

2.1 Aortitis and Large Vessel Vasculitis

2.1.1 Overview

The aorta is the main artery in the human body and runs from the heart down the centre of the chest and abdomen. Blood enters it from the heart via the aortic valve and from there blood is distributed through the body via other arteries and then returned to the heart through veins. Aortitis refers to inflammatory conditions affecting the aortic wall that cannot be explained by atherosclerosis alone [1, 4, 5]. It can arise as an isolated condition or in association with other infectious or non-infectious diseases, where non-infectious causes are more common [1, 6]. Large vessel vasculitis (LVV) is the main type of non-infectious aortitis and can affect the aorta and any of its branches (Figure 2.1). Giant cell arteritis (GCA) and Takayasu arteritis (TAK) are the most common types of LVV [1, 7]. Immunoglobulin-G4-related disease (IgG4-RD) commonly has peri-aortic involvement and is sometimes classified as a LVV rarer sub-type [8, 9]. The causes and processes of aortitis can overlap with peri-aortitis where inflammation extends past the aortic wall to the periaortic space [5].

GCA was initially considered a condition that affected branches of the external carotid and vertebral arteries but modern imaging techniques have shown large vessel involvement in up to 83% of patients [10–12]. GCA is now further categorised as cranial GCA

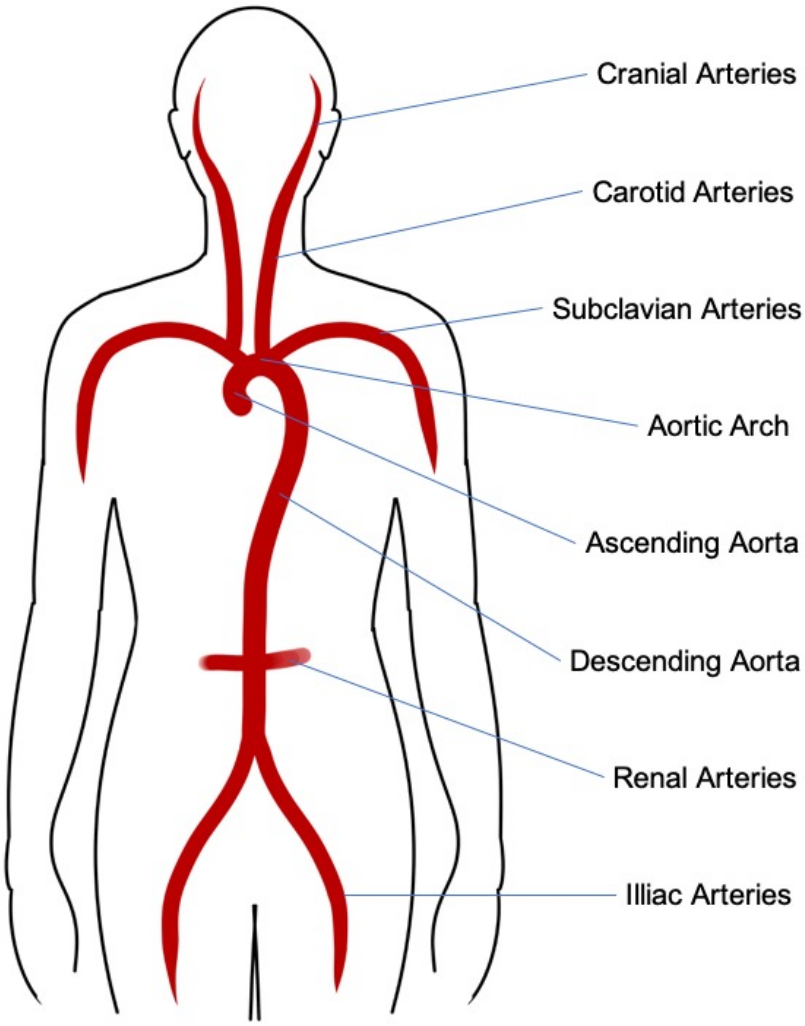


Figure 2.1: Arteries of the body most commonly affected by large vessel vasculitis

or large vessel GCA although both can be present [13, 14]. To be called aortitis, large vessel involvement including the aorta would need to be present. GCA affects women more than men, is most prevalent in people with northern european heritage, and is rarely found in patients younger than 50 years old [15, 16]. Symptoms vary depending on the vascular territories affected but aortic involvement often presents with systemic symptoms, such as fever, weight loss and lethargy [1]. Cranial GCA symptoms include jaw claudication (pain associated with reduced blood flow during chewing), headache, scalp tenderness and vision loss [6, 17, 18]. Polymyalgia Rheumatica (PMR) is an overlapping inflammatory disorder that occurs along side about 40% of GCA cases adding symptoms of of myalgia, arthralgia and stiffness [15]. The risk of aortic aneurysm formation is increased in GCA, along with greater aneurysm growth rate and risk of dissection [19].

TAK presents similarly to GCA in terms of constitutional symptoms in the early stages. It also predominantly affects women. The key difference is age of diagnosis which generally occurs between the ages of 20 - 40 years old in TAK [20]. However TAK is much rarer than GCA and this, along with subtle initial symptoms can cause a delay in diagnosis [21] meaning some studies have populations with an average age of diagnosis above 40 years old [22].TAK is often referred to as the pulse-less disease due to ischemic complications in later stages from stenosis and occlusion, which occur more frequently compared to GCA [21].

Both GCA and TAK are rare with the exact prevalence hard to define for this reason. One retrospective study found a mean incidence rate for GCA of 112 people per 1 million [23, 24], and another study found a mean rate for TAK of 0.8 people per million[25]. These are both UK based studies and the prevalence varies with location and race [16].

Regardless of the cause of aortitis it can lead to significant morbidity [26, 27]. A major cause of morbidity is vascular complications such as stenosis, stroke,sight loss, aortic dilatation and aneurysm formation [28, 29]. Morbidity can also be attributed to side-effects of treatment as described in Section 2.1.2 [30]. There can be a diagnostic delay in GCA, particularly if symptoms such as headaches are absent, and in TAK due to subtle initial symptoms. This can lead to decreases in quality of life such as loss of sight or stroke, particularly in GCA.

Mortality is significantly higher in infectious aortitis [1, 31]. Exact mortality rates in

non-infectious aortitis are varied and conflicting but a general higher mortality rate has been observed in those with non-infectious aortitis compared to similar demographics without it [32, 33]. This contradicts earlier studies [34]. There is also a shown trend of aortitis increasing mortality of conditions such as aneurysms and cardiovascular disease when compared to patients without aortitis [1, 35]. It is worth noting that due to the higher prevalence of GCA there is a natural bias toward those patients in aortitis mortality studies. The younger age demographic affected by TAK means patients with TAK have a higher mortality rate compared to similar controls who are statistically less likely to die of other causes making the increase in mortality more noticeable [29].

2.1.2 Treatment

Treatment for aortitis usually includes glucocorticoids and/or immunosuppressants [36, 37]. GCs can induce remission but relapses during or after discontinuation of treatment are common resulting in prolonged treatment. This increases the risk of glucocorticoid related toxicity which can lead to infection, diabetes, cardiovascular disease, osteoporosis or fracture [38–41]. It is recommended that glucocorticoids start at a higher dose, 40-60mg, and are gradually tapered to 15-20mg within a few months and 5-10mg in a year [37]. This is to balance the risks of toxicity and flare on discontinuation.

Other treatments include immunosuppressants, such as methotrexate, and more recently Tocilizumab [42]. Which treatment is used can depend on disease activity state. Tocilizumab in particular is mostly used to treat relapsing or refractory disease with methotrexate used as an alternative [37].

Treatment of aortitis also includes treatment of drug side effects and complications from aortitis or LVV itself. For example therapies can be applied to mitigate risks of osteoporosis or infections. Interventions may be required in the case of complications such as stroke, vision loss, aneurysm or stenosis in limbs.

Treatment for GCA and TAK are generally similar but different strategies may be adopted based on age, symptoms or vascular territories and the risk of complications associated with these factors.

Accurate, and quick diagnosis of aortitis is essential in order to treat correctly and

minimise patient discomfort.

2.1.3 Diagnosis

Reliable diagnosis of aortitis and its underlying cause is essential to disease management [1]. While infectious aortitis is much less common than non-infectious causes, it is essential to exclude it to prevent the incorrect use of immunosuppression in treatment. It is essential to provide the correct treatment but any delays in providing this treatment can risk complications.

There are several diagnostic pathways a patient may go through, starting with symptoms arising or incidental histological (Figure 2.2) or radiological findings (Figures 2.3, 2.4, and 2.5[1]). A combination of laboratory tests, imaging and review of clinical signs and symptoms are used to then confirm diagnosis and determine the underlying cause [1]. Diagnosis of aortitis remains challenging as laboratory markers (C reactive protein (CRP) and erythrocyte sedimentation rate (ESR)) and aortitis symptoms are non specific and not always elevated [2, 3]. Other common tests such as temporal artery biopsy are also invasive, can give false negatives and are not always feasible if the inflamed vascular territory cannot be accessed.

Imaging plays an important role in diagnosis as it can help overcome some of these problems [1, 14]. Imaging techniques for LVV include ultrasound (US) of the temporal and axillary arteries, magnetic resonance imaging and computed tomography with or without angiography (MR(A) and CT(A))(Figure 2.3, and 2.4), and Positron Emission Tomography (PET) (Figure 2.5) [14, 45]. US, MR(A) and CT(A) are considered anatomical imaging and can be used to assess structural changes caused by aortitis. Wall thickening can be used to assess inflammation but is non-specific since it can also represent vascular damage. Wall thickening can be identified with MR(A), CT(A), and US [1, 45, 46]. Besutti *et al.* set a threshold of 3mm to qualify as wall thickening [47]. Muto *et al.* found a mean wall thickness of 3.8mm in LVV patients compared to 2.6mm in controls [48]. PET

[1]Scale of radiological images in Figures unclear/varied. Chang *et al.* found the average proximal descending aortic diameter to be 24.8 ± 3.4 mm in a general hospital population [43]. Mensel *et al.* found the average aortic wall thickness in a general population to be 1.26mm in females and 1.36mm in males [44].

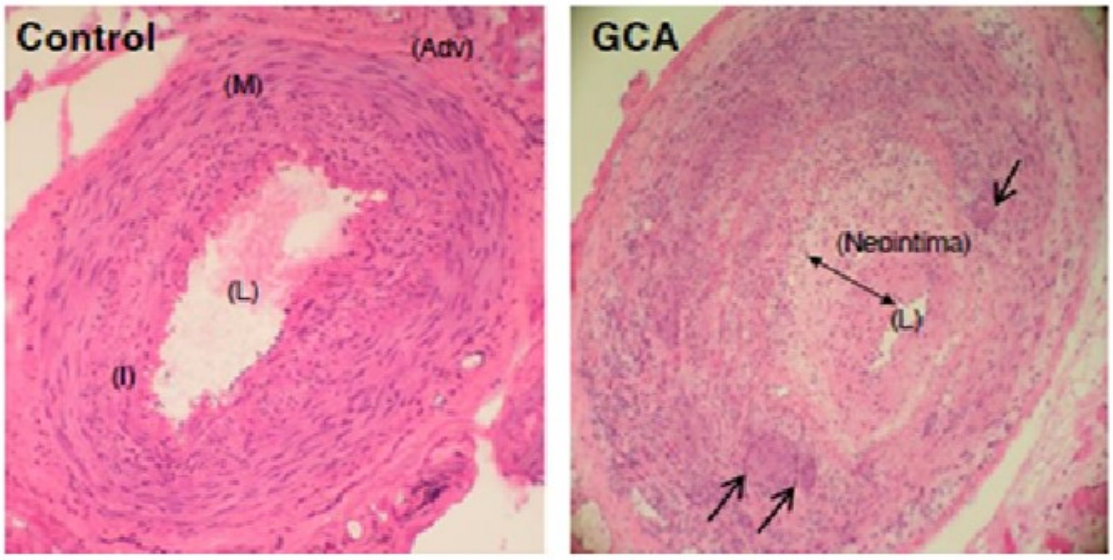


Figure 2.2: Histological analysis of temporal artery biopsies from a control patient (left) and a LVV(GCA) patient (right). Image from Planas-Rigol *et al.* is licensed under Creative Commons Attribution License [51]. Exact scale unclear but Gajree *et al.* found a mean temporal artery biopsy diameter of 2.35mm [52].

imaging is used to observe functional processes within the body and in the case of LVV highlight inflammation. It is not used for anatomical observation but PET activity can be a precursor for angiographic changes [49]. Due to improvements in vascular imaging more precise characterization of LVV and its sub-types has been achieved. While FDG PET-CT is a useful imaging tool for LVV it can also highlight other vascular conditions, mainly atherosclerosis. The two conditions can appear similarly in FDG PET-CT but a review by Nienhuis *et al.* showed that the high intensity diffuse pattern of LVV can be used with some success to differentiate between them [50]. Along side this calcification in the arteries is visible by the accompanying CT.

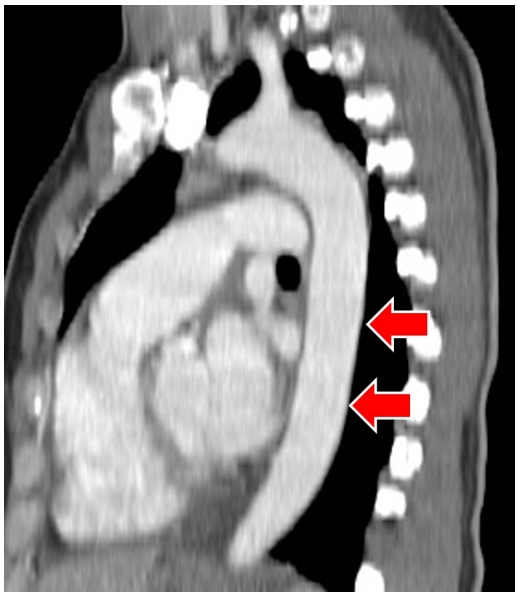
2.2 PET

Positron emission tomography (PET) is a molecular imaging technique that is used to observe functional processes within the body [53, 54]. Patients receive radioactive tracers, known as radiotracers, usually by injection, and the radiotracers distribute throughout the body similarly to their non-radioactive counterparts. The tracer contains a radioisotope in place of another constituent of the molecule. This radioisotope emits a positron which annihilates with a nearby electron producing two gamma rays 180°from one another (Figure 2.6), and the detection of pairs of gamma rays produce three-dimensional PET images. The distribution of the radiotracer is used to determine if the functional processes are working as expected. PET is used in the diagnosis and monitoring of several conditions. As PET scans are used to image processes they are often paired with CT or Magnetic resonance imaging (MRI) to provide anatomical reference [55].

2.2.1 PET Technical Aspects

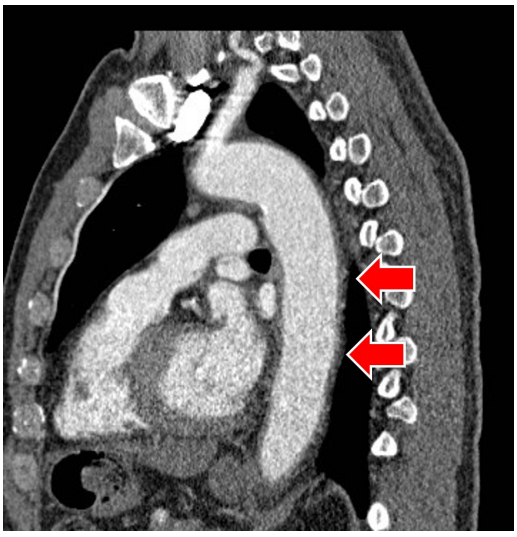
The fundamental steps of PET imaging are set out in Figure 2.6.

Atoms have a central positively charged nucleus and surrounding electron shells. The emitted positron originates from the nucleus [56]. Once emitted a positron will travel until it meets it's anti-particle, the electron. Small interactions with other particles and their electric and magnetic fields will cause the path of the positron to not be straight.



(a) Computed tomography angiography of a healthy aorta showing minimal wall thickness

(b) Computed tomography angiography of an aorta with atherosclerotic plaques and a co-occuring aneurysm



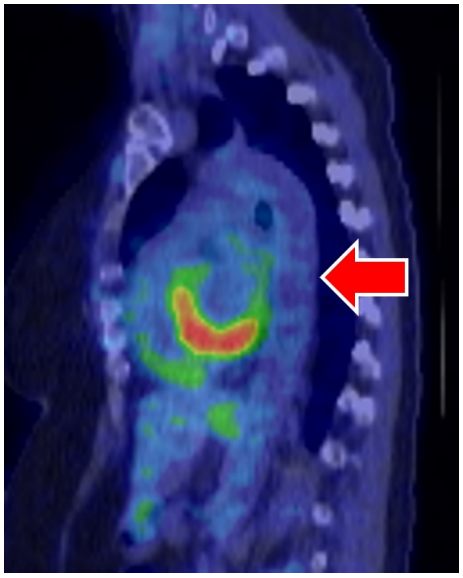
(c) Computed tomography angiography of large vessel vasculitis showing wall thickening

Figure 2.3: Computed tomography angiography of a) healthy aorta (aortic arch and descending aorta), b) atherosclerotic aorta (descending aorta), and c) large vessel vasculitis (aortic arch and descending aorta).

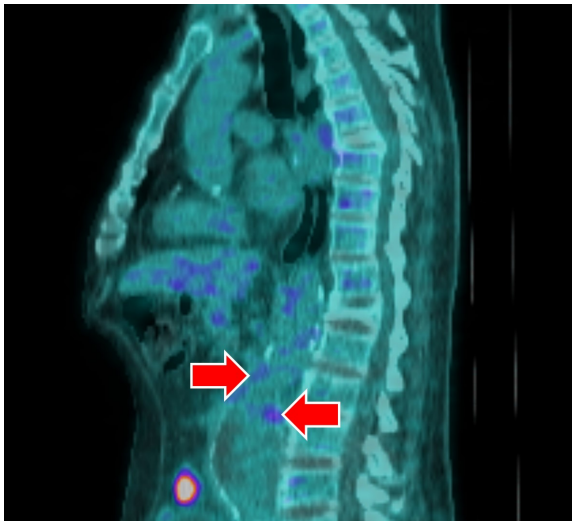
(a) Magnetic resonance angiography pre-treatment

(b) Magnetic resonance angiography post-treatment showing less intense signal in the vessel wall

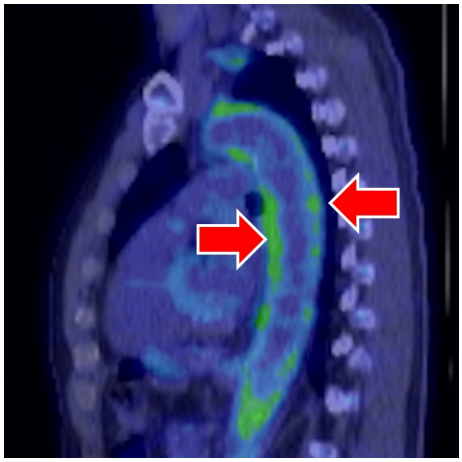
Figure 2.4: Magnetic resonance angiography of aortic arch and descending aorta in a) pre-treatment large vessel vasculitis, b) post-treatment large vessel vasculitis



(a) Positron emission tomography imaging of a healthy aorta



(b) Positron emission tomography imaging of ahtersclerosis showing active hotspots



(c) Positron emission tomography imaging of large vessel vasculitis showing a diffuse uptake pattern

Figure 2.5: Positron emission tomography imaging of a) a healthy aorta (aortic arch and descending aorta), b) atherosclerosis (descending aorta), c) large vessel vasculitis (aortic arch and descending aorta)

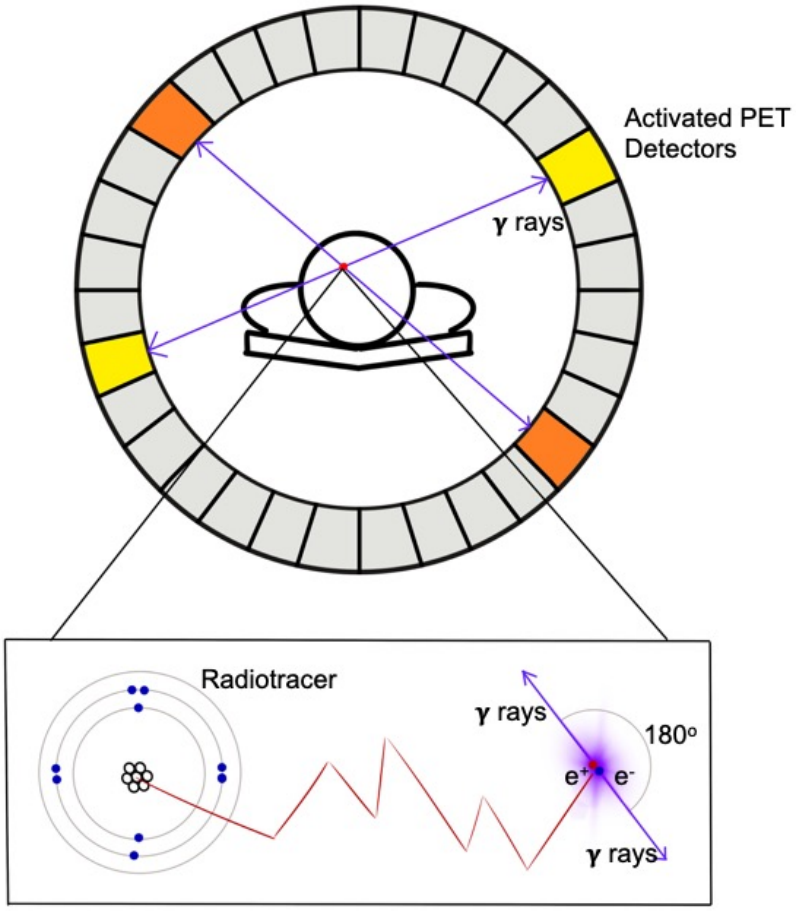


Figure 2.6: Positron emission tomography (PET) scanning - emitted positron travels until annihilation with electron producing anti-parallel gamma (γ) rays detected by PET detectors. Detected coincidence events (one example in yellow and another in orange) are detected and used to build an image representing radiotracer distribution.

The measured positron range varied depending on the radiotracer used, detector size and system diameter [57]. [18F]-Fluorodeoxyglucose ([^{18}F]-FDG), a widely used radiotracer and used for aortitis assessment, had a maximum radius of 2.4mm in water in a recent study [58]. This annihilation produces two gamma rays with an energy of 511keV at 180°from each other [59].

A PET detectors main components are scintillation crystals and photodetectors[59]. Scintillation crystals in the PET detectors detect the gamma photons, absorb those photons, and convert them to visible or UV light. This light is detected by photodetectors producing an electrical signal [60, 61]. Four important characteristics of scintillation crystals are the stopping power, decay constant, energy resolution and light output. Higher stopping power results in a shorter travel distance in crystal. This is favourable for a better detection efficiency. The decay constant dictates how long a scintillation flash occurs. Shorter flashes are better to allow for more photons to be detected. A precise energy resolution is beneficial so scattered photons can be distinguished and discounted. Finally, a high light output is preferable for better detection [59, 61].

Photo-multiplier tubes (PMTs) are a common photodetector. They take incoming light photons and produce electrons that are accelerated and amplified. The current produced is proportional to the energy deposited into crystal by pet photons. To determine the location of the detected photon many small PMTs or position sensitive PMTs are used. Crystal size is a key factor in determining resolution of PET images (approx crystal size 3mm). PET has an inherent limitation on its spatial resolution due to positron range and hardware features - especially in the detectors [62]. In this thesis a common voxel size of 4x4x4mm was used. PET can also appear blurrier than CT or MR as fewer photons can be collected per image [59]. Functions such as PSF (point spread function) can be used to improve spatial resolution [63].

Coincidences are defined as two detected gamma photons registered as from the same emission. Coincidences can be split into true coincidences and background events. Background events can be further split into randoms and scatter. Random coincidences are when two detected photons inside the time window were not from the same emission. Scatter occurs when two photons were from the same emission but were scattered so not detected along the same line of response. The produced gamma rays can scatter in

multiple ways with human tissue but the most relevant is Compton Scatter. This results in direction changes and attenuation. Approximately 7cm of human tissue will half the number of photons detected in a straight line from where they were emitted. [59].

Further errors can be introduced by the fact simultaneous emission does not mean both photons will be detected at the same time. Although two gamma photons are released at 180°from each other along the line of response it usually occurs closer to one side. A detector can also have timing uncertainty caused by decay time of the scintillation crystals and the PMT processing time. These factors together makes a time window where detections could be same emission. Time-of-flight (TOF) PET imaging allows for better localisation of signal due to pico-second differences in incident photon arrival times [56, 64–66] (Equation 2.1).

$$location = \frac{speed\,of\,light \times (time\,2 - time\,1)}{2} \tag{2.1}$$

All of the detected coincidences passing through a single point are represented as a sinogram. Several corrections are applied to these either before or during the reconstruction of an image. Attenuation correction is applied as photons are attenuated on their way out of the body attenuation correction needs to be applied to give a better representation of radiotracer distribution in the body. Without it areas such as the lungs would look disproportionately active as they are a less dense tissue. Attenuation is determined for all lines of response using CT. As discussed earlier PET is also influenced by scatter and randoms. Many scatter correction techniques have been proposed such as convolution subtraction [67] and Monte-Carlo modelling [68] , Single Scatter Simulation(SSS) [69], model based [70]. Within this thesis model-based and SSS were most commonly used by scanner manufacturers. Model based scatter correction and SSS are similar with SSS being simpler [71]. Both calculate the mean number of scattered coincidences in the data using information such as the attenuation map, emission data, the mechanism of Compton Scattering, and information regarding scanner geometry and detector systems

to calculate the mean number of scattered coincidences. Finally random coincidences can estimated and subtracted using either singles rate estimation or delayed window method. Singles rate calculates the mean random coincidence rate based on the coincidence time window and the single photon rate for the two detectors involved in a coincidence detection. The delayed window method calculates coincidences at a delayed time (larger than the coincidence time window) as well and uses the theory that the random coincidences will be the same in both the delayed time window and the original to calculate the random coincidences [72].

Reconstruction is process of assembling the collected detections into an image of the distribution of radiotracer. As this is a computationally heavy task iterative algorithms are used to reconstruct [73]. Iterative algorithms make an initial estimate and this is compared to measured projections from the sinogram. Discrepancies are corrected and new estimate is made. This process repeats until the two converge [73]. MLEM (Maximum- likelihood expectation maximization) is an example of an iterative reconstruction method [73]. It was first described by Shepp and Vardi in 1982 [74] and as indicated by it's name, it maximizes the likelihood function to improve the estimated image. The image contrast, resolution and noise are improved as a function of the number of iterations [75]. OSEM (ordered subsets expectation maximization) is a progression from MLEM and is faster to apply [63, 76]. It divides the projections into subsets - usually equally distributed around the field of view. MLEM is applied to each subset and the result of each subset is the starting value for the following subset [77]. Methods incorporating different components can also be used, such as PSF-TOF OSEM [63]. The process of reconstructing the image can also introduce noise due to the non-negative restraint. TOF has improved image quality in this respect but it is still limited and the issue is amplified in older scanners. Boundaries between areas in PET of high and low activity can be unclear due to the partial volume effect leading to segmentation and quantification errors.

Using one scanner from this thesis as an example, GE Healthcare Discovery 710 has the following traits. It is made of a full cylindircal arrangement of detectors, the patient port is 70cm in diameter, it has an axial field of view of 15.7cm and the scan range is 2m. The acquisition is in 3D and uses a step and shoot approach to capturing images. The scintillation crystals are 4.2mm x 6.3mm x 25 mm and there are 13,824 in total. This is

followed by 256 PMTs. The dose of PET radiotracers is given in units of becquerels (Bq) where 1 Bq is one decay per second. Doses are usually of the order of magnitude of MBq. The sensitivity of a scanner can be given in counts per second (cps) per bequerel with Discovery 710 being quoted as around 7 cps/kBq. It has a coincidence window of 4.9ns and a TOF resolution of 544 ps [64, 78]. For acquisitions in this thesis it uses Model based scatter correction, Singles random correction and a voxel size of 3.65mmx3.65mmx3.27mm in resolution.

Developments are consistently being made to improve PET image quality [79, 80]. Artificial intelligence (AI) enhancement is being explored , for example improvements to attenuation correction that reduce artifacts [81] and reconstruction either directly or in-directly [82]. Scanner hardware is also progressing, for example total body PET, improves the field of view and sensitivity allowing for faster acquisition times, lower doses of radiotracer and a more complete picture of the radiotracer distribution [83, 84].

2.2.2 PET Quantification

PET can be semi-quantitatively analysed using standardized uptake value (SUV) (Eqn. 2.2). The measurement gives the ratio between the concentration of radiotracer in a pixel and the concentration of radiotracer across the entire body. Doing so normalizes the measured signal for variations in body size and injected dose and allows comparison between patients and different time points of the same patient. Abnormally high or low SUV values can signify illness or disease and variation in SUV over time can be used to monitor progression [85]. SUV can be measured per pixel but it is more common for the mean (SUV_{mean}) or maximum SUV (SUV_{max}) of a region of interest (ROI) to be determined. Other SUV metrics can be used to summarise as well as described below. SUV is commonly used as it is a useful parameter, is straightforward to calculate, reproducible, and accounts for the patient's body weight thus removing some variability [86]. However, it is affected by respiratory motion, the blood glucose level of patients and the body fat percentage [87]. These factors can introduce an element of variation between PET scans of the same patient. Physiological variation can be harder to control for and quantifying the exact extent of the variation has given varied results. Guidelines from PERCIST (PET

Response Criteria in Solid Tumors) suggest a change in SUV of at least 30% is required to ensure the change is due to a response in treatment rather than physiological variation [88, 89].

Technical factors such as image processing and scanner variability can also influence results [90–92]. The SUV_{mean} is vulnerable to the variations in the segmentation of the ROI but avoids issues with noise while the (SUV_{max}) has issues with noise but not with segmentation. Therefore, an awareness of these factors is important in the design of experiments.

$$SUV = \frac{radioactivity\,concentration}{injection\,dose\,(MBq)\,/\,patient's\,weight\,(kg)} \tag{2.2}$$

- SUV 90th Percentile – 90% of the voxels SUV value's fall below this number
- SUV mean (SUV_{mean}) – the mean SUV value in the ROI
- SUV maximum (SUV_{max}) - the maximum SUV value in the ROI
- (SUV_x) (x=50, 60, 70, 80, 90) - mean of the voxels that are equal or greater than x% of SUV maximum

2.2.3 PET in LVV

PET-CT imaging plays a key role in diagnosis of LVV due to its ability to detect inflammation early and non-invasively (Figure 2.7) [93]. Both of these features minimise discomfort and prevent possible complications evolving from physical changes in the arterial anatomy that may have otherwise been required for an imaging diagnosis [1, 94]. FDG PET-CT is recommended for early diagnosis in LVV by European Alliance of Associations for Rheumatology (EULAR, formerly called European League Against Rheumatism) [53]. It is also recommended in cranial GCA but with a lesser priority as PET imaging of the cranial arteries is difficult but not impossible [95]. The high uptake seen in Figure 2.7 in the brain, bladder and spine is normal physiological activity. Brain uses a high amount of energy relative to other body parts so will take up more FDG. Radiotracer that has been used and is ready to be excreted will gather in the bladder. Spinal

Figure 2.7: A more indepth depiction of positron emission tomography - computed tomography in large vessel vasculitis . A - coronal fused FDG PET-CT; B - sagittal fused FDG PET-CT; C - axial fused FDG PET-CT. Arrows indicate areas of high FDG Uptake in aorta (ascending aorta, aortic arch and descending aorta) and subclavian arteries, suggesting probable LVV.

uptake is not as fully explained but is well documented and can occur in healthy and diseased spines. Relatively high uptake is especially noted in the spinal cord around T11 and T12 and the main theory for this is poor clearance of tracer from nearby arteries. Some vertebrae uptake is also documented as bone marrow can uptake FDG [96].

2.2.3.1 FDG Radiotracer

FDG is a commonly used PET radiotracer which is a glucose analogue and is metabolised similarly by the body ((Figure 2.8 and 2.9). PET-CT scans using FDG highlight areas of high glycolytic uptake which in turn indicates vascular wall inflammation in LVV although it is not specific to this alone [94, 97, 98]. The link between FDG uptake and inflammation is due to several factors. Firstly, the greater tissue permeability means more FDG available at sites of inflammation. The glycolytic pathways are also augmented due to cytokine release, increase in glucose transporter proteins and increase in hexokinase. Chronic inflammation leads to more monocytes and macrophages with high glucose uptake [99]. Due to the Warburg Effect, activated immune cells require higher amounts of energy which they generate using aerobic glycolysis, therefore taking up more glucose and FDG [100].

FDG is produced by electrophilic or nucleophilic fluorination and the [^{18}F] radioisotope is manufactures in a cyclotron with the irradiation of [^{18}O] with a proton [101, 102]. It is recommended that FDG is stored at room temperature in a shielded container [103]. In most cases FDG should also be used within approximately 12 hours of synthesis. This is due to the half life of [^{18}F] which is 109.7 min [101]. Similar to most radiotracers it is administered by injection up to an hour before imaging. Doses of FDG can vary but are usually based on patient weight, e.g. 5 MBq/kg body weight [104–106].

FDG PET-CT imaging is mostly used at the diagnosis stage of LVV and less so for monitoring because treatment with glucocorticoids reduces its sensitivity. Multiple studies have demonstrated this is the case but it has also been shown that there is a 3 day window where the sensitivity remains high and that it does not decrease significantly until 10 days of treatment [105–107]. This allows treatment to be started in urgent cases without delaying for imaging.

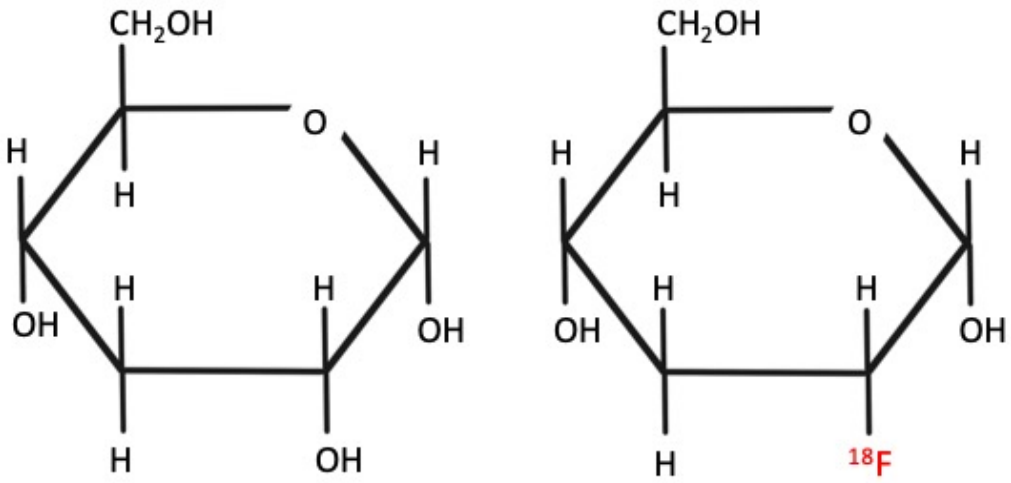


Figure 2.8: A comparison of the structures of a glucose molecule (left) and [^{18}F]-Fluorodeoxyglucose (FDG) molecule (Right).

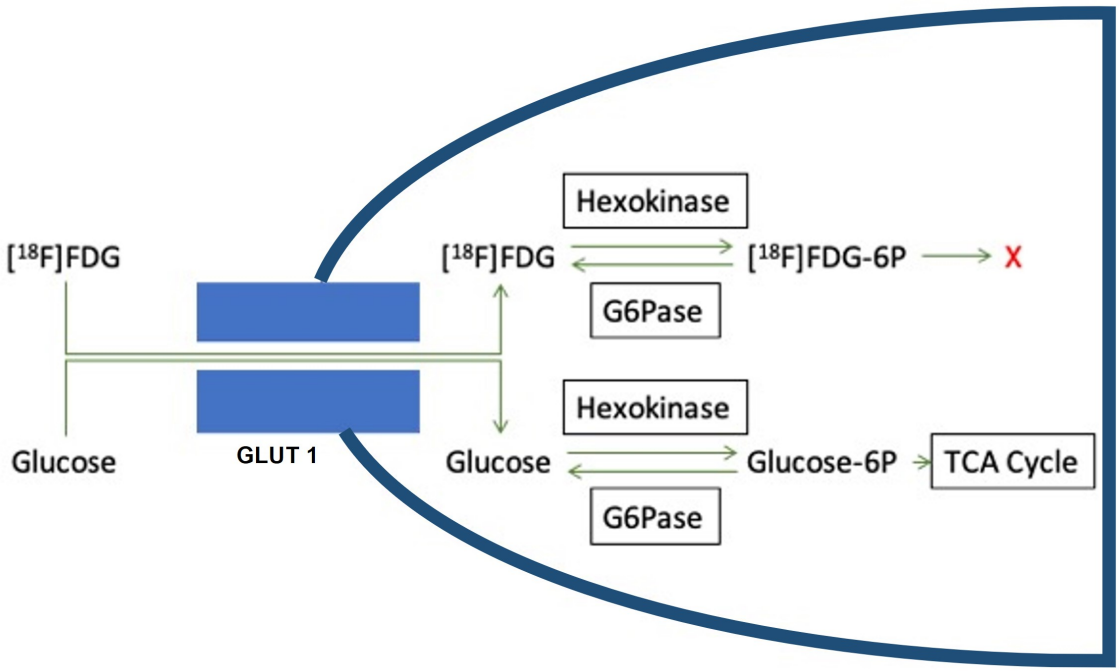


Figure 2.9: [^{18}F]-Fluorodeoxyglucose (FDG) metabolism in the body. The kinetics of these metabolic pathways occur faster than the half-life decay of FDG allowing for imaging. Key - TCA = Tricarboxylic acid , [^{18}F]-FDG = [^{18}F]-Fluorodeoxyglucose, G6Pase = glucose 6-phosphatase, 6P = 6-phosphate

2.2.3.2 Current PET Analysis in LVV

FDG PET-CT images in patients with suspected LVV are often qualitatively assessed using EANM/SNMMI imaging guidelines where FDG activity in the vessel wall is graded compared to background liver uptake using the following criteria [53]:

0: no uptake (less than mediastinum)

1: low-grade uptake (less than liver)

2: intermediate-grade uptake (equal to liver), (possible LVV)

3: high-grade uptake (greater than liver), (positive active LVV)

Grading can be extrapolated further to encompass several vascular territories by using the PET vascular activity score (PETVAS) [108, 109]. The same grading criteria is applied to nine territories and summation of the individual scores gives the PETVAS. PETVAS has been shown to be successful in PET activity measurement, treatment monitoring for LVV and predicting relapse [108–110]. Although grading and PETVAS are conducted by imaging specialists, visual assessment can be subjective and inconsistent [53, 111–113]. PETVAS is also a more involved process than grading alone which may inhibit wide-spread adoption. It is viewed as easier to apply than semiquantitative metrics such as SUV measurements and easy to interpret. Dashora *et al.* found that semi-quantitative measurements such as SUV were more reliable, and less vulnerable to assessors interpretation [114]. Laffon and Marthan agreed and suggested that SUV measurements could be made easier by manufacturers [115].

SUV based parameters can be determined (Eq. 2.2) for LVV analysis but defining an ROI to measure SUV is complicated as vascular wall inflammation does not always have a clear edge and is spread diffusely. Some of the issues with analysing LVV PET could be addressed using automated quantitative analysis with radiomics [114].

2.3 Radiomic Analysis

Radiomics is a medical imaging analytical technique which involves extraction of a large number of quantitative parameters, also referred to as radiomic features, to build decision making tools to aid in diagnosis, prognosis and to better understand disease [113, 116–

118]. The field of radiomics has rapidly expanded over the last ten years. The term 'radiomics' was first used in 2012 [119] but studies extracting quantitative handcrafted features from medical images were conducted prior to this date [120–122]. More recently the term 'imiomics' has been presented to mean the combination of whole body imaging data and non-imaging data [123].

While in many cases medical imaging is assessed qualitatively, there are some commonly used quantitative features such as SUV metrics, volume and grading [80]. Radiomics includes these features but expands on them with more complex descriptors of the shape and spatial relationships between individual voxels. Most radiomic features require complex calculations and are not visually appreciated [124].

The use of radiomics has been extensively studied in oncology but less so in cardiovascular applications [125]. Similarly, there have been a larger number of studies evaluating the use of radiomic features derived from CT and MRI rather than PET [126]. This may in part relate to relatively smaller numbers of patients and scans being acquired limiting the size of potential datasets for analysis. However, small datasets are no longer as much of an obstacle as they once were as adjustments can be made and larger datasets are becoming easier to acquire [126–128]. In general most radiomic studies have forgone typical calculations of statistical power to determine the sample size required. This is likely due to the difficulty in many studies to acquire large datasets and the fact the number of extracted features is often larger or similar to the number of patients [129]. Instead approaches such as reducing dimensionality or adjusting p values, i.e. with the Bonferonni Correction, are often conducted. Further establishment of PET in cardiovascular applications and in this case aortitis and LVV could lead to the development of clinical decision support tools [130].

2.3.1 Radiomic Workflow

While feature extraction is the defining step in radiomics analysis, several elements of the process from initial image acquisition through to the final diagnostic and predictive modelling have been shown to influence results [131, 132]. Radiomic workflows are established to extract a large range of features in a systematic way as set out in Figure

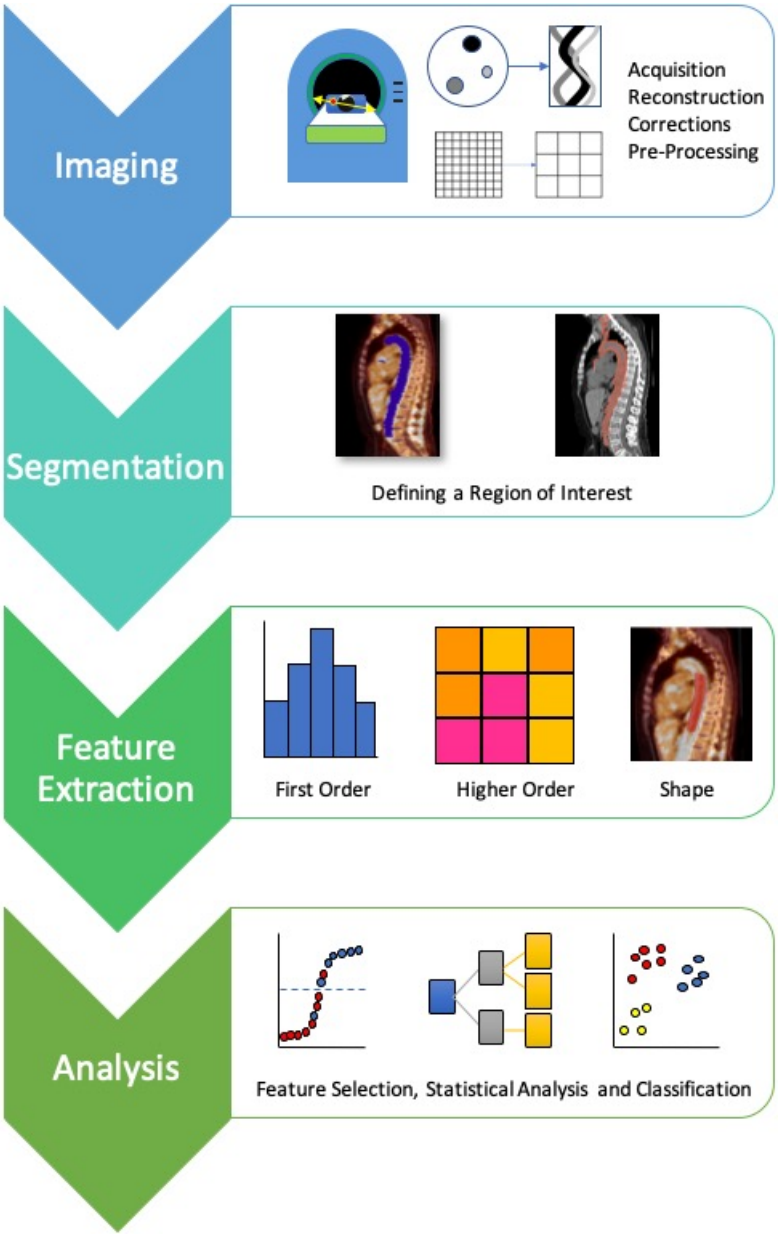


Figure 2.10: The key steps in a radiomic workflow. While radiomics is defined as the extraction of large quantities of data, the methodology used for each of the described steps must be considered in the workflow to gain reproducible and robust results.

2.10.

Artificial Intelligence can be utilised at several points in the radiomic workflow. AI has several listed definitions but essentially is building computer programs that can conduct logical processes, show intelligence, and perform tasks similarly or better than a human [124]. Within AI there are several branches but of most note in medical imaging is machine learning (ML). ML involves using data to learn how to conduct a task rather than using a set of rules. Deep Learning (DL) sits within ML and uses multi-layered neural networks composed of artificial neurons to emulate how a human brain works. In this thesis the most commonly used DL technique used is a convolutional neural network (CNN). CNNs are a type of artificial neural network that are made up of several layers of interconnected 'neurons' or processing units. A defining feature of CNNs is the convolutional layer whose purpose is to extract features from an input. While CNNs have been used on several types of input they have had great success in image analysis and in the context of medical image analysis have been used for several tasks such as diagnosis, segmentation and image quality improvement. An example architecture of a CNN can be seen in Figure 2.11. Following the convolutional layer the pooling layer downsamples the extracted features most often with maximum or average pooling. This process is multi-layered in order to learn several features and is also repeated several times with convolutions acting on feature maps from previous convolutional layers to learn both large scale and small scale features. Activation functions are also added either at the end or throughout CNNs to transform non-linear data. Finally is the fully connected layer connected to each unit of the previous layer, arranges then in a 1D arrary and every value contributes to the final output. In the case of classification this will be the probability of the input being in a given category. For example in Figure 2.11 categories A, B, C and D and given a probability on a scale of 0-1 of the input being in that category. The CNN is trained using known labelled data and minimising a loss function by altering parameters such as weighting in the network [125, 133, 134].

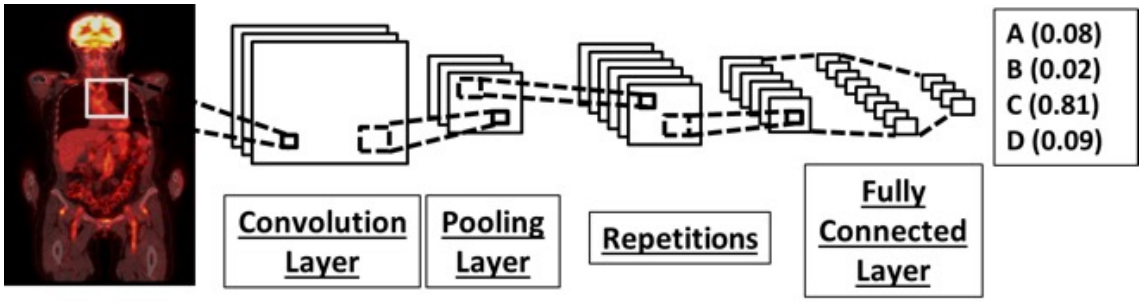


Figure 2.11: Example of a convolutional neural network architecture.

2.3.2 Image Acquisition, Reconstruction and Pre-Processing

The steps preceding feature extraction alter the input image so cannot be overlooked. These include the acquisition of the image including all scanner settings; corrections for attenuation and scatter; reconstruction; and pre-processing of the images. In many cases these steps have been optimised for visual analysis so may not produce the best radiomic results [80]. However, the need for consistent methodologies means the visually optimised images must be used.

The method of image acquisition includes several potential areas causing variation. For example, time per bed position has some effect but little in comparison to other factors [135]. Factors such as time between injection and scanning have a much greater effect. In the case of performance of PET-CT in patients with suspected LVV the recommended image acquisition protocol provided in guidelines by EANM/SNMMI is generally adhered to in clinical practice [53]. Some steps in image acquisition vary depending on the application or scanner manufacturer. For example, image noise can also be reduced with longer scan times but this is not always clinically feasible [136].

Common reconstruction methods in PET include the more traditional filter back projection, progressing to iterative algorithms and more recently AI [82, 137]. Different combinations of corrections - such as scatter, attenuation and randoms - can also be applied by scanners and have been shown to improve both visual and quantitative analysis but often differ between scanners [80]. Reconstruction methods vary between different scanners and several studies have demonstrated that this has a large influence on the variability of features [122, 131, 136, 138]. Numerous studies have been conducted to determine how large an influence these factors have and while there is some agreement there is no consensus about the optimal settings which are application dependent [139]. Every factor discussed has a proven influence so intuitively standardization of these steps is required to improve the performance of diagnostic and predictive models built with radiomics.

There are several explored methods to harmonize or standardize data used in radiomics and these fall into two categories - image-based and feature-based [140, 141]. Image based methods include, standardization of imaging protocols, applying deep learning to the image directly to harmonize, or processing the raw image data from the sensor-

level to ensure corrections and reconstructions are all identical [142]. Generally imaging protocols are established for prospective studies but a large percentage of radiomic studies are retrospective and strict inclusion and exclusion criteria based on acquisition and reconstruction can reduce the dataset to a stage where no significant results can be determined. While a standardized clinical protocol would be ideal, any future clinical use of a model relies on being useful in realistic data. Therefore if an optimised radiomic imaging protocol is not clinically feasible or accepted the final model risks not being clinically transferable. If the study purpose is to find underlying information about a condition and not about clinical adoption then this may not be as important a consideration. The trade off between standardization and the amount of data leads to a large amount of variation in radiomic methodologies. DL image based harmonization consists mostly of general adversarial networks (GAN) and convolutional neural networks (CNN) to recreate the input image with more similar properties to the reference dataset particularly in terms of voxel size and noise properties. This has mostly been explored in MRI and CT though [143, 144]. Some studies have explored this approach in PET but further research is required [145]. It has also not been fully explored if this approach reduces the quantitative information available more than conventional filtering and voxel size resampling [140].

Feature based techniques are utilised to allow the inclusion of numerous datasets taken with different acquisition and reconstruction methodologies. Firstly, radiomic features with high variation across scanners can be excluded from a study [135]. Normalization or re-scaling of features can also be applied but it is argued that this is too simplistic in many cases [140]. Some studies have explored DL for normalization, using it to learn and apply both linear and non-linear transformations to collected feature data [140].

ComBat harmonization is widely used and has been shown to be effective in several independent studies [146]. It retrospectively standardizes radiomic features from PET images obtained using different protocols by removing the centre effect whilst keeping patient-related features [147, 148]. This method was first developed by Johnson *et al.* [149] for adjusting for batch effects in microarray data. Fortin *et al.* adapted it for application to medical imaging [150] and Orlhac *et al.* applied it to PET radiomics [148]. The effectiveness of ComBat was further verified by Da-Ano *et al.* [146] who also suggested improvements to the method.

Each feature is expressed using:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i \varepsilon_{ij} \tag{2.3}$$

α = average value for feature
X_{ij} = design matrix for the covariates of interest
β = vector of regression coefficients for each covariate
γ_i = additive effect of scanner i
δ_i = multiplicative scanner effect
ε_{ij} = error term that encompasses site specific factors and overall follows a normal distribution with mean of zero
y_{ij} = value of each feature y measured in VOI (Volume of Interest) j and scanner i

The method aims to estimate γ_i and δ_i using empirical Bayes estimates. The harmonization determines a transformation for each feature separately, based on the batch effect observed on feature values. Batches are defined in this scenario as images taken with the same imaging protocol. This can not always be achieved though either due to too little data from each protocol or unknown details about how images were acquired or reconstructed. Grouping images by scanner or centre is common.

Pre-processing steps are also used to minimize non-pathological variations in the data. These can include spatial resampling, filtering and gray level discretization but not all steps will necessarily be applied and in cases where the acquisition methods were similar pre-processing may not be necessary. In some cases pre-processing is conducted after segmentation to allow for more accurate delineation [151]. Providing the mask is spatially re-sampled using the same method the effect on the result should be minimal. Shafiq-ul-Hassan *et al.* found these pre-processing steps greatly reduced the radiomic features dependencies on acquisition parameters and in some cases almost removed the effects of different scanners [152]. As these steps are to prepare an image for feature extraction the decisions concerning method design will be more influenced by the purpose of the experiment and application. For example, filtering to smooth the image has been shown to improve repeatability [136] due to reduced noise. However, using this method should be carefully considered though as it removes information and depending on the

size and shape of the ROI it may provide very little gain or even be detrimental.Shiri *et al.* determined that voxel resampling size had the biggest impact with 56% of features having a coefficient of variation larger than 20% [135]. Similar results were confirmed by several others [138]. Spatial resampling has the added benefit of making textural features independent of direction [151].

The discretization of intensity values is often conducted prior to feature extraction. This is the process of dividing a continuous spectrum into discreet groups or 'bins' and is necessary for the calculation of several radiomic features and also makes the method less computationally expensive. This has a similar effect as smoothing the image with filtering so has similar benefits and disadvantages. The alterations to pixel values are more controlled and tangible in this method. When exploring the effect of discretization on radiomic features Leijenaar *et al.* found that the intraclass correlation coefficient for every tested radiomic feature was low when two different discretization methods were used showing that using different methods can make the features incomparable [153].

2.3.3 Image Segmentation

When utilising radiomic analysis a region of interest (ROI) is normally defined to analyse. Analysis of the full imaging volume can be conducted but it is often computationally expensive and appropriate only in selected scenarios. As segmentation in the case of radiomics defines the region to be analysed it is logical that the method to segment must be accurate and reproducible. For example, Gallivanone *et al.* found in their PET-CT study using a phantom that only 20% of features were stable when the segmentation method was varied [154]. Similarly, Altazi *et al.* demonstrated that 13% of features were insensitive to whether a computer assisted or manual segmentation method was used [155].

Similarly practical aspects of the segmentation method must be considered, mainly whether the process is manual or automated, or a combination of the two. Manual methods are slow, require a large amount of expert human input and are vulnerable to inter/intra-observer variation. However, it can be more accurate than automatic methods especially in low resolution or non-contrast enhanced imaging, which is often the case in PET-CT, or if the shape and structure of the ROI is abnormal. Automating the segmen-

tation is generally faster and more reproducible allowing for analysis of larger datasets and more significant results [156]. This approach may struggle with images that deviate from normal anatomy or image processing, and are not high resolution or high contrast. Compromises can be found with semi-automated methods

When selecting or developing a segmentation method for a radiomic workflow it needs to be tested against a ground truth. Producing a ground truth segmentation can require a significant time commitment from experts for both segmentation and validating the segmentations of others. This is the case for both manual and automated segmentations. There are several metrics to evaluate the quality of a segmentation method when comparing it to a reference or ground truth. One of the most popular is the Dice similarity coefficient (DSC) which quantifies overlap (Eqn. 2.4, Fig. 2.12). The DSC does have some limitations. Firstly, is it's relaince on size of ROI/VOI. Errors in smaller structures disproportionately influence the metric compared to larger structures. Therefore in some applications, a final average DSC may not give a true representation of segmentation quality. Similarly, due to it's reliance on size it can favour oversemgentation to undersegmentation, and does not treat sensitivity and specificity equally making its usefulness application dependent. DSC also does not include any information regarding shape unlike boundary based metrics, nor does it account for a segmentation methods ability to approximate location as no priority is given to the centre of an object [157]. While a few different metrics exist for evaluating segmentation DSC was still used in this thesis as it allowed for comparison to other published methods and the size of the aorta was sufficiently large and consistent.

$$DSC = \frac{2|A \bigcap B|}{|A| + |B|} \tag{2.4}$$

Aortic and other major arterial anatomical segmentation using contrast-enhanced CT or MRA is well established, but there is no consensus on segmentation methods when using unenhanced CT - the NHS clinical standard in PET-CT imaging used as the anatomical reference for PET. In low-dose CT it is harder to segment the major arteries, both manually or automatically. Methods for low dose CT have been established but have limitations [158]. Manually, expert knowledge and experience is required to segment the

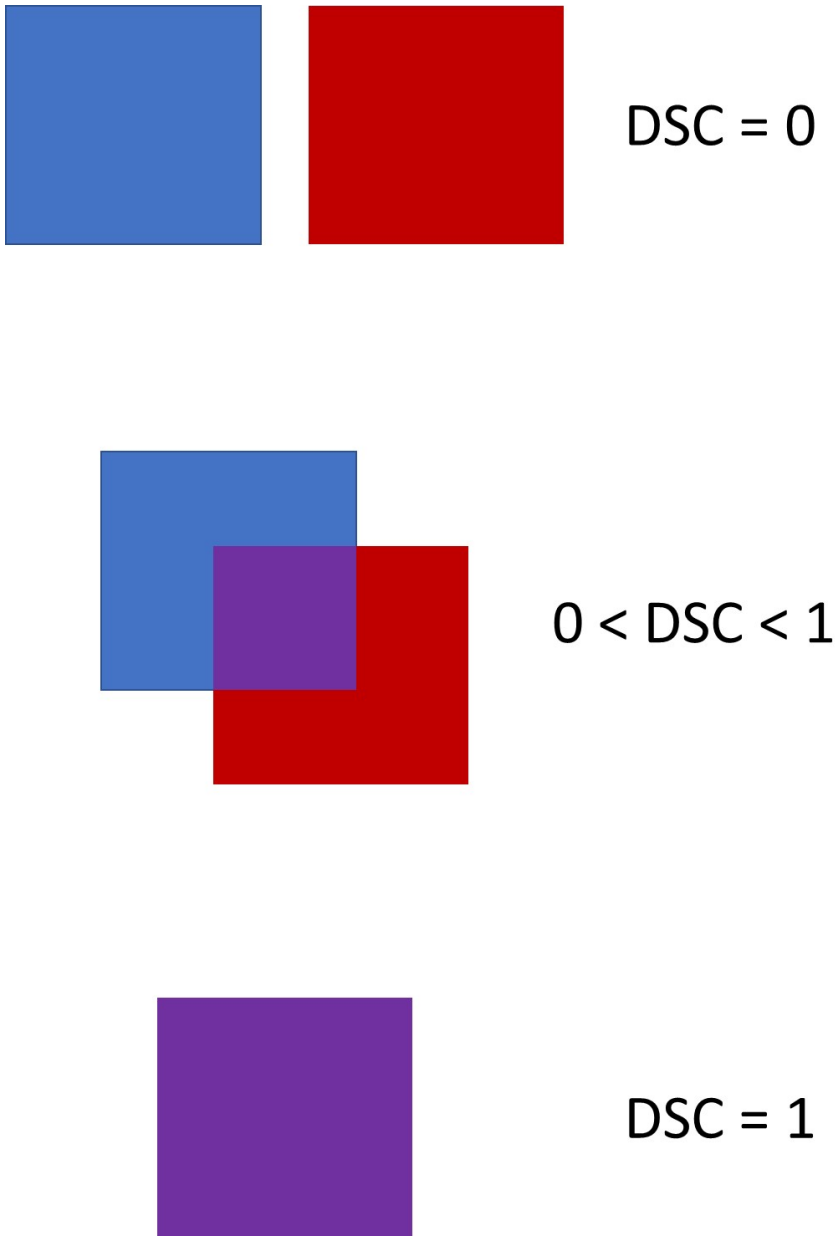


Figure 2.12: The Dice Similarity Coefficient is used for evaluating agreement between segmentations. It is based on the ratio between overlapped area to the total area of the two compared segmentations with one being a perfect agreement and zero being no overlap.

arteries other than the aorta due to lower image quality. Automated methods that rely on manual segmentations as reference are limited in the same way. Many automated methods that do not require labelled input, such as circle detection and growth methods, are inhibited by low or no edge contrast making the arteries indistinguishable [159, 160]. In many cases the aorta can be segmented in PET-CT but not the other arteries. The exception to this is threshold-based segmentation on PET vessel wall activity but this cannot be used for control cases as their lower uptake would mean they lay below any useful threshold and would not be delineated.

When manual segmentation can be conducted either on the aorta alone or all relevant arteries this can be performed on a sample set to then establish other semi-automated or automated methods [161]. For example, atlas based or deformable methods can work but AI methods are more commonly used [162]. A common example of an AI based within segmentation applications is a (CNN) that makes pixel-wise predictions (Figure 2.11). Traditional CNN architectures have been used for segmentation but another popular architecture is U shaped and is referred to as U-Nets [163]. U-Nets consist of a prediction part and a synthesis part to construct a segmentation. Using DL for segmentation is popular and has proven repeatedly to work. However, it requires a large manually segmented dataset and can be computationally expensive to run [164–167].

Currently most established artery segmentation methods use the entire artery including the lumen [162, 168]. Segmentations using lumen removal could allow more specific radiomic analysis of the vessel wall, removing redundant information and noise, but this is more difficult to implement particularly in low dose CT or PET [156, 169]. Several aortic wall segmentation methods utilised CTA or MRA so were not conducted using low-dose non-contrast enhanced CT as was used in this thesis [170]. Piri *et al.* implemented lumen removal but did so using a predefined thickness of wall which is not precise and does not account for physiological variation [156]. The diffuse nature of LVV uptake makes it more challenging to either manually segment active areas or decide on an appropriate thresholding value for PET without introducing bias [80]. Thresholding can be used to isolate the aortic wall but selection of a cut-off point that does not make controls redundant is complicated. Creating an automated method to select a threshold value based on SUV of the lumen is viable but as wall activity can be both higher and lower than this

value (Section 2.2.3.2) this would only segment areas of activity meaning only LVV cases could be studied and not controls. Thresholding also introduces problems with respect to volume of the ROI. Several radiomic values are highly correlated with volume and may only give additional information in ROIs with a volume above $45cm^3$ [80, 171]. While there is some leniency in this cut-off value, thresholding the aorta would likely inhibit radiomic evaluation as the diffuse uptake pattern would create several hot spots smaller than this volume. Excluding small ROIs would remove a large amount of textural information.

2.3.4 Radiomic Feature Extraction

Most radiomic features, other than shape, are formed based on the idea that every pixel represents a value, SUV in the case of PET. There are hundreds of defined radiomic features and potentially thousands when individual researchers alterations are considered [80]. General definitions of each type or category of radiomic feature are described below but it is essential that radiomic studies give reproducible definitions of their included features or refer to a set of standardized definitions such as the Image Biomarker Standardization Initiative (IBSI) [139, 172]. Other radiomic features aside from those discussed have been formulated, such as fractal-based features, but they currently have much less evidence of clinical utility and have not been standardized by the IBSI.

While individuals can develop their own methods for feature extraction, it is more common to use software or coding packages. There are several open source options such as PyRadiomics, LifeX, IBEX and MaZda [173–175], and reviews vary with their results when exploring which is the most popular [176, 177] . Being open source allows for more wider dissemination of the method and easier transfer to a clinical use. Using established software or packages reduces variation in extracted features - although does not eliminate it, allows for comparison with other studies, and in most cases ensures compliance with standardized definitions e.g. IBSI. Variation between these different softwares and packages can still reduce reproducibility so full details of this step is advised [174].

2.3.4.1 Conventional Features

Conventional features refer to quantitative parameters that are well established in medical image analysis independently of the growth in the field of radiomics. Some conventional features are used in radiomics, either alone or alongside other radiomic features. While established parameters are often useful for the diagnosis and monitoring of conditions, other radiomic features may provide additional useful information and prove to be more accurate and reproducible.

Vessel diameter is commonly used to identify stenosis, dilation and aneurysms all of which may be present in LVV patients [178–180]. Depending on application either internal, external or both diameters may be measured [181, 182]. In some modalities such as low-dose CT the vessel wall can not be distinguished making only one measurable diameter. Diameter is easily extracted from most images but requires a sufficient resolution which is not always possible in PET or low-dose CT especially for smaller vessels. Volume is similar as another shape related measurement. Different methods for grading PET scans based on activity are described in Section 2.2.3 and are a form of conventional features. An example of a feature commonly extracted from PET is SUV (Eq. 2.2) which can be used to identify abnormal FDG uptake such as inflamed tissue making it a useful indicator of LVV. SUV can be determined per voxel but is normally summarised for a ROI by the metrics discussed in Section 2.2.2.

2.3.4.2 First Order Features

First order features are derived from voxel intensity values but exclude the spatial relationships between voxels. They describe the intensities within the ROI and how they are distributed which is commonly expressed as a histogram where the intensity values are binned either by a defined bin width or number of bins. First order features describe properties of the distribution such as averages and standard deviation, skewness and kurtosis, and uniformity and randomness (Figure 2.13). Although first order features provide diagnostic and prognostic information, they exclude information concerning the spatial relationship between voxels so are inherently limited.

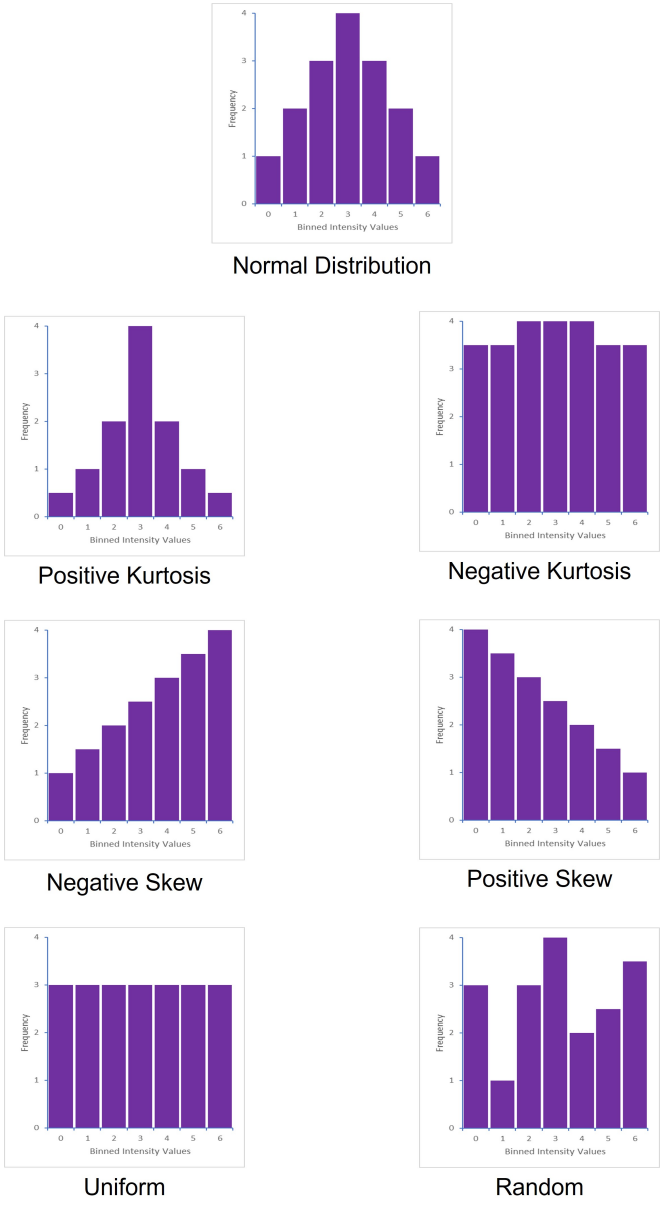
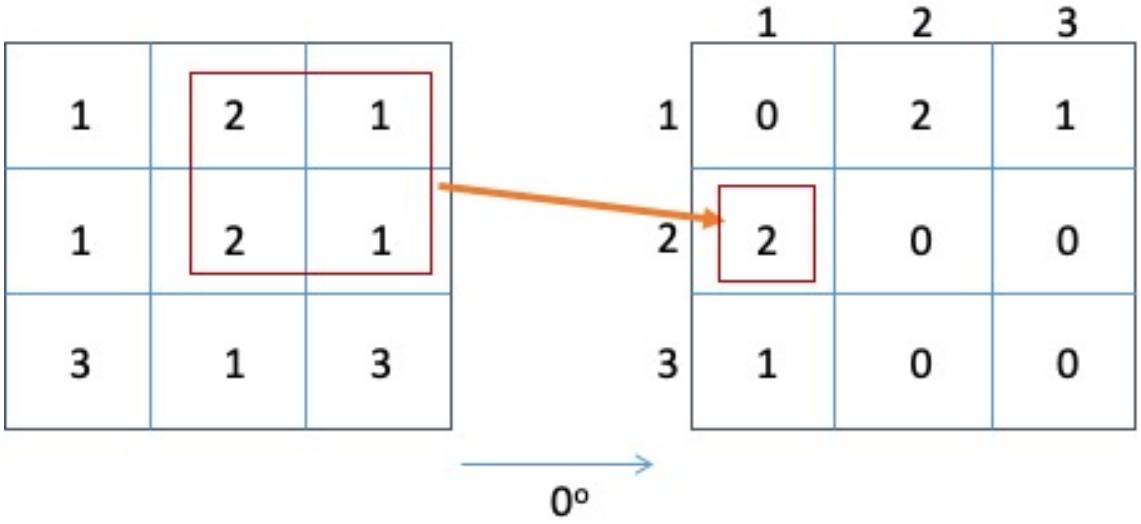


Figure 2.13: Visualization of histogram-based features. In PET the pixel intensity/SUV values are binned to convert from a continuous to a categorical measurement. Kurtosis is defined by the 'peakedness' of the histogram. Skew is a measurement of how much the histogram lays to the left or right. A fully uniform histogram would be when each bin has an equal frequency. A histogram with high randomness or entropy occurs when the values follow no specific distribution and occur at random.

Figure 2.14: An example of the Gray-Level Co-occurrence Matrix (GLCM). Each element of a GLCM determines how many times a pair of intensity values occur in neighbouring voxels for a given direction. In this case 1 appears directly to the left (0^o) of 2 twice.



2.3.4.3 Second and Higher Order Features

Second and higher order features explore the spatial relationships and patterns between voxels. Second order features express the relationship between two voxels while higher order features express the relationship between three more voxels. It is important to include spatial information as its utility is already well established in several conditions. In the case of LVV, the distribution of FDG uptake is known to visually appear different to other aortic conditions such as atherosclerosis proving the diagnostic potential of these features.

Gray-level co-occurrence matrices (GLCMs) express how often an intensity value i (columns) occurs in a neighbouring voxel to intensity value j (rows) (Figure 2.14). Only four directions are required when looking at the relationship in 2D (0°, 45°, 90°and 135°) as the remaining angles are accounted for by the GLCM in other voxels. There are several equivalents in higher order features such as a Gray-level run length matrix (GLRLM) which assesses how many voxels are next to each other with the same value (run length), and the Gray-level zone length matrix which provides information of the size of homogenous zones. The matrices discussed contain a lot of information but several parameters have been formulated to either summarises the matrices or extract desired information.

2.3.4.4 Shape-based Features

Some shape-based features are intuitive and can be considered conventional features. More intricate shape-based features have been crafted and are measured as well in radiomics. They include features such as compactness, sphericity and ratios of pairs of shape-based features. Shape-based features are independent from intensity values but can be combined with them to form other features such as Total Legion Glycolisis (TLG) in PET. Shape-based features can be measured in both 2 and 3 dimensions.

The comprehensibility of shape-based features makes them easier to translate to clinical practice. They are less reliant on intensity, so are less affected by external factors but can be complicated to measure due to natural patient variation and there are not always established methods to normalize [135]. A couple of examples of shape based features can be seen in Figure 2.15.

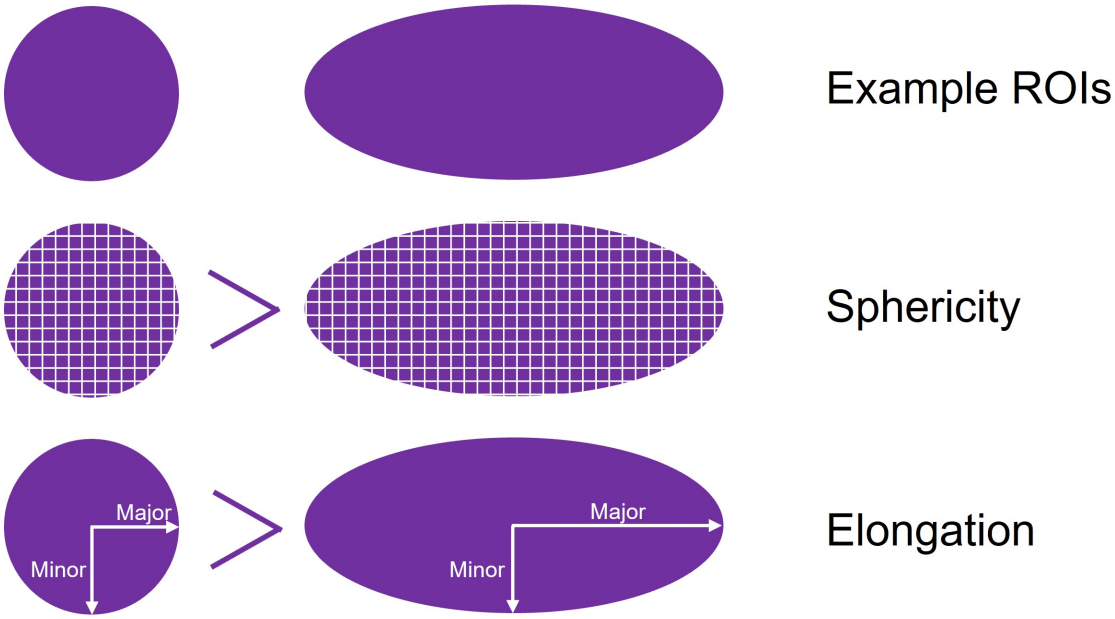


Figure 2.15: A representation of two shape based radiomic features, sphericity and elongation. Sphericity measures the roundness of a ROI and equals one when it is a perfect sphere. As spheres have the smallest possible surface area for a given volume it is calculated using both surface area and volume. Elongation is based on the relationship between the major (largest) and minor (second largest) axis in an ellipsoid that encapulates the ROI. An elongation of one represents a sphere and an elongation of zero is a maximally elongated object / straight line in 1D.

2.3.5 Feature Selection and Analysis

Once the radiomic features have been extracted they are analysed to find relationships with clinical data such as diagnosis, outcomes, and treatment response.

The number of extracted features often reach into the hundreds so feature selection is important to reduce noise, redundant information, and to minimize the risk of Type 1 (false discovery) errors. The methods used to select features can be driven by the application but some of the most common are to use Principal Component Analysis (PCA), eliminate features based on individual diagnostic and predictive performance, remove those that correlate highly with other features, and use feature selection methods that are built into the classification method [147].

PCA is a method based on linear algebra that reduces the dimensionality of a dataset. Extracted radiomic features have as many dimensions as there are features and PCA combines these dimensions to form principal components that retain a high amount of data variation but eliminates redundant information [183].

One intuitive method is to remove features that are not useful for diagnosis. This involves analyzing the diagnostic or predictive performance of each feature individually and eliminated those that do not perform well. This method is useful when combinations of features will be used in the final modelling but if features are used individually this arguably is vulnerable to Type 1 error [147].

As one purpose of feature selection is to remove redundant information it is sensible to remove highly correlated features. Many features are related to the same underlying values or patterns. Keeping only the highest performing diagnostic or predictive features and eliminating those correlated to them is a increasingly utilized method in feature selection [147].

Finally, some diagnostic and predictive models and algorithms discussed later have feature selection embedded so these steps can be combined.

Once the radiomic feature dataset has been reduced, the diagnostic and predictive utility of the remaining extracted features is determined if not already conducted as part of feature selection. In some cases, this is performed using traditional statistical analysis - such as the Mann Whitney U test, Pearson's Correlation, Elastic Net Regularisation and

Cox Regression - but ML classifiers have grown in popularity. Traditional statistical tests can be used to determine if the radiomic features are statistically significantly different between two or more populations, i.e. positive for a condition and healthy control. To use this information for diagnostic or predictive purposes a cut-off or threshold point is determined. However, this method is similar to simple ML classifiers such as linear or logistic regression leading to these classifiers and others becoming well established for building diagnostic or predictive models [147].

ML classifiers are easily applied with current software and programming packages, and can give extensive information about the relationship between radiomic features and clinical data. There are several types of ML classifier but most fall within three main categories based on the underlying principals used for classification (Figure 2.16). Firstly, some use regression analysis to predict a result. Examples of classifiers in this category include linear regression, logistic regression and to some extent support vector machine (Figure 2.16a). Support vector machine also lies within the second category that uses clustering to find similarities in the data (Figure 2.16b). This technique is used more in unlabeled data. Lastly decision trees can be constructed to classify patients based on the values in their radiomic features (Figure 2.16c) [147].

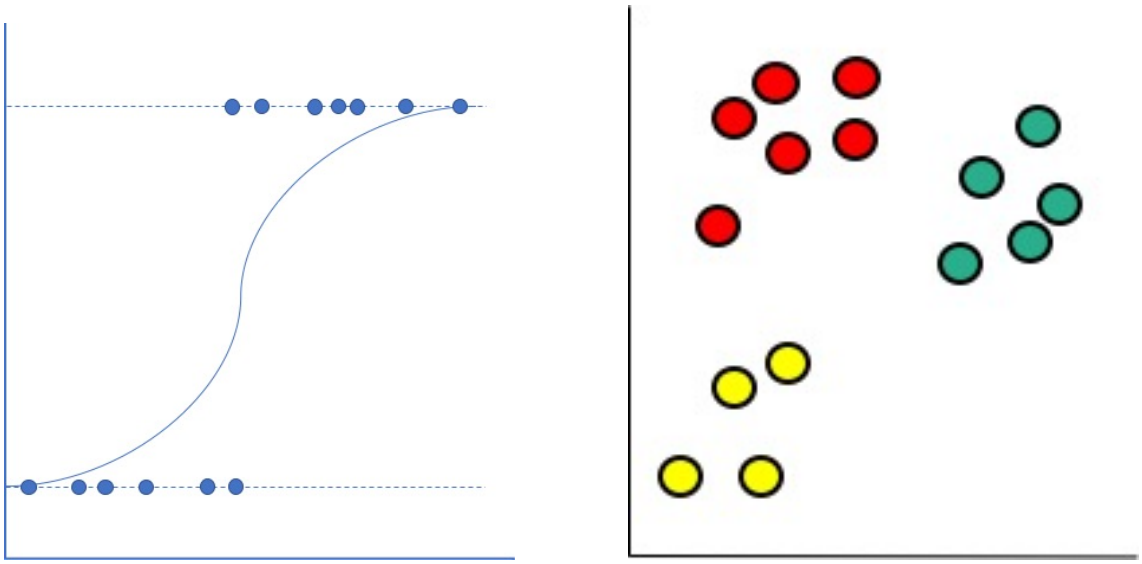
There are several metrics for evaluating the ability of radiomic feature and ML classifier combinations to diagnose or predict outcomes. Accuracy is most well known and while informative is vulnerable to imbalanced datasets and treats false positives and false negatives as equal errors where in a medical scenario this may not be the case. Class imbalance can be mitigated using balanced accuracy but still treats sensitivity and specificity equally. Several other metrics for evaluation are used but most common in LVV diagnosis and imaging is the Area Under the Receiver Operating Characteristic (ROC) Curve. The ROC curve plots the false positive rate along the x axis and the true positive rate along the y-axis at several predicition probability thresholds. It shows the trade-off between increased true positives, which is desired, and increased false positives, which should be minimised. The larger the area under the ROC curve (AUC) the higher the classification, or in this case diagnostic or predictive, ability of the classifier [184]. Major benefits of AUC as a performance metric is that it is not vulnerable to changes in the aortitis:control ratio and can with stand a large imbalance in the dataset - around 100:1 [185]. Another

significant benefit of ROC curves is that visualising the trade off between sensitivity and specificity at several thresholds allows threshold selection that is more suitable for the given application [186]. AUC is considered a robust metric and good summary of classifier performance [187, 188]. However, while AUC is widely adopted in diagnostics it is not as relevant for many clinicians. Firstly, it treats sensitivity and specificity equally which is not reflective of many clinical applications. Many clinicians prefer metrics that reflect what a diagnostic test means for an individual patient which AUC does not encompass [189]. In this regard using AUC for building a diagnostic method can be useful for determining a optimum threshold but may not be the most suitable metric to summarise a finalized model. In this thesis AUC was used to summarise models and compare but more clinical input would be required to determine the best trade off between sensitivity and specificity. One additional limitation of AUC relevant to this thesis is the comparison of qualitative and quantitative ROC curves. Confidence scales are reproducibly calculated in quantitative methods but less so in qualitative diagnostics[190]. Therefore, comparisons should be considered carefully.

False discovery is prevalent in radiomic analysis due to high ratios of radiomic features to patients, multiple non-biological factors influencing the dataset and poor understanding of some of the complex statistics. Chalkidou, O'Doherty and Marsden used statistical corrections to disapprove all fifteen studies they analysed which had reported statistically significant results [191]. Furthermore, when image derived parameters were replaced with random variables they found 10% came out as significant predictors. This makes multi-centre validation essential to determine which discoveries are generalisable and transferable [80, 192]. Insufficient data, over training, or too many radiomic features - reinforcing the need for feature selection - can also lead ML classifiers to under or over-fit their models [151].

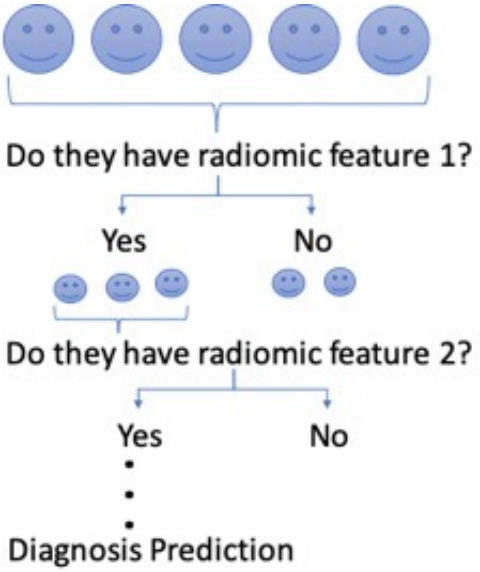
2.4 Challenges and Outlook

Despite promising results, radiomics has yet to be adopted in clinical practice even in oncology where the field is more established due to several factors discussed in this chap-



(a) Logistic Regression

(b) K Nearest Neighbours



(c) Decision Tree

Figure 2.16: Three examples of machine learning classifiers and how they distinguish different categories from one another. There are several types of machine learning classifiers but most work similarly to these three with small alterations in the method. a) Logistic regression models the probability of an outcome based on an input variable, in this case a radiomic feature. b) K nearest neighbours defines clusters based on training data and then assigns new inputs to a cluster. c) Decision trees use a list of conditions to categorise the input.

ter. The vulnerability of radiomic features to image reconstruction, image acquisition, segmentation methods and feature calculation are increasingly recognized and mitigation strategies to minimize variation are being developed. Similarly, the analytical process is complex resulting in low reproducibility, several opportunities for variation and lack of translation into clinical practice. Some methods require human input that can be time-consuming, and introduce bias and variability [193]. A common limitation listed in papers is a small cohort size and a trade-off between a small cohort size or a non-homogeneous cohort. Single centre studies are also a common limitation as multi-centre studies are difficult to establish and introduce more variations [151].

Progress has been made in tackling all these issues [80]. The need for method standardization is widely accepted and several initiatives have been established. For example, the IBSI had standardized the definitions of 169 radiomic features which are widely used in different open source software and coding packages. When studies do deviate from the definitions the deviations can be easily described using it as a reference point in a way that was often overlooked previously. Many journals also require a set of guidelines to be adhered to for radiomic studies such as the STARD or TRIPOD guidelines [194, 195]. Standardization has also helped reduce the variation in human input and the growth of AI, ML and DL has allowed for more automation in the radiomic pipeline making the process more manageable. Automation has made large cohort sizes easier to analyse leading to more significant results. Large datasets are not always possible to acquire but an increase in open-source data and large online data collections such as the UK Biobank and Scottish Medical Imaging Service have helped mitigate this issue. Where variations in the data acquisition and processing are the issue data harmonization has been proven to remove centre effects whilst keeping patient-related features and outperforms similar techniques [148, 196]. Techniques such as distributed learning can help overcome problems with data sharing in multi-centre studies [197]. This is where a model is trained on several datasets that cannot be shared by each data owner individually and then the model updates alone are shared [198]. This is a useful technique but has some disadvantages. Firstly, some people question if some sensitive data can still be extracted from shared model updates [199]. Secondly, there is debate about how to best parallelize the model and then create a coherent model after all training is conducted. There can be some prac-

tical difficulties as well such as ensuring implementation is consistent, recreating exact programming environments and avoiding software incompatibilities, and differences in computational abilities and hardware . Finally, despite not sharing sensitive data there is still a need to protect intellectual property around the model's design [200].

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 [124, 147, 201]. 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 do not always produce better results than other well-established methods. There is growing interest in combining handcrafted radiomic features with DL as combining both methods can provide complimentary information. For example DL based features tend to characterise local features while radiomics characterise over a larger structure. So far only a small number of studies have explored this method but have shown promising results [202, 203].

2.5 Project Justification

The purpose of this project was to investigate the diagnostic utility of radiomic analysis in PET-CT imaging of LVV. Firstly, the feasibility had to be determined and a methodological pipeline established. Following this, the findings needed to be validated and the overall method automated to improve the likelihood of clinical acceptance. The overall aim of the project was to develop an automated decision-support tool for a more objective and standardized assessment of aortitis [130]. The project focused around diagnosis of LVV and more specifically aortitis due to imaging data availability and limitations in the segmentation method. Further vascular territories and applications of radiomics could be explored in future work.

Chapter 3

Experiment Set 1 - Method Development

The purpose of this chapter is to evaluate the potential utility of radiomic features extracted from FDG PET-CT for improving the accuracy of detecting active aortitis. The methodological framework established combines radiomic features and ML classifiers to develop a prototype and rigorous semi-automated decision-making tool for a more objective and standardized assessment of aortitis. The key steps in a radiomic methodology are set out in Figure 3.1. As this is an initial analysis, at this stage the study is concentrated on data from a single centre and makes use of manual segmentation of the ROI. Further validation and automation are conducted in the following chapter.

3.1 Methods

3.1.1 Patient Selection

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 from a single institution, Leeds Teaching Hospitals NHS Trust, between January 2011 and December 2019. The ground truth diagnoses for all patients and controls were confirmed by a consultant rheumatologist with 17 years' experience

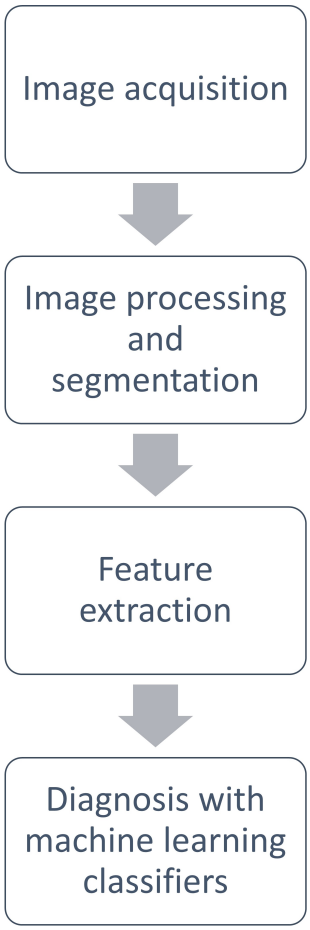


Figure 3.1: The key stages in the methodological framework established in this chapter.

of vasculitis (supervisor 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 (supervisor AFS) with more than 15 years' 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 (n=50) and controls (n=25) were extracted from the institutional PACS (Picture Archiving and Communication System) and pseudoanonymised.

3.1.2 Imaging Protocol

FDG PET-CT scans were acquired using a standard protocol: images were acquired from the upper thighs to the skull vertex in the supine position[53, 204, 205]. 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. Imaging was acquired on three different scanners during the study period including a 64-slice Gemini TF64 scanner (Philips Healthcare, Best, Netherlands; n=29), a 64-slice Discovery 690 scanner (GE Healthcare, Chicago, IL, USA; n=12) or a 64-slice Discovery 710 scanner (GE Healthcare, Chicago, IL, USA; n=34). Each scanner used iterative reconstruction, CT for attenuation correction, applied scatter and randoms correction. Image reconstruction parameters for the different scanners are shown in Table 3.1. Acquisition and reconstruction parameters were the same for all patients within each scanner.

Scanner	Reconstruction	Scatter Correction	Randoms Correction	Matrix	Voxel size in mm^3
Philips Gemini TF64	BLOB-OS-TF	SS-SIMUL	DLYD	144	$4 \times 4 \times 4$
GE Healthcare Discovery 690/710	VPFX	Model based	Singles	192	$3.65 \times 3.65 \times 3.27$

Table 3.1: 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

3.1.3 Segmentation

The entire aorta was manually segmented using 3D Slicer [1] on the baseline FDG PET-CT scan of each patient [206, 207]. Segmentation was conducted by a single observer (candidate LD, physics and engineering researcher, limited experience) under supervision of supervisor AFS. 3D Slicer was selected for segmentation as it is open source, has intuitive graphical user interface and comprehensive user documentation and support. Segmentation was conducted in the axial plane of the CT image on every slice with no filters or interpolation applied. Segmentation started from the aortic valve to the point of bifurcation. Once completed the ROI was exported in both the PET and CT resolution. An initial batch (n=15) of segmented volumes were validated against those performed by a clinical radiologist with 3 years' experience (acknowledged PA) to confirm inter-observer concordance. Segmentation was conducted in the axial plane of the CT image on every slice with no filters or interpolation applied. Segmentation started from the aortic valve to the point of bifurcation. Once completed the ROI was exported in both the PET and CT resolution.

[1] Version 4.10.2, https: //www.slicer.org/

Intra-observer variation was not evaluated as over the timeline of this study the observer conducting segmentation (candidate LD) gained experience making a comparison of the two sets un-informative. Supervision during this time ensured segmentations were of high enough quality to proceed. DSC (Eqn. 2.4) was used for contour comparison. The PET images and segmented masks were then resampled to a 4mm isotropic voxel size to ensure a uniform sampling across the entire cohort. This voxel size was selected as it was the lowest resolution of our 3 scanners.

3.1.4 Feature Extraction

Pyradiomics (Version 3.0.1, https: //www.radiomics.io/pyradiomics.html) was used to extract 102 radiomic features from the entire 3D volume of the segmented aorta in the PET images [175]. Pyradiomics complies with the IBSI standards for most radiomics features and SUV metrics; any minor deviations are clearly described in their documentation (https://pyradiomics.readthedocs.io/en/latest/). All features available through Pyradiomics were used. The SUV bin width was set to 0.075 in the Pyradiomics parameter input file. This bin width was selected by finding the max SUV value in the ROIs and dividing it by 64, a commonly used bin number in radiomics. No additional filters were used, and all other parameters were left as default. Five SUV features not included in Pyradiomics (SUVx) were calculated separately and added to the radiomic features dataset using Python packages Numpy (Version 1.18.1) and Simple ITK (Version 2.01). Full definitions of each radiomic feature are described in the Pyradiomics documentation. The SUV metrics are defined in the previous chapter.

Extracted radiomic features and SUV metrics were harmonized using the ComBat method (neuroCombat, Version 0.2.7) (Section 2.3.2). A list of all radiomic features and SUV features (107 in total) used is provided in Supplemental Material 6.1.1. SUV metrics were used instead of target-to-blood pool ratio as it not commonly used with aortitis and liver is a more common reference point as discussed in Section 4.1.4.

3.1.5 Qualitative Grading of Vessel Wall FDG Activity

A radiologist (supervisor AFS) reanalyzed all scans and documented the vascular uptake score based on EANM/SNMMI guidelines [53]:

0: no uptake ($\leq$ mediastinum)

1: low-grade uptake ($<$ liver)

2: intermediate-grade uptake (= liver), (possible aortitis)

3: high-grade uptake ($>$ liver), (positive active aortitis)

3.1.6 SUV Metrics and Radiomic Feature Diagnostic Utility Analysis

The diagnostic utility of a range of commonly used SUV metrics and extracted radiomic features was evaluated using two methods. Firstly, the Mann Whitney U test was used. The *p* value for significance was adjusted using Bonferroni correction (p = 0.05 / number of features) to reduce the risk of false discovery (type 1 error) related to multiple testing.

The second method of evaluating feature diagnostic utility was to use ML classifiers. Logistic Regression (LR) classifiers were trained with each feature individually (Sci-kit Learn Version 0.23.2). First the hyperparameters for each feature were tuned using the Sci-kit Learn function GridSearchCV where every combination of hyperparameters provided to the function are tested to find the optimal set.

Stratified 5-fold cross validation was used for both hyperparameter tuning and training of all final ML algorithms meaning the ratio of patients to controls in each fold was equal to the ratio in the total population. The AUC and the accuracy ($\frac{correct\ predictions}{all\ predictions}$) were both used to select the best performing hyperparameters. The tuned hyperparameters for each feature were then used to train the final logistic regression model for that feature and the overall diagnostic utility was determined using the mean accuracy and mean AUC from stratified 5-fold cross validation. Confidence intervals were determined using the standard error of the five testing AUCs and accuracies. Only cross validation scores are reported in this study as splitting the data into training and test samples would be inappropriate for the sample size available [208]. In this report of model development we therefore focus on internal validation.

3.1.7 Radiomic Fingerprint Building

Many radiomic features can be extracted but not all of the derived features may provide useful information [124]. Several radiomic features can be clustered together to achieve a higher diagnostic performance than single features. However, using all available features retains a large amount of redundant information and creates noise in the final diagnostic model. Therefore, Fingerprints of a smaller number of features were built to reduce the noise of the larger dataset while retaining the useful information provided. Three Fingerprints were built using the methods described below.

3.1.7.1 Performance Criteria and Correlation

The first method involved selecting features with high individual diagnostic utility. For Fingerprint A, features had to meet the following criteria: AUC$\geq$0.5, accuracy$\geq$0.7, Mann Whitney U test p value$\leq \frac{0.05}{n}$ where n = number of features (n=107). Features were filtered based on their evaluation results from section 3.1.4 using Python package Pandas (Version 1.1.4). Features which met these criteria formed Fingerprint A. Fingerprint B was generated by removing highly correlated features from Fingerprint A: for each pair of features, if the correlation coefficient was greater than 0.9, the feature with the lower AUC was removed.

3.1.7.2 PCA

The number of features can be reduced using 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) and the number of principal components needed to account for 90% of the information were retained. These principal components formed the third Fingerprint, C.

3.1.8 Machine Learning

Once relevant features had been determined and feature sets reduced,the resulting radiomics Fingerprints were used as an input for a ML algorithm to be used collectively to diagnose active aortitis [209–211]. In order to determine the best ML algorithm for distinguishing aortitis nine different classifiers were built, trained and tested using Sci-kit Learn (Version 0.23.2): support vector machine, random forest, passive aggressive, LR, k nearest neighbours, perceptron, multi-layered perceptron, decision tree and gaussian process classification. The nine ML classifiers were trained on the radiomics Fingerprints using the same methodology described in section 3.1.7 for logistic regressiontraining on individual radiomics features. The best classifier for each Fingerprint was determined using the mean AUC of each classifier with a minimum mean accuracy of 80% or 70% if that was not possible. The tuned hyperparameters for each of the ten classifiers for Fingerprint A, B and C can be found in Supplemental Material 6.1.2.1, 6.1.2.2 and 6.1.2.3, respectively.

3.1.9 The Utility of Harmonization

As stated above, harmonization was applied to all methods and any presented results used harmonized data unless stated otherwise. However, the developed methodology was repeated without harmonization to determine the effect. The effect of harmonization was evaluated with the Mann Whitney U test. It was used to evaluate whether two populations – the feature distribution for scanner x and y- were different populations ($p < 0.05$). Scanner 1 (GE Discovery 710) was compared to scanner 2 (Phillips Gemini TF64), 2 to 3 (GE Discovery 690) and 1 to 3 before and after harmonization for each of the 107 features (radiomics features and SUV metrics). The effect of emitting harmonization was also examined on the performance of all diagnostic models discussed above.

3.2 Results

3.2.1 Patient Characteristics

In total 75 participants were included, 50 of whom had a FDG PET-CT scan indicating active aortitis (Table 4.3). The age of the patients and female predominance reflects the typical demographic of patients with LVV, the commonest cause of which is GCA. The sensitivity of FDG PET-CT is significantly reduced within a few days of starting glucocorticoid treatment so doses were zero at the time of scanning unless stated otherwise [106]. CRP (C-reactive Protein) and ESR (Erythrocyte sedimentation rate) are biomarkers of inflammation.

3.2.2 Segmentation

The manual segmentation method was shown to be reproducible and accurate when compared to those performed by an experienced radiologist. Inter-observer variability scored an average DSC of 0.91 (0.90-0.92 95% Confidence Interval (CI)).

3.2.3 Qualitative Grading

Guidelines advocate qualitative grading of PET-CT scans based on FDG activity in the aortic wall relative to the liver [53]. Table 3.3 shows the grades assigned by an experienced radiologist on retrospective review of the images. Note that the single aortitis patient who was graded as 1 rather than 3 was taking 25mg of prednisolone at the time reducing the sensitivity of FDG PET-CT.

3.2.4 Diagnostic Utility of SUV Metrics

All SUV metrics evaluated, except SUV_{min} and SUV 10^{th} percentile, fulfilled the criteria based on the Mann-Whitney U test that there was a statistically significant difference between the mean metric value for the aortitis and control group (Bonferroni-corrected $p< 0.00047$). Figure 3.2a demonstrates the performance of SUV features in an logistic

regression classifier where higher accuracy and AUC are preferable and indicate good diagnostic utility. The performance of all SUV metrics in logistic regression classifiers and in the Mann Whitney U test can be viewed in Supplemental Material 6.1.3.

3.2.5 Diagnostic Utility of Radiomic Features

In the Mann Whitney U test 65 of 107 radiomic features fulfilled the criteria based on the Mann-Whitney U test that there was a statistically significant difference between the mean feature value for the aortic LVV and control group (Bonferroni-corrected $p<0.00047$). Furthermore, their p values were 2-3 orders of magnitude smaller than SUV features (Supplemental Material 6.1.3). The five-best performing radiomic features in terms of AUC, when used individually in an logistic regression classifier, are shown in Figure 3.2b. The performance of all individual radiomic features in logistic regression classifiers and in the Mann Whitney U test can be viewed in Supplemental Material 6.1.3.

3.2.6 Correlation between SUV Metrics and Best Performing Radiomic Features

Table 3.3 displays the correlation matrix of SUV metrics and the best performing radiomics features. It shows an intuitive split between the two groups but also emphasizes that GLSZM Size Zone Non-Uniformity Normalized is only weakly correlated to other well performing radiomics features.

3.2.7 Radiomic Feature Fingerprint Building and Machine Learning

Fingerprint A was based on passing minimum thresholds of diagnostic performance metrics. For this fingerprint the best performing ML classifier was the support vector machine with an accuracy of 82.7% (71.5-93.9% 95% CI) and an AUC of 0.86 (0.68-1.00 95% CI). The ROC curve is shown in Figure 3.6a. The diagnostic performance thresholds were set at equal to or higher than a Machine Learning model that classified at random. Therefore, all features included in Fingerprint A will hold some diagnostic utility. Minimum accuracy is adjusted to 0.7 from 0.5 due to the imbalanced dataset.

Table 3.2: A description of patient demographics
Key - Large Vessel Vasculitis (LVV), Giant Cell Arteritis (GCA), Takayasu's arteritis (TAK), IgG4 related disease (IgG4) and Retroperitoneal Fibrosis (RPF), Not Applicable (n/a), CRP (C-reactive Protein), ESR(Erythrocyte sedimentation rate)

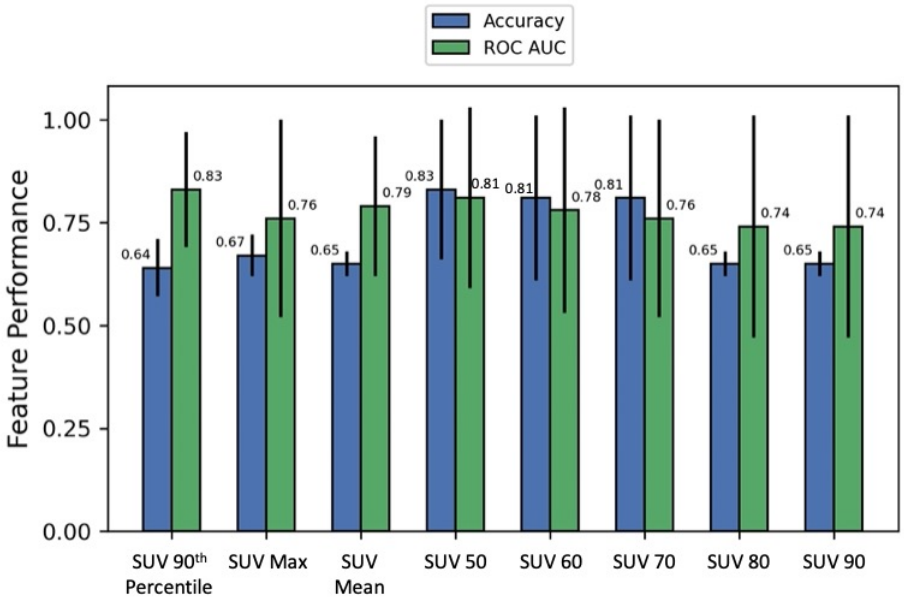
Characteristic	Aortitis	Controls
Participants	50	25
Age at time of scan, years - median (range)	60 (41-84)	68 (37-82)
Sex (male/female)	17/33	13/12
LVV type	GCA: 37, TAK: 4, IgG4 or RPF: 4, Misc: 5	n/a
Prednisolone dose (at time of scan, mg - median (range))	0 (0-40)*	0 (0-60)
Polymyalgic symptoms	yes (n=15), no (n=24), not known (n=11)	n/a
Cranial Symptoms	yes (n=11), no (n=25), not know (n=14)	n/a
Claudication	yes (n=12), no (n=25), not known (n=13)	n/a
CRP (mg/L) - median (range)	39 (5-164), not performed (n=8), not known (n=1)	n/a
ESR (mm/Hr) - median (range)	54 (0-143), not performed (n=32), not known (n=3)	n/a
Blood Glucose (mmol/L) - median (range)	5.7 (4.2-9.9)	5.9 (4.2-12.0)

*12 Aortitis Patients were taking presnisolone at the time of scanning at the following doses: $<$ 5mg (n=7), 20mg (n=1), 25mg (n=2), 40mg (n=2)

Table 3.3: Grading of patient dataset based on the EANM/SNMMI guidelines [53]

Grade	No. of Scans (LVV)	No. of Scans (Control)	Ground Truth Diagnosis of Aortitis	Ground Truth Diagnosis of No Aortitis (Control)
0	0	25	0	25
1	1	0	0	0
2	0	0	0	0
3	49	0	50	0

(a) SUV metrics



(b) 5-best performing radiomic features

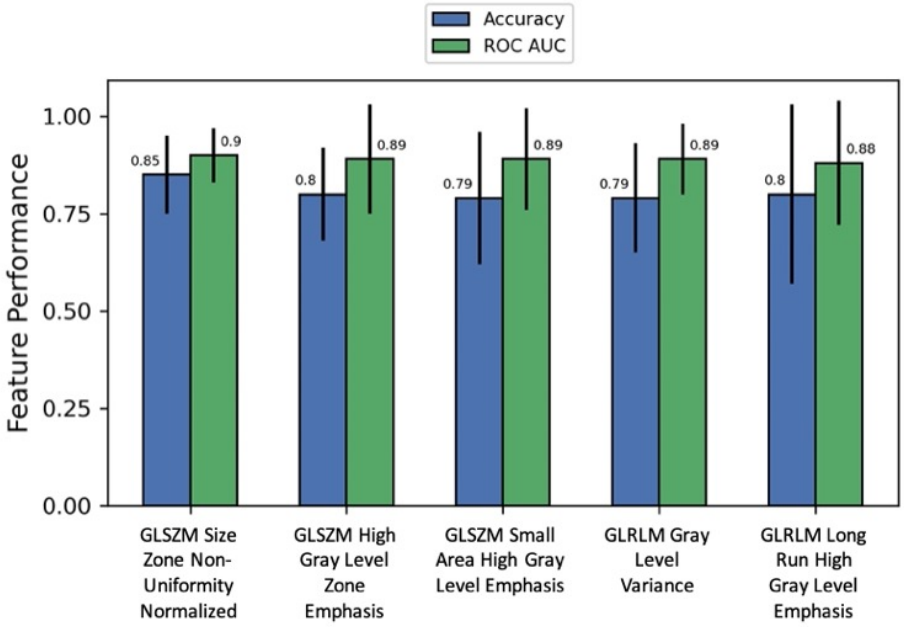


Figure 3.2: Diagnostic utility of SUV metrics and the 5-best performing radiomic features for distinguishing active aortitis after harmonization.

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*).

https://doi.org/10.1007/s12350-022-02927-4

	GLSZM Size Zone Non-Uniformity Normalized	GLSZM High Gray Level Zone Emphasis	GLSZM Small Area High Gray Level Emphasis	GLRLM Gray Level Variance	GLRLM Long Run High Gray Level Emphasis	SUV 90th Percentile	SUV Maximum	SUV Mean	SUV 50	SUV 60	SUV 70	SUV 80	SUV 90
GLSZM Size Zone Non-Uniformity Normalized	1.00	0.48	0.49	0.42	0.44	0.28	0.15	0.13	0.15	0.15	0.14	0.14	0.14
GLSZM High Gray Level Zone Emphasis	0.48	1.00	0.99	0.86	0.95	0.91	0.58	0.73	0.57	0.54	0.52	0.53	0.56
GLSZM Small Area High Gray Level Emphasis	0.49	0.99	1.00	0.90	0.91	0.87	0.61	0.67	0.60	0.57	0.55	0.55	0.58
GLRLM Gray Level Variance	0.42	0.86	0.90	1.00	0.70	0.72	0.55	0.41	0.56	0.53	0.50	0.49	0.52
GLRLM Long Run High Gray Level Emphasis	0.44	0.95	0.91	0.70	1.00	0.92	0.49	0.84	0.46	0.44	0.43	0.43	0.47
SUV 90th Percentile	0.28	0.91	0.87	0.72	0.92	1.00	0.47	0.93	0.45	0.43	0.41	0.41	0.45
SUV Maximum	0.15	0.58	0.61	0.55	0.49	0.47	1.00	0.39	0.98	0.99	0.99	0.99	1.00
SUV Mean	0.13	0.73	0.67	0.41	0.84	0.93	0.39	1.00	0.35	0.34	0.33	0.33	0.37
SUV 50	0.15	0.57	0.60	0.56	0.46	0.45	0.98	0.35	1.00	0.99	0.98	0.98	0.98
SUV 60	0.15	0.54	0.57	0.53	0.44	0.43	0.99	0.34	0.99	1.00	0.99	0.99	0.99
SUV 70	0.14	0.52	0.55	0.50	0.43	0.41	0.99	0.33	0.98	0.99	1.00	1.00	0.99
SUV 80	0.14	0.53	0.55	0.49	0.43	0.41	0.99	0.33	0.98	0.99	1.00	1.00	1.00
SUV 90	0.14	0.56	0.58	0.52	0.47	0.45	1.00	0.37	0.98	0.99	0.99	1.00	1.00

Figure 3.3: Correlation matrix of the best performing radiomic features and SUV metrics (harmonized). Key - *SUV* (*standardized uptake value*), *GLCM* (*Gray-Level Co-Occurrence Matrix*), *GLRLM* (*Gray-Level Run Length Matrix*), and *GLSZM* (*Gray-Level Size Zone Matrix*). This figure by Duff *et al.* is licensed under CC BY 4.0. https://doi.org/10.1007/s12350-022-02927-4

Fingerprint B was built using the same thresholds but also removed highly correlated features. For this fingerprint the best performing ML classifier was random forest with an accuracy of 84.0% (72.8-95.2% 95% CI) and an AUC of 0.91 (0.80-1.00 95% CI). The ROC curve is shown in Figure 3.6b. The results are not sensitive to the correlation threshold. Varying the threshold between 70-95% (generally considered range for high correlation) shows almost no variation in the best results. Some variation can be seen in the machine learning models that do not perform well but these would not be utilised in a final analytical pipeline so are not considered important.

Six principal components were produced to account for 90% of the information in the original dataset. These principal components were used in Fingerprint C. The best performing ML classifier was support vector machine with an accuracy of 82.7% (71.5-93.9% 95% CI) and an AUC of 0.87 (0.74-1.00 95% CI). The ROC curve is shown in Figure 3.6c.

The performance of all ML classifiers with Fingerprints A, B and C can be viewed in Tables 3.5a, 3.5b, 3.5c, respectively.

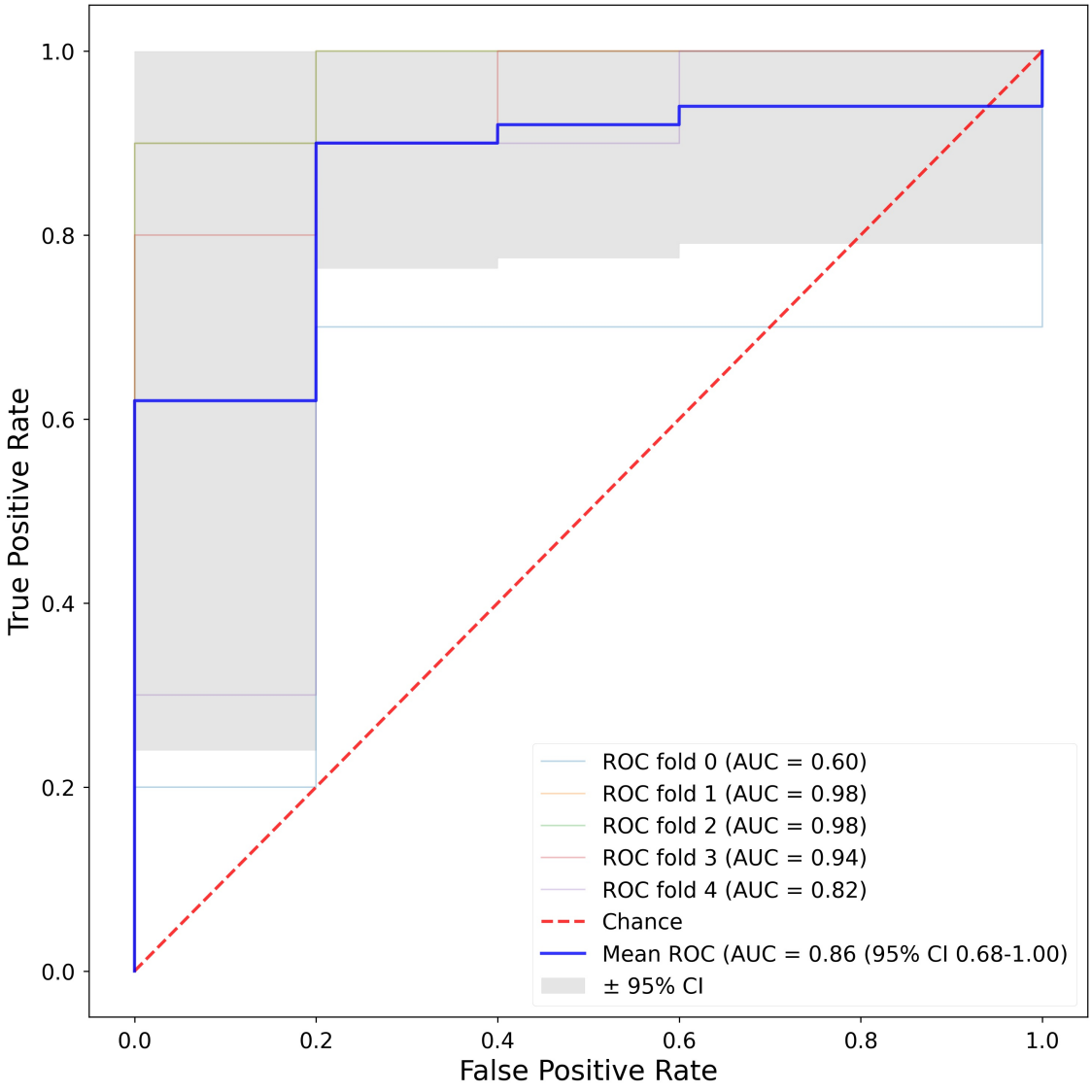
3.2.8 The Utility of Harmonization

The Mann-Whitney U test was used to evaluate the effect of harmonization. The null hypothesis was defined as both feature distributions (before and after) being from the same population. The average p value increased in all cases as did the number of features where the null hypothesis was accepted (Table 3.4). When the two GE scanners were compared with the Mann-Whitney U test, we found sufficient difference that we chose to analyse them separately rather than combining the two GE scanners into a single batch.

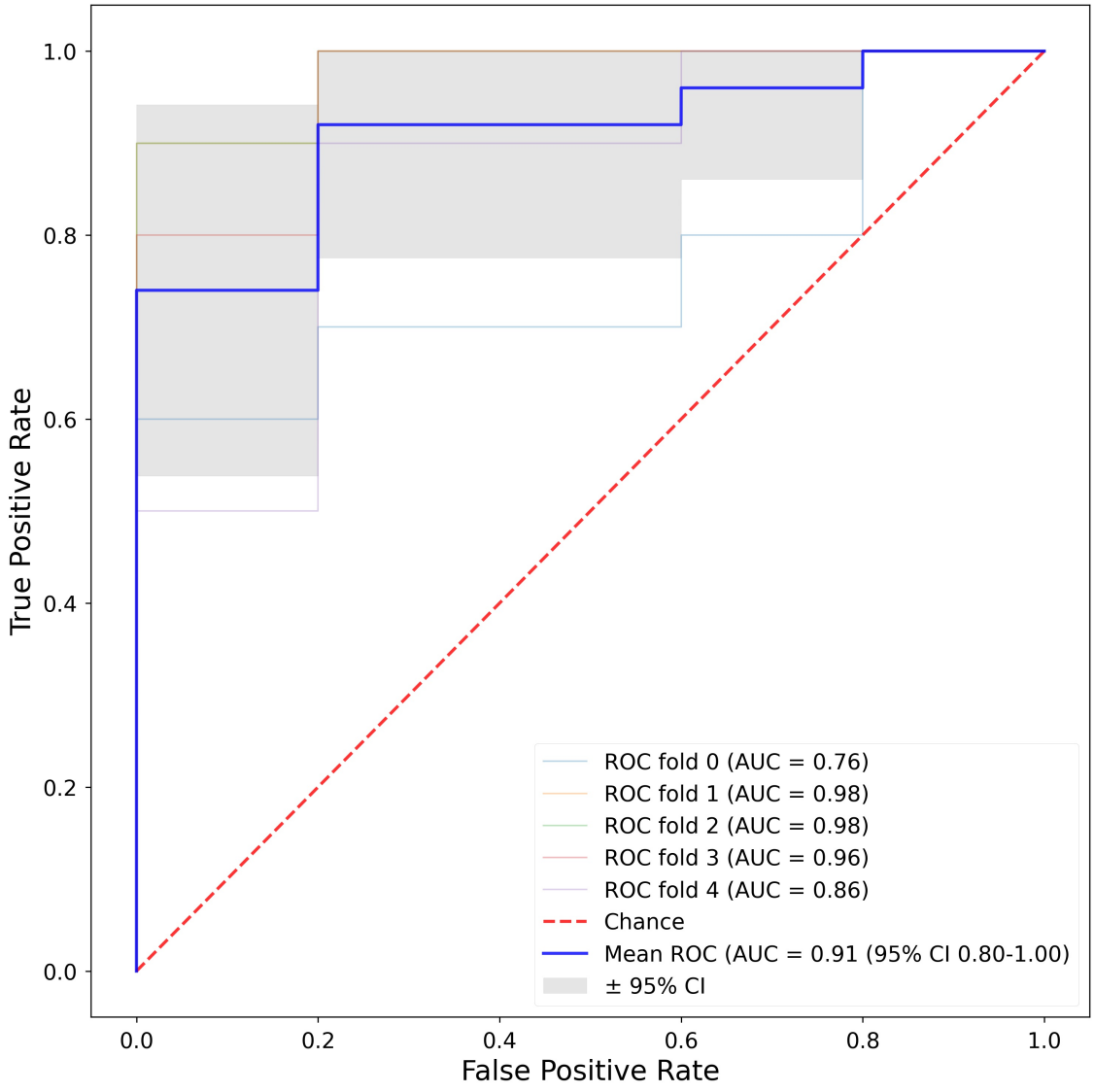
Figures 3.5a and 3.5b show the accuracy and AUC of non-harmonized SUV metrics and radiomics features respectively. The main difference between the two sets of results is a different set of radiomics features being ranked in the top five however overall performance of each feature was similar. The confidence intervals are too large to determine if there is a significant difference. No noticeable decrease in diagnostic utility along with the results from Table 3.4 justify keeping harmonization as a step in the proposed methodology to improve the potential for generalizability. Further investigation into the utility of

Table 3.4: Mann Whitney U test results when feature distributions were compared before and after harmonization. Key – Scanner 1 = GE Discovery 710, Scanner 2 = Phillips Gemini TF64, Scanner 3 = GE Discovery 690

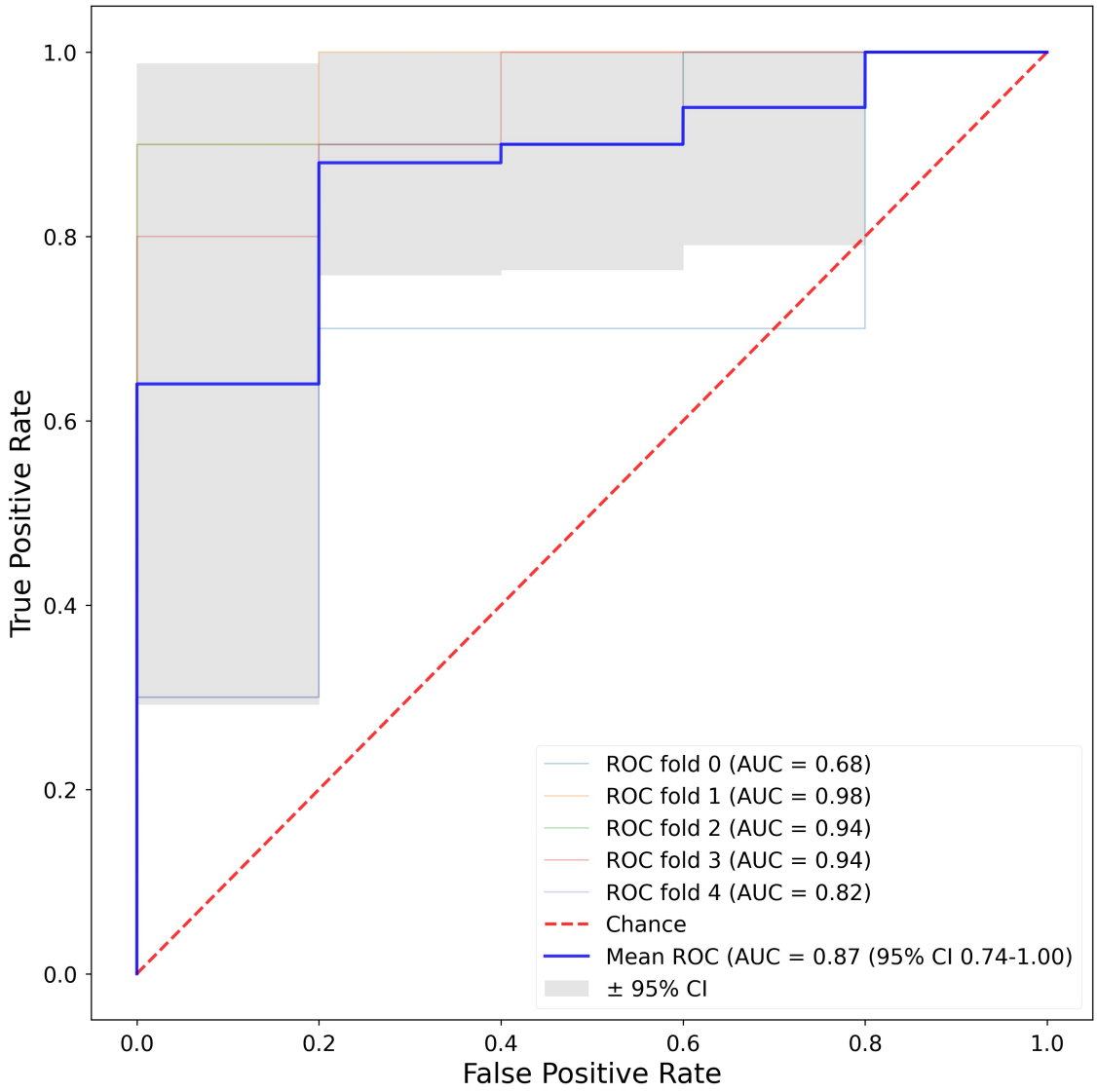
	Before Harmonization			After Harmonization		
Scanners compared	1 vs 2	2 vs 3	1 vs 3	1 vs 2	2 vs 3	1 vs 3
Number of features where the null hypothesis was accepted (out of 107)	52	97	66	81	99	85
Average p value	0.148	0.224	0.144	0.199	0.230	0.182



(a) Fingerprint A - Support Vector Machine Learning Classifier



(b) Fingerprint B - Random Forest Classifier



(c) Fingerprint C - Support Vector Machine Classifier

Figure 3.4: ROC curves of the best performing machine learning classifier trained on Fingerprints A, B and C. This figure by Duff *et al.* is licensed under CC BY 4.0. https://doi.org/10.1007/s12350-022-02927-4

Table 3.5: All fingerprint ML classifier results. Key - ML = Machine Learning, ACC = Accuracy , CI = Confidence Interval, AUC = Area Under the Receiver Operating Characteristic Curve.

ML Type	ACC	ACC CI ($\pm$)	AUC	AUC CI ($\pm$)
Random Forest	0.760	0.170	0.838	0.187
Logistic Regression	0.787	0.122	0.804	0.114
Support Vector Machine	0.827	0.112	0.864	0.179
Decision Tree	0.813	0.142	0.810	0.168
Gaussain Process Classifier	0.333	0.000	0.500	0.000
Perceptron	0.627	0.193	0.772	0.288
Passive Aggressive	0.680	0.122	0.772	0.098
Neural Net	0.760	0.112	0.808	0.123
K Nearest Neighbour	0.800	0.209	0.826	0.179

(a) Fingerprint A ML Classifier Results

ML Type	ACC	ACC CI (±)	AUC	AUC CI (±)
Random Forest	0.840	0.112	0.908	0.107
Logistic Regression	0.693	0.112	0.868	0.116
Support Vector Machine	0.813	0.097	0.860	0.158
Decision Tree	0.760	0.134	0.806	0.168
Gaussain Process Classifier	0.480	0.110	0.500	0.000
Perceptron	0.640	0.066	0.720	0.207
Passive Aggressive	0.667	0.052	0.884	0.163
Neural Net	0.640	0.154	0.642	0.247
K Nearest Neighbour	0.827	0.112	0.822	0.181

(b) Fingerprint B ML Classifier Results

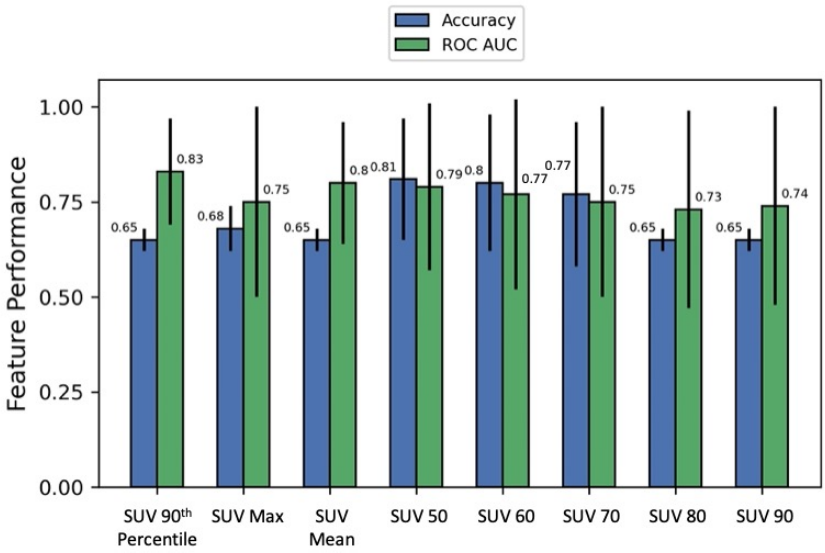
ML Type	ACC	ACC CI (±)	AUC	AUC CI (±)
Random Forest	0.827	0.084	0.866	0.104
Logistic Regression	0.747	0.184	0.796	0.177
Support Vector Machine	0.827	0.112	0.872	0.137
Decision Tree	0.720	0.169	0.724	0.203
Gaussain Process Classifier	0.813	0.152	0.852	0.137
Perceptron	0.747	0.177	0.800	0.174
Passive Aggressive	0.733	0.074	0.792	0.105
Neural Net	0.827	0.124	0.852	0.122
K Nearest Neighbour	0.773	0.124	0.792	0.198

(c) Fingerprint C ML Classifier Results

harmonization will be conducted in the multi-centre validation in the following chapter but a likely reason for its lack of impact in these results is that all images were preprocessed to the same voxel size removing a major source of variability [141]. With further prepossessing, such as filtering, the need for harmonization may be further reduced.

When the three fingerprints were built using non-harmonized features there was no significant change to results. A slight improvement can be seen in Fingerprint A when the data is not harmonized and it is the only fingerprint where a different classifier is the highest ranked (random forest instead of support vector machine). It is of interest that random forest is the best classifier for Fingerprint B in both harmonized and non-harmonized cases meaning the non-harmonized Fingerprint A may be more similar to Fingerprint B. Overall, there is not enough evidence to select non-harmonized or harmonized as the superior method so both results were retained.

(a) Diagnostic utility metrics of of SUV metrics for distinguishing active aortitis using logistic regression classifiers - before harmonization



(b) Diagnostic utility metrics of the 5-best performing radiomic features for distinguishing active aortitis using logistic regression classifiers - before harmonization

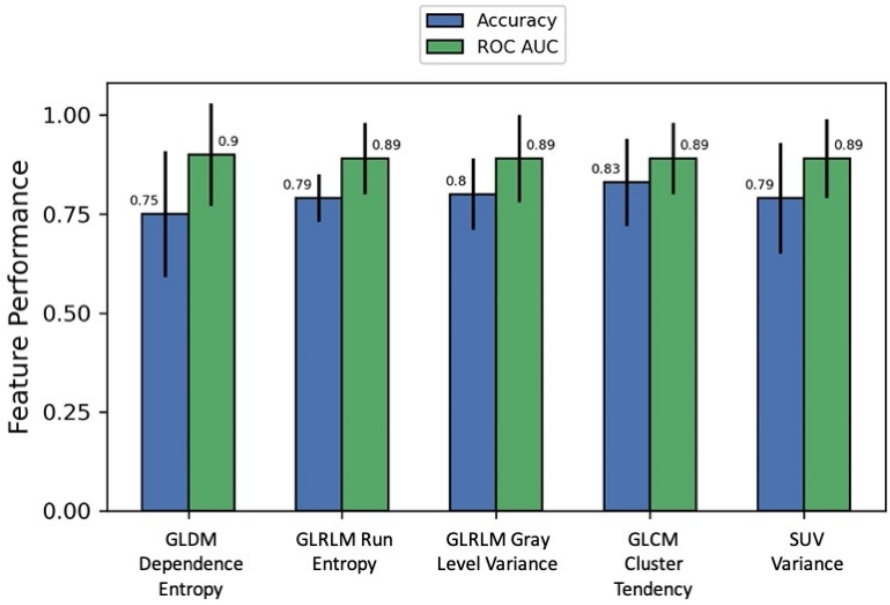
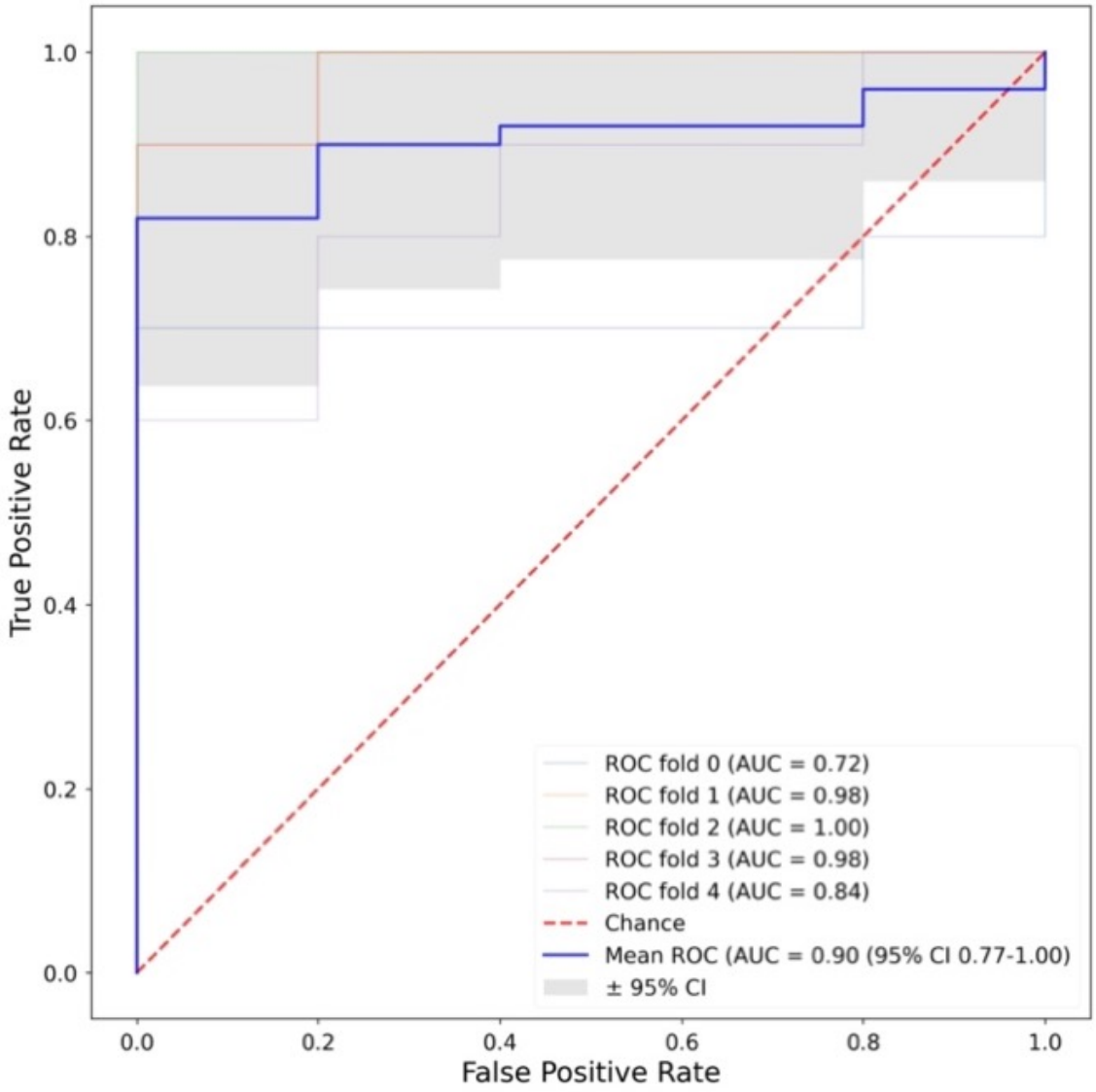
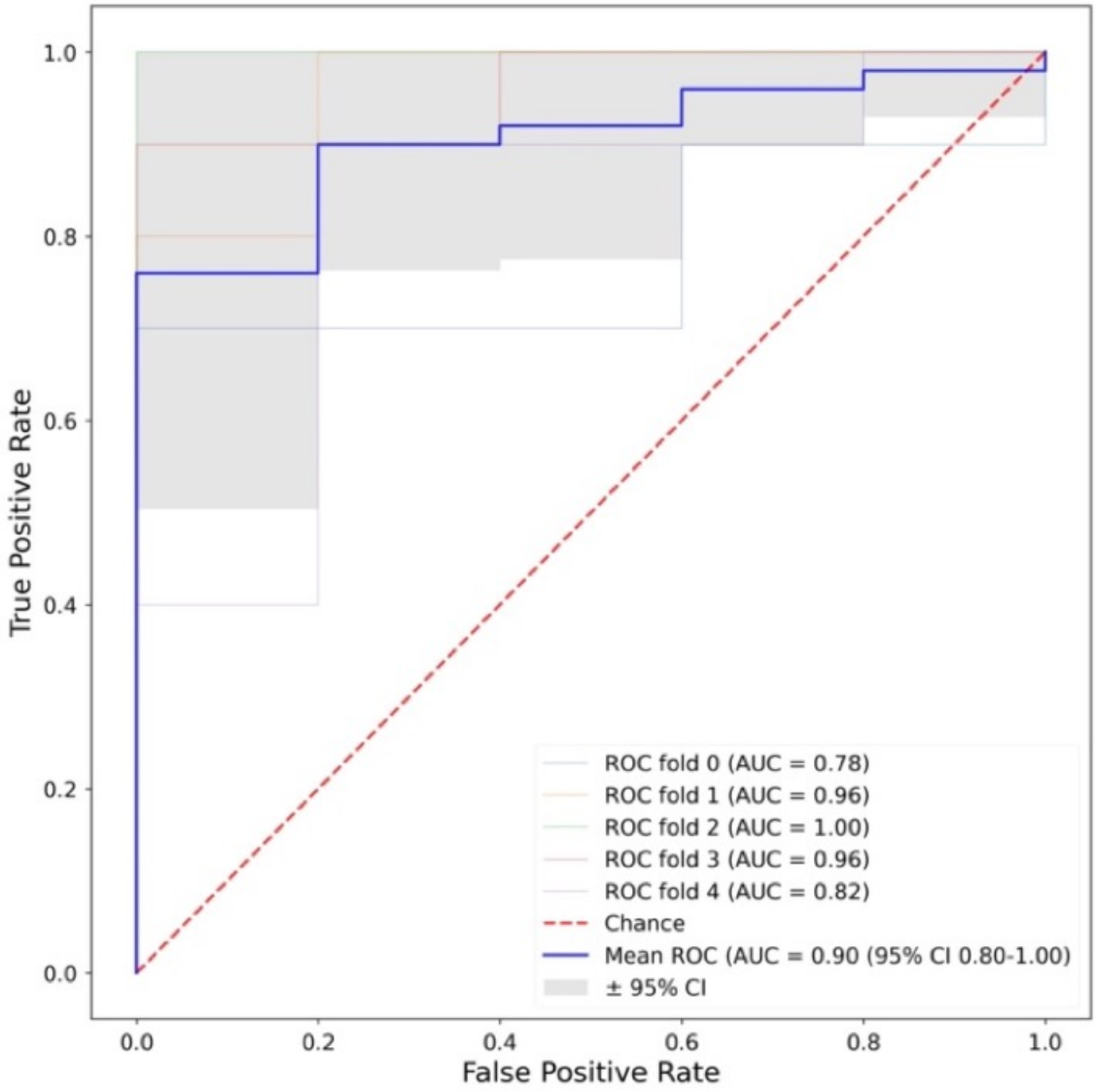


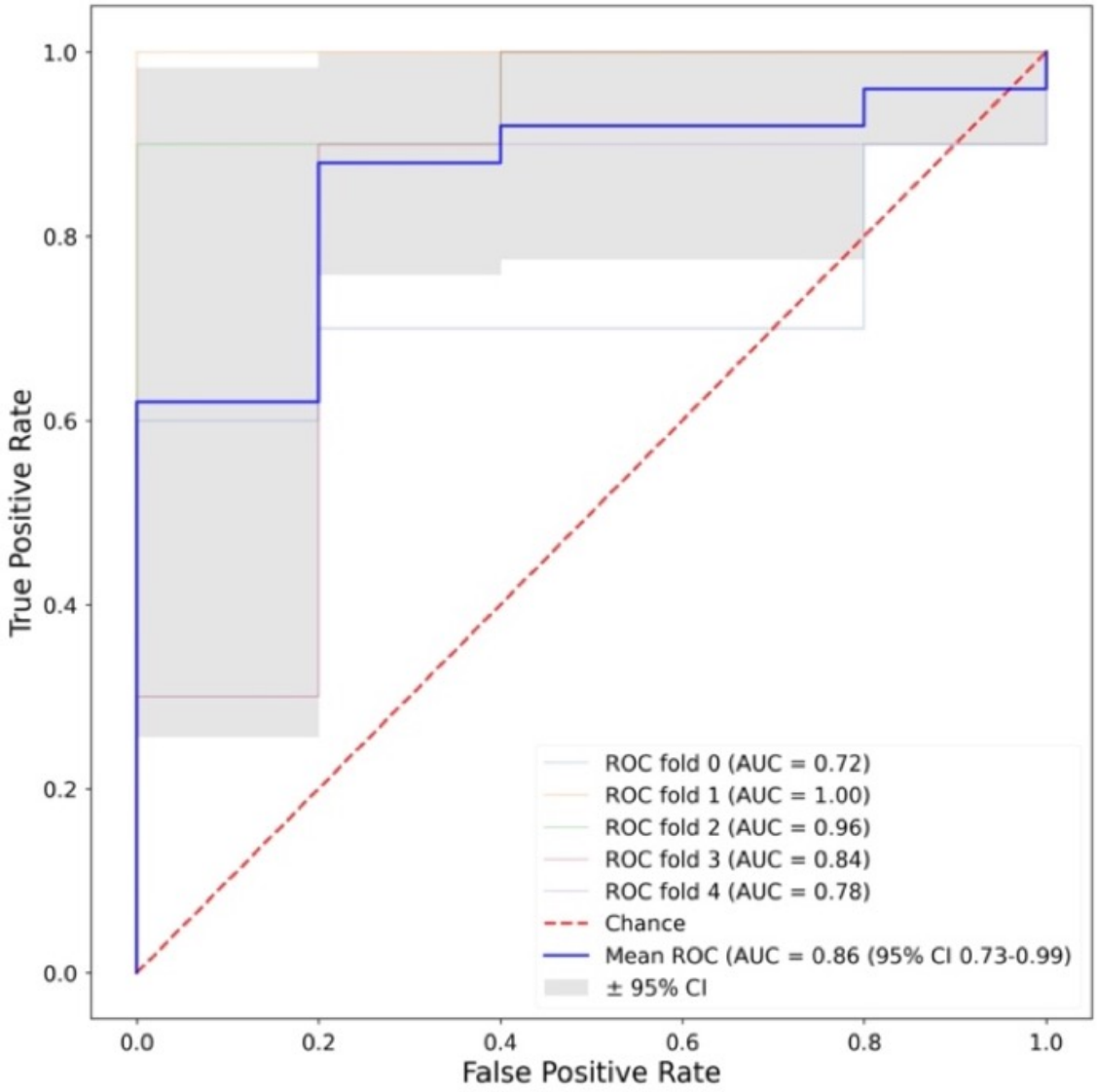
Figure 3.5: Diagnostic utility of SUV metrics and the 5-best performing radiomic features for distinguishing active aortitis - before harmonization. Key - *SUV(standardized uptake value)*, *GLDM* (*Gray-Level Dependence Matrix*), *GLCM* (*Gray-Level Co-Occurrence Matrix*), *GLRLM* (*Gray-Level Run Length Matrix*), and *GLSZM* (*Gray-Level Size Zone Matrix*). This figure by Duff *et al.* is licensed under CC BY 4.0.
https://doi.org/10.1007/s12350-022-02927-4



(a) Fingerprint A - Random Forest Classifier – non-Harmonized



(b) Fingerprint B - Random Forest Classifier - non-Harmonized



(c) Fingerprint C - Support Vector Machine Classifier - non-Harmonized

Figure 3.6: ROC curves of the best performing machine learning classifier trained on Fingerprints A, B and C when the radiomic features were not harmonized. This figure by Duff *et al.* is licensed under CC BY 4.0.
https://doi.org/10.1007/s12350-022-02927-4

Table 3.6: Summary of the best diagnostic performance of each method. 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*)

Method				
Qualitative Assessment AUC - Literature [53]	0.81-0.98			
	Harmonized		Non-harmonized	
	AUC	AUC 95% CI (±)	AUC	AUC 95% CI (±)
SUV Feature - SUV 90th percentile	0.83	0.14	0.83	0.14
Radiomic Feature - GLSZM Size Zone Non Uniformity Normalized (harmonized) / GLDM Dependence Entropy(non-harmonized)	0.9	0.07	0.9	0.13
Fingerprint A	0.86	0.18	0.9	0.13
Fingerprint B	0.91	0.11	0.9	0.11
Fingerprint C	0.87	0.14	0.86	0.14

3.2.9 Summary of Diagnostic Performance

A summary of the diagnostic performance of each method is shown in Table 3.6. The AUC range presented for qualitative assessment were determined by a meta-analysis exploring the diagnostic accuracy of FDG PET-CT imaging in LVV [53]. In the case of SUV metrics and radiomic features the best individual feature was determined by their AUC but with a minimum accuracy of 70%. The best SUV metric and radiomic feature for distinguishing aortitis was SUV 90^{th} percentile and GLSZM High Gray Level Zone Emphasis, respectively.

3.3 Discussion

The purpose of this set of experiments was to develop a methodological framework to support AI-assisted diagnosis of active aortitis, using ML classifiers trained with radiomic features from FDG PET-CT. The best performing individual radiomic feature (GLSZM Size Zone Non-Uniformity Normalized) had an AUC of 0.9 (0.83-0.97 95% CI), similar to the current clinical standard of qualitative assessment by an experienced nuclear medicine physician (AUC=0.81-0.98 [53]). The three fingerprints performed similarly to the best-performing individual radiomic features. Of particular promise is Fingerprint B with an

AUC of 0.91 (0.80-1.00 95% CI). This was the highest AUC of any of the proposed methods and taking the confidence intervals into account performs very similar to the current standard of qualitative assessment [53]. This method has potential to be used as an automated quantitative analysis tool alongside standard clinical assessment towards a more rapid, objective, and standardized evaluation of aortitis to guide therapeutic choices.

Visual scores assigned using the EANM/SNMMI guidelines [53] showed good agreement with ground truth diagnoses. As borderline cases weren't used in our analysis all but one case was graded as either 0 or 3 meaning there was no uptake or high-grade uptake respectively. One case was graded as 1 (low grade uptake) but this reduced signal was a result of prednisolone treatment (25mg daily) which dimishes PET sensitivity. A similar scoring system based on arterial uptake across different regions was proposed by Grayson *et al.* named PET Vascular Activity Score (PETVAS) [113]. PETVAS is not routinely used in clinical practice as it is time consuming. Kang *et al.* showed that PETVAS is superior to $\mathrm{SUV}_{\mathrm{max}}$, but it is unclear if it is better than a single visual score assigned using the EANM/SNMMI guidelines [109].

The diagnostic utility of semi-quantitative measurements using SUV, which are widely utilised in PET, was compared against other features for detecting active aortitis. *SUV* 90^{th} *percentile* (90% of voxels in the ROI are less than this value) performed best with an AUC of 0.83, only slightly below the best performing radiomic features. Overall SUV metrics demonstrated some utility for distinguishing aortitis from controls when measured with Mann Whitney U and logistic regression classifier testing but radiomic features were superior. The performance of SUV_{max} is disproportionately affected by noise. This may suggest why it did not perform well but *SUV* 90^{th} *percentile* did despite it also being based on high intensity values [212]. *SUV* 90^{th} *percentile* gives the value for which 90% of SUV values lie beneath meaning outliers created by noise are removed. Similarly, atherosclerosis can be associated with FDG activity and although patients and controls with a large amount of atherosclerotic plaque were removed from the cohort some degree of the condition is present in the relevant age group [54]. Together, these two factors may have lowered the diagnostic utility of SUV_{max}. The ability to reliably distinguish aortitis from atherosclerosis will need to be considered in any automated diagnostic methods and in any longitudinal studies conducted in the future. SUV_x also relies on SUV_{max}. In

particular, SUV_{50} perform s better than other SUV_x metrics, probably because it covers a larger percentage of the voxels, so the effect of noise and bright patches is mitigated. *SUV* 90^{th} percentile is a result of high activity over a larger volume so it is more resistant to small focal areas of high activity. SUV_{mean} and SUV_{50} would likely perform better if only active tissue had been included in the ROI rather than the whole aorta.

All SUV features explored were found to be significantly different between patients with active aortitis and controls using the Mann Whitney U Test. Radiomic features shown to have the highest diagnostic utility focus mainly on high gray levels and heterogeneity. The *GLSZM Size Zone Non-Uniformity Normalized* was the best radiomic featureaccording to AUC and performed also well in terms of accuracy and the Mann Whitney U test. Its value is higher in active aortitis than controls, which means there is more heterogeneity in zone size volumes in aortic LVV imaging. This is an expected finding as it reflects greater metabolic activity in the aortic wall of patients with active aortitis than in controls. However, since it has superior performance compared to SUV parameters which are based on metabolic activity, this is a potentially useful new association. The importance of high gray values and zones, and heterogeneity is further emphasised in other radiomic features with high diagnostic utility. The addition of heterogeneity, encompassing spacial relationship between voxels, may help explain why radiomic features outperformed SUV metrics which focus on voxel values alone.

Chapter 4

Experiment Set 2 - Method Automation and Validation

In the previous chapter a methodological framework for assisting the diagnosis of active aortitis using radiomic analysis of FDG PET-CT was established [213]. In this chapter the aim was to continue this work by developing, testing and validating with multi-centre data an automated radiomic analysis pipeline to assist the diagnosis of active aortitis. 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. By validating with external data and automating the process this addresses two key barriers to clinical transferability [214]. Places where the methodology varies from the previous chapter is explained.

An updated version of this chapter has now been published in Biomolecules [215]. In this paper reviewers comments were addressed altering the numerical results slightly but the conclusions remained the same. The most significant change was an improvement in Fingerprint B.

4.1 Method

4.1.1 Patient Selection

Figure 4.1 demonstrates the distribution of the imaging cohorts - training, test and validation. The data acquired from Leeds Teaching Hospitals NHS Trust was split into training and test (80:20) datasets. The training dataset was used to train ML models including optimisation 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.

4.1.1.1 Training and Test Patient Dataset

The training and test dataset was procured from Leeds Teaching Hospitals NHS Trust. The same collating procedure, determination of the ground truth diagnosis, and inclusion and exclusion criteria were used as described in the previous chapter. The only additional criteria applied was that only aortitis caused by GCA and TAK were included in order to homogenize the dataset. The overall dataset is mostly the same as the data used in the previous chapter but some additional cases were added. The collated data was then split into the training and test dataset (80:20) unlike the previous chapter.

4.1.1.2 Validation Patient Dataset

To evaluate multi-centre transferability, a validation dataset was formed using data from external institutions. Data from patients recruited to the UK GCA consortium (REC Ref. 05/Q1108/28) [216] with suspected aortitis, and had FDG PET-CT scans performed as part of routine clinical care at Alliance Medical Ltd (AML) centres in England was extracted from the organizational PACS (IntelePACS Version 4, Intelerad 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

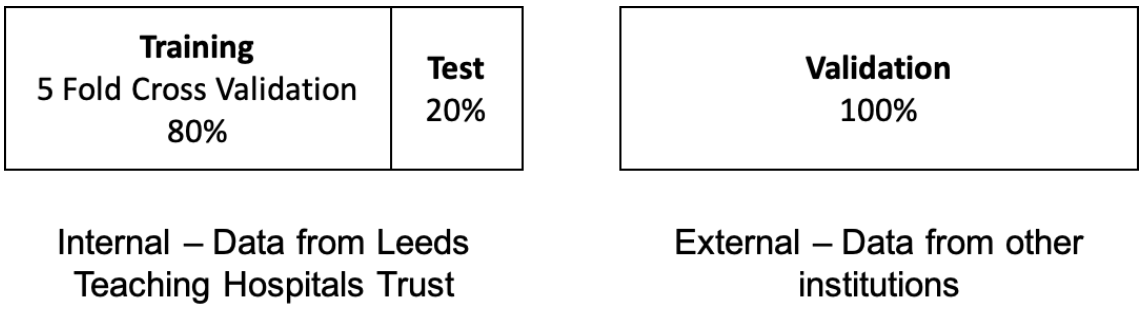


Figure 4.1: The distribution of datasets into training, test and validation cohorts

by data from the PITA (PET Imaging of Giant Cell and Takayasu Arteritis) (REC approval: 19/EE/0043 Clinical trials registration: NCT04071691) study in the University of Cambridge and Imperial College London.

4.1.2 Imaging Protocol

FDG PET-CT scans from all three cohorts were acquired using the imaging protocol described in the previous chapter. Nine scanners from 3 different manufacturers were used. Table 4.1 describes the acquisition parameters in further detail.

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)[1].

[1] www.slicer.org/wiki/Documentation/Nightly/Modules/SUVFactorCalculator - Accessed September 2021

Table 4.1: 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 x 4.00 x 4.00
Discovery 710	VPFX, QCFX, or VPHD	Model based	Singles	192	3.65 x 3.65x 3.27
Discovery 690	VPFX or VPFX	Model-based	Singles	193	3.65 x 3.65x 3.28
Discovery MI DR	VPFX, QCFX, or VPHD	Model-based	SING	256	2.73 x 2.73 x 3.27
Discovery ST	OSEM	Convolution subtraction	DLYD	128	4.69x 4.69 x 3.27
Discovery STE	OSEM	Convolution subtraction	SING	128	5.47 x 5.47 x 3.27
Biograph6 TruePoint	OSEM2D 4i8s	Model-based	DLYD	168	4.07 x 4.07 x 3.00
Biograph 6	OSEM2D 4i8s	Model-based	DLYD	168	4.07 x 4.07 x 3.00
Biograph64 mCT	PSF+TOF 2i21s or OSEM3D 2i24s	Model-based	DLYD	200	4.07 x 4.07 x 3.00

4.1.3 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 in the previous chapter 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 [2]) [206, 207]. The CT component was used as the main reference as it provides more anatomical information but the result was checked against the PET scan.

[2]https: //www.slicer.org/

The DSC where values range from 0 to 1 where a higher value is a higher degree of similarity (Eqn. 2.4) [217] was used to evaluate segmentation quality.

The PET and CT components, and segmented masks were then resampled to a 4mm 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 3 scanners in the training and test dataset meaning downsampling alone was applied. A lower resolution was present in multi-centre data collected later (5.47 mm) but the 4mm 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 was manually checked to ensure 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 4.2. Training was undertaken on ARC4, 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 minute 12 seconds.

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 (Eqn. 2.4). 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 leaky rectified linear unit (ReLU). Convolution stride was 1 and pooling stride was 2. Kernel size was 3x3 for convolution and 2x2 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 and other parts of the body that were 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 CNN outputs did not fully fill the aorta or encapsualte the aorta wall so a dilation filter was applied using Simple ITK (Version 2.01). The segmented slices were then reassembled into 3D volumes for use in feature extraction (Section 4.1.5).

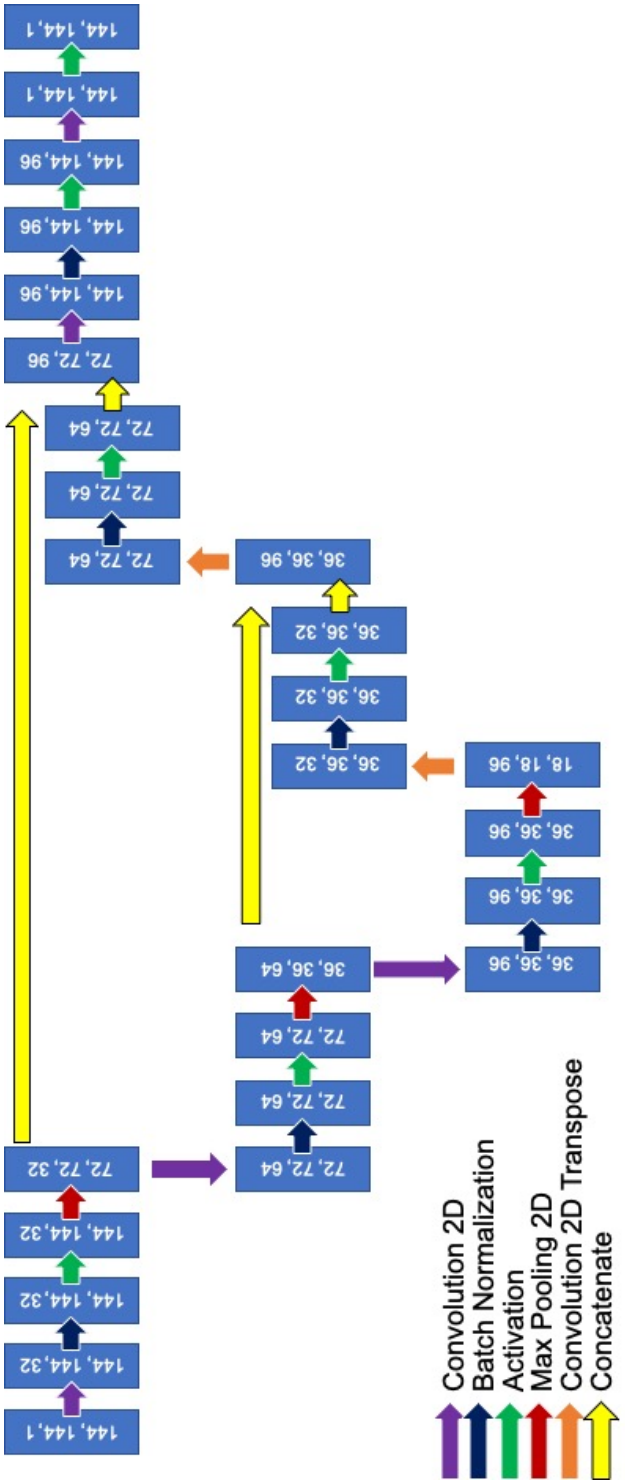


Figure 4.2: Architecture of convolutional neural network (CNN) used to segment the aorta. An explanation of each component can be found in section 2.3.1.

4.1.4 Qualitative Grading of Vessel Wall FDG Activity

All scans were evaluated based on EANM/SNMMI guidelines [53] 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)

4.1.5 Feature Extraction

Radiomic features encompass a large number of quantitative parameters. This includes, but is not limited to, SUV metrics. SUV metrics will be referred to when studied separately from radiomic features. SUV metrics were used instead of target-to-blood pool ratio as liver is a more common reference point as discussed in Section 4.1.4.

Radiomic features (n=102) were extracted with Pyradiomics [3]. 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 [218]. In most cases Pyradiomics is broadly compliant with the IBSI standards but deviates in some cases as described in their documentation

[3] Version 3.0.1, radiomics.io/pyradiomics

[4]. This will affect some of the extracted features in this study where the feature relies 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 Supplemental Material 6.2.1.

4.1.6 Harmonization

The ComBat method (neuroCombat, Version 0.2.7) was used to reduce the effect of different imaging protocols on radiomic features [148–150]. These factors cannot be standardized retrospectively without reducing the size of the dataset, so harmonization is recommended to minimize the effect [148]. The overall dataset (training, test and validation combined) was grouped in batches as shown in Table 4.2 based on similar imaging protocol parameters. A less thorough comparison between non-harmonized and harmonized radiomic results was conducted in this chapter due to much smaller populations in each batch making comparisons such as the Mann Whitney U Test not applicable. A brief comparison of the effect on final diagnostic ability was conducted.

[4]https://pyradiomics.readthedocs.io/en/latest/faq.html

Table 4.2: Distribution of participants across scanners

	Training		Test		Validation		Harmonization Batch
Scanner	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 TruePoint	0	0	0	0	5	2	4
Biograph64 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

4.1.7 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 [147]. As the benchmark AUCs for qualitative assessment of PET-CT in suspected aortitis quoted in the literature are 0.81-0.98 [53], 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% was discounted.

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 4.1.5) they were separated and com-

pared 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 optimisation method was changed to BayesSearchCV from GridSearchCV from the previous chapter as it more thoroughly searches the parameter options. 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.

4.1.8 Forming Radiomic Fingerprints

Individually radiomic features (including SUV metrics) can be used as metrics but when used collectively they can provide complimentary information to improve diagnostic performance [219]. Using all or most extracted radiomic features can introduce a significant amount of redundant information and creates noise in the diagnostic model [124]. Therefore, radiomic fingerprints were created with the extracted radiomic features (including SUV metrics). Three radiomic fingerprints were built using the methods described below.

4.1.8.1 Fingerprint A - Performance Criteria and Correlation

Fingerprint A was produced by selecting features with high individual diagnostic utility based on their training dataset performance in section 4.1.7 : AUC$\geq$0.5, balanced accuracy$\geq$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.

4.1.8.2 Fingerprint B - 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). Fingerprint B was formed with principal components needed to account for at least 90% of variance in the radiomic data.

4.1.8.3 Fingerprint C - Random Forest

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.

4.1.9 Diagnostic Utility of Fingerprints

The diagnostic utility of radiomic fingerprints A and B were evaluated using the same methodology described in section 4.1.7 but additional ML classifiers were tested alongside logistic regression [209–211]. Ten different ML classifiers were built, trained, and tested: support vector machine, random forest, passive aggressive, LR, k nearest neighbours, perceptron, multi-layered perceptron, decision tree, stochastic gradient descent and gaussian process classification.

Fingerprint C was evaluated as in section 4.1.7 using only the Random Forest Classifier as it uses the embedded feature selection in this ML classifier.

4.2 Results

4.2.1 Patient Characteristics

Overall, 114 participants were included in the training, test and validation datasets collectively (Table 4.3). 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 glucocorticoid treatment, so glucocorticoid (prednisolone) doses were zero at the time of scanning unless stated otherwise [106]. CRP and ESR are laboratory markers of inflammation used in clinical care.

Less clinical data was available for the validation dataset but as shown in Table 4.3 the gender distribution, LVV Type, prednisolone dose, CRP, ESR, blood glucose and median age of all datasets are similar.

4.2.2 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 dilation filters added, 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 minute 12 seconds.

An example of the CNN segmentations is shown in Figure 4.3.

4.2.3 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 [53]. Table 4.4 shows the grades assigned to the training, test and validation cohorts respectively by an experienced radiologist on retrospective review of the images.

4.2.4 Diagnostic Utility of Individual SUV Metrics and Radiomic Features

Figure 4.4 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. SUV 90th percentile performed better consistently across all three cohorts with a validation AUC of 0.8 and an balanced accuracy of 62%.

The five-best performing radiomic features, when used individually in an logistic regression classifier, are shown in Figure 4.5. Performance was based on validation AUC but a minimum balanced accuracy of 50% had to be met across the training, testing and

Table 4.3: Patient demographics

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)

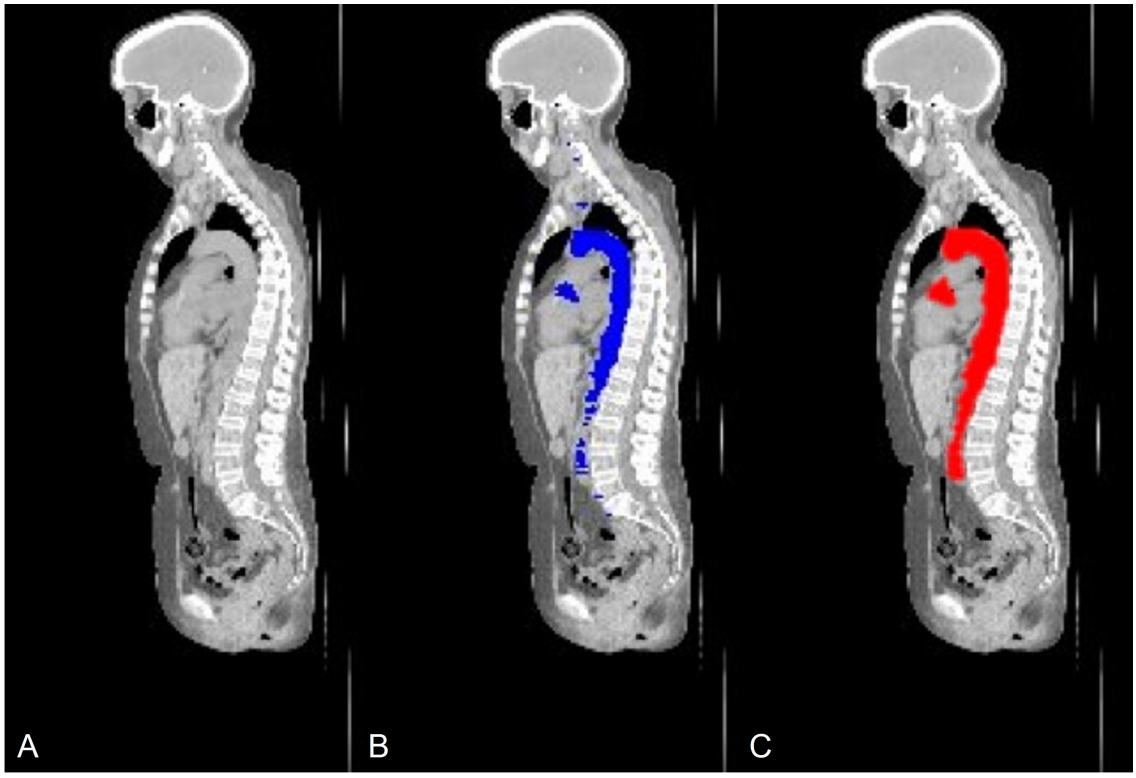


Figure 4.3: An example segmentation produced by the automated method. A) The reference CT scan B) The original output, C) The output filtered to remove pixels not part of the largest segmentation and then a dilation filter set to one pixel was applied.

Table 4.4: Grading of patient dataset based on EANM/SNMMI guidelines [53]

	Training		Test		Validation	
Grade	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

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 4.5 suggests heterogeneity is an important characteristic in distinguishing aortitis from controls. The first two features are based on energy which is based on voxel intensities. High intensities in PET images was already known as a feature of LVV. The following three features indicate the importance of heterogeneity. GLRLM Run Entropy characterises the randomness in run lengths (number of voxels of the same gray level in a run) and gray levels. GLCM Sum Entropy is the summation of the differences in neighbouring voxels intensity values. GLDLM Dependence Non-Uniformity quanitifies how dissimilar dependencies are throughout the ROI.

The performance of all individual radiomic features and SUV metrics in logistic regression classifiers, and in all three cohorts, are listed in Supplemental Material 6.2.1.

4.2.5 Diagnostic Utility of Fingerprints

Fingerprint A was based on minimum thresholds of diagnostic performance for each feature and a maximum correlation to other features. While Perceptron and Passive Aggressive classifiers produced higher validation AUCs their accuracies were low ($\leq$ 50%) in either one or both of the testing and validation cohorts. Logistic Regression performed more consistently across the training, testing and validation cohorts in both AUC and balanced accuracy (Figure 4.6), suggesting this method may have multi-centre transferability. Random Forest also performed well in all three cohorts. The performance of all explored ML classifiers is shown in Table 4.5.

Fingerprint B was based on PCA. For this fingerprint the best validation AUC was achieved by a Passive Aggressive Classifier with a validation AUC=0.65 (Figure 4.7). In the case of Fingerprint B it was not possible to filter the model performances based on balanced accuracy as all produced an balanced accuracy of $\leq$ 50% in at least one cohort. Overall, the performance in the testing and validation cohorts was poor demonstrating that this method is not generalizable or transferable. The performance of all explored ML classifiers is shown in Table 4.6.

Fingerprint C used the feature selection that is intrinsically part of Random Forest

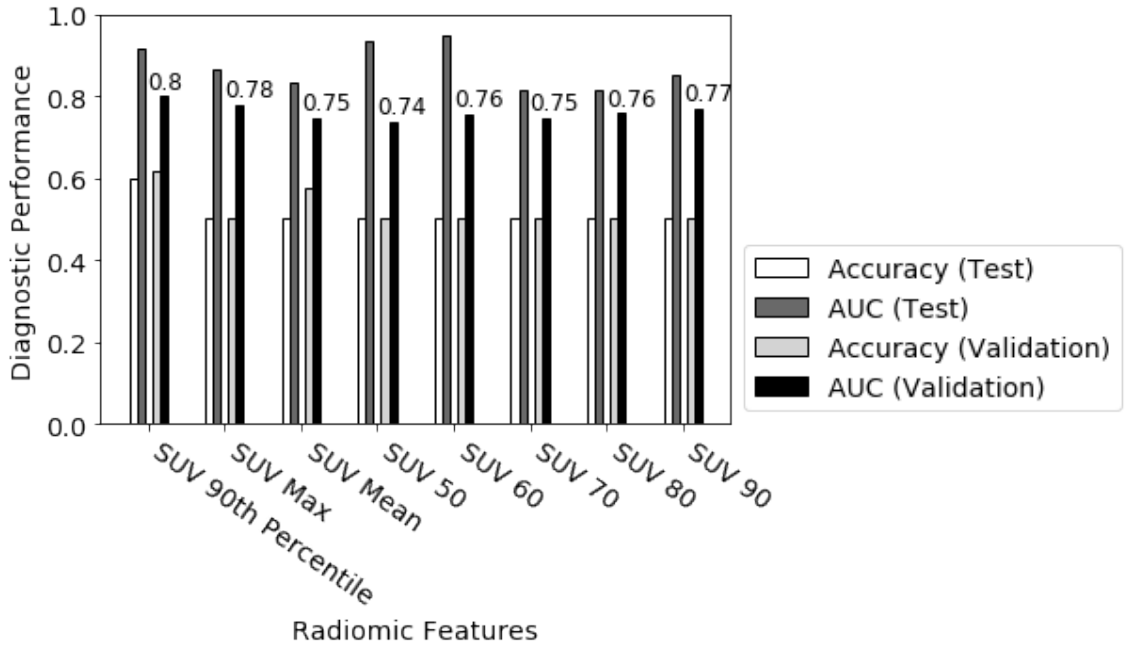


Figure 4.4: The diagnostic utility, expressed as balanced accuracy and AUC, of individual SUV metrics to demonstrate where current semi-quantitative methods stand

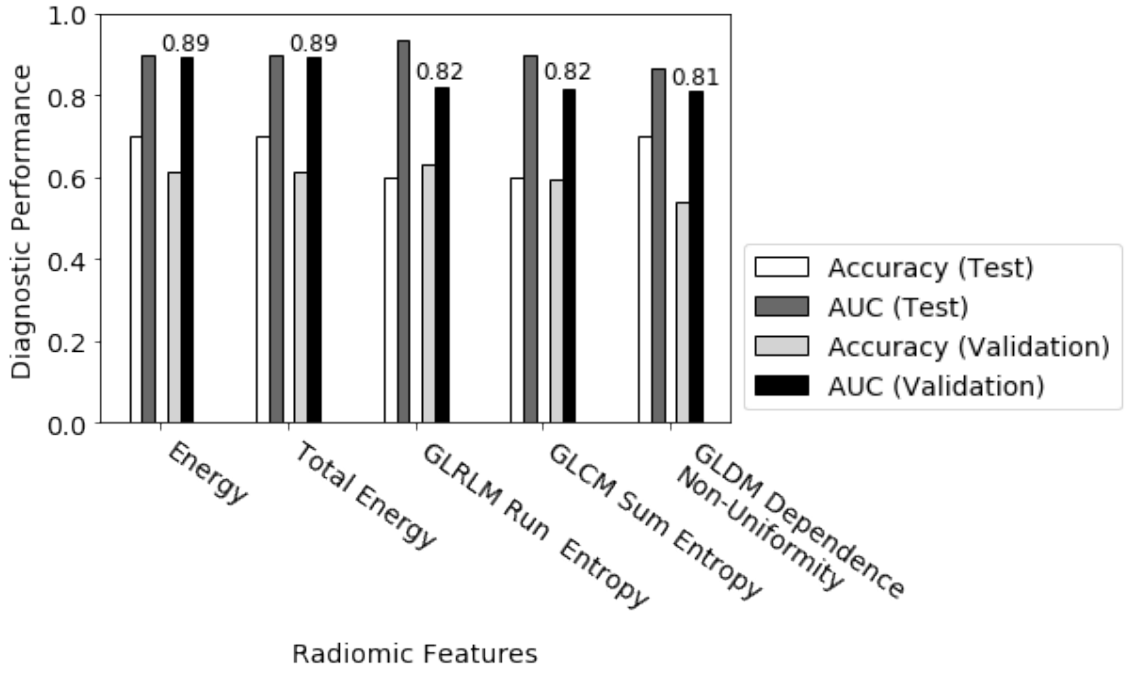


Figure 4.5: Diagnostic utility of the five highest performing individual radiomic features - performance ranked by validation AUC with an balanced accuracy above 50%

classification and did not include any other ML classifiers. This method produced good results in both the testing (balanced accuracy = 66%, AUC = 0.90) and validation (balanced accuracy = 75%, AUC = 0.88) cohorts, demonstrating Fingerprint C is a promising method for the diagnosis of aortitis. Figure 4.8 displays the ROC curves for Fingerprint C. As this method only used Random Forest it was not tested in all ML classifiers.

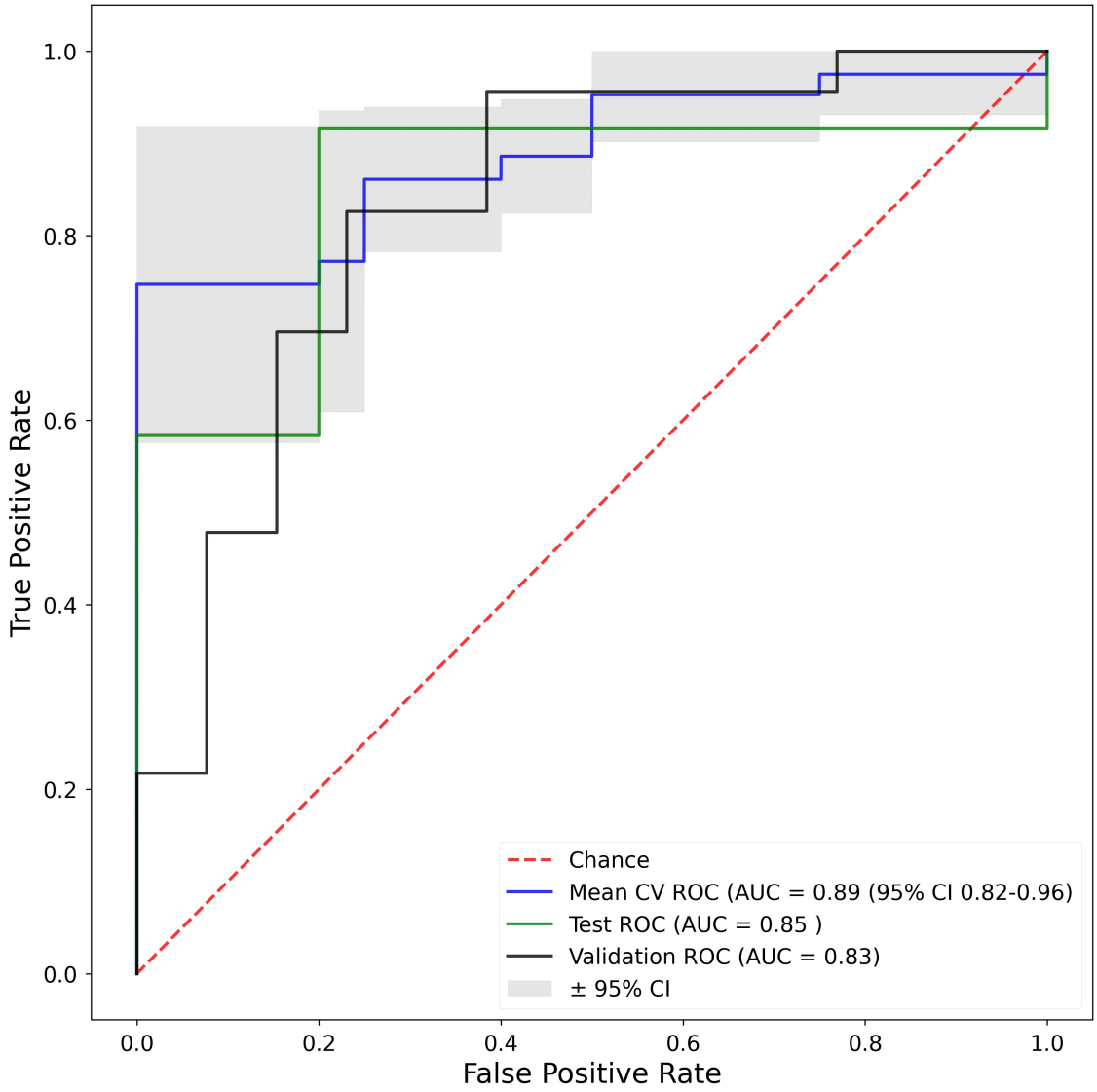


Figure 4.6: ROC curves of the best performing (by validation AUC and minimum accuracies) machine learning classifier trained on Fingerprint A (high performing individual features with highly correlated features removed) - Logistic Regression.
Key : Mean CV ROC - Mean cross validation ROC from training dataset, Test ROC - ROC from test dataset, Validation ROC - ROC from validation dataset.

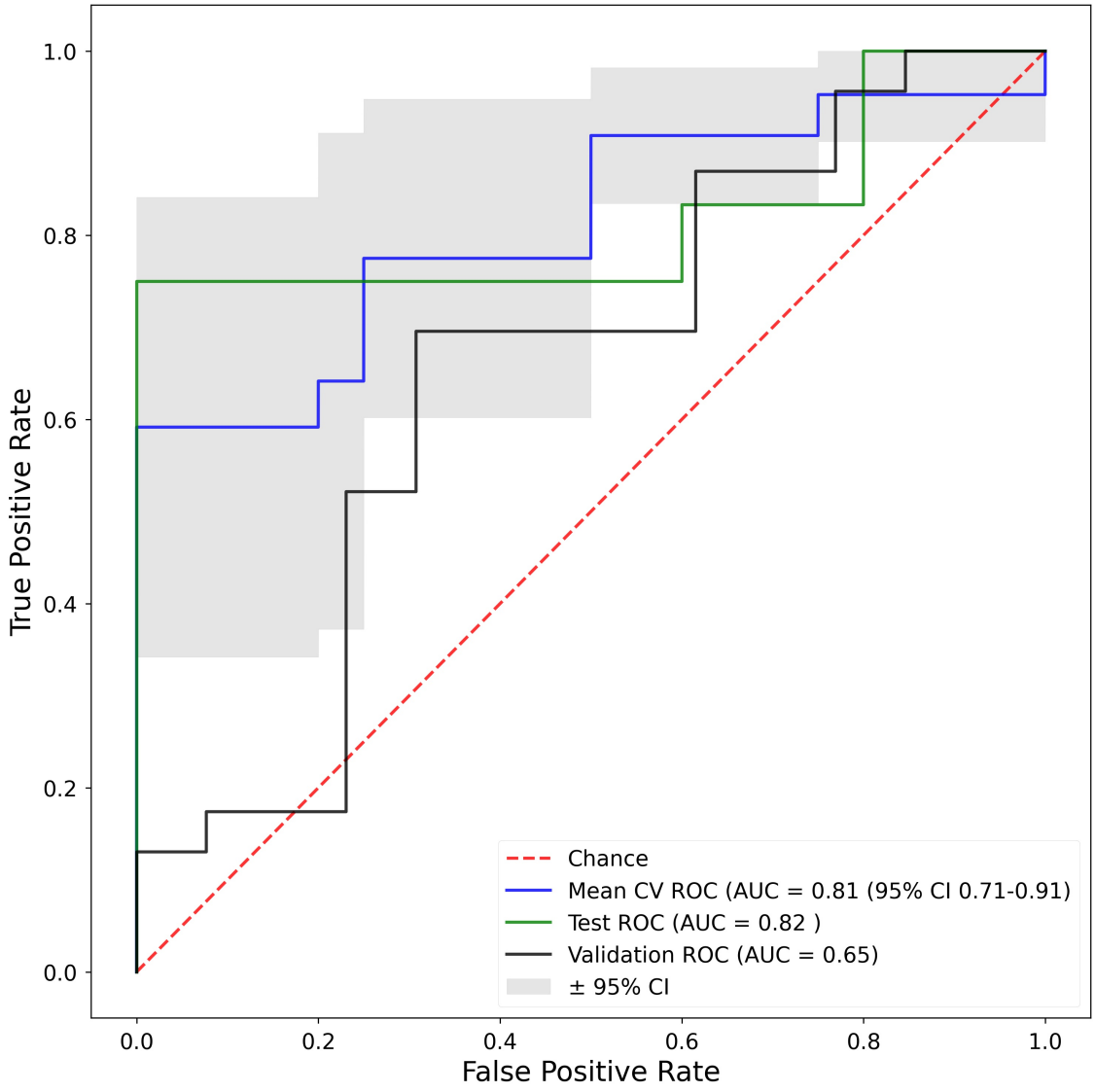


Figure 4.7: ROC curves of the best performing machine learning classifier (by validation AUC) trained on Fingerprint B (PCA)- Passive Aggressive

Key : Mean CV ROC - Mean cross validation ROC from training dataset, Test ROC - ROC from test dataset, Validation ROC - ROC from validation dataset

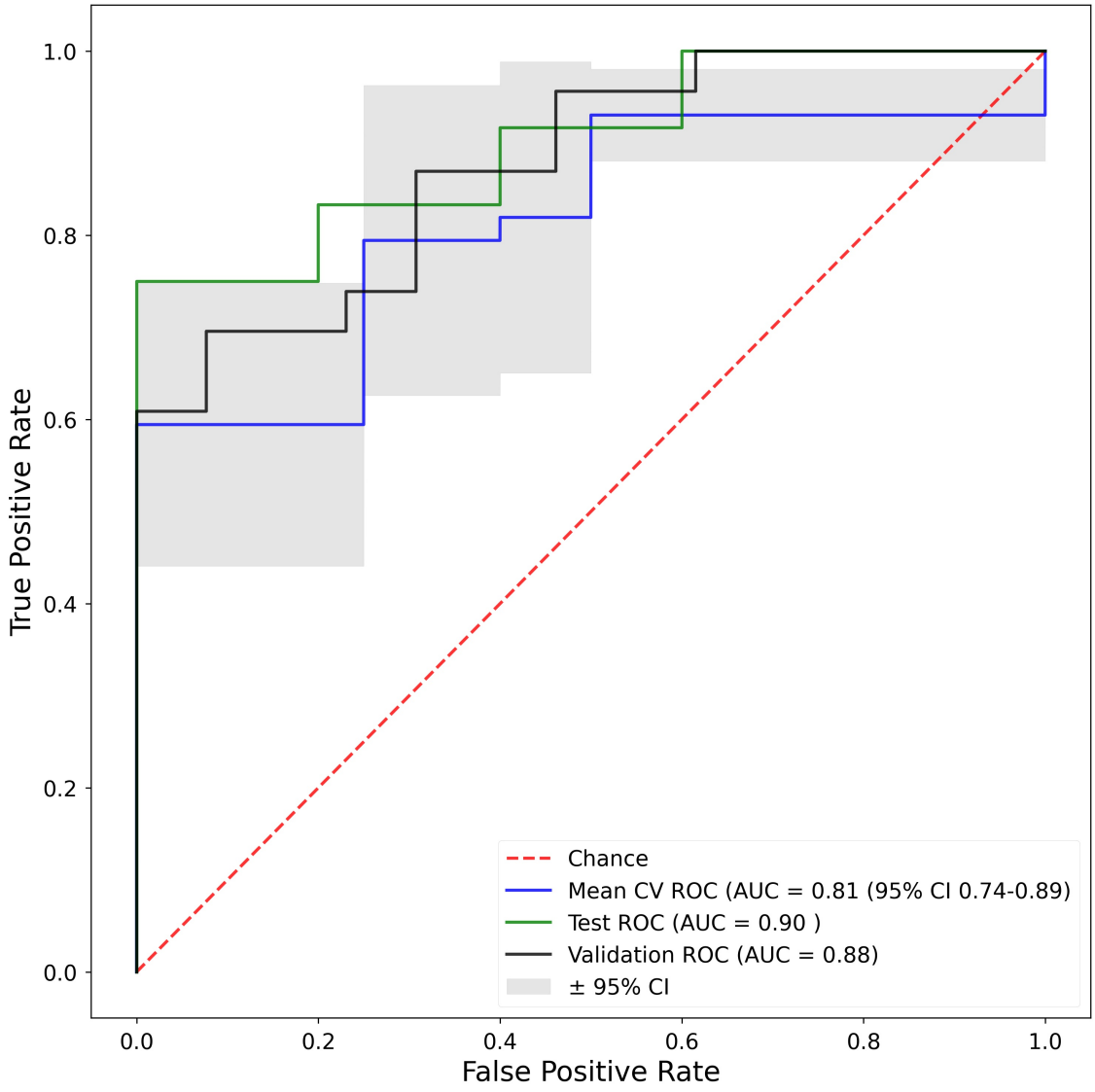


Figure 4.8: ROC curves of the Random Forest classifier in Fingerprint C

Key : Mean CV ROC - Mean cross validation ROC from training dataset, Test ROC - ROC from test dataset, Validation ROC - ROC from validation dataset

Table 4.5: Diagnostic performance of Fingerprint A - after harmonization

ML Type	Accuracy Training	Accuracy CI	AUC Training	AUC CI	Accuracy Test	AUC Test	Accuracy Validation	AUC Validation
Perceptron	0.50	(0.5-0.5)	0.82	(0.78-0.87)	0.50	0.90	0.50	0.89
Passive Aggressive	0.58	(0.44-0.72)	0.86	(0.82-0.9)	0.50	0.92	0.58	0.86
Logistic Regression	0.68	(0.51-0.84)	0.89	(0.81-0.96)	0.76	0.85	0.79	0.83
Random Forest	0.71	(0.63-0.78)	0.84	(0.8-0.88)	0.72	0.87	0.73	0.83
Decision Tree	0.74	(0.65-0.82)	0.76	(0.7-0.82)	0.92	0.92	0.72	0.77
K Nearest Neighbours	0.73	(0.62-0.84)	0.79	(0.69-0.89)	0.72	0.78	0.66	0.74
Stochastic Gradient Descent	0.69	(0.55-0.84)	0.69	(0.55-0.84)	0.50	0.50	0.58	0.58
Neural Network	0.65	(0.54-0.76)	0.77	(0.63-0.92)	0.50	0.60	0.50	0.58
Gaussian Process	0.50	(0.5-0.5)	0.50	(0.5-0.5)	0.50	0.50	0.50	0.50

Table 4.6: Diagnostic performance of Fingerprint B - after harmonization

ML Type	Accuracy Training	Accuracy CI	AUC Training	AUC CI	Accuracy Test	AUC Test	Accuracy Validation	AUC Validation
Passive Aggressive	0.81	(0.72-0.91)	0.81	(0.71-0.91)	0.50	0.82	0.54	0.65
Perceptron	0.68	(0.57-0.78)	0.74	(0.68-0.81)	0.50	0.82	0.54	0.64
K Nearest Neighbours	0.73	(0.65-0.8)	0.77	(0.72-0.82)	0.50	0.60	0.50	0.57
Random Forest	0.77	(0.7-0.83)	0.81	(0.7-0.92)	0.60	0.58	0.49	0.54
Logistic Regression	0.75	(0.71-0.79)	0.82	(0.74-0.89)	0.50	0.50	0.54	0.54
Stochastic Gradient	0.76	(0.72-0.81)	0.79	(0.72-0.86)	0.50	0.50	0.51	0.51
Support Vector Machine	0.75	(0.68-0.82)	0.84	(0.73-0.95)	0.50	0.50	0.50	0.50
Decision Tree	0.65	(0.61-0.69)	0.70	(0.65-0.74)	0.50	0.50	0.50	0.50
Gaussian Process	0.75	(0.66-0.84)	0.84	(0.76-0.92)	0.50	0.50	0.50	0.50
Neural Network	0.72	(0.64-0.79)	0.81	(0.7-0.92)	0.50	0.50	0.50	0.50

4.2.6 Comparison of Selected Features

Table 4.7 shows the features selected in Fingerprint A and the top 10 features (by feature importance) in Fingerprint C. As Fingerprint B used PCA and produces new components it is not simple to directly compare them. Table 4.7 demonstrates that heterogeneity is important in distinguishing aortitis from controls. This is confirmed by earlier results in Section 4.2.4

Table 4.7: 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*)

Features Selected in:	
Fingerprint A	**Fingerprint C (10 most important)**
SUV 90th Percentile	GLCM Joint Energy
GLCM Difference Variance	Energy
GLCM InverseVariance	Shape - Major Axis Length
GLRLM Run Entropy	GLSZM Small Area Emphasis
Energy	Skewness
GLSZM ZonePercentage	Gray Level Variance
GLDM LargeDependenceEmphasis	SUV 80
GLSZM SmallAreaHighGrayLevelEmphasis	GLRLM Run Percentage
GLSZM SizeZoneNonUniformity	GLDM Dependence Variance
GLRLM RunLengthNonUniformity	GLCM Difference Variance

4.2.7 Comparison of Results from Different Segmentation Methods

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 the results given in Table 4.8 are the mean AUC values for cross validation in training. These results show comparable performance for both segmentation methods. In the case of individual SUV metrics and Radiomic Features the confidence intervals narrowed while maintaining similar mean values. This suggests and automated method causes less variation in individual quantitative measurements. Fingerprint A and C both dropped in performance and both used the same classifier. One potential reason for this could be the receding ROI wall from the aorta edge shown in Figure 4.3. While the DSC determined the segmentations were similar it does not account for boundary variations which may be a useful addition for future aortitis work as the vessel wall is essential. Fingerprint B improved in performance but also changed the best performing classifier and was shown to be less reliable in multi-centre studies so the effect of segmentation is hard to conclude.

4.2.8 Summary of Results

Table 4.9 summarises 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 an balanced accuracy $\leq$50% were removed. While the displayed results ranked the best in each method by the given criteria, none were significantly better than each other ($p > 0.05$, DeLong's Algorithm [220, 221]).

4.3 Discussion

This study presents an automated pipeline to assist diagnosis of active aortitis using radiomic analysis and ML. The main component in automation was aortic segmentation

Table 4.8: Comparison of 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 (manual , GLCM Difference Variance (Automated)	0.91 (0.84-0.98)	0.89 (0.87-0.91)
Fingerprint A - Random Forest (both)	0.91 (0.80-1.0)	0.85 (0.81-0.89)
Fingerprint B - Random Forest (Manual), Support Vector Machine (Automated)	0.88 (0.81-0.95)	0.91 (0.84-0.98)
Fingerprint C - Random Forest (both)	0.86 (0.78-0.94)	0.81 (0.74-0.89)

Table 4.9: Summary of results

Method	Training Accuracy mean (95% CI)	Training AUC mean (95% CI)	Test Accuracy	Test AUC	Validation Accuracy	Validation AUC
Qualitative Assessment - Literature [53]	0.81-0.98 Overall AUC	-	-	-	-	-
SUV Feature - SUV 90th Percentile	0.62 (0.51-0.72)	0.85 (0.81-0.90)	0.60	0.92	0.620	0.80
Radiomic Feature - Energy	0.64 (0.55-0.74)	0.83 (0.78-0.88)	0.70	0.90	0.62	0.89
Fingerprint A - Logistic Regression	0.68 (0.51-0.84)	0.89 (0.81-0.96)	0.76	0.85	0.79	0.83
Fingerprint B - Passive Aggressive	0.81 (0.72-0.91)	0.81 (0.71-0.91)	0.50	0.82	0.54	0.65
Fingerprint C - Random Forest	0.77 (0.7-0.83)	0.81 (0.74-0.89)	0.66	0.90	0.75	0.88

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, therefore similar to the benchmark AUCs for qualitative assessment of PET-CT in suspected aortitis [53], and a (balanced) accuracy > 0.5 in all three cohorts. When compared to the performance using manual segmentations (Table 4.8) comparable results were achieved. Automating the method reduces the likelihood of inter and intra observer variability, increasing reproducibility [222]. It also makes routine clinical adoption a more realistic proposition.

Some 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 this could be more reliable [87]. Several individual radiomic features produced high AUC values and met the minimum balanced accuracy values. In particular features based on heterogeneity performed well across all three cohorts with the highest validation AUC coming from Energy (AUC = 0.89).The features selected in Fingerprint A and C further demonstrate that heterogeneity is important in distinguishing aortitis (Table 4.7).

Most ML classifiers when given the proposed radiomic Fingerprints A and B as input did not show generalizability in the test and validation cohorts. In particular, Fingerprint B, where PCA was used to reduce the number of features, achieved an balanced accuracy $\leq 50\%$ in at least one cohort in every ML classifier. Fingerprint A performed better with two ML classifiers (Logistic Regression and Random Forest) achieving AUC and balanced accuracy values above the minimum thresholds stated earlier. However, Fingerprint C (Random Forest) performed the best achieving higher AUC values in both the test and validation cohorts (AUC = 0.9 and 0.88 respectively) and higher balanced accuracy values than either other fingerprint.

Chapter 5

Synopsis

5.1 Overview of Results

This project presents an automated pipeline to assist diagnosis of active aortitis using radiomic analysis of PET-CT and classification with ML.

The purpose of the first set of experiments was to develop a methodological framework and explore the feasibility of the radiomic pipeline for this application. Overall, promising results were attained. The best performing individual radiomic feature GLSZM Size Zone Non-Uniformity Normalized had an AUC of 0.9 (0.83-0.97 95% CI), similar to the current clinical benchmark of qualitative assessment by an experienced radiologist or nuclear medicine physician (AUC=0.81-0.98) [53]. The three fingerprints, groups of selected radiomic features, performed similarly to the best-performing individual RFs. Of particular promise was Fingerprint B, features with high individual diagnostic performance and removal of highly correlated features, with an AUC of 0.91 (0.80-1.00 95% CI).

The diagnostic utility of semi-quantitative measurements using SUV were compared against other radiomics features for detecting active aortitis. The best performing SUV based feature was SUV 90^{th} percentile, which gives the value that 90% of SUV values in the ROI are equal or less than. It has similarities to SUV_{peak} in the sense it is related to SUV_{max} but altered in order to reduce the effect of noise. Unlike SUV_{peak} it does not

incorporate spatial location in it's definition. SUV 90^{th} percentile scored an AUC of 0.83, only slightly below the best performing radiomic features when confidence intervals are included. All SUV features explored were found to be significantly different between patients with active aortitis and controls using the Mann Whitney U Test. Overall SUV metrics had some diagnostic utility in Mann Whitney U and logistic regression classifier testing but did not perform as well as RFs.

Radiomic features shown to have the highest diagnostic utility focus mainly on high gray levels and heterogeneity. The GLSZM Size Zone Non-Uniformity Normalized was the best RF according to AUC and performed well in terms of accuracy and the Mann Whitney U test. Its value is higher in active aortitis than controls, which means there may be more heterogeneity in zone size volumes in aortic LVV imaging. This is an expected finding as it reflects greater metabolic activity in the aortic wall of patients with active aortitis than in controls. However, since it has superior performance compared to SUV parameters which are based on metabolic activity, this is a potentially useful new association. The importance of high gray values and zones, and heterogeneity is further emphasised in other radiomic features with high diagnostic utility. The addition of heterogeneity, encompassing spacial relationship between voxels, may help explain why radiomic features outperformed SUV metrics which focus on voxel values alone. They also demonstrated better transferability to the validation cohort suggesting heterogeneity also provides more robustness.

With the feasibility of radiomic analysis in LVV PET determined and a methodology established, the next set of experiments aimed to validate the results with multi-centre data. The secondary aim was to automate the process to facilitate use of the pipeline making routine clinical adoption a more realistic proposition.

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, therefore similar to the benchmark AUCs for qualitative assessment of PET-CT in suspected aortitis [53], and a (balanced) accuracy > 0.5 in all three cohorts. When compared to the performance using manual segmentations (Table 4.8) comparable results were achieved. Automating the method reduces the likelihood of inter and intra observer

variation, increasing reproducibility [222]. Other automated segmentation methods were considered but a CNN was selected due to more successful and reproducible results both in the published literature and our own experiments [223].

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 [224]. Some diagnostic performance was shown when validating the proposed methods in multi-centre data. SUV metrics performed well in training cohorts but did not demonstrate good transferability to the testing or validation cohort from other institutions. SUV 90^{th} percentile demonstrated the most diagnostic utility from all the explored SUV metrics and did so in all three cohorts (validation AUC = 0.8). 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 this could be more reliable [87]. 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 'Energy' (AUC = 0.89).The features selected in Fingerprint A and C further demonstrate that heterogeneity is important in distinguishing aortitis (Table 4.7).

Most ML classifiers when given the proposed radiomic Fingerprints A and B as input did not show generalizability in the test and validation cohorts. In particular, Fingerprint B, where PCA was used to reduce the number of features, achieved an accuracy $\leq 50\%$ in at least one cohort in every ML classifier. Fingerprint A performed better with two ML classifiers (Logistic Regression and Random Forest) achieving AUC and accuracy values above the minimum thresholds stated earlier. However, Fingerprint C (Random Forest) performed the best achieving higher AUC values in both the test and validation cohorts (AUC=0.9 and 0.88 respectively) and high accuracy values (66% and 0.75% respectively). This demonstrates good generalizability and transferability which are important prerequisites for clinical use [224]. Fingerprint C is different to the others as it uses embedded feature selection in the ML classifier(i.e. Random Forest) rather than preselecting features. A similar result was reported by Da-ano *et al.* [225].

The key points from the two sets of experiments are that heterogeneity-based features show the greatest potential in distinguishing aortitis from controls, that the pipeline can

be automated, and that a few diagnostic models performed well in external data demonstrating potential for transferability and generalizability.

A key point discussed in numerous radiomic studies and reviews is the need for standardized methodology. This allows for reproducibility which is a common limitation of radiomic studies. TRIPOD reporting guidelines were adhered to in this project to ensure transparency of methodological details [194]. 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.* [139]. Deviations from these definitions are discussed in the user documentation [1] and accompanying publication [175]. IBSI found that even with well-defined rules there was still a lot of internal variation due lack of communication, different interpretations of the same rules and variation in workstations so care still needs to be given to standardization after a methodological pipeline is established and reported [139]. There is no clear consensus on the optimal settings for some steps in the radiomic workflow. This is intentional as they can be situation dependent [139]. Providing sufficient detail is given in any publications or standard operating procedure, reproducibility is still achievable within the same clinical application.

In the case of this thesis the code will be made open access in the near future to maximise transparency and promote reproducibility.

The cohort size is reasonably large, especially when sub-optimal sample size is a common limitation in radiomic studies. This is especially the case in the second set of experiments. It is difficult to directly compare our sample size to other radiomic studies as aortitis is less common than cancers, aortitis is PET imaged less than cancers, and PET is generally used less than CT or MRI. The importance of a good sample size is also becoming more well known so reviews and meta-analyses of sample sizes in radiomic studies quickly become outdated. In our final experiments we reached a sample size of 114 patients in total which splits into 74 aortitis cases and 40 controls. It is not uncommon for oncological radiomic studies to have greater than one hundred or very occasionally thousands of cases [176, 226, 227]. However, it has been recognised in a recent position

[1] https://pyradiomics.readthedocs.io/en/latest/features

paper by European Association of Nuclear Medicine and the European Association of Cardiovascular Imaging that applications of AI in cardiovascular imaging with PET are much less common. Cardiovascular radiomic studies with CT or MRI with a sample size in the early hundreds have become more frequent [228] but similar sample sizes using PET have only been achieved recently, are still rare, and studied more common conditions such as coronary artery disease [229, 230]. Sample size is an important consideration because over-fitting and type 1 errors are prone when smaller cohorts are used [191, 231]. Bonferroni correction and feature reduction were used to reduce the risk of these errors but over-fitting is still possible. The cohort size in the case of this project is sufficient to have reasonable confidence in the results but more data would solidify the conclusions drawn, and open up other applications of radiomics in LVV such as treatment response prediction, outcome prediction and classification by cause of LVV. As LVV and aortitis are relatively rare when compared to many oncological conditions, establishing a multi-centre repository of imaging data in order to promulgate clinical translation is a key future aim.

There remains debate as to the validity of harmonization with ComBat [118, 232, 233]. Orlhac *et al.* stated that ComBat is only appropriate in situations described in their guide [234]. As this study uses data from several institutions and scanners, harmonization was deemed necessary, although the effect on the data was not confirmed. Papadimitroulas *et al.* described several other alternatives to ComBat but also concluded that ComBat performed well overall [147]. ComBat requires a minimum number of cases per batch to produce a transform to be applied to all cases in that batch. Individual cases can be harmonized if they are part of a pre-defined batch but new centres would need to form a new transform and essentially calibrate a radiomic method that uses ComBat if they wanted to use the radiomic method. This could be a hurdle to smaller centres. Standardization of imaging protocols was not feasible as this was a retrospective study with insufficient data to exclude patients based on a unified imaging protocol. All steps after reconstruction were kept consistent as this has been proven to have a significant effect [235]. This included voxel size resampling to uniform voxel size which may partially explain the lack of impact of harmonization [141]. In future prospective work a standardized protocol would be possible [236, 237], but may not be implemented if the radiomic

pipeline was ever utilised in a clinical setting. Other image based approaches, as opposed to feature based approaches, are likely more feasible in a clinical setting as it can be applied in centres with a small number of cases but more exploration would be required to determine the best method [140, 141].

5.2 Outlook

Feature selection was conducted in this project but one method that was overlooked was removing non-reproducible features. Retrospectively it was determined that the best performing and most relevant features appeared to do well across all datasets but further analysis is warranted. However, this was achieved somewhat as any method that did not perform well across all three cohort was discounted, just not prior to feature selection. There is a risk of over-fitting if feature selection is done on test and validation data as well. Repeatability of the results from variaitons in methods is worth exploring further as well but not essential as good results were achieved with the current method and any observed change in results due to method alterations would be expected and not a sign that features should be removed. Removing features excessively can lead to a large loss in information and providing future users follow the same method reproducibility due to unforeseen alterations in the method are not a limitation of this project.

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 optimisation 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 [139, 194]. 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 [139, 151].

The automated segmentation method used in the second set of experiments in this project could be further refined. Other automated segmentation methods were attempted but did not work well on our dataset. It would be preferable to only analyse the aortic wall, but a segmentation method which reliably distinguishes wall from the lumen at non-contrast enhanced CT has not been developed to the extent that it reliably worked on the patient cohort in this study. Other segmentation methods or adaptations of the current method may be explored going forward. For example, threshold-based segmentation of PET data was initially discounted as a method as it was conventional in LVV to use the liver SUV as a cut off value and this method would not be suitable for normal controls. Another common threshold used in clinical PET evaluation is mediastinal blood pool activity which would work in all patients. It was initially too difficult to automatically segment but a potential method for measuring blood pool activity in this project's data has been solidified recently. Advantages to reintroducing thresholding would be more specific feature extraction as the lumen would no longer be in the ROI. An alternative route would be further study of confirmed aortitis cases without normal controls, for example, to evaluate the utility of radiomics for prognostication/prediction of vascular calcification but due to the low event rate this would likely require a large data cohort. 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.

There are other radiotracers explored for use in LVV PET imaging or other inflammatory conditions. While not widely used they may provide additional insights into LVV or facilitate more straight-forward segmentation of imaging data [93, 156]. Other tracers include copper or gallium labelled DOTATE - a peptide that targets somatostatin receptors which can be found on macrophages [238, 239], and immuno-PET approaches such zirconium labelled antibodies [240], and [11C]-PK11195 which selectively bonds to translocator proteins in activated macrophages [241, 242]. Applying the method developed in this thesis to PET images acquired with other radiotracers would likely result in reduced diagnostic utility. Due to different radiotracer distribution patterns the features

would probably not be transferable. However, the process used to develop the models could be used in the same way meaning any progression in PET imaging for aortitis and LVV would not make radiomic analysis obsolete. Steps such as retraining the models, feature selection and possibly segmentation would need to be repeated but the radiomic pipeline presented could be adapted to accommodate changes.

The next stage in this project would be to test this pipeline on the whole spectrum of those presenting with suspected aortitis; atherosclerosis is common in this age group [54]. Another group has reported promising results using SUV metrics so there is potential for further success with radiomics [243]. Patients with atherosclerosis were excluded from the study cohort, for both aortitis patients and controls, as atherosclerosis can also cause increased FDG uptake in the vessel wall [244–246]. 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. Going forward it would be clinically useful to determine if radiomics can differentiate LVV from atherosclerosis.

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 glucocorticoids reduce its sensitivity [105–107]. 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 [85]. 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.

In future prospective imaging studies of LVV and aortitis multi-modality approaches could be expanded upon. While FDG PET-CT utilises low dose CT for anatomical information, CTA or MRA provide much clearer images and are currently used to detect anatomical changes such as aneurysms and stenosis. Using these imaging techniques within the same acquisition would allow better direct comparison between the charac-

teristics of both images and would be synergistic in quantitative analysis. The clearer images would improve segmentation of the other major arteries expanding the analysis from aortitis to LVV as a whole. Functional MRI techniques could also add to metabolic characterisation of LVV.

All of the issues discussed remain important but there have been concerted efforts to tackle these by the international radiomics research community happening in parallel during the conduct of this project [80]. Therefore, when undertaking radiomics research it is essential to remain abreast of the literature and routinely review relevant emerging research studies as this field of study remains dynamic and continues to evolve with stepwise improvements in quality and reproducibility of more recently published work . As more of the discussed limitations are resolved or minimised, in particular standardization and procuring large datasets, more in-depth radiomic analysis can be conducted and combined with fields such as digital pathology and genomics [124]. This is already being explored but is still in the initial stages [247]. These developments are exciting as it should facilitate a deeper understanding of the pathophysiological basis of several medical conditions and in the case of PET-CT the utility of advanced imaging analysis for non-invasive monitoring of disease. Incorporation of other clinical data will require more consistent reporting in patient records and the practicalities of implementing this are not trivial [126].

If larger datasets are acquired for LVV including sufficient controls further applications of radiomics in LVV and aortitis could be explored. For example, cohort size prevented significant results being obtained when classification by cause of aortitis was attempted. However, this would be a useful progression in order to guide treatment [1]. Establishing a repository containing imaging and clinical meta-data would help facilitate studies such as these and would open up opportunities to make tools to aid prediction of outcomes, treatment response, relapse or refractory risk and the likelihood of adverse events as all of these have too low an occurrence rate for the current dataset to produce significant results. Larger datasets would also although for further exploration of deep learning applications.

5.3 Conclusion

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. Some of the proposed methods demonstrated good diagnostic performance across the training, testing and validation datasets showing that a radiomic pipeline can be generalizable and transferable. Similarly, it was shown that radiomic features outperform conventional SUV metrics but that radiomic fingerprints only perform slightly better if not the same when used in machine learning classifiers. This is important knowledge gained as diagnosis if aortitis can be difficult and is vulnerable to intra- and inter-observer variability. These findings could be used to build an automated clinical decision tool which would facilitate objective and standardized assessment regardless of observer experience.

References

1. Pugh D, Grayson P, Basu N, and Dhaun N. Aortitis: recent advances, current concepts and future possibilities. Heart 2021;107:1620–9.
2. Parikh M, Miller NR, Lee AG, Savino PJ, Vacarezza MN, 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;113:1842–5.
3. Monach PA. Biomarkers in vasculitis. Current opinion in rheumatology 2014;26:24.
4. Gornik HL and Creager MA. Aortitis. Circulation 2008;117:3039–51.
5. Stone JR, Bruneval P, Angelini A, Bartoloni G, Basso C, Batoroeva L, Buja LM, Butany J, d'Amati G, Fallon JT, 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. Cardiovascular Pathology 2015;24:267–78.
6. Marvisi C, Buttini EA, and Vaglio A. Aortitis and periaortitis: the puzzling spectrum of inflammatory aortic diseases. La Presse Médicale 2020;49:104018.
7. Monti S, Águeda AF, Luqmani RA, Buttgereit 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.

8. Wallace ZS, Zhang Y, Perugino CA, Naden R, Choi HK, and Stone JH. Clinical phenotypes of IgG4-related disease: an analysis of two international cross-sectional cohorts. Annals of the rheumatic diseases 2019;78:406–12.

9. Peng L, Zhang P, Li J, Liu Z, Lu H, Zhu L, Wang X, Teng F, Li X, Guo H, et al. IgG4-related aortitis/periaortitis and periarteritis: a distinct spectrum of IgG4-related disease. Arthritis Research & Therapy 2020;22:1–11.

10. Horton B. An undescribed form of arteritis of temporal vessels. Proc Staff Mayo Clin Proc 1932;7:700–1.

11. Muratore F, Kermani TA, Crowson CS, Green AB, Salvarani C, Matteson EL, and Warrington KJ. Large-vessel giant cell arteritis: a cohort study. Rheumatology 2015;54:463–70.

12. Koster MJ, Matteson EL, and Warrington KJ. Large-vessel giant cell arteritis: diagnosis, monitoring and management. Rheumatology 2018;57:ii32–ii42.

13. Kebed DT, Bois JP, Connolly HM, Scott CG, Bowen JM, Warrington KJ, Makol A, Greason KL, Schaff HV, and Anavekar NS. Spectrum of aortic disease in the giant cell arteritis population. The American Journal of Cardiology 2018;121:501–8.

14. Dejaco C, Ramiro S, Duftner C, Besson FL, Bley TA, Blockmans D, Brouwer E, Cimmino MA, Clark E, Dasgupta B, et al. EULAR recommendations for the use of imaging in large vessel vasculitis in clinical practice. Annals of the rheumatic diseases 2018;77:636–43.

15. Salvarani C, Cantini F, and Hunder GG. Polymyalgia rheumatica and giant-cell arteritis. The Lancet 2008;372:234–45.

16. Maksimowicz-McKinnon K, Clark TM, and Hoffman GS. Takayasu arteritis and giant cell arteritis: a spectrum within the same disease? Medicine 2009;88:221–6.

17. González-Gay MA, Garcıa-Porrúa C, Llorca J, Hajeer AH, Brañas F, Dababneh A, González-Louzao C, Rodriguez-Gil E, Rodrıguez-Ledo P, and Ollier W. Visual manifestations of giant cell arteritis. Trends and clinical spectrum in 161 patients. Medicine 2000;79:283–92.

18. Gonzalez-Gay MA, Barros S, Lopez-Diaz MJ, Garcia-Porrua C, Sanchez-Andrade A, and Llorca J. Giant cell arteritis: disease patterns of clinical presentation in a series of 240 patients. Medicine 2005;84:269–76.

19. Robson JC, Kiran A, Maskell J, Hutchings A, Arden N, Dasgupta B, Hamilton W, Emin A, Culliford D, and Luqmani RA. The relative risk of aortic aneurysm in patients with giant cell arteritis compared with the general population of the UK. Annals of the rheumatic diseases 2015;74:129–35.

20. Richards BL, March L, and Gabriel SE. Epidemiology of large-vessel vasculidities. Best Practice & Research Clinical Rheumatology 2010;24:871–83.

21. Zaldivar Villon ML, Rocha JAL de la, and Espinoza LR. Takayasu arteritis: recent developments. Current Rheumatology Reports 2019;21:1–10.

22. Goel R, Chandan JS, Thayakaran R, Adderley NJ, Nirantharakumar K, and Harper L. Cardiovascular and renal morbidity in Takayasu arteritis: a Population-Based retrospective cohort study from the United Kingdom. Arthritis & Rheumatology 2021;73:504–11.

23. Petri H, Nevitt A, Sarsour K, Napalkov P, and Collinson N. Incidence of giant cell arteritis and characteristics of patients: data-driven analysis of comorbidities. Arthritis care & research 2015;67:390–5.

24. Sharma A, Mohammad AJ, and Turesson C. Incidence and prevalence of giant cell arteritis and polymyalgia rheumatica: a systematic literature review. In: *Seminars in Arthritis and Rheumatism*. Vol. 50. 5. Elsevier. 2020:1040–8.

25. Watts R, Al-Taiar A, Mooney J, Scott D, and MacGregor A. The epidemiology of Takayasu arteritis in the UK. Rheumatology 2009;48:1008–11.

26. Clifford AH, Arafat A, Idrees JJ, Roselli EE, Tan CD, Rodriguez ER, Svensson LG, Blackstone E, Johnston D, Pettersson G, et al. Outcomes among 196 patients with noninfectious proximal aortitis. Arthritis & Rheumatology 2019;71:2112–20.

27. Ferfar Y, Morinet S, Espitia O, Agard C, Vautier M, Comarmond C, Desbois AC, Domont F, Fouret PJ, Redheuil A, et al. Long-Term outcome and prognosis factors of isolated aortitis. Circulation 2020;142:92–4.

28. Tomasson G, Peloquin C, Mohammad A, Love TJ, Zhang Y, Choi HK, and Merkel PA. Risk for cardiovascular disease early and late after a diagnosis of giant-cell arteritis: a cohort study. Annals of internal medicine 2014;160:73–80.

29. Koster MJ, Warrington KJ, and Matteson EL. Morbidity and mortality of large-vessel vasculitides. Current rheumatology reports 2020;22:1–13.

30. Buttgereit F, Matteson EL, Dejaco C, and Dasgupta B. Prevention of glucocorticoid morbidity in giant cell arteritis. Rheumatology 2018;57:ii11–ii21.

31. Foote EA, Postier RG, Greenfield RA, and Bronze MS. Infectious aortitis. Current Treatment Options in Cardiovascular Medicine 2005;7:89–97.

32. Barra L, Pope JE, Pequeno P, Gatley JM, and Widdifield J. Increased mortality for individuals with Giant Cell Arteritis: a population-based study. Arthritis Care & Research 2021.

33. Therkildsen P, Nielsen BD, Thurah A de, Hansen IT, Nørgaard M, and Hauge EM. All-cause and cause-specific mortality in patients with giant cell arteritis: a nationwide, population-based cohort study. Rheumatology 2022;61:1195–203.

34. Hill CL, Black RJ, Nossent JC, Ruediger C, Nguyen L, Ninan JV, and Lester S. Risk of mortality in patients with giant cell arteritis: a systematic review and meta-analysis. In: *Seminars in arthritis and rheumatism*. Vol. 46. 4. Elsevier. 2017:513–9.

35. Uddhammar A, Eriksson AL, Nyström L, Stenling R, and Rantapää-Dahlqvist S. Increased mortality due to cardiovascular disease in patients with giant cell arteritis in northern Sweden. The Journal of Rheumatology 2002;29:737–42.

36. Muratore F, Pipitone N, and Salvarani C. Standard and biological treatment in large vessel vasculitis: guidelines and current approaches. Expert review of clinical immunology 2017;13:345–60.

37. Hellmich B, Agueda A, Monti S, Buttgereit F, De Boysson H, Brouwer E, Cassie R, Cid MC, Dasgupta B, Dejaco C, et al. 2018 Update of the EULAR recommendations for the management of large vessel vasculitis. Annals of the rheumatic diseases 2020;79:19–30.

38. Wilson JC, Sarsour K, Collinson N, Tuckwell K, Musselman D, Klearman M, Napalkov P, Jick SS, Stone JH, and Meier CR. Serious adverse effects associated with glucocorticoid therapy in patients with giant cell arteritis (GCA): a nested case–control analysis. In: *Seminars in arthritis and rheumatism*. Vol. 46. 6. Elsevier. 2017:819–27.

39. Wu J, Keeley A, Mallen C, Morgan AW, and Pujades-Rodriguez M. Incidence of infections associated with oral glucocorticoid dose in people diagnosed with polymyalgia rheumatica or giant cell arteritis: a cohort study in England. Cmaj 2019;191:E680–E688.

40. Wu J, Mackie SL, and Pujades-Rodriguez M. Glucocorticoid dose-dependent risk of type 2 diabetes in six immune-mediated inflammatory diseases: a population-based cohort analysis. BMJ Open Diabetes Research and Care 2020;8:e001220.

41. Pujades-Rodriguez M, Morgan AW, Cubbon RM, and Wu J. Dose-dependent oral glucocorticoid cardiovascular risks in people with immune-mediated inflammatory diseases: A population-based cohort study. PLoS medicine 2020;17:e1003432.

42. Stone JH, Tuckwell K, Dimonaco S, Klearman M, Aringer M, Blockmans D, Brouwer E, Cid MC, Dasgupta B, Rech J, et al. Trial of tocilizumab in giant-cell arteritis. New England Journal of Medicine 2017;377:317–28.

43. Chang HW, Kim SH, Hakim AR, Chung S, Kim DJ, Lee JH, Kim JS, Lim C, and Park KH. Diameter and growth rate of the thoracic aorta—analysis based on serial computed tomography scans. Journal of Thoracic Disease 2020;12:4002.

44. Mensel B, Quadrat A, Schneider T, Kühn JP, Dörr M, Völzke H, Lieb W, Hegenscheid K, and Lorbeer R. MRI-based determination of reference values of thoracic aortic wall thickness in a general population. European radiology 2014;24:2038–44.

45. Choe YH, Han BK, Koh EM, Kim DK, Do YS, and Lee WR. Takayasu's arteritis: assessment of disease activity with contrast-enhanced MR imaging. American Journal of Roentgenology 2000;175:505–11.

46. Tso E, Flamm SD, White RD, Schvartzman PR, Mascha E, and Hoffman GS. Takayasu arteritis: utility and limitations of magnetic resonance imaging in diagnosis and treatment. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology 2002;46:1634–42.

47. Besutti G, Muratore F, Mancuso P, Ferrari M, Galli E, Spaggiari L, Monelli F, Casali M, Versari A, Boiardi L, et al. Vessel inflammation and morphological changes in patients with large vessel vasculitis: a retrospective study. RMD open 2022;8:e001977.

48. Muto G, Yamashita H, Takahashi Y, Miyata Y, Morooka M, Minamimoto R, Kubota K, Kaneko H, Kano T, and Mimori A. Large vessel vasculitis in elderly patients: early diagnosis and steroid-response evaluation with FDG-PET/CT and contrast-enhanced CT. Rheumatology international 2014;34:1545–54.

49. Quinn KA, Ahlman MA, Alessi HD, LaValley MP, Neogi T, Marko J, Novakovich E, and Grayson PC. Association of 18F-Fluorodeoxyglucose Positron Emission Tomography and Angiographic Progression of Disease in Large-Vessel Vasculitis. Arthritis & Rheumatology 2022.

50. Nienhuis PH, Praagh GD van, Glaudemans AW, Brouwer E, and Slart RH. A review on the value of imaging in differentiating between large vessel vasculitis and atherosclerosis. Journal of personalized medicine 2021;11:236.

51. Planas-Rigol E, Corbera-Bellalta M, Espıgol-Frigolé G, Terrades-Garcia N, Alba M, et al. Giant-Cell Arteritis: Immunopathogenic Mechanisms Involved in Vascular Inflammation and Remodeling. J Vasc 1: 103. doi: 10.4172/2471-9544.100103 Page 2 of 7 J Vasc ISSN: 2471-9544 JOV, an open access journal Volume 1• Issue 2• 100103. dose glucocorticoids, neutrophils have lower membrane expression of integrin CD11b, are less adhesive to endothelial cells and are able to suppress T cell proliferation. These abnormalities revert when 2016:3.

52. Gajree S, Borooah S, Dhillon N, Goudie C, Smith C, Aspinall P, and Dhillon B. Temporal artery biopsies in south-east Scotland: a five year review. The Journal of the Royal College of Physicians of Edinburgh 2017;47:124–8.

53. Slart RH 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. European journal of nuclear medicine and molecular imaging 2018;45:1250–69.

54. Slart RH, Glaudemans AW, Gheysens O, Lubberink M, Kero T, Dweck MR, 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. European Journal of Nuclear Medicine and Molecular Imaging 2020:1–24.

55. Ehman EC, Johnson GB, Villanueva-Meyer JE, Cha S, Leynes AP, Larson PEZ, and Hope TA. PET/MRI: where might it replace PET/CT? Journal of Magnetic Resonance Imaging 2017;46:1247–62.

56. Basu S, Kwee TC, Surti S, Akin EA, Yoo D, and Alavi A. Fundamentals of PET and PET/CT imaging. Annals of the New York Academy of Sciences 2011;1228:1–18.

57. Levin CS and Hoffman EJ. Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution. Physics in Medicine & Biology 1999;44:781.

58. Conti M and Eriksson L. Physics of pure and non-pure positron emitters for PET: a review and a discussion. EJNMMI physics 2016;3:1–17.

59. Schmitz RE, Alessio AM, Kinahan PE, Mason NS, and Lin EC. The physics of PET/Ct scanners. PET and PET/CT: a clinical guide 2005;3.

60. Berger A. How does it work?: Positron emission tomography. BMJ: British Medical Journal 2003;326:1449.

61. Humm JL, Rosenfeld A, and Del Guerra A. From PET detectors to PET scanners. European journal of nuclear medicine and molecular imaging 2003;30:1574–97.

62. Moses WW. Fundamental limits of spatial resolution in PET. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 2011;648:S236–S240.

63. Rapisarda E, Bettinardi V, Thielemans K, and Gilardi M. Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET. Physics in medicine & biology 2010;55:4131.

64. Slomka PJ, Pan T, Berman DS, and Germano G. Advances in SPECT and PET hardware. Progress in Cardiovascular Diseases 2015;57:566–78.

65. Surti S and Karp JS. Update on latest advances in time-of-flight PET. Physica Medica 2020;80:251–8.

66. Slomka PJ, Pan T, and Germano G. Recent advances and future progress in PET instrumentation. In: *Seminars in nuclear medicine*. Vol. 46. 1. Elsevier. 2016:5–19.

67. Bailey DL and Meikle SR. A convolution-subtraction scatter correction method for 3D PET. Physics in Medicine & Biology 1994;39:411.

68. Levin CS, Dahlbom M, and Hoffman EJ. A Monte Carlo correction for the effect of Compton scattering in 3-D PET brain imaging. IEEE transactions on nuclear science 1995;42:1181–5.

69. Watson CC, Newport D, and Casey ME. A single scatter simulation technique for scatter correction in 3D PET. Three-dimensional image reconstruction in radiology and nuclear medicine 1996:255–68.

70. Ollinger JM. Model-based scatter correction for fully 3D PET. Physics in Medicine & Biology 1996;41:153.

71. Polycarpou I, Thielemans K, Manjeshwar R, Aguiar P, Marsden PK, and Tsoumpas C. Comparative evaluation of scatter correction in 3D PET using different scatter-level approximations. Annals of nuclear medicine 2011;25:643–9.

72. Brasse D, Kinahan PE, Lartizien C, Comtat C, Casey M, and Michel C. Correction methods for random coincidences in fully 3D whole-body PET: impact on data and image quality. Journal of nuclear medicine 2005;46:859–67.

73. Słomski A, Rudy Z, Bednarski T, Białas P, Czerwiński E, Kapłon Ł, Kochanowski A, Korcyl G, Kowal J, Kowalski P, et al. 3D PET image reconstruction based on the maximum likelihood estimation method (MLEM) algorithm. Bio-Algorithms and Med-Systems 2014;10:1–7.

74. Shepp LA and Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE transactions on medical imaging 1982;1:113–22.

75. Miller TR and Wallis JW. Clinically important characteristics of maximum-likelihood reconstruction. Journal of Nuclear Medicine 1992;33:1678–84.

76. Boellaard R, Van Lingen A, and Lammertsma AA. Experimental and clinical evaluation of iterative reconstruction (OSEM) in dynamic PET: quantitative characteristics and effects on kinetic modeling. Journal of Nuclear Medicine 2001;42:808–17.

77. Hudson HM and Larkin RS. Accelerated image reconstruction using ordered subsets of projection data. IEEE transactions on medical imaging 1994;13:601–9.

78. Bettinardi V, Presotto L, Rapisarda E, Picchio M, Gianolli L, and Gilardi M. Physical Performance of the new hybrid PET/CT Discovery-690. Medical physics 2011;38:5394–411.

79. Aide N, Lasnon C, Kesner A, Levin CS, Buvat I, Iagaru A, Hermann K, Badawi RD, Cherry SR, Bradley KM, et al. New PET technologies–embracing progress and pushing the limits. European journal of nuclear medicine and molecular imaging 2021;48:2711–26.

80. Hatt M, Le Rest CC, Tixier F, Badic B, Schick U, and Visvikis D. Radiomics: data are also images. Journal of Nuclear Medicine 2019;60:38S–44S.

81. Lee JS. A review of deep-learning-based approaches for attenuation correction in positron emission tomography. IEEE Transactions on Radiation and Plasma Medical Sciences 2020;5:160–84.

82. Reader AJ, Corda G, Mehranian A, Costa-Luis C da, Ellis S, and Schnabel JA. Deep learning for PET image reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences 2020;5:1–25.

83. Zhao YM, Li YH, Chen T, Zhang WG, Wang LH, Feng J, Li C, Zhang X, Fan W, and Hu YY. Image quality and lesion detectability in low-dose pediatric 18F-FDG scans using total-body PET/CT. European journal of nuclear medicine and molecular imaging 2021;48:3378–85.

84. Nadig V, Herrmann K, Mottaghy FM, and Schulz V. Hybrid total-body pet scanners—current status and future perspectives. European journal of nuclear medicine and molecular imaging 2022;49:445–59.

85. Van der Geest K, Treglia G, Glaudemans A, Brouwer E, Sandovici M, Jamar F, Gheysens O, and Slart R. Diagnostic value of [18F] FDG-PET/CT for treatment monitoring in large vessel vasculitis: a systematic review and meta-analysis. European journal of nuclear medicine and molecular imaging 2021;48:3886–902.

86. Lodge MA. Repeatability of SUV in oncologic 18F-FDG PET. Journal of Nuclear Medicine 2017;58:523–32.

87. Praagh GD van, Nienhuis PH, Jong DM de, Reijrink M, Geest KSM van der, Brouwer E, Glaudemans AWJM, Sinha B, Willemsen ATM, and Slart RHJA. Toward Reliable Uptake Metrics in Large Vessel Vasculitis Studies. Diagnostics 2021;11.

88. Wahl RL, Jacene H, Kasamon Y, and Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. Journal of nuclear medicine 2009;50:122S–150S.

89. Schwartz J, Humm J, Gonen M, Kalaigian H, Schoder H, Larson S, and Nehmeh S. Repeatability of SUV measurements in serial PET. Medical physics 2011;38:2629–38.

90. Fletcher J and Kinahan P. PET/CT standardized uptake values (SUVs) in clinical practice and assessing response to therapy. NIH Public Access 2010;31:496–505.

91. Adams MC, Turkington TG, Wilson JM, and Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. American Journal of Roentgenology 2010;195:310–20.

92. Eskian M, Alavi A, Khorasanizadeh M, Viglianti BL, Jacobsson H, Barwick TD, Meysamie A, Yi SK, Iwano S, Bybel B, et al. Effect of blood glucose level on standardized uptake value (SUV) in 18F-FDG PET-scan: a systematic review and meta-analysis of 20,807 individual SUV measurements. European journal of nuclear medicine and molecular imaging 2019;46:224–37.

93. Iking J, Staniszewska M, Kessler L, Klose JM, Lückerath K, Fendler WP, Herrmann K, and Rischpler C. Imaging inflammation with positron emission tomography. Biomedicines 2021;9:212.

94. Pelletier-Galarneau M and Ruddy TD. PET/CT for diagnosis and management of large-vessel vasculitis. Current cardiology reports 2019;21:34.

95. Nielsen BD, Hansen IT, Kramer S, Haraldsen A, Hjorthaug K, Bogsrud TV, Ejlersen JA, Stolle LB, Keller KK, Therkildsen P, et al. Simple dichotomous assessment of cranial artery inflammation by conventional 18F-FDG PET/CT shows high accuracy for the diagnosis of giant cell arteritis: a case-control study. European Journal of Nuclear Medicine and Molecular Imaging 2019;46:184–93.

96. Patel PY, Dalal I, and Griffith B. [18F] FDG-PET Evaluation of Spinal Pathology in Patients in Oncology: Pearls and Pitfalls for the Neuroradiologist. American Journal of Neuroradiology 2022;43:332–40.

97. Lee SW, Kim SJ, Seo Y, Jeong SY, Ahn BC, and Lee J. F-18 FDG PET for assessment of disease activity of large vessel vasculitis: A systematic review and meta-analysis. Journal of Nuclear Cardiology 2019;26:59–67.

98. Veeranna V, Fisher A, Nagpal P, Ghosh N, Fisher E, Steigner M, Creager MA, Dorbala S, and Di Carli MF. Utility of multimodality imaging in diagnosis and follow-up of aortitis. Journal of Nuclear Cardiology 2016;23:590–5.

99. Vaidyanathan S, Patel C, Scarsbrook A, and Chowdhury F. FDG PET/CT in infection and inflammation—current and emerging clinical applications. Clinical radiology 2015;70:787–800.

100. Chen Z, Liu M, Li L, and Chen L. Involvement of the Warburg effect in non-tumor diseases processes. Journal of cellular physiology 2018;233:2839–49.

101. Jacobson O, Kiesewetter DO, and Chen X. Fluorine-18 radiochemistry, labeling strategies and synthetic routes. Bioconjugate chemistry 2015;26:1–18.

102. Yu S. Review of 18F-FDG synthesis and quality control. Biomedical imaging and intervention journal 2006;2.

103. Fludeoxyglucose F 18 injection (FDG): Uses, dosage, side effects, interactions, warning. 2022. URL: https://www.rxlist.com/fludeoxyglucose-drug.htm#indications.

104. Walter MA, Melzer RA, Schindler C, Müller-Brand J, Tyndall A, and Nitzsche EU. The value of [18 F] FDG-PET in the diagnosis of large-vessel vasculitis and the assessment of activity and extent of disease. European journal of nuclear medicine and molecular imaging 2005;32:674–81.

105. Nielsen BD, Gormsen LC, Hansen IT, Keller KK, Therkildsen P, and Hauge EM. 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. European journal of nuclear medicine and molecular imaging 2018;45:1119–28.

106. Fuchs M, Briel M, Daikeler T, Walker UA, Rasch H, Berg S, Ng QK, 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. European journal of nuclear medicine and molecular imaging 2012;39:344–53.

107. Stellingwerff MD, Brouwer E, Lensen KJD, Rutgers A, Arends S, Van Der Geest KS, Glaudemans AW, and Slart RH. Different scoring methods of FDG PET/CT in giant cell arteritis: need for standardization. Medicine 2015;94.

108. Grayson PC, Alehashemi S, Bagheri AA, Civelek AC, Cupps TR, Kaplan MJ, Malayeri AA, Merkel PA, Novakovich E, Bluemke DA, et al. 18F-fluorodeoxyglucose–positron emission tomography as an imaging biomarker in a prospective, longitudinal cohort of patients with large vessel vasculitis. Arthritis & Rheumatology 2018;70:439–49.

109. Kang F, Han Q, Zhou X, Zheng Z, Wang S, Ma W, Zhang K, Quan Z, Yang W, Wang J, et al. Performance of the PET vascular activity score (PETVAS) for qualitative and quantitative assessment of inflammatory activity in Takayasu's arteritis patients. European journal of nuclear medicine and molecular imaging 2020:1–11.

110. Quinn KA, Dashora H, Novakovich E, Ahlman MA, and Grayson PC. Use of 18F-fluorodeoxyglucose positron emission tomography to monitor tocilizumab effect on vascular inflammation in giant cell arteritis. Rheumatology 2021;60:4384–9.

111. Mackie SL, Dejaco C, Appenzeller S, Camellino D, Duftner C, Gonzalez-Chiappe S, Mahr A, Mukhtyar C, Reynolds G, De Souza AWS, et al. British Society for Rheumatology guideline on diagnosis and treatment of giant cell arteritis. Rheumatology 2020;59:e1–e23.

112. Versari A, Pipitone N, Casali M, Jamar F, and Pazzola G. Use of imaging techniques in large vessel vasculitis and related conditions. The quarterly journal of nuclear medicine and molecular imaging: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of 2018;62:34–9.

113. Grayson PC, Alehashemi S, Bagheri AA, Civelek AC, Cupps TR, Kaplan MJ, Malayeri AA, Merkel PA, Novakovich E, Bluemke DA, et al. Positron emission tomography as an imaging biomarker in a prospective, longitudinal cohort of patients with large vessel vasculitis. Arthritis & rheumatology (Hoboken, NJ) 2018;70:439.

114. Dashora HR, Rosenblum JS, Quinn KA, Alessi H, Novakovich E, Saboury B, Ahlman MA, and Grayson P. Comparing semi-quantitative and qualitative methods of vascular FDG-PET activity measurement in large-vessel vasculitis. Journal of Nuclear Medicine 2021.

115. Laffon E and Marthan R. On Semi-quantitative Methods for Assessing Vascular 18FDG-PET Activity in Large-Vessels Vasculitis. Journal of Nuclear Medicine 2021.

116. Dellavedova L, Carletto M, Faggioli P, Sciascera A, Del Sole A, Mazzone A, and Maffioli L. The prognostic value of baseline 18 F-FDG PET/CT in steroid-naıve large-vessel vasculitis: introduction of volume-based parameters. European journal of nuclear medicine and molecular imaging 2016;43:340–8.

117. Motwani M. Hiding beyond plain sight: Textural analysis of positron emission tomography to identify high-risk plaques in carotid atherosclerosis. 2019.

118. Hatt M, Le Rest CC, 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*. Vol. 51. 2. Elsevier. 2021:126–33.

119. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 2012;48:441–6.

120. O'Sullivan F, Roy S, and Eary J. A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. Biostatistics 2003;4:433–48.

121. El Naqa I, Grigsby PW, Apte A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern recognition 2009;42:1162–71.

122. Galavis PE, Hollensen C, Jallow N, Paliwal B, and Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta oncologica 2010;49:1012–6.

123. Strand R, Malmberg F, Johansson L, Lind L, Sundbom M, Ahlström H, and Kullberg J. A concept for holistic whole body MRI data analysis, Imiomics. PloS one 2017;12:e0169966.

124. Visvikis D, Le Rest CC, Jaouen V, and Hatt M. Artificial intelligence, machine (deep) learning and radio (geno) mics: definitions and nuclear medicine imaging applications. European Journal of Nuclear Medicine and Molecular Imaging 2019:1–8.

125. Duff L and Tsoumpas C. Automated diagnosis and prediction in cardiovascular diseases using tomographic imaging.

126. Visvikis D, Lambin P, Beuschau Mauridsen K, Hustinx R, Lassmann M, Rischpler C, Shi K, and Pruim J. Application of artificial intelligence in nuclear medicine and molecular imaging: a review of current status and future perspectives for clinical translation. European journal of nuclear medicine and molecular imaging 2022:1–12.

127. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, and He Q. A comprehensive survey on transfer learning. Proceedings of the IEEE 2020;109:43–76.

128. Shorten C and Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of big data 2019;6:1–48.

129. Park JE, Park SY, Kim HJ, and Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. Korean journal of radiology 2019;20:1124–37.

130. Slart RH, Williams MC, Juarez-Orozco LE, Rischpler C, Dweck MR, Glaudemans AW, Gimelli A, Georgoulias P, Gheysens O, Gaemperli O, et al. Position paper of the EACVI and EANM on artificial intelligence applications in multimodality cardiovascular imaging using SPECT/CT, PET/CT, and cardiac CT. European journal of nuclear medicine and molecular imaging 2021;48:1399–413.

131. Lovinfosse P, Visvikis D, Hustinx R, and Hatt M. FDG PET radiomics: a review of the methodological aspects. Clinical and Translational Imaging 2018;6:379–91.

132. 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.

133. Chen L, Li S, Bai Q, Yang J, Jiang S, and Miao Y. Review of image classification algorithms based on convolutional neural networks. Remote Sensing 2021;13:4712.

134. Sarvamangala D and Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence 2022;15:1–22.

135. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, and Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. European radiology 2017;27:4498–509.

136. Pfaehler E, Beukinga RJ, Jong JR de, Slart RH, Slump CH, Dierckx RA, and Boellaard R. Repeatability of 18F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. Medical physics 2019;46:665–78.

137. Reader AJ and Zaidi H. Advances in PET image reconstruction. PET clinics 2007;2:173–90.

138. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, Tham IW, and Townsend D. Impact of image reconstruction settings on texture features in 18F-FDG PET. Journal of nuclear medicine 2015;56:1667–73.

139. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295:328.

140. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A, and Lambin P. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. Journal of Personalized Medicine 2021;11:842.

141. Da-Ano R, Visvikis D, and Hatt M. Harmonization strategies for multicenter radiomics investigations. Physics in Medicine & Biology 2020;65:24TR02.

142. Gatys LA, Ecker AS, and Bethge M. Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016:2414–23.

143. Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D, and Jaouen V. Standardization of multicentric image datasets with generative adversarial networks. In: *IEEE Nuclear Science Symposium and Medical Imaging Conference 2019*. 2019.

144. Modanwal G, Vellal A, Buda M, and Mazurowski MA. MRI image harmonization using cycle-consistent generative adversarial network. In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol. 11314. SPIE. 2020:259–64.

145. Shah J, Gao F, Li B, Ghisays V, Luo J, Chen Y, Lee W, Zhou Y, Benzinger TL, Reiman EM, et al. Deep residual inception encoder-decoder network for amyloid PET harmonization. Alzheimer's & Dementia 2022.

146. Da-Ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Castelli J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Scientific Reports 2020;10:1–12.

147. Papadimitroulas P, Brocki L, Chung NC, Marchadour W, Vermet F, Gaubert L, Eleftheriadis V, Plachouris D, Visvikis D, Kagadis GC, et al. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. Physica Medica 2021;83:108–21.

148. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, and Buvat I. A postreconstruction harmonization method for multicenter radiomic studies in PET. Journal of Nuclear Medicine 2018;59:1321–8.

149. Johnson WE, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27.

150. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 2018;167:104–20.

151. Timmeren JE van, Cester D, Tanadini-Lang S, Alkadhi H, and Baessler B. Radiomics in medical imaging—"How-to" guide and critical reflection. Insights into Imaging 2020;11:1–16.

152. Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, Abdalah MA, Schabath MB, Goldgof DG, Mackin D, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Medical physics 2017;44:1050–62.

153. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, and Lambin P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5:1–10.

154. Gallivanone F, Interlenghi M, D'Ambrosio D, Trifirò G, and Castiglioni I. Parameters influencing PET imaging features: a phantom study with irregular and heterogeneous synthetic lesions. Contrast media & molecular imaging 2018;2018.

155. Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, Biagioli MC, and Moros EG. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. Journal of applied clinical medical physics 2017;18:32–48.

156. Piri R, Edenbrandt L, Larsson M, Enqvist O, Nøddeskou-Fink AH, Gerke O, and Høilund-Carlsen PF. Aortic wall segmentation in 18F-sodium fluoride PET/CT scans: Head-to-head comparison of artificial intelligence-based versus manual segmentation. Journal of Nuclear Cardiology 2022;29:2001–10.

157. Reinke A, Tizabi MD, Sudre CH, Eisenmann M, Rädsch T, Baumgartner M, Acion L, Antonelli M, Arbel T, Bakas S, et al. Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642 2021.

158. Sedghi Gamechi Z, Bons LR, Giordano M, Bos D, Budde RP, Kofoed KF, Pedersen JH, Roos-Hesselink JW, and Bruijne M de. Automated 3D segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced CT. European radiology 2019;29:4613–23.

159. Rahman H, Rahman S, and Din F. Automatic segmentation of the aorta in cardiac medical images. The Nucleus 2017;54:90–6.

160. Florez E, Fatemi A, Claudio PP, and Howard CM. Emergence of radiomics: novel methodology identifying imaging biomarkers of disease in diagnosis, response, and progression. SM journal of clinical and medical imaging 2018;4.

161. Jin Y, Pepe A, Li J, Gsaxner C, Zhao Fh, Kleesiek J, Frangi AF, and Egger J. Ai-based aortic vessel tree segmentation for cardiovascular diseases treatment: status quo. arXiv preprint arXiv:2108.02998 2021.

162. Noothout JM, De Vos BD, Wolterink JM, and Išgum I. Automatic segmentation of thoracic aorta segments in low-dose chest CT. In: *Medical Imaging 2018: Image Processing*. Vol. 10574. International Society for Optics and Photonics. 2018:105741S.

163. Ronneberger O, Fischer P, and Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015:234–41.

164. Mohammadi S, Mohammadi M, Dehlaghi V, and Ahmadi A. Automatic segmentation, detection, and diagnosis of abdominal aortic aneurysm (AAA) using convolutional neural networks and hough circles algorithm. Cardiovascular Engineering and Technology 2019;10:490–9.

165. Jha D, Riegler MA, Johansen D, Halvorsen P, and Johansen HD. Doubleu-net: A deep convolutional neural network for medical image segmentation. In: *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE. 2020:558–64.

166. Bonechi S, Andreini P, Mecocci A, Giannelli N, Scarselli F, Neri E, Bianchini M, and Dimitri GM. Segmentation of Aorta 3D CT Images Based on 2D Convolutional Neural Networks. Electronics 2021;10:2559.

167. Hahn LD, Baeumler K, and Hsiao A. Artificial intelligence and machine learning in aortic disease. Current Opinion in Cardiology 2021;36:695–703.

168. Xie Y, Padgett J, Biancardi AM, and Reeves AP. Automated aorta segmentation in low-dose chest CT images. International journal of computer assisted radiology and surgery 2014;9:211–9.

169. Egger J, Freisleben B, Setser R, Renapuraar R, Biermann C, and O'Donnell T. Aorta segmentation for stent simulation. arXiv preprint arXiv:1103.1773 2011.

170. Subasic M, Loncaric S, and Sorantin E. Region-based deformable model for aortic wall segmentation. In: *3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the*. Vol. 2. IEEE. 2003:731–5.

171. Brooks FJ and Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. Journal of Nuclear Medicine 2014;55:37–42.

172. Hatt M, Vallieres M, Visvikis D, and Zwanenburg A. IBSI: an international community radiomics standardization initiative. 2018.

173. Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, Pellot-Barakat C, Soussan M, Frouin F, and Buvat I. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer research 2018;78:4786–9.

174. Korte JC, Cardenas C, Hardcastle N, Kron T, Wang J, Bahig H, Elgohari B, Ger R, Court L, Fuller CD, et al. Radiomics feature stability of open-source software evaluated on apparent diffusion coefficient maps in head and neck cancer. Scientific reports 2021;11:1–11.

175. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin JC, Pieper S, and Aerts HJ. Computational radiomics system to decode the radiographic phenotype. Cancer research 2017;77:e104–e107.

176. Song J, Yin Y, Wang H, Chang Z, Liu Z, and Cui L. A review of original articles published in the emerging field of radiomics. European journal of radiology 2020;127:108991.

177. Pfaehler E, Zhovannik I, Wei L, Boellaard R, Dekker A, Monshouwer R, El Naqa I, Bussink J, Gillies R, Wee L, et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. Physics and imaging in radiation oncology 2021;20:69–75.

178. Lorbeer R, Grotz A, Dörr M, Völzke H, Lieb W, Kühn JP, and Mensel B. Reference values of vessel diameters, stenosis prevalence, and arterial variations of the lower limb arteries in a male population sample using contrast-enhanced MR angiography. PLoS One 2018;13:e0197559.

179. Samarzija K, Milosevic P, Jurjevic Z, and Erdeljac E. Grading of carotid artery stenosis with computed tomography angiography: whether to use the narrowest diameter or the cross-sectional area. Insights into Imaging 2018;9:527–34.

180. Olson SL, Wijesinha MA, Panthofer AM, Blackwelder WC, Upchurch GR, Terrin ML, Curci JA, Baxter BT, and Matsumura JS. Evaluating growth patterns of abdominal aortic aneurysm diameter with serial computed tomography surveillance. JAMA surgery 2021;156:363–70.

181. Polak JF, Kronmal RA, Tell GS, O'Leary DH, Savage PJ, Gardin JM, Rutan GH, and Borhani NO. Compensatory increase in common carotid artery diameter: relation to blood pressure and artery intima-media thickness in older adults. Stroke 1996;27:2012–5.

182. Thapar A, Cheal D, Hopkins T, Ward S, Shaloub J, and Yusuf S. Internal or external wall diameter for abdominal aortic aneurysm screening? The Annals of The Royal College of Surgeons of England 2010;92:503–5.

183. Olive DJ. Principal component analysis. In: *Robust multivariate analysis*. Springer, 2017:189–217.

184. Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006;27:861–74.

185. Leevy JL, Khoshgoftaar TM, Bauder RA, and Seliya N. A survey on addressing high-class imbalance in big data. Journal of Big Data 2018;5:1–30.

186. Klawonn F, Höppner F, and May S. An alternative to ROC and AUC analysis of classifiers. In: *Advances in Intelligent Data Analysis X: 10th International Symposium, IDA 2011, Porto, Portugal, October 29-31, 2011. Proceedings 10*. Springer. 2011:210–21.

187. Provost F and Fawcett T. Robust classification for imprecise environments. Machine learning 2001;42:203–31.

188. Ling CX, Huang J, Zhang H, et al. AUC: a statistically consistent and more discriminating measure than accuracy. In: *Ijcai*. Vol. 3. 2003:519–24.

189. Halligan S, Altman DG, and Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. European radiology 2015;25:932–9.

190. Harrington MB. Some methodological questions concerning receiver operating characteristic (ROC) analysis as a method for assessing image quality in radiology. Journal of digital imaging 1990;3:211–8.

191. Chalkidou A, O'Doherty MJ, and Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PloS one 2015;10:e0124165.

192. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, Boellaard R, Bohndiek SE, Brady M, Brown G, et al. Imaging biomarker roadmap for cancer studies. Nature reviews Clinical oncology 2017;14:169–86.

193. Vallières M, Zwanenburg A, Badic B, Le Rest CC, Visvikis D, and Hatt M. Responsible radiomics research for faster clinical translation. 2018.

194. Collins GS, Reitsma JB, Altman DG, and Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery 2015;102:148–58.

195. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clinical chemistry 2015;61:1446–52.

196. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, and Liu C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one 2011;6:e17238.

197. Jochems A, Deist TM, Van Soest J, Eble M, Bulens P, Coucke P, Dries W, Lambin P, and Dekker A. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. Radiotherapy and Oncology 2016;121:459–67.

198. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RT, Jochems A, Miraglio B, Townend D, and Lambin P. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO clinical cancer informatics 2020;4:184–200.

199. Michael JR, Stanley C, Adamson R, and Kotevska O. Addressing the Limitations to Distributed Learning Containing Sensitive Data.

200. Verbraeken J, Wolting M, Katzy J, Kloppenburg J, Verbelen T, and Rellermeyer JS. A survey on distributed machine learning. Acm computing surveys (csur) 2020;53:1–33.

201. Ford RA, Price W, and Nicholson I. Privacy and accountability in black-box medicine. Mich. Telecomm. & Tech. L. Rev. 2016;23:1.

202. Wang S, Dong D, Li L, Li H, Bai Y, Hu Y, Huang Y, Yu X, Liu S, Qiu X, et al. A deep learning radiomics model to identify poor outcome in COVID-19 patients with underlying health conditions: A multicenter study. IEEE Journal of Biomedical and Health Informatics 2021;25:2353–62.

203. Wennmann M, Klein A, Bauer F, Chmelik J, Grözinger M, Uhlenbrock C, Lochner J, Nonnenmacher T, Rotkopf LT, Sauer S, et al. Combining deep learning and radiomics for automated, objective, comprehensive bone marrow characterization from whole-body MRI: a multicentric feasibility study. Investigative Radiology 2022;57:752–63.

204. Brown P, Zhong J, Frood R, Currie S, Gilbert A, Appelt A, Sebag-Montefiore D, and Scarsbrook A. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. European journal of nuclear medicine and molecular imaging 2019;46:2790–9.

205. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, Verzijlbergen FJ, Barrington SF, Pike LC, Weber WA, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. European journal of nuclear medicine and molecular imaging 2015;42:328–54.

206. Kikinis R, Pieper SD, and Vosburgh KG. 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. In: *Intraoperative imaging and image-guided therapy*. Springer, 2014:277–89.

207. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic resonance imaging 2012;30:1323–41.

208. Steyerberg EW. Validation in prediction research: the waste by data splitting. Journal of clinical epidemiology 2018;103:131–3.

209. Langs G, Röhrich S, Hofmanninger J, Prayer F, Pan J, Herold C, and Prosch H. Machine learning: from radiomics to discovery and routine. Der Radiologe 2018;58:1–6.

210. Nappi C and Cuocolo A. The machine learning approach: Artificial intelligence is coming to support critical clinical thinking. 2020.

211. Shrestha S and Sengupta PP. Machine learning for nuclear cardiology: The way forward. 2019.

212. Lodge MA, Chaudhry MA, and Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. Journal of Nuclear Medicine 2012;53:1041–7.

213. Duff L, Scarsbrook AF, Mackie SL, Frood R, Bailey M, Morgan AW, and Tsoumpas C. A methodological framework for AI-assisted diagnosis of active aortitis using Radiomic analysis of FDG PET–CT Images: Initial analysis. Journal of Nuclear Cardiology 2022.

214. Hatt M, Lucia F, Schick U, and Visvikis D. Multicentric validation of radiomics findings: challenges and opportunities. EBioMedicine 2019;47:20–1.

215. Duff LM, Scarsbrook AF, Ravikumar N, Frood R, Praagh GD van, Mackie SL, Bailey MA, Tarkin JM, Mason JC, Geest KS van der, et al. An Automated Method for Artifical Intelligence Assisted Diagnosis of Active Aortitis Using Radiomic Analysis of FDG PET-CT Images. Biomolecules 2023;13:343.

216. LIDA. Target. Last accessed 1st April 2022. 2022. URL: https://lida.leeds.ac.uk/target/.

217. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

218. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin JC, Pieper S, and Aerts HJ. Computational radiomics system to decode the radiographic phenotype. Cancer research 2017;77:e104–e107.

219. Xing H, Hao Z, Zhu W, Sun D, Ding J, Zhang H, Liu Y, and Huo L. Preoperative prediction of pathological grade in pancreatic ductal adenocarcinoma based on 18F-FDG PET/CT radiomics. EJNMMI research 2021;11:1–10.

220. Sun X and Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters 2014;21:1389–93.

221. DeLong ER, DeLong DM, and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988:837–45.

222. Larue RT, Defraene G, De Ruysscher D, Lambin P, and Van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. The British journal of radiology 2017;90:20160665.

223. López-Linares K, Garcıa I, Garcıa-Familiar A, Macıa I, and Ballester MAG. 3D convolutional neural network for abdominal aortic aneurysm segmentation. arXiv preprint arXiv:1903.00879 2019.

224. Sollini M, Antunovic L, Chiti A, and Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. European journal of nuclear medicine and molecular imaging 2019:1–17.

225. Ronrick Da, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Pradier O, Schick U, et al. Pre-selecting radiomic features based on their robustness to changes in imaging properties of multicentre data: impact on predictive modelling performance compared to ComBat harmonization of all available features. 2021.

226. Schlett C, Hendel T, Weckbach S, Reiser M, Kauczor H, Nikolaou K, Günther M, Forsting M, Hosten N, Völzke H, et al. Population-based imaging and radiomics: rationale and perspective of the German National Cohort MRI Study. In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 188. 07. © Georg Thieme Verlag KG. 2016:652–61.

227. Traverso A, Wee L, Dekker A, and Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. International Journal of Radiation Oncology* Biology* Physics 2018;102:1143–58.

228. Fei JL, Pu CL, Xu FY, Wu Y, and Hu HJ. Progress in radiomics of common heart disease based on cardiac magnetic resonance imaging. Journal of Molecular and Clinical Medicine 2021;4:29–38.

229. Juarez-Orozco LE, Martinez-Manzanera O, Zant FM van der, Knol RJ, and Knuuti J. Deep learning in quantitative PET myocardial perfusion imaging: a study on cardiovascular event prediction. Cardiovascular Imaging 2020;13:180–2.

230. Kwiecinski J, Kolossvary M, Tzolos E, Meah M, Adamson P, Joshi N, Williams M, Van Beek E, Berman D, Maurovich-Horvat P, et al. 18F-sodium fluoride positron emission tomography and coronary plaque radiomics derived from computed tomography angiography for prediction of myocardial infarction. European Heart Journal 2022;43:ehac544–212.

231. Yip SS and Aerts HJ. Applications and limitations of radiomics. Physics in Medicine & Biology 2016;61:R150.

232. Ibrahim A, Primakov S, Barufaldi B, Acciavatti RJ, Granzier RW, Hustinx R, Mottaghy FM, Woodruff HC, Wildberger JE, 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.

233. Orlhac F and 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.

234. Orlhac F, Eertink JJ, Cottereau AS, Zijlstra JM, Thieblemont C, Meignan MA, Boellaard R, and Buvat I. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. Journal of Nuclear Medicine 2021.

235. Bettinelli A, Marturano F, Avanzo M, Loi E, Menghi E, Mezzenga E, Pirrone G, Sarnelli A, Strigari L, Strolin S, and Paiusco M. A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. Radiology 0000;0. PMID: 35230182:211604.

236. Carles M, Fechter T, Martı-Bonmatı L, Baltas D, and Mix M. Experimental phantom evaluation to identify robust positron emission tomography (PET) radiomic features. EJNMMI physics 2021;8:1–17.

237. Oliveira C, Amstutz F, Vuong D, Bogowicz M, Hüllner M, Foerster R, Basler L, Schröder C, Eboulet EI, Pless M, et al. Preselection of robust radiomic features does not improve outcome modelling in non-small cell lung cancer based on clinical routine FDG-PET imaging. EJNMMI research 2021;11:1–12.

238. Toner YC, Ghotbi AA, Naidu S, Sakurai K, Leent MM van, Jordan S, Ordikhani F, Amadori L, Sofias AM, Fisher EL, et al. Systematically evaluating DOTATATE and FDG as PET immuno-imaging tracers of cardiovascular inflammation. Scientific reports 2022;12:1–15.

239. Jensen JK, Madsen JS, Jensen ME, Kjaer A, and Ripa RS. [64Cu] Cu-DOTATATE PET metrics in the investigation of atherosclerotic inflammation in humans. Journal of Nuclear Cardiology 2022:1–15.

240. Jiemy WF, Heeringa P, Kamps JA, Laken CJ van der, Slart RH, and Brouwer E. Positron emission tomography (PET) and single photon emission computed tomography (SPECT) imaging of macrophages in large vessel vasculitis: current status and future prospects. Autoimmunity reviews 2018;17:715–26.

241. Pugliese F, Gaemperli O, Kinderlerer AR, Lamare F, Shalhoub J, Davies AH, Rimoldi OE, Mason JC, and Camici PG. Imaging of vascular inflammation with [11C]-PK11195 and positron emission tomography/computed tomography angiography. Journal of the American College of Cardiology 2010;56:653–61.

242. Lamare F, Hinz R, Gaemperli O, Pugliese F, Mason JC, Spinks T, Camici PG, and Rimoldi OE. Detection and quantification of large-vessel inflammation with 11C-(R)-PK11195 PET/CT. Journal of Nuclear Medicine 2011;52:33–9.

243. Espitia O, Schanus J, Agard C, Kraeber-Bodéré F, Hersant J, Serfaty JM, and Jamet B. Specific features to differentiate Giant cell arteritis aortitis from aortic atheroma using FDG-PET/CT. Scientific Reports 2021;11:1–10.

244. Zerizer I, Tan K, Khan S, Barwick T, Marzola MC, Rubello D, and Al-Nahhas A. Role of FDG-PET and PET/CT in the diagnosis and management of vasculitis. European journal of radiology 2010;73:504–9.

245. Soussan M, Nicolas P, Schramm C, Katsahian S, Pop G, Fain O, and Mekinian A. Management of large-vessel vasculitis with FDG-PET: a systematic literature review and meta-analysis. Medicine 2015;94.

246. Tatsumi M, Cohade C, Nakamoto Y, and Wahl RL. Fluorodeoxyglucose uptake in the aortic wall at PET/CT: possible finding for active atherosclerosis. Radiology 2003;229:831–7.

247. Tomaszewski MR and Gillies RJ. The biological meaning of radiomic features. Radiology 2021;298:505–16.

Chapter 6

Appendix

6.1 Experiment Set 1

6.1.1 Individual radiomic featuresand SUV Metrics Tuned Hyperparameters

In the case a hyperparameter is not listed the default for the given Sci-kit Learn version number was used. 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*)

	Feature	Params
0	GLSZM Small Area High Gray Level Emphasis	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
1	GLSZM High Gray Level Zone Emphasis	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 750, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}

2	GLRLM Run Entropy	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 1250, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}
3	GLSZM Size Zone Non Uniformity Normalized	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
4	firstorder Mean Absolute Deviation	{'C': 4, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 2, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
5	GLDM Dependence Entropy	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
6	GLRLM Gray Level Variance	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 250, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
7	GLDM Small Dependence High Gray Level Emphasis	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 5000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.001}
8	firstorder Entropy	{'C': 2, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}

9	firstorder Variance	{'C': 1.5, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
10	GLCM Cluster Tendency	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 5000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.001}
11	GLCM Contrast	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
12	GLCM Sum Squares	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 3000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
13	GLDM Gray Level Variance	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
14	GLCM Difference Variance	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 500, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}
15	GLSZM Small Area Emphasis	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}

16	GLSZM Size Zone Non Uniformity	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 750, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
17	GLCM Sum Entropy	{'C': 2, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
18	GLRLM High Gray Level Run Emphasis	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 250, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.001}
19	GLRLM Short Run High Gray Level Emphasis	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 250, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}
20	GLSZM Gray Level Variance	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 1250, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
21	GLCM Difference Entropy	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
22	GLRLM Long Run High Gray Level Emphasis	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 250, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 1e-05}

23	firstorder Energy	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
24	firstorder Total Energy	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
25	SUV 50	{'C': 3, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
26	firstorder Robust Mean Absolute Deviation	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
27	GLCM Autocorrelation	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 1750, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
28	GLCM Joint Entropy	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
29	GLDM High Gray Level Emphasis	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 3000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}

30	firstorder Interquartile Range	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
31	GLCM Inverse Variance	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
32	GLCM Difference Average	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
33	firstorder Range	{'C': 4, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 2, 'max_iter': 100, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
34	GLCM Joint Average	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
35	GLCM Sum Average	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 5, 'max_iter': 2500, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
36	GLSZM Small Area Low Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

37	SUV 60	{'C': 2, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 2, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
38	firstorder 90 Percentile	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
39	firstorder Maximum	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 100, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
40	GLCM Id	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
41	GLDM Dependence Non Uniformity	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
42	GLDM Small Dependence Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
43	GLRLM Gray Level Non Uniformity Normalized	{'C': 4, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 1e-07}

44	GLSZM Zone Percentage	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
45	firstorder Uniformity	{'C': 2, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
46	GLCM Idm	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
47	firstorder Root Mean Squared	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
48	firstorder Mean	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
49	GLDM Dependence Non Uniformity Normalized	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
50	GLSZM Zone Entropy	{'C': 4, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 2000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}

51	SUV 70	{'C': 1.5, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}
52	GLRLM Run Length Non Uniformity Normalized	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
53	SUV 90	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
54	SUV 80	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
55	GLRLM Run Percentage	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
56	GLRLM Short Run Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
57	GLDM Large Dependence Emphasis	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': False, 'intercept_scaling': 2, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.001}

58	GLSZM Gray Level Non Uniformity	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
59	firstorder Median	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
60	GLRLM Run Length Non Uniformity	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
61	firstorder 10 Percentile	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
62	GLDM Dependence Variance	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
63	GLDM Small Dependence Low Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
64	shape Surface Volume Ratio	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

65	GLCM Imc2	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
66	shape Flatness	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
67	GLCM Cluster Prominence	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
68	GLCM Cluster Shade	{'C': 4, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 1e-07}
69	shape Mesh Volume	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
70	shape Voxel Volume	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
71	shape Least Axis Length	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

72	firstorder Skewness	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
73	shape Surface Area	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
74	GLSZM Large Area High Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
75	GLRLM Long Run Emphasis	{'C': 4, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': False, 'intercept_scaling': 2, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
76	GLDM Gray Level Non Uniformity	{'C': 1.5, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 5000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
77	GLCM M C C	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
78	GLSZM Zone Variance	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

79	shape Sphericity	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
80	GLRLM Gray Level Non Uniformity	{'C': 4, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 1e-07}
81	GLDM Large Dependence High Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
82	GLCM Idmn	{'C': 1, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': False, 'intercept_scaling': 3, 'max_iter': 5000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
83	GLCM Correlation	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
84	GLSZM Large Area Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
85	GLDM Large Dependence Low Gray Level Emphasis	{'C': 4, 'class_weight': None, 'dual': True, 'fit_intercept': False, 'intercept_scaling': 3, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}

86	shape Elongation	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
87	GLSZM Large Area Low Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
88	GLRLM Short Run Low Gray Level Emphasis	{'C': 4, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 5, 'max_iter': 10000, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 1e-07}
89	shape Maximum2 D Diameter Slice	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
90	shape Major Axis Length	{'C': 3, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 5000, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
91	shape Minor Axis Length	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
92	GLCM Idn	{'C': 3, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 2, 'max_iter': 10, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}

93	firstorder Kurtosis	{'C': 3, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': False, 'intercept_scaling': 1, 'max_iter': 50, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
94	shape Maximum2 D Diameter Column	{'C': 4, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': False, 'intercept_scaling': 1, 'max_iter': 100, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
95	firstorder Minimum	{'C': 2, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
96	GLSZM Gray Level Non Uniformity Normalized	{'C': 4, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 1, 'max_iter': 50, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
97	shape Maximum3 D Diameter	{'C': 4, 'class_weight': 'balanced', 'dual': True, 'fit_intercept': False, 'intercept_scaling': 1, 'max_iter': 100, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
98	shape Maximum2 D Diameter Row	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
99	GLCM Imc1	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

100	GLCM Joint Energy	{'C': 3, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 2000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
101	GLCM Maximum Probability	{'C': 1.5, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 100, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-06}
102	GLDM Low Gray Level Emphasis	{'C': 1, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': False, 'intercept_scaling': 4, 'max_iter': 5000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
103	GLRLM Low Gray Level Run Emphasis	{'C': 2, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 2000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
104	GLSZM Low Gray Level Zone Emphasis	{'C': 3, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 2000, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.01}
105	GLRLM Long Run Low Gray Level Emphasis	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
106	GLRLM Run Variance	{'C': 1, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 10, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}

6.1.2 Fingerprint Tuned Hyperparameters

In the case a hyperparameter is not listed the default for the given Sci-kit Learn version number was used. Key - ML = Machine Learning, Params = parameters, rf = Random Forest, lgr = Logistic Regression, svm = Support Vector Machine, dt = Decision Tree, gpc = Gaussain Process Classifier, perc = Perceptron, pasagr = Passive Aggressive, nnet = Neural Network, kneigh = K Nearest Neighbours

6.1.2.1 Fingerprint A

ML Type	Params
rf	{'bootstrap': True, 'max_depth': 100, 'max_features': None, 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 500, 'random_state': 1}
lgr	{'C': 3, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter': 50, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.0001}
svm	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf', 'random_state': 1}
dt	{'ccp_alpha': 0.1, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'auto', 'max_leaf_nodes': 10, 'min_impurity_decrease': 0.1, 'min_samples_leaf': 1, 'min_samples_split': 2, 'random_state': 1, 'splitter': 'best'}
gpc	{'max_iter_predict': 100, 'random_state': 1}
perc	{'alpha': 0.1, 'fit_intercept': False, 'max_iter': 1000, 'penalty': 'None', 'random_state': 1}
pasagr	{'C': 1.0, 'average': True, 'fit_intercept': True, 'max_iter': 500, 'random_state': 1}
nnet	{'alpha': 0.0001, 'hidden_layer_sizes': (10,), 'max_iter': 200, 'random_state': 1}
kneigh	{'algorithm': 'brute', 'leaf_size': 10, 'n_neighbors': 5, 'weights': 'uniform'}

6.1.2.2 Fingerprint B

ML Type	Params
rf	{'bootstrap': True, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 1000, 'random_state': 1}
lgr	{'C': 1, 'dual': True, 'fit_intercept': True, 'intercept_scaling': 4, 'max_iter': 1500, 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear', 'tol': 1e-07}
dt	{'ccp_alpha': 0.0, 'criterion': 'gini', 'max_depth': 50, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 10, 'random_state': 1, 'splitter': 'random'}
gpc	{'max_iter_predict': 100, 'random_state': 1}
perc	{'alpha': 0.01, 'fit_intercept': True, 'max_iter': 500, 'penalty': 'l2', 'random_state': 1}
pasagr	{'C': 1.0, 'average': True, 'fit_intercept': True, 'max_iter': 500, 'random_state': 1}
nnet	{'alpha': 0.001, 'hidden_layer_sizes': (5,), 'max_iter': 2000, 'random_state': 1}
kneigh	{'algorithm': 'brute', 'leaf_size': 50, 'n_neighbors': 5, 'weights': 'uniform'}
svm	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf', 'random_state': 1}

6.1.2.3 Fingerprint C

	ML Type	Params
0	rf	{'bootstrap': False, 'max_depth': 100, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 10, 'random_state': 1}
1	lgr	{'C': 4, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 3, 'max_iter': 2500, 'penalty': 'l1', 'random_state': 1, 'solver': 'liblinear', 'tol': 0.1}
2	svm	{'C': 2, 'gamma': 'auto', 'kernel': 'rbf', 'random_state': 1}
3	dt	{'ccp_alpha': 0.0, 'criterion': 'gini', 'max_depth': 50, 'max_features': 'auto', 'max_leaf_nodes': 10, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 5, 'random_state': 1, 'splitter': 'random'}
4	gpc	{'max_iter_predict': 100, 'random_state': 1}
6	perc	{'alpha': 0.01, 'fit_intercept': True, 'max_iter': 500, 'penalty': 'l1', 'random_state': 1}
7	pasagr	{'C': 1.0, 'average': False, 'fit_intercept': True, 'max_iter': 500, 'random_state': 1}
8	nnet	{'alpha': 0.1, 'hidden_layer_sizes': (2000,), 'max_iter': 2000, 'random_state': 1}
9	kneigh	{'algorithm': 'ball_tree', 'leaf_size': 50, 'n_neighbors': 5, 'weights': 'distance'}

6.1.3 Individual Radiomic Features and SUV Metrics Results

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLSZM Size Zone Non Uniformity Normalized	0.853	0.097	0.896	0.066	5.82e-08
GLSZM High Gray Level Zone Emphasis	0.787	0.110	0.892	0.140	4.01e-08
GLSZM Small Area High Gray Level Emphasis	0.773	0.154	0.892	0.129	4.83e-08
GLRLM Gray Level Variance	0.760	0.066	0.888	0.087	8.92e-08
GLRLM Long Run High Gray Level Emphasis	0.800	0.228	0.884	0.163	3.92e-07
GLDM Dependence Entropy	0.800	0.117	0.884	0.107	2.75e-08
GLRLM Short Run High Gray Level Emphasis	0.827	0.186	0.884	0.147	1.94e-07
firstorder Variance	0.773	0.099	0.884	0.088	1.07e-07
GLDM Gray Level Variance	0.800	0.091	0.884	0.088	1.07e-07

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLSZM Small Area Emphasis	0.800	0.128	0.884	0.105	2.61e-07
GLRLM High Gray Level Run Emphasis	0.827	0.186	0.880	0.146	1.94e-07
GLDM Small Dependence High Gray Level Emphasis	0.760	0.099	0.880	0.122	1.36e-07
GLRLM Run Entropy	0.787	0.097	0.880	0.097	4.83e-08
GLCM Cluster Tendency	0.800	0.091	0.880	0.077	1.07e-07
GLCM Sum Squares	0.800	0.052	0.876	0.076	1.28e-07
GLDM High Gray Level Emphasis	0.827	0.186	0.872	0.141	3.11e-07
GLCM Autocorrelation	0.813	0.184	0.868	0.150	5.84e-07
firstorder Mean Absolute Deviation	0.787	0.097	0.864	0.095	1.83e-07
GLCM Sum Average	0.800	0.181	0.860	0.159	1.08e-06
GLCM Joint Average	0.653	0.033	0.860	0.159	1.08e-06
GLCM Sum Entropy	0.800	0.052	0.860	0.096	2.76e-07
GLSZM Size Zone Non Uniformity	0.813	0.062	0.856	0.096	4.93e-07
firstorder Entropy	0.787	0.062	0.856	0.102	3.29e-07

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLCM Difference Variance	0.787	0.122	0.844	0.109	4.65e-07
GLCM Joint Entropy	0.653	0.033	0.844	0.108	1.02e-06
firstorder Robust Mean Absolute Deviation	0.667	0.000	0.844	0.103	1.68e-06
GLSZM Gray Level Variance	0.787	0.033	0.840	0.099	8.20e-07
GLCM Difference Entropy	0.653	0.033	0.840	0.114	1.51e-06
GLCM Contrast	0.800	0.105	0.840	0.116	5.52e-07
firstorder Total Energy	0.667	0.000	0.840	0.116	1.02e-06
firstorder Energy	0.667	0.000	0.840	0.116	1.02e-06
GLDM Dependence Non Uniformity	0.667	0.000	0.836	0.135	2.20e-06
GLCM Inverse Variance	0.667	0.000	0.832	0.122	3.20e-06
GLCM Difference Average	0.707	0.066	0.828	0.124	2.20e-06
firstorder 90 Percentile	0.653	0.033	0.828	0.139	3.04e-06
GLSZM Small Area Low Gray Level Emphasis	0.667	0.000	0.820	0.112	2.59e-06

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLSZM Zone Percentage	0.667	0.000	0.816	0.113	1.10e-05
firstorder Interquartile Range	0.653	0.033	0.816	0.106	4.17e-06
GLCM Id	0.667	0.000	0.812	0.126	1.28e-05
SUV 50	0.773	0.186	0.808	0.217	7.37e-06
GLDM Small Dependence Emphasis	0.667	0.000	0.808	0.118	9.03e-06
firstorder Uniformity	0.667	0.000	0.808	0.105	4.88e-06
GLRLM Gray Level Non Uniformity Normalized	0.333	0.000	0.808	0.105	6.65e-06
firstorder Root Mean Squared	0.653	0.033	0.804	0.185	8.16e-06
GLSZM Zone Entropy	0.573	0.134	0.800	0.108	6.65e-06
GLCM Idm	0.667	0.000	0.800	0.143	1.90e-05
firstorder Mean	0.653	0.033	0.788	0.170	1.72e-05
firstorder Range	0.720	0.081	0.780	0.200	5.16e-05
SUV 60	0.773	0.226	0.780	0.251	2.54e-05
firstorder Median	0.653	0.033	0.780	0.164	9.31e-05

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLRLM Run Length Non Uniformity Normalized	0.667	0.000	0.780	0.127	6.79e-05
GLSZM Gray Level Non Uniformity	0.667	0.000	0.768	0.108	1.16e-04
GLRLM Run Percentage	0.667	0.000	0.768	0.128	8.90e-05
GLRLM Run Length Non Uniformity	0.667	0.000	0.768	0.177	1.27e-04
shape Surface Volume Ratio	0.667	0.000	0.768	0.243	1.39e-04
firstorder Maximum	0.693	0.066	0.764	0.240	8.51e-05
SUV 70	0.787	0.212	0.760	0.239	8.90e-05
GLRLM Short Run Emphasis	0.667	0.000	0.756	0.116	1.33e-04
GLDM Dependence Non Uniformity Normalized	0.667	0.000	0.756	0.136	1.21e-04
SUV 80	0.653	0.033	0.744	0.266	2.22e-04
SUV 90	0.653	0.033	0.744	0.266	1.72e-04
GLDM Large Dependence Emphasis	0.347	0.033	0.740	0.164	5.47e-04

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
shape Voxel Volume	0.667	0.000	0.732	0.237	1.10e-03
shape Mesh Volume	0.667	0.000	0.732	0.237	1.14e-03
shape Flatness	0.667	0.000	0.724	0.241	5.93e-04
GLDM Dependence Variance	0.653	0.033	0.724	0.179	1.71e-03
firstorder 10 Percentile	0.653	0.033	0.720	0.203	8.74e-04
GLDM Small Dependence Low Gray Level Emphasis	0.667	0.000	0.716	0.188	4.86e-04
shape Least Axis Length	0.667	0.000	0.716	0.256	1.90e-03
GLCM Imc2	0.667	0.000	0.716	0.157	8.74e-04
shape Surface Area	0.667	0.000	0.704	0.237	3.68e-03
GLCM Cluster Shade	0.600	0.091	0.696	0.066	3.10e-03
firstorder Skewness	0.667	0.052	0.692	0.136	5.63e-03
shape Sphericity	0.667	0.000	0.680	0.242	7.25e-03
GLDM Gray Level Non Uniformity	0.387	0.097	0.664	0.163	1.14e-02
GLRLM Long Run Emphasis	0.333	0.000	0.632	0.244	2.18e-02

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLDM Large Dependence High Gray Level Emphasis	0.667	0.000	0.628	0.175	3.93e-02
GLCM Correlation	0.667	0.000	0.608	0.207	4.33e-02
GLCM M C C	0.667	0.000	0.604	0.159	4.43e-02
shape Elongation	0.667	0.000	0.596	0.195	1.07e-01
shape Minor Axis Length	0.667	0.000	0.580	0.181	1.65e-01
shape Maximum2 D Diameter Slice	0.667	0.000	0.572	0.190	1.60e-01
shape Major Axis Length	0.333	0.000	0.568	0.231	1.29e-01
GLSZM Gray Level Non Uniformity Normalized	0.333	0.000	0.536	0.089	6.65e-06
GLCM Imc1	0.667	0.000	0.536	0.148	3.74e-01
GLCM Cluster Prominence	0.547	0.192	0.532	0.245	2.70e-03
shape Maximum2 D Diameter Column	0.333	0.000	0.520	0.223	4.73e-01
shape Maximum2 D Diameter Row	0.667	0.000	0.506	0.213	4.24e-01

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
shape Maximum3 D Diameter	0.333	0.000	0.504	0.258	4.84e-01
GLRLM Low Gray Level Run Emphasis	0.333	0.000	0.500	0.000	1.11e-02
GLDM Low Gray Level Emphasis	0.333	0.000	0.500	0.000	1.18e-02
GLCM Idn	0.667	0.000	0.500	0.000	3.08e-01
GLCM Maximum Probability	0.333	0.000	0.500	0.000	2.13e-04
GLCM Joint Energy	0.333	0.000	0.500	0.000	3.93e-03
GLSZM Low Gray Level Zone Emphasis	0.333	0.000	0.500	0.000	4.95e-03
firstorder Kurtosis	0.667	0.000	0.492	0.167	4.93e-01
GLCM Idmn	0.400	0.166	0.468	0.206	1.86e-01
GLRLM Gray Level Non Uniformity	0.400	0.166	0.464	0.254	3.48e-02
GLRLM Short Run Low Gray Level Emphasis	0.467	0.203	0.460	0.434	2.67e-05
firstorder Minimum	0.667	0.000	0.432	0.167	2.66e-01
GLSZM Zone Variance	0.400	0.117	0.352	0.149	3.07e-02

Table 6.2: Individual radiomic features and SUV Metrics Results : Key - *CI (Confidence Interval)*, *AUC (Area Under the Receiver Operating Characteristic Curve)*, *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	Accuracy	Accuracy CI	ROC AUC	ROC AUC CI	*p* value
GLSZM Large Area Emphasis	0.400	0.117	0.336	0.203	1.13e-01
GLSZM Large Area High Gray Level Emphasis	0.400	0.117	0.332	0.213	2.56e-02
GLRLM Long Run Low Gray Level Emphasis	0.653	0.033	0.324	0.193	8.60e-02
GLSZM Large Area Low Gray Level Emphasis	0.400	0.117	0.324	0.196	1.27e-01
GLRLM Run Variance	0.653	0.033	0.320	0.178	1.75e-02
GLDM Large Dependence Low Gray Level Emphasis	0.400	0.117	0.320	0.194	1.07e-01

6.2 Experiment Set 2

6.2.1 ML Parameters and Diagnostic Performance of Individual SUV metrics and Radiomic Features

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	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order Energy	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.642	0.096	0.826	0.050	0.700	0.900	0.615	0.893
first order TotalEnergy	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.642	0.096	0.826	0.050	0.700	0.900	0.615	0.893

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Dependence Entropy	('C', 2.1005202681605333), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 4071), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08938020829083303)	0.500	0.000	0.737	0.106	0.500	0.850	0.500	0.870
first order Uniformity	('C', 1.7802480372728924), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 2127), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02392007563908262)	0.500	0.000	0.739	0.221	0.500	0.783	0.500	0.849

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm Run Entropy	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 4750), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.603	0.110	0.831	0.084	0.600	0.933	0.632	0.819
glcm Sum Entropy	('C', 3.993769314626932), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 5592), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00021969320512658575)	0.578	0.092	0.803	0.063	0.600	0.900	0.594	0.816
gldm Dependence NonUniformity	('C', 3.8408563166271104), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1569), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0004258049224511096)	0.665	0.126	0.721	0.125	0.700	0.867	0.538	0.813

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm Run-Length-NonUniformity	('C', 3.5740797936701934), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 9395), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07777955013689973)	0.539	0.068	0.683	0.170	0.500	0.850	0.538	0.806
first order 90Percentile	('C', 2.718646137921694), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 8730), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08536276523234584)	0.617	0.107	0.854	0.046	0.600	0.917	0.615	0.799
glcm Sum-Squares	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.748	0.095	0.819	0.111	0.658	0.883	0.704	0.796

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order Entropy	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10000), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.578	0.092	0.803	0.070	0.600	0.900	0.594	0.796
glrlm Run-Variance	('C', 1.5258121739648165), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 7507), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06657014158248468)	0.500	0.000	0.822	0.073	0.500	0.683	0.500	0.796
glszm Gray Level Variance	('C', 2.127618551416409), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7726), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.036778331570860544)	0.539	0.068	0.810	0.114	0.500	0.867	0.500	0.796

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm Zone Entropy	('C', 1.2839430530160425), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 5449), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.003741177274838674)	0.500	0.000	0.673	0.135	0.500	0.950	0.500	0.796
glszm Size-Zone NonUniformity	('C', 1.0004183610598405), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 79), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.028501415351397944)	0.603	0.103	0.798	0.027	0.758	0.900	0.671	0.796

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm ClusterTendency	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.728	0.087	0.819	0.111	0.758	0.883	0.704	0.789
glcm DifferenceVariance	('C', 2.113537527362806), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 9276), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06055994223970045)	0.768	0.085	0.847	0.056	0.758	0.917	0.704	0.789
glcm DifferenceAverage	('C', 1.7285617514498897), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 7279), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03162024999032157)	0.637	0.102	0.815	0.039	0.700	0.883	0.671	0.789

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order MeanAbsoluteDeviation	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.626	0.082	0.830	0.104	0.600	0.867	0.726	0.786
glcm Contrast	('C', 1.273776071729263), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 4522), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 4.7896309831593e-05)	0.734	0.107	0.831	0.043	0.758	0.900	0.726	0.786
glszm HighGray Level Zone Emphasis	('C', 2.487532345055786), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 321), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05719484501798079)	0.628	0.125	0.809	0.073	0.500	0.933	0.709	0.786

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm SmallArea HighGray Level Emphasis	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.1)	0.642	0.109	0.814	0.081	0.700	0.967	0.747	0.779
first order Maximum	('C', 2.406061918574112), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 6133), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.008718682280560601)	0.500	0.000	0.718	0.063	0.500	0.867	0.500	0.776
gldm SmallDependence HighGray Level Emphasis	('C', 3.5820753121303586), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 8617), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04935620839082142)	0.642	0.109	0.803	0.082	0.500	0.867	0.594	0.773

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order RootMean-Squared	('C', 1.5327054833890021), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 344), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05669142719368856)	0.603	0.132	0.817	0.083	0.500	0.883	0.577	0.773
glcm Joint Entropy	('C', 1.3927742769764304), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5626), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07748328569448332)	0.578	0.092	0.820	0.035	0.600	0.917	0.555	0.773

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Difference Entropy	('C', 2.6195762846485184), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 4457), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05288857799872605)	0.553	0.093	0.815	0.039	0.600	0.883	0.555	0.769
SUV 90	('C', 2.4244968771926634), ('dual', False), ('fit intercept', False), ('intercept scaling', 1), ('max iter', 1722), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.033909349677505876)	0.500	0.000	0.675	0.074	0.500	0.850	0.500	0.769

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldmcargeDependence Emphasis	(’C’, 2.5352247856333654), (’dual’, False), (’fit intercept’, True), (’intercept scaling’, 2), (’max iter’, 7622), (’penalty’, ’l2’), (’random state’, 1), (’solver’, ’liblinear’), (’tol’, 0.07978938765199757)	0.559	0.064	0.816	0.065	0.658	0.650	0.632	0.769
first order Interquartil-eRange	(’C’, 2.7766313751946816), (’dual’, False), (’fit intercept’, True), (’intercept scaling’, 1), (’max iter’, 10000), (’penalty’, ’l1’), (’random state’, 1), (’solver’, ’liblinear’), (’tol’, 0.04856253189274283)	0.635	0.110	0.820	0.041	0.600	0.883	0.671	0.766

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Gray Level Variance	('C', 3.6396129799025188), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 9627), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0009406748589996944)	0.587	0.055	0.836	0.113	0.558	0.867	0.666	0.766
glcm Idm	('C', 1.36824291492517), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 5772), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.023090099089280777)	0.500	0.000	0.611	0.120	0.500	0.850	0.500	0.766

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Imc1	('C', 2.888312473667224), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 6922), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07073965149829738)	0.500	0.000	0.518	0.100	0.500	0.417	0.500	0.766
glcm Imc2	('C', 1.8226093327930175), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 5466), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.016846207468831566)	0.500	0.000	0.544	0.040	0.500	0.550	0.500	0.763

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order Variance	('C', 3.854998757422557), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1246), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0998769147597789)	0.539	0.068	0.836	0.113	0.500	0.867	0.478	0.763
glrlm Gray Level Variance	('C', 3.334445713160191), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7864), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05915104676294525)	0.550	0.088	0.836	0.113	0.500	0.867	0.594	0.759

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order Range	('C', 2.255078222495345), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 4481), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.018335711167007784)	0.525	0.044	0.689	0.075	0.500	0.850	0.500	0.759
SUV 80	('C', 2.7327112335961634), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 6452), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09619248818230378)	0.500	0.000	0.670	0.078	0.500	0.817	0.500	0.759

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order RobustMean-AbsoluteDevi-ation	('C', 3.9089933705144815), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1573), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.030044861885312284)	0.614	0.124	0.809	0.065	0.600	0.867	0.632	0.756
SUV 60	('C', 1.2395197537743265), ('dual', False), ('fit in-tercept', False), ('inter-cept scaling', 2), ('max iter', 9100), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.010505875632877961)	0.500	0.000	0.707	0.064	0.500	0.950	0.500	0.756

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm MCC	('C', 3.981512763672209), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 3128), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06580703815293194)	0.500	0.000	0.514	0.085	0.500	0.467	0.500	0.754
glszm SizeZone NonUniformityNormalized	('C', 3.9948777017611143), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 799), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08280778956235492)	0.500	0.000	0.804	0.072	0.500	0.800	0.500	0.746

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
SUV 70	('C', 1.780282005988736), ('dual', False), ('fit intercept', False), ('intercept scaling', 5), ('max iter', 9067), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04823710111198745)	0.500	0.000	0.675	0.074	0.500	0.817	0.500	0.746
first order Mean	('C', 3.2831405881361713), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 8189), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05090148535753912)	0.589	0.156	0.828	0.087	0.500	0.833	0.577	0.746

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Small-Dependence Emphasis	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.637	0.094	0.814	0.051	0.517	0.783	0.594	0.742
glszm Small-Area Emphasis	('C', 1.7721645038605833), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 6054), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.005225511696897342)	0.500	0.000	0.776	0.093	0.500	0.800	0.500	0.742

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Correlation	('C', 3.16725277727334), ('dual', False), ('fit intercept', False), ('intercept scaling', 1), ('max iter', 661), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0441786617964023)	0.500	0.000	0.544	0.092	0.500	0.400	0.500	0.736
SUV 50	('C', 1.1585973852260885), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 7719), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07910532963541077)	0.500	0.000	0.745	0.044	0.500	0.933	0.500	0.736

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Large Dependence Low Gray Level Emphasis	('C', 3.4944666036345877), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 7229), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09935679154711934)	0.525	0.044	0.751	0.126	0.500	0.733	0.615	0.732
glrlm ShortRunHighGray Level Emphasis	('C', 1.270884315070445), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 70), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09525369448065293)	0.617	0.083	0.777	0.062	0.500	0.717	0.709	0.729

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape MeshVolume	('C', 1.3433300133497719), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 8217), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.026421331836750907)	0.500	0.000	0.622	0.202	0.500	0.800	0.500	0.729
shape VoxelVolume	('C', 1.7188383250667447), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 5939), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06484907769886163)	0.500	0.000	0.622	0.202	0.500	0.800	0.500	0.729

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Dependence NonUniformityNormalized	('C', 1.1840941552706046), ('dual', False), ('fit intercept', False), ('intercept scaling', 5), ('max iter', 6382), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.018382870678376382)	0.500	0.000	0.805	0.051	0.500	0.800	0.500	0.729
glrlm HighGray Level RunEmphasis	('C', 2.4151318684104424), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 1846), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.048899543371909375)	0.617	0.083	0.772	0.062	0.500	0.683	0.687	0.726

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm Zone Percentage	('C', 3.7783288591737643), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 351), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.004178404335664326)	0.648	0.101	0.820	0.049	0.558	0.767	0.594	0.726
glrlm Run-Length-NonUniformityNormalized	('C', 2.318621828168132), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 6317), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06628191689545848)	0.500	0.000	0.761	0.029	0.500	0.800	0.500	0.726

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Dependence Variance	('C', 1.7329457130246644), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5970), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09357822552748749)	0.584	0.041	0.773	0.067	0.558	0.650	0.533	0.722
first order Median	('C', 1.8627553518641826), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 520), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.025922873784291426)	0.589	0.156	0.807	0.109	0.500	0.817	0.517	0.722

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm RunPercentage	('C', 2.5962089176715737), ('dual', False), ('fit intercept', False), ('intercept scaling', 5), ('max iter', 3757), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09372292284948522)	0.500	0.000	0.749	0.043	0.500	0.750	0.500	0.722
glrlm ShortRunEmphasis	('C', 1.9315098666459505), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 263), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05482878855058506)	0.500	0.000	0.744	0.042	0.500	0.733	0.500	0.722

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm HighGray Level Emphasis	('C', 3.313068378779602), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5070), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.031126375369455256)	0.617	0.083	0.767	0.059	0.500	0.667	0.687	0.719
glrlm LongRunHighGray Level Emphasis	('C', 1.2987071459092818), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 3215), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06962530794166485)	0.578	0.083	0.751	0.067	0.500	0.567	0.649	0.699

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Inverse-Variance	('C', 3.9904752621068065), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 9062), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00574814302948315)	0.648	0.123	0.838	0.040	0.700	0.833	0.594	0.699
first order Skewness	('C', 3.0261540451098115), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 8735), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07964719755117543)	0.500	0.000	0.600	0.145	0.500	0.900	0.500	0.699

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Autocorrelation	('C', 1.020476756565701), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 1829), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0005161803109579847)	0.578	0.083	0.762	0.050	0.500	0.583	0.610	0.696
shape Surface Area	('C', 1.281478434716215), ('dual', False), ('fit intercept', False), ('intercept scaling', 5), ('max iter', 7142), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07116623351470397)	0.500	0.000	0.583	0.202	0.500	0.783	0.500	0.692

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm SumAverage	('C', 3.4988989501982863), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.578	0.089	0.751	0.063	0.500	0.583	0.632	0.686
glcm JointAverage	('C', 3.86267464731058), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 4569), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03725735171503304)	0.553	0.057	0.751	0.063	0.500	0.583	0.538	0.686
first order 10Percentile	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 10000), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.589	0.110	0.801	0.119	0.558	0.817	0.594	0.672

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm Large Area High-Gray Level Emphasis	('C', 3.737622435453167), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 1144), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08813865923597523)	0.489	0.019	0.507	0.239	0.500	0.533	0.500	0.642
shape Sphericity	('C', 2.0147735877220985), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 6111), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08224642113319439)	0.500	0.000	0.727	0.186	0.500	0.567	0.500	0.635

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
first order Kurtosis	('C', 2.354557796341528), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 9519), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06130108266788335)	0.503	0.066	0.673	0.070	0.500	0.667	0.473	0.629
glszm Gray Level NonUniformity	('C', 2.027525284016723), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.578	0.066	0.671	0.156	0.517	0.767	0.538	0.619

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape Elongation	('C', 2.412468187953435), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 4739), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.044882924744715884)	0.500	0.000	0.554	0.209	0.500	0.767	0.500	0.619
gldm Large Dependence HighGray Level Emphasis	('C', 2.5858057655141624), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 6243), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02957702360812368)	0.500	0.000	0.508	0.070	0.500	0.350	0.500	0.548

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape Flatness	('C', 1.4155060575255303), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 2147), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.02064999828200791)	0.500	0.000	0.564	0.134	0.500	0.683	0.500	0.545
first order Minimum	('C', 3.2236623870291203), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 2753), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03554504012980045)	0.500	0.000	0.614	0.051	0.500	0.750	0.500	0.545

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Cluster-Shade	('C', 3.5323639824016), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 8397), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.00582326313270637)	0.500	0.000	0.697	0.143	0.458	0.600	0.490	0.528
shape Maximum 2D Diameter Slice	('C', 4.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 10000), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.525	0.044	0.641	0.105	0.500	0.167	0.538	0.515

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape MinorAxis Length	('C', 3.0935459192232453), ('dual', False), ('fit intercept', False), ('intercept scaling', 2), ('max iter', 2978), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07429424849579275)	0.500	0.000	0.501	0.193	0.500	0.950	0.500	0.515
glcm JointEnergy	('C', 1.119566634948437), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 3261), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.06197017115976267)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Maximum Probability	('C', 3.7908176819826496), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 2421), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08030721108621787)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500
glrlm Gray Level NonUniformityNormalized	('C', 2.6738260217931975), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 5965), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.05221354580348003)	0.500	0.000	0.572	0.127	0.500	0.500	0.500	0.500

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm Low Gray Level RunEmphasis	('C', 2.1527563581342966), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 9216), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08861477857658649)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500
glrlm Short-RunLow Gray Level Emphasis	('C', 2.735333087682047), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 3344), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.01911736349328676)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Low Gray Level Emphasis	('C', 1.6636850358454844), ('dual', False), ('fit intercept', True), ('intercept scaling', 1), ('max iter', 3588), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03192994381506042)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500
gldm Small-Dependence Low Gray Level Emphasis	('C', 3.547587586047396), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 1048), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.055716091258290204)	0.500	0.000	0.500	0.000	0.500	0.500	0.500	0.500

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Idmn	('C', 2.623710119932979), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5830), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.012547567344942752)	0.500	0.000	0.404	0.108	0.500	0.383	0.500	0.492
glcm Idn	('C', 3.321449344055166), ('dual', False), ('fit intercept', True), ('intercept scaling', 2), ('max iter', 8774), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.016655547049025523)	0.500	0.000	0.365	0.030	0.500	0.433	0.500	0.485

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Cluster-Prominence	('C', 1.5365478573415725), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 8398), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.09115728001329215)	0.500	0.000	0.668	0.177	0.500	0.550	0.500	0.465
shape Maximum 3DDiameter	('C', 1.0532486949437028), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 2146), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1.917278642786953e-05)	0.584	0.080	0.632	0.062	0.458	0.583	0.495	0.465

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape Least Axis Length	('C', 1.433233290480333), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 2288), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.013509930881730366)	0.500	0.000	0.557	0.212	0.500	0.750	0.500	0.462
shape Maximum 2D Diameter Column	('C', 1.0), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 10), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 1e-07)	0.520	0.035	0.574	0.122	0.458	0.650	0.478	0.421

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
shape Maximum 2D Diameter Row	('C', 2.2244780607035213), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 6179), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0626244749905593)	0.564	0.049	0.694	0.095	0.458	0.567	0.533	0.395
shape MajorAxis Length	('C', 1.295302605548843), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 2932), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.049623324401062055)	0.500	0.000	0.296	0.185	0.500	0.517	0.500	0.378

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm Gray Level NonUniformity	('C', 3.7507175756847277), ('dual', False), ('fit intercept', False), ('intercept scaling', 4), ('max iter', 9736), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.022530046847114297)	0.489	0.019	0.319	0.093	0.500	0.533	0.500	0.375
glszm Large Area Low Gray Level Emphasis	('C', 1.3671707821025645), ('dual', False), ('fit intercept', False), ('intercept scaling', 1), ('max iter', 9096), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08828288882515187)	0.500	0.000	0.234	0.131	0.500	0.233	0.485	0.328

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm Large Area Emphasis	('C', 2.777304232871118), ('dual', False), ('fit intercept', False), ('intercept scaling', 1), ('max iter', 7821), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03171048642863875)	0.500	0.000	0.201	0.161	0.500	0.250	0.485	0.311
glszm Zone Variance	('C', 1.1529793554402141), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5119), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.045052159029819086)	0.500	0.000	0.206	0.162	0.500	0.250	0.485	0.311

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
gldm Gray Level NonUniformity	('C', 3.8635854671811907), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 684), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.0455194511749416)	0.523	0.049	0.336	0.052	0.458	0.567	0.457	0.301
glrlm LongRunLow Gray Level Emphasis	('C', 3.407163375733295), ('dual', False), ('fit intercept', True), ('intercept scaling', 5), ('max iter', 8564), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08112240953500788)	0.500	0.000	0.343	0.216	0.500	0.217	0.500	0.278

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glcm Id	('C', 1.6682209042923852), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 5383), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.04803659518559741)	0.500	0.000	0.173	0.066	0.500	0.133	0.500	0.234
glszm Low Gray Level Zone Emphasis	('C', 2.0214023937431733), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 2769), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.07729620982245787)	0.500	0.000	0.269	0.255	0.500	0.117	0.500	0.234

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glszm SmallArea Low Gray Level Emphasis	('C', 1.5455932030988664), ('dual', False), ('fit intercept', True), ('intercept scaling', 4), ('max iter', 5448), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.08757151157263013)	0.500	0.000	0.263	0.155	0.500	0.183	0.500	0.211
shape Surface VolumeRatio	('C', 1.8107979056026746), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 9606), ('penalty', 'l1'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03554878545439961)	0.500	0.000	0.305	0.179	0.500	0.350	0.500	0.207

Feature	ML Parameters	ACC Train	ACC CI	AUC Train	AUC CI	ACC Test	AUC Test	ACC Val	AUC Val
glrlm LongRunEmphasis	('C', 1.6637332212875222), ('dual', False), ('fit intercept', False), ('intercept scaling', 3), ('max iter', 7232), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.003529302207978712)	0.500	0.000	0.200	0.060	0.500	0.317	0.500	0.197
glszm Gray Level NonUniformityNormalized	('C', 1.7955217151831298), ('dual', False), ('fit intercept', True), ('intercept scaling', 3), ('max iter', 6159), ('penalty', 'l2'), ('random state', 1), ('solver', 'liblinear'), ('tol', 0.03730938139661626)	0.500	0.000	0.323	0.259	0.500	0.167	0.500	0.137