

Predicting limb loss in patients with diabetic foot disease on magnetic resonance imaging using deep learning models

A thesis submitted to Imperial College London for the degree of Doctor of Philosophy by

Manal Ahmad

Section of Vascular Surgery
Department of Surgery and Cancer
Imperial College London

Supervisors

Professor Alun Huw Davies

*Professor of Vascular Surgery & NIHR Senior Clinical Investigator
Academic Section of Vascular Surgery, Department of Surgery and Cancer,
Imperial College London & Imperial College NHS Healthcare Trust*

Professor Andrea Rockall

*Clinical Chair of Radiology, Section of Radiology, Department of Surgery and Cancer,
Imperial College London
Honorary Consultant Radiologist, Imperial College NHS Healthcare Trust*

Mr. Joseph Shalhoub

*Consultant Vascular Surgeon, Imperial College NHS Healthcare Trust
Honorary Clinical Senior Lecturer, Department of Surgery and Cancer,
Imperial College London*

I. Statement of originality

The material presented in this thesis is the original work of the author. Any information, results or figures reproduced from the work of others have the original source stated, and permission has been sought. The use of generative artificial intelligence (AI) has been utilised as part of the thesis work.

II. Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

III. Acknowledgements

I would like to dedicate this to my parents, Komal and Suhail, and my sister, Basma. Thank you for being a constant source of unwavering support, for believing in me, and for pushing me to be the best version of myself. I hope I can make you proud one day.

This PhD would not have been possible without the incredible mentorship and guidance from Professor Alun Davies and Professor Andrea Rockall. I am so grateful for the opportunity to explore the world of academia whilst being able to pursue my side quests. Despite my multitudes of questions and learning curves, for their belief that I had potential to do all of this. Their guidance, support, and encouragement — from academic direction to countless discussions about both research and life, were invaluable. I would also like to thank Mr Joseph Shalhoub for teaching me so much, both clinically and academically, and for always taking the time to listen and offer thoughtful guidance. I am also grateful to Dr Dimitri Amiras for his advice and support throughout this project.

I never truly appreciated how many people it takes to complete a PhD until I embarked on my own journey. To name a few, I would like to thank Susana Murphy and Cheuk Fung Wong from the Research, Governance and Integrity team; Lesley Honeyfield, Alastair Jackson, and Liam Greenshields from the Imaging Research Department and, Mohammad Al Sad and Xingfeng Li from the Imaging Computational team.

I am also thankful to the entire vascular team at Imperial for their patience, kindness and teaching over the years. I feel incredibly blessed for all the wonderful friends I have made along the way. Last, but by no means least, I would like to thank the patients, without whom this work would not have been possible.

IV. Abstract

Introduction

Diabetic foot disease (DFD) is a complex disease and is associated with lower limb amputation. Magnetic resonance imaging (MRI) is commonly used in patients with DFD. Early detection of surrogate markers on MRI may help to predict the risk of limb loss.

Aim

To explore the use of deep learning models to identify potential predictors on MRI for limb loss in patients with DFD.

Methods

Four methodological approaches were undertaken:

1. 419 MRI reports were retrospectively reviewed to identify salient imaging features, including collections, ulceration, osteomyelitis, diabetic myopathy/sarcopenia, and Charcot foot. Multiple deep learning models were applied to predict amputation outcomes.
2. Two-dimensional MRI images of the foot at the base of the first metatarsal were classified as mild, moderate, or severe sarcopenia. 824 three-dimensional MRI scans images were manually annotated, with inter-rater reliability assessed on a subset of 50 images. Following preprocessing and augmentation, 1740 images were used to train Vision Transformer and ResNet architectures.
3. MRI scans were manually segmented and modelled using nnU-Net, with a pseudo fat fraction calculated from muscle and plantar fat signal intensities.
4. Sarcopenia grading and pseudo fat fraction were combined with clinical metadata to identify predictors for limb loss and death, and composite outcomes.

Results

XGBoost was the only statistically significant report-based model, achieving 63.4% accuracy ($p < 0.001$). Vision Transformer and ResNet-50/101 achieved accuracies of up to 80.7% for sarcopenia classification, with good inter-rater agreement ($\kappa = 0.8$). nnU-Net segmentation had a mean DICE score 0.4. Sarcopenia severity was not a significant predictor of major amputation or death; however, mild sarcopenia was associated with the composite outcome of all amputations and death (OR 1.368, $p = 0.0377$).

Conclusion

These findings suggest that while sarcopenia severity alone is not a robust surrogate marker for limb loss in DFD, MRI-derived characteristics combined with clinical variables may improve risk stratification.

[297 words]

V. Table of Contents

I.	Statement of originality	2
II.	Copyright Declaration	2
III.	Acknowledgements	3
IV.	Abstract	4
V.	Table of Contents	5
VI.	List of Tables	9
VII.	List of Figures	9
VIII.	Abbreviations	15
Chapter 1: Introduction		17
1.1	<i>Diabetes</i>	17
1.1.1	History of Diabetes	17
1.1.2	Definition of Diabetes	18
1.1.3	Diagnosis of Diabetes	19
1.1.4	Epidemiology of Diabetes	19
1.2	<i>Complications of Diabetes</i>	20
1.3	<i>Diabetic Foot Disease</i>	24
1.3.1	Anatomy of the foot	24
1.3.2	Intrinsic and extrinsic muscles of the foot	24
1.3.3	Plantar fat pad	28
1.3.4	Pathophysiology of Diabetic Foot Disease	28
1.3.5	Risk assessment	31
1.3.6	Diabetic foot ulcer diagnosis and classification	33
1.4	<i>Management of Diabetic Foot Disease</i>	37
1.5	<i>Amputations</i>	39
1.5.1	Impact of amputations	40
1.5.2	Diabetes and amputations	40
1.5.3	Morbidity and mortality associated with amputations	41
1.5.4	Amputations and litigation	41
1.6	<i>Cost of diabetes and diabetes related complications</i>	41
1.7	<i>Risk prediction</i>	42
1.7.1	Classification and risk prediction models	43
1.7.2	Predicting the risk of lower limb amputation	44
1.8	<i>Imaging</i>	46
1.8.1	Magnetic Resonance Imaging (MRI)	47
1.8.2	Imaging the Foot and Ankle	48
1.8.3	MRI in Diabetic Foot Disease	49
1.8.4	Cost of MRIs	51
1.9	<i>Sarcopenia</i>	52

1.9.1 Impact of Sarcopenia	52
1.9.2 Diagnosing Sarcopenia	53
1.9.3 Diabetes and Sarcopenia	56
1.9.4 Sarcopenia and diabetes related complications	56
1.10 Sarcopenia in the foot on magnetic resonance imaging in patients with diabetes	58
1.10.1 Stereological total volume measurement	61
1.10.2 Soft tissue compositional analysis	62
1.10.3 Segmentation based volume measurements	62
1.10.4 Five-point atrophy scale	63
1.10.5 Goutallier Grading	63
1.11 Artificial Intelligence	63
1.11.1 Machine Learning Models	64
1.11.2 Deep Learning Models	64
1.11.3 Radiomics and AI	65
1.11.4 Evaluating the performance of convolutional neural networks	66
1.11.5 Challenges in the use of artificial intelligence for research	69
1.12 Research priority and rationale for this study	70
1.13 Aims of this thesis	71
Chapter 2: Methodology	73
2.1 Introduction	73
2.2 Patient and Public Involvement (PPI)	73
2.2.1 Survey questionnaire	74
2.2.2 Open forum discussion	76
2.2.3 Diabetic Foot clinics	77
2.2.4 Emerging themes from PPI	78
2.3 Power calculation and sample size	78
2.4 External peer review	79
2.5 Sponsorship and intellectual property	80
2.6 Ethics approval	80
2.7 Consent	80
2.8 Inclusion criteria	80
2.9 Exclusion criteria	81
2.10 Site Selection	81
2.11 Patient identification	81
2.11.1 MRI identification and storage	81
2.11.2 Patient metadata	83
2.12 Data storage	86
2.13 Data cleaning	87
2.13.1 Metadata cleaning	87
2.13.2 Segmentations and cleaning the imaging data	88
2.14 Dataset	89
Chapter 3: Risk prediction using MRI reports	91
3.1 Introduction	91
3.1.1 Random Forest	91
3.1.2 Extreme Gradient Boosting	94
3.1.3 Multilayer perceptron model	95

3.2 Aim	96
3.3 Methodology	96
3.4 Summary of MRI reports	97
3.4.1 Feature presence	97
3.4.2 Summary of outcomes	98
3.4.3 Kruskal-Wallis Test	99
3.5 Logistic Regression	100
3.6 Random Forest	103
3.6.1 Method	103
3.6.2 Results	104
3.7 Extreme Gradient Boosting	109
3.7.1 Method	109
3.7.2 Results	110
3.8 Multilayer Perceptron	114
3.8.1 Method	114
3.8.2 Results	114
3.9 Composite Analysis	116
3.9.1 Data re-labelling	117
3.9.2 No amputation versus all amputation group	117
3.9.2.1 Random Forest (no amputation versus all amputation group)	117
3.9.2.2 Extreme Gradient Boosting (no amputation versus all amputation group)	120
3.9.2.3 Multilayer Perceptron (no amputation versus all amputation group)	122
3.9.3 Major lower limb amputation versus death	124
3.9.3.1 Random Forest (major lower limb amputation versus death)	124
3.9.3.2 Extreme Gradient Boosting (major lower limb amputation versus death)	127
3.9.3.3 Multilayer Perceptron (major lower limb amputation versus death)	129
3.9.4 All amputations versus death	131
3.9.4.1 Random Forest (major lower limb amputation versus death)	131
3.9.4.2 Extreme Gradient Boosting (major lower limb amputation versus death)	134
3.9.4.3 Multilayer Perceptron (major lower limb amputation versus death)	137
3.9.5 No amputation versus major lower limb amputation and death	139
3.9.5.1 Random Forest (no amputation versus major lower limb amputation and death)	139
3.9.5.2 Extreme Gradient Boosting (no amputation versus major lower limb amputation and death)	143
3.9.5.3 Multilayer Perceptron (no amputation versus MLLA and death combined)	145
3.9.6 No amputation versus all amputation and death	147
3.9.6.1 Random Forest (no amputation versus all amputation and death)	147
3.9.6.2 Extreme Gradient Boosting (no amputation versus all amputation and death)	151
3.9.6.3 Multilayer Perceptron (no amputation versus all amputations and death combined)	153
3.10 Discussion	156
Chapter 4: Applying AI models to MRI images of the foot	160
4.1 Introduction	160
4.1.1 Vision Transformer (ViT)	160
4.1.2 Convolutional Neural Networks (CNN)	164
4.1.3 Residual Network (ResNet)	171
4.1.4 U-Net and no-new-U-Net (nnU-Net)	174
4.2 Aims	177
4.3 Methodology	177
4.3.1 Reliability and congruence in measurements	177
4.3.2 Methodology - ViT and ResNet	180
4.3.2.1 Creating the dataset	180
4.3.2.2 Data pre-processing and labelling	181

4.3.2.3 Data Augmentation	182
4.3.2.4 Data split and training the models	183
4.3.2.5 Results	184
4.3.2.5.1 Results [ViT]	186
4.3.2.5.1 Results [ResNet]	190
4.3.3 Methodology - nnU-Net	198
4.3.3.1 Segmentation rules	199
4.3.3.2 Segmentation and checking	200
4.3.3.4 Running nnU-Net	201
4.3.3.5 Results [nnU-Net]	203
4.4 Discussion	209
Chapter 5: Summary of patient metadata	211
5.1 Patient characteristics	211
5.2 Diabetes related and other co-morbidities	212
5.2.1 Diabetes-related	212
5.2.2 Cardiovascular disease	213
5.2.3 Cerebrovascular disease	213
5.2.4 Respiratory disease	213
5.2.5 Other medical and social conditions	213
5.3 Peripheral arterial disease	215
5.4 Biochemical markers	216
5.5 MRI related and outcomes	217
Chapter 6: Risk prediction	219
6.1 Sarcopenia measurements and outcomes	219
6.1.1 Sarcopenia and all amputations as an outcome	219
6.1.2 Composite outcomes: sarcopenia severity and major lower limb amputations plus death	220
6.1.3 Composite outcomes: sarcopenia severity and all amputations and death	221
6.1.4 Fat percentage values and all amputations as an outcome	222
6.1.5 Composite outcomes: pseudo fat fraction values and major lower limb amputations plus death	224
6.1.5 Composite outcomes: pseudo fat fraction values and all amputations and death	225
6.2 Identifying risk factors	225
6.2.1 Lasso regression for selecting salient predictors for LLAs	226
6.2.2 Ridge regression for selecting salient predictors for LLAs	228
6.2.3 Random Forest for selecting salient predictors for LLAs	229
6.2.4 Extreme gradient boosting for selecting salient predictors for LLAs	232
6.2.5 LASSO and ridge regression for composite outcomes – (MLLA + death) and (all amputations + death)	232
6.2.6 Random Forest for composite outcomes – MLLA + Death and All amputations + Death	234
6.3 Estimating survival	237
6.3.1 Univariate and Multivariate Cox proportional hazard	237
6.4 Discussion	241
Chapter 7: Final Discussion and conclusion	243
7.1 Introduction (to the end)	243
7.2 Summary of results and reflection	244
7.3 Limitations	247
7.4 Real-world implications and scope for future work	248
7.5 Conclusion	248

References	249
Appendix	266
Appendix 1 - R Code for Kruskal-Wallis Test	266
Appendix 2 – Setting up Jupyter notebook workspace and downloading Python and Sci-kit learn + running logistic regression	267
Appendix 3 - R Code for Random Forest	272
Appendix 4 - R Code for Extreme Gradient Boosting	275
Appendix 5 - R Code for intraclass correlation and Bland-Altman plot	278
Appendix 6 – R Code for analysing the nnU-Net output	279
Appendix 7 – nnU-NET output	281
Appendix 8 – R code for logistic regression and sarcopenia severity	284
Appendix 9 – R code for logistic regression and pseudo fat fraction	285
Appendix 10 – R code for feature importance with patient metadata using LASSO, Ridge regression and Random Forest	287
Appendix 11 – R code for Cox proportional hazard	290
Appendix 12 – PhD related funding	290
Appendix 13 – Publications during the PhD	291
Directly related to the PhD	291
Other publications during the PhD	291
Appendix 14 – Oral presentations from PhD related work	292
Appendix 15 – Poster presentations from PhD related work	294

VI. List of Tables

Table 1: Extrinsic muscles of the foot and their functions. Adapted from Ahmad et al (2024)	27
Table 2: Intrinsic muscles of the foot and their functions. Adapted from Ahmad et al (2024)	28
Table 3: Risk stratification of developing DFU or requiring an amputation (Adapted from UK NICE Guidelines on Diabetic Foot Problems: diagnosis and management (2015) (37))	32
Table 4: IWGDF Risk stratification and screening (Adapted from IWGDF Guidelines on prevention and management of Diabetic Foot Disease (2023))	33
Table 5: SINBAD Classification system (Adapted from the IWGDF 2023 Classification Guidelines (37))	34
Table 6: WiFi Classification system – Wound description; TMA: Transmetatarsal Amputation (Adapted from the IWGDF 2023 Classification Guidelines (37))	36
Table 7: WiFi Classification system – Ischaemia; TcPO ₂ : Transcutaneous Partial Oxygen Pressure; TP: Toe Pressure; mmHg – millimeters of mercury (Adapted from the IWGDF 2023 Classification Guidelines (37))	36
Table 8: WiFi Classification system – Infection; ABI: Ankle-Brachial Index; PaCO ₂ : Partial Pressure of Carbon Dioxide; SIRS: Systemic Inflammatory Response Syndrome; TcPO ₂ : Transcutaneous Partial Oxygen Pressure; TP: Toe Pressure (Adapted from the IWGDF 2023 Classification Guidelines (37))	37
Table 9: Summary of anatomical amputation levels - adapted from Van Netten 2024 (19)	40
Table 10: MRI costs	52
Table 11: Summary of diagnostic criteria for sarcopenia. Adapted from Voulgaridou et al (65). 5TSST – 5 time stand-to-sit test; 6MGS – 6 m gait speed; ALM – Appendicular lean mass; ASM – Appendicular skeletal mass; BIA – Bioimpedance Analysis; BMI – Body mass index; CT – Computed tomography; DXA – Dual-energy X-ray absorptiometry; Kg – Kilogram; m – Meter; MRI – Magnetic resonance imaging; s – seconds; SCTP – Stair climb power test; SPPB – Short Physical Performance Battery; TUG – Time Up and Go	55
Table 12: Summary of sarcopenia on MRI in the foot. Adapted from Ahmad et al (29)	61

Table 13: Responses from the PPI survey questionnaire	75
Table 14: Categories and details of patient metadata collected	86
Table 15: Summary of MRI reported features	97
Table 16: Kruskal-Wallis results for MRI reported features on outcomes.	100
Table 17: Logistic Regression for MRI reports	102
Table 18: Random Forest training set confusion matrix (class 0 = no amputation, class 1=minor amputation, class 2= major amputation)	105
Table 19: Random Forest test set confusion matrix (class 0 = no amputation, class 1=minor amputation, class 2= major amputation)	105
Table 20: Random Forest performance table	107
Table 21: Feature importance (Random Forest)	107
Table 22: XGB Hyperparameters used (Round 1)	110
Table 23: XGB hyperparameters used (improved accuracy)	111
Table 24: XGB confusion matrix for the test set. Class 0 - no amputation, Class I- minor amputation and Class II – major amputation.	111
Table 25: XGB feature selection; OM – osteomyelitis, DMS – diabetic myopathy and sarcopenia.	112
Table 26: Confusion matrix for Multilayer Perceptron Model testing set	116
Table 27: Random Forest training set confusion matrix (class 0 = no amputation, class 1= all amputation)	118
Table 28: Random Forest test set confusion matrix (class 0 = no amputation, class 1= all amputation)	118
Table 29: RaF performance table (no amputation versus all amputation group)	119
Table 30: Extreme gradient boosting - confusion matrix. No amputation versus all amputation group.	121
Table 31: Extreme gradient boosting - feature selection. No amputation versus all amputation group. OM- osteomyelitis	121
Table 32: Multilayer perceptron confusion matrix (no amputation versus all amputations)	123
Table 33: Random Forest training set confusion matrix. Major lower limb amputation (MLLA) and death	125
Table 34: Random Forest test set confusion matrix	125
Table 35: Random Forest performance table (major lower limb amputation versus death)	126
Table 36: Extreme gradient boosting confusion matrix of the test set. Major lower limb amputation (MLLA) versus death.	127
Table 37: Extreme gradient boosting feature selection. Major lower limb amputation versus death. OM- osteomyelitis.	127
Table 38: MLP confusion matrix. Major lower limb amputation (MLLA) versus Death.	129
Table 39: Random Forest training set confusion matrix (all amputation versus death)	132
Table 40: Random Forest test set confusion matrix (all amputations versus death)	132
Table 41: Random Forest performance table (all amputations versus death)	133
Table 42: Random Forest feature importance (all amputations versus death)	134
Table 43: Extreme Gradient Boosting. Test set confusion matrix. All amputations versus death.	135
Table 44: Random Forest performance table. All amputations versus death.	136
Table 45: Extreme gradient boosting. Feature importance. All amputation versus death; OM- Osteomyelitis, DMS- Diabetic Myopathy and Sarcopenia.	136
Table 46: Multilayer Perceptron. Test set confusion matrix. All amputations versus death.	138
Table 47: Random Forest training set confusion matrix (class 0 = no amputation, class 1= MLLA and death).	140
Table 48: Random Forest test set confusion matrix (class 0 = no amputation, class 1= MLLA and death).	140
Table 49: RaF performance table (no amputation versus MLLA and death combined).	141
Table 50: RaF feature importance – Class 0 (no amputation), Class 1 (MLLA and death combined).	142
Table 51: Extreme gradient boosting - confusion matrix. No amputation versus MLLA and death combined.	144
Table 52: Extreme gradient boosting - feature selection. No amputation versus MLLA and death combined. OM – osteomyelitis, DMS – diabetic myopathy/ sarcopenia.	144
Table 53: Multilayer perceptron confusion matrix (no amputation versus MLLA and death combined).	146
Table 54: Random Forest training set confusion matrix (class 0 = no amputation, class 1= all amputation and death).	148
Table 55: Random Forest test set confusion matrix (class 0 = no amputation, class 1= All amputations and death).	148
Table 56: RaF performance table (no amputation versus all amputations and death combined).	149
Table 57: RaF feature importance – Class 0 (no amputation), Class 1 (all amputation and death combined).	150
Table 58: Extreme gradient boosting - confusion matrix. No amputation versus all amputation and death combined.	152

Table 59: Extreme gradient boosting - feature selection. No amputation versus all amputation and death combined. OM – osteomyelitis, DMS – diabetic myopathy/ sarcopenia. _____	152
Table 60: Multilayer perceptron confusion matrix (no amputation versus all amputation and death combined). _____	155
Table 61: Summary of applied models and results; Amp- Amputation, AUC – Area under the curve, DMS – Diabetic myopathy/Sarcopenia, MLLA – major lower limb amputation, MLP – Multilayer perceptron, OM – osteomyelitis, Raf – Random Forest, v – versus, XGB – Extreme Gradient Boosting, ** - statistically significant. _____	158
Table 62: Summary of ResNet versions. Partially adapted from He et al [168]. _____	174
Table 63: Data pre-processing steps. _____	182
Table 64: Vision Transformer (ViT) summary of model metrics. _____	187
Table 65: Vision Transformer (ViT) confusion matrix of the test set. _____	190
Table 66: ResNet-18, ResNet-34, ResNet-50, ResNet-101 summary of model metrics. _____	191
Table 67: Test set confusion matrix ResNet-18 _____	196
Table 68: Test set confusion matrix ResNet-34. _____	196
Table 69: Test set confusion matrix ResNet-50. _____	197
Table 70: Test set confusion matrix ResNet-101. _____	197
Table 71: nnU-Net test set performance output. _____	205
Table 72: Patient characteristics (Table A). _____	212
Table 73: Patient characteristics (Table B). Kg – kilograms. _____	212
Table 74: Patient characteristics - diabetes related. CKD – chronic kidney disease, PCI – percutaneous coronary intervention. _____	215
Table 75: Patient characteristics - peripheral arterial disease. CFA – common femoral artery, PFA – profunda femoris artery, SFA – superficial femoral artery. _____	216
Table 76: Patient characteristics - biochemical markers. _____	216
Table 77: Patient characteristics - MRI/episode related. _____	217
Table 78: Patient outcomes. TMT – transmetatarsal, BKA – Below knee amputation, AKA – Above knee amputation. _____	218
Table 79: Random Forest feature importance incorporating patient metadata. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction. _____	230
Table 80: Ridge regression feature selection. Major lower limb amputation and death combined. CKD – chronic kidney disease, CVA – cerebrovascular accident, PD – peritoneal dialysis, TIA – transient ischaemic attack. The minus sign indicates it decreased the likelihood of the outcome. _____	233
Table 81: Ridge regression feature selection. All amputation and death combined. CKD – chronic kidney disease, CVA – cerebrovascular accident, CVD – cardiovascular disease, HHD- handheld doppler, PD – peritoneal dialysis, TIA – transient ischaemic attack. The minus sign indicates it decreased the likelihood of the outcome. _____	234
Table 82: RaF feature importance MLLA and Death. HbA1c – glycated haemoglobin, Hb – haemoglobin, CRP- C-reactive protein, CVA – cerebrovascular accident, PAD – peripheral arterial disease. _____	236
Table 83: Cox proportional hazard model output and hazard ratios for risk of amputation (minor and major combined). Statistically significant results in bold. HR – Hazard ratio, HbA1c – glycated haemoglobin, CRP – C-reactive protein, PAD- peripheral arterial disease, SE(CE) – standard error of the coefficient/log hazard ratio. _____	239

VII. List of Figures

Figure 1: Complications of Diabetes.....	20
Figure 2: Pathophysiological processes implicated in diabetes related complications.....	21
Figure 3: Osteal anatomy of the foot	25
Figure 4: a – Plantar view of the 1st layer of intrinsic muscles (flexor digitorum brevis, abductor hallucis, abductor digiti minimi); b – Plantar view of the 2nd layer of intrinsic muscles (adductor hallucis [oblique and transverse head], flexor hallucis brevis.....	25
Figure 5: Medial anatomical view of the foot demonstrating the flexor hallucis longus, extensor hallucis longus and flexor digitorum longus).....	26

Figure 6: Lateral anatomical view of the foot demonstrating the tibialis anterior, extensor digitorum longus, extensor hallucis longus, extensor digitorum brevis, peroneus [longus, brevis and tertius]. (From Gray's Anatomy (1918) – available in public domain).....	26
Figure 7: Development of diabetic foot disease	29
Figure 8a: Coronal view of foot at the base of the 1st Metatarsal on MRI (normal foot).....	49
Figure 9b: Coronal view of foot at the base of the 1st Metatarsal on MRI (normal foot – annotated).....	50
Figure 10: Coronal view of foot following amputation of the 2nd toe.....	50
Figure 11: Coronal view of the foot demonstrating extensive oedema and fat infiltration in the muscle	51
Figure 12: Schematic representation of a coronal view of the normal foot without sarcopenia. Adapted from Ahmad et al (29).....	58
Figure 13: Schematic representation of a coronal view of the foot with mild and early sarcopenia. Adapted from Ahmad et al (29).....	59
Figure 14: Schematic representation of a coronal view of the foot with moderate sarcopenia. Adapted from Ahmad et al (29).....	59
Figure 15: Schematic representation of a coronal view of the foot with severe sarcopenia. Adapted from Ahmad et al (29).....	60
Figure 16: Illustration of Cavalieri's principle	62
Figure 17: Subtypes of artificial intelligence	64
Figure 18: Calculating the DICE score for segmentations.....	68
Figure 19: Calculating the Jaccard index.....	69
Figure 20: XNAT Directory for the project with pseudonymised scan folders	83
Figure 21: Schematic representation of a pixel versus a voxel	88
Figure 22: CONSORT diagram of final dataset	90
Figure 23: Structure of a decision tree. Adapted from IBM [130].	92
Figure 24: Schematic representation of Random Forest. The purple arrows demonstrate the pathway the DT uses to reach the decision. The dark blue circles demonstrate the options chosen at each point in the decision-making process. The light blue circles indicate the decision which was not chosen. Adapted from IBM [130]... ..	93
Figure 25: Schematic representation of Extreme Gradient Boosting structure. The purple arrows demonstrate the pathway the DT uses to reach the decision. The dark blue circles demonstrate the options chosen at each point in the decision-making process. The light blue circles indicate the decision which was not chosen. Adapted from Kavlakoglu [136].	94
Figure 26: Structure of a multilayer perceptron model. Adapted from Krizhevsky et al [140].	95
Figure 27: Anatomical distribution of Osteomyelitis as per MRI reports	98
Figure 28: Bar chart demonstrating the corresponding outcomes for all included cases.	98
Figure 29: Anaconda Navigator desktop main page with Jupyter notebook highlighted in red.....	101
Figure 30: Visual representation of logistic regression coefficients	102
Figure 31: Confusion Matrix.....	104
Figure 32: Random Forest mean decrease in accuracy and mean decrease in Gini.....	108
Figure 33: Random Forest ROC-AUC.....	108
Figure 34: XGB visual chart demonstrating feature importance; OM – osteomyelitis, DMS – diabetic myopathy and sarcopenia.	113
Figure 35: ROC-AUC for Extreme Gradient Boosting	113
Figure 36: MLP output. This is a visual representation of the MLP. The coloured boxed on the left indicate the predictors. The grey ovals in the centre show the number of hidden layers. The boxes on the right are the outcome classes. The lines in blue and grey show the connections to each hidden layer. Synaptic weights which are indicative of the strength of a connection i.e. greater than or less than zero. It is the equivalent of an action potential between neurons. If it is significant threshold value, it then signals to the next neuron. The blue lines represent a synaptic weight of less than zero. The grey lines represent a synaptic weight greater than zero. Outcome 0 (no amputation), Outcome 1 (minor amputation), Outcome 2 (major amputation).	115
Figure 37: ROC AUC Multilayer Perceptron	116
Figure 38: Random Forest ROC-AUC. No amputation versus amputation.....	120
Figure 39: Extreme Gradient Boosting - Feature importance. No amputation versus All amputation group; OM- Osteomyelitis.....	121
Figure 40: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1(all amputations)	122
Figure 41: Multilayer perceptron output (no amputation versus all amputation). Class 0 (no amputation), Class 1 (all amputations).....	123
Figure 42: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus all amputations (class1)	124

Figure 43: Random Forest ROC-AUC. Major lower limb amputation versus death - Class 0 (major lower limb amputation), Class 1 (death).....	126
Figure 44: Extreme Gradient Boosting feature importance. Major lower limb amputation versus death; OM- Osteomyelitis.....	128
Figure 45: ROC-AUC Extreme Gradient Boosting. Class 0 (major lower limb amputation), Class 1 (death).	128
Figure 46: MLP output. Major lower limb amputation versus death	130
Figure 47: ROC-AUC for MLP. Class 0 (major lower limb amputation), Class 1 (death).....	131
Figure 48: Random Forest feature importance (all amputations versus death).....	133
Figure 49: Random Forest ROC-AUC. All amputations versus death.	134
Figure 50: Extreme gradient boosting. Feature importance (visual). All amputation versus death; OM- Osteomyelitis, DMS- Diabetic Myopathy and Sarcopenia.	136
Figure 51: Extreme Gradient Boosting ROC-AUC. Class 0 (all amputations), Class 1 (death).	137
Figure 52: Multilayer perceptron output. Class 0 (all amputations) and Class 1 (death)	138
Figure 53: Multilayer perceptron ROC-AUC. Class 0 (all amputations), Class 1 (death)	139
Figure 54: RaF feature importance (no amputation versus MLLA and death combined).	142
Figure 55: Random Forest ROC-AUC. No amputation versus MLLA and death. MLLA- major lower limb amputation.	143
Figure 56: Extreme Gradient Boosting - Feature importance. No amputation versus MLLA and death combined.	144
Figure 57: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1 (MLLA and death combined).	145
Figure 58: Multilayer perceptron output (no amputation versus MLLA and death combined). Class 0 (no amputation), Class 1 (MLLA and death).	146
Figure 59: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus MLLA and death combined (class1).	147
Figure 60: RaF feature importance (no amputation versus all amputation and death combined).	150
Figure 61: Random Forest ROC-AUC. No amputation versus all amputations and death.	151
Figure 62: Extreme Gradient Boosting - Feature importance. No amputation versus all amputation and death combined.	152
Figure 63: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1 (all amputations and death combined).	153
Figure 64: Multilayer perceptron output (no amputation versus all amputation and death combined). Class 0 (no amputation), Class 1 (all amputation and death).	154
Figure 65: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus all amputation and death combined (class1).	155
Figure 66: Transformer model architecture. As printed in the original paper by Vaswani et al [157].	161
Figure 67: Architecture of Vision Transformer. Adapted from Dosovitskiy et al to demonstrate its use for this thesis [159]. CLS - classification token.	163
Figure 68: Imaging data stored as a numerical matrix on a computing device and a kernel/filter. Adapted from IBM to demonstrate its use in this thesis [164].	166
Figure 69: Schematic representation of a convolution. The kernel scans across the entire image to produce a convolution. Adapted from IBM to demonstrate its use in this thesis [164].	167
Figure 70: Feature maps through various convolutions. Adapted from LeCun et al to demonstrate its use for this thesis [162].	168
Figure 71: Schematic representation of a convolutional neural network. Adapted from LeCun et al and IBM to demonstrate its use in this thesis [162, 164].	169
Figure 72: Schematic representation of normal CNN learning (A) versus residual learning/skip connections (B). Adapted from He, Brunton [168, 169].	172
Figure 73: Schematic representation of a ResNet block. Adapted from He, Brunton [168, 169]. Conv – convolution, norm – batch normalisation, ReLU – rectified linear unit.	172
Figure 74: ResNet structural blocks within CNNs. Not adapted.	173
Figure 75: U-Net structure as proposed by Ronneberger et al. As printed in the original publication [165].	176
Figure 76: nnU-Net functions as described by Isensee et al. As printed in the original paper [110].	177
Figure 77: Grading of sarcopenia.	178
Figure 78: Bland-Altman plot. The green and pink lines are indicative of the upper and lower limits of agreement i.e. the standard deviation. The points are within these bands which demonstrate consistent differences within those bounds.	180

Figure 79: Collating the images to prepare the dataset. Full area of the foot was available for segmentation in each included case.	181
Figure 80: Image annotation of each MRI into three categories of sarcopenia severity - mild, moderate or severe.....	182
Figure 81: Data augmentation strategies prior to using ML models.	183
Figure 82: Vision Transformer (ViT) model performance metrics (automated).....	187
Figure 83: Vision Transformer (ViT) model performance metrics (adjusted).	188
Figure 84: Vision Transformer (ViT) graph demonstrating training loss. Number of epochs (x-axis) and the training loss (y-axis).....	188
Figure 85: Vision Transformer (ViT) graph demonstrating validation accuracy. Number of epochs (x-axis) and validation accuracy value (y-axis).	189
Figure 86: ResNet-18 and ResNet-34 model performance metrics (automated and adjusted).....	192
Figure 87: ResNet-50 and ResNet-101 model performance metrics (automated and adjusted).	193
Figure 88: Graphs demonstrating training loss for ResNet-18, ResNet-34, ResNet-50 and ResNet-101. Number of epochs (x-axis) and the training loss (y-axis).	194
Figure 89: Graphs demonstrating the validation accuracy for ResNet-18, ResNet-34, ResNet-50 and ResNet-101. Number of epochs (x-axis) and validation accuracy (y-axis).	195
Figure 90: XNAT Viewer with the mask tool (purple box).	198
Figure 91: XNAT Viewer with the manual segmenting tool.	199
Figure 92: Early segmentation variations.	199
Figure 93: Region of interest segmented in red. The purple box on the top right provides the volume and the teal box in the lower right corner provides the area, mean signal intensity and the standard deviation.	200
Figure 94: XNAT viewer with details of the segmented area. The purple box provides the volume, and the teal box provides the area, mean signal intensity and the standard deviation.	201
Figure 95: Final segmentation of the muscle (in red) and the plantar fat (in green).	201
Figure 96: JSON file for nnU-Net.	202
Figure 97: Image viewer prior to running nnU-Net.	203
Figure 98: Bar chart of number of cases with low, medium and high DICE scores.	206
Figure 99: Scatterplot demonstrating the relationship between precision (x-axis) and recall (y-axis).	207
Figure 100: Scatterplot demonstrating the relationship between DICE scores (x-axis) and precision (y-axis). ..	207
Figure 101: Scatterplot demonstrating the relationship between DICE scores (x-axis) and recall (y-axis).	208
Figure 102: MRI scan variations.	208
Figure 103: Outcomes per sarcopenia severity group. 0 = no amputation, 1= amputation.	220
Figure 104: Outcomes per sarcopenia severity group. 0 = no amputation, 1= major lower limb amputation and death combined.	221
Figure 105: Outcomes per sarcopenia severity group. 0 = no amputation, 1= all amputation and death combined.	222
Figure 106: Logistic regression (Pseudo fat fraction and predicted probability of all amputations).	223
Figure 107: Visual chart demonstrating the pseudo fat fraction values and the outcomes. 0 = no amputation, 1 = all amputations.	224
Figure 108: Logistic regression. Pseudo fat fraction and the probability of major lower limb amputation and death as a combined outcome.	224
Figure 109: Logistic regression. Pseudo fat fraction and the probability of all amputations and death as a combined outcome.	225
Figure 110: Cross validation error with LASSO.	228
Figure 111: Visual chart of Random Forest feature importance incorporating patient metadata. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction.	231
Figure 112: ROC-AUC Random Forest using metadata.	231
Figure 113: Random Forest feature importance. Visual chart of Random Forest feature importance incorporating patient metadata for composite outcomes. Left chart – Major lower limb amputation and death as the outcome. Right chart – all amputations and death as the outcome. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction.	235
Figure 114: Kaplan – Meier curve. Sarcopenia severity (grouped into mild, moderate and severe) and survival.	240
Figure 115: Kaplan- Meier curve - C-reactive protein (divided into tertiles) and survival.	241

VIII. Abbreviations

4MGS	4 m gait speed
5TSST	5-time stand-to-sit test
6MGS	6 m gait speed
ABI	Ankle-Brachial Index
ABPI	Ankle Brachial Pressure Index
AD	Anno Dominius
AGEs	Advanced Glycation End-Products
AI	Artificial Intelligence
AKA	Above Knee Amputation
ALM	Appendicular lean mass
ANN	Artificial Neural Network
ANZSSFR	Australian and New Zealand Society for Sarcopenia and Frailty Research
ASM	Appendicular skeletal mass
AWGS	The Asian Working Group for Sarcopenia
BC	Before Christ
KA	Below Knee Amputation
BMI	Body mass index
CE	Cross-entropy
CI	Confidence Interval
CLS	Classification Token
CONSORT	Consolidated Standards of Reporting Trials
CNN	Convolutional neural network
CRP	C-Reactive Protein
CUDA	Compute Unified Device Architecture
DFD	Diabetic Foot Disease
DFU	Diabetic Foot Ulcer
DIAFORA	Diabetic Foot Risk Assessment
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DT	Decision Tree
DXA	Dual-energy X-ray absorptiometry
ESVS	European Society of Vascular Surgeons
EWGSOP2	European Working Group on Sarcopenia in older people
FNIH	Foundation for the National Institutes of Health Sarcopenia Project
GD	Gestational Diabetes
GPU	Graphics processing unit
HbA1c	Glycated Haemoglobin
ICC	intraclass correlation co-efficient
ICHT	Imperial College Healthcare NHS Trust
IDSA	Infectious Disease Society
IL-6	Interleukin-6
IWGDF	International Working Group on the Diabetic Foot
JSON	JavaScript Object Notation
Kg	Kilogram
LASSO	Least Absolute Shrinkage and Selection Operator

LLA	Lower Limb Amputation
MICE	Multivariate Imputation by Chained Equations
ML	Machine Learning
MLP	Multilayer perceptron
mmol/L	millimoles per litre
mmol/mol	millimoles per mole
MRI	Magnetic Resonance Imaging
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NiftI	Neuroimaging Informatics Technology Initiative
NIHR	National Institute for Health and Care Research
NLRP3	NLR Family Pyrin Domain Containing 3
nnU-Net	No-new-U-Net
PACS	Picture Archiving and Communication System
PAD	Peripheral Arterial Disease
PODUS	Prediction of Diabetic Foot UlcerationS
PaCO₂	Partial Pressure of Carbon Dioxide
RaF	Random Forest
RAM	Random Access Memory
ResNet	Residual Network
RF	Radiofrequency
ROC-AUC	receiver-operating characteristic-area under the curve
ROS	Reactive Oxygen Species
RPM	Risk prediction models
SCTP	Stair climb power test
SIGN	Scottish Intercollegiate Guidelines Network
SINBAD	Site, Ischaemia, Neuropathy, Bacterial Infection, Depth
SIRS	Systemic Inflammatory Response Syndrome
SPPB	Short Physical Performance Battery
SPSS	Statistical package for the social sciences
SVS	Society fo Vascular Surgeons
T1D	Type 1 Diabetes
T2DM	Type 2 Diabetes / Diabetes Mellitus
TE	Echo Time
TGF-β1	Transforming Growth Factor β1
TMT	Transmetatarsal Amputation
TNF-α	Tumour Necrosis Factor Alpha
TP	Toe Pressure
TR	Repetition time
TUG	Time Up and Go
TcPO₂	Transcutaneous Partial Oxygen Pressure
UK	United Kingdom
UTWCS	University of Texas Wound Classification System
WIFI	Wound, Ischaemia and Foot Infection
XGB	Extreme Gradient Boosting
XNAT	eXtensible Neuroimaging Archive Toolkit
ViT	Vision Transformer

Chapter 1: Introduction

1.1 Diabetes

1.1.1 History of Diabetes

The concept of Diabetes dates back several millennia. The Ebers Papyrus, written around 1552 B.C. describes a condition whereby the body needs “to eliminate urine which is too plentiful” [1]. It appears in sanskrit texts from the ancient Indus Valley civilisations who described a group of ailments in individuals presenting with *madhumeha* or “honey urine” and extreme thirst [2]. The ancient Chinese textbook *Huángdì Nèijīng* describes a condition *xiāo kě* or “wasting thirst” characterised by excessive thirst, hunger, urine and weight loss [2, 3].

In classical antiquity, it is, at times, attributed to the now lost texts of Demetrius of Apamea [2]. The term Diabetes appears to have been first coined by Aretaeus of Cappadocia, a Greek physician, around 150 A.D. [4, 5]. The etymological origin of the word arises from the Ionic term *διαβαίνω* (Diabaino) which broadly translates as ‘to siphon’ or ‘run through’. This was characterised by Aretaeus’ observation in those effected as water failed to remain within the body. He described it as a “rare and wonderful” affliction causing incessant thirst, polyuria and emaciation and described the gradual “melting down and passing out of the flesh and limbs along with the water” [5]. These key symptoms are later further described in ancient classical texts by Aulus Cornelius Celsus (25B.C. – 50 A.D.), Rufus of Ephesus (98 – 117 A.D.), Galen (130-201A.D.), Oribasius (320-403 A.D.), Avicenna (980-1037 A.D.) and Paracelsus (1493-1541).

Thomas Willis (1621 -1675) modernised the current understanding of these symptoms in 1674 by describing previously described symptoms in the *Pharmaceutice Rationalis* by suggesting that it was due to a “distemper of the blood” [6]. The role of the pancreas in Diabetes is first attributed to Thomas Cawley who, in 1794, first noted that the pancreas was “hard and scirrhus” on autopsy in a patient with diabetes [7]. In 1890, Joseph Von Mering (1849-1908) and Oskar Minkowski (1858-1931) pioneered the idea that diabetes was

related to removal of the pancreas after they noted dogs developed similar symptoms once they had undergone a total pancreatectomy [4]. It would however, be over another hundred years before the connection between glucose metabolism and the role of the pancreatic beta cells was outlined by John MacLeod in 1913 [8].

Shortly after this, Diabetes and its treatment made major strides culminating in the invention of synthetic insulin and earning Frederick Banting (1891-1941) and John Macleod (1876-1935) the Nobel prize in physiology in 1923 [4, 8]. The history of Diabetes can be divided into the pre- and post- insulin era. Prior to the discovery of Insulin in 1921, Diabetes was often fatal with a high morbidity and mortality associated with it. The advent of insulin and its incorporation into clinical practice has improved the lives of millions over the last century.

1.1.2 Definition of Diabetes

Diabetes is a chronic and complex set of autoimmune, metabolic conditions characterised by high levels of circulating blood sugars known as hyperglycemia as a result of partial or total insulin insufficiency i.e. insulin resistance [9]. Diabetes is further characterised into Type 1 Diabetes (T1DM), Type 2 Diabetes or Diabetes Mellitus (T2DM) and Gestational Diabetes (GD).

T1DM is due to the autoimmune and or inflammatory complete destruction of pancreatic beta islet cells. It is also known as insulin dependent diabetes. T2DM, which is the most common type and accounts for over 90% of all diabetes, is a metabolic disorder with beta cell dysfunction and insulin resistance of varying severity [9, 10]. Gestational diabetes is associated with glucose intolerance at the onset or first recognition of pregnancy [9, 11]. Diabetes can also be caused by viral infections associated with beta cell destruction, certain genetic syndromes such as Prader-Willi and Down's Syndrome and can be drug induced [9, 10].

The global incidence and prevalence of diabetes has grown exponentially over the last fifty years and is reflective of the complex interplay between inherent genetic factors in combination with external influences [12]. These include personal, social and environmental

influences including sedentary lifestyles, poor dietary choices, smoking and obesity surmounted by an overall global aging population [13].

1.1.3 Diagnosis of Diabetes

Diabetes presents with a series of classical signs and symptoms including polyuria, polydipsia, polyphagia, blurred vision, fatigue and weight loss. The current diagnostic criteria for diabetes is defined by the UK National Institute for Health and Care Excellence (NICE) by persistent hyperglycemia and biochemical markers [11]. These include the presence of one or more of the following:

1. Glycated haemoglobin (HbA1c) reading greater than or equal to 48mmol/mol.
2. Fasting plasma glucose of 7.0 mmol/L or more.
3. Random plasma glucose of 11.1 mmol/L or more along with symptoms.
4. Two-hour plasma glucose (following administration of 75g of oral glucose) of 11.1mmol/L or more.

Impaired glucose tolerance, the precursor to diabetes, is defined by a fasting plasma glucose of >7.0mmol/L and a two hour plasma glucose between 7.8-11.1mmol/L [10].

1.1.4 Epidemiology of Diabetes

Recent global population studies report an alarming 589 million adults between the ages of 20-79 were living with diabetes in 2024 [10]. One in four of these are over the age of 65 [10]. This number has more than quadrupled since 1990 [14]. The overall prevalence of diabetes is predicted to rise exponentially to 853 million by 2050 [10]. It is also estimated that a further 252 million are likely to have diabetes but have yet to be diagnosed [10].

In the United Kingdom, the number of patients with diabetes has continued to grow and is reflective of an overall global pattern. A reported 4.3 million patients are living with diabetes, 850,000 are yet to be diagnosed and another 2.4 million are at risk of developing diabetes in the UK [15, 16].

The International Diabetes Federation Diabetes Atlas (11th Edition) reported an almost equal prevalence of diabetes amongst women and men (10.9% versus 11.3%) [10]. Consistent reports show individuals from low- and middle- income countries are likely to be more effected but future projections predict a 92% increase in Middle Eastern and North African countries and a 142% increase in Africa in the overall prevalence of Diabetes [10, 14]. Early diagnosis and management is crucial in reducing complications but is often curtailed by barriers to accessing care, lack of education and limited capacities of healthcare systems already under pressure [10].

1.2 Complications of Diabetes

Diabetes is associated with a plethora of chronic multiorgan complications. Diabetes and its complications have led to 3.4 million deaths in 2024 which is representative of 9.3% of global deaths from all causes [10]. Diabetes related complications have an adverse impact on all organs in the body. It is associated with cardiovascular and cerebrovascular disease, metabolic dysregulation, end organ damage and are summarised in Figure 1 below:

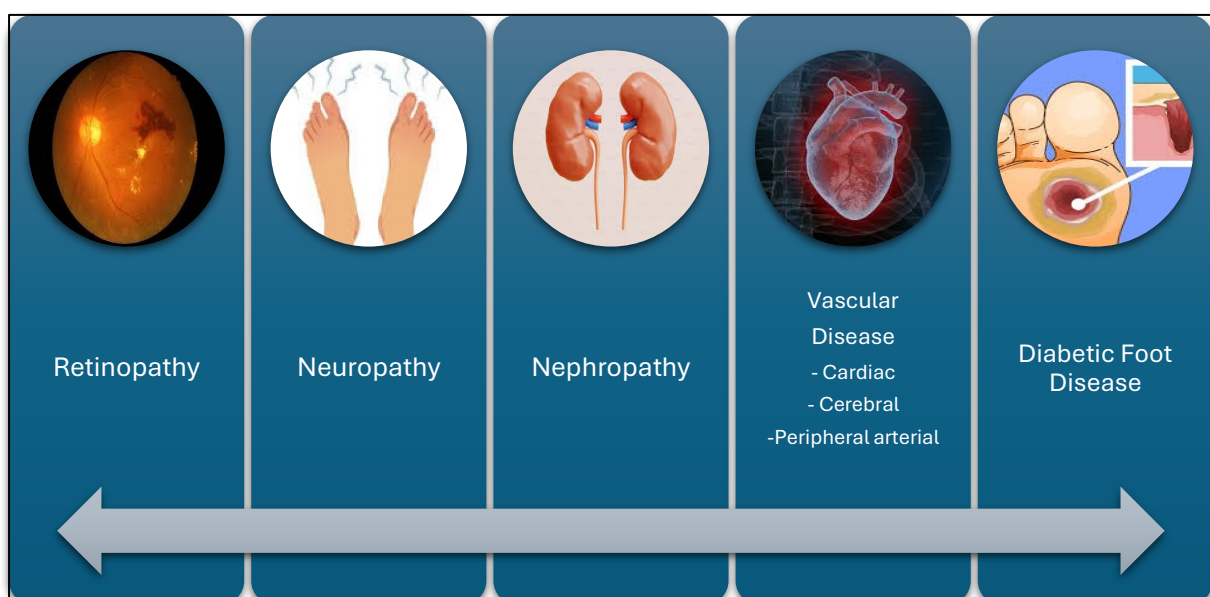


Figure 1: Complications of Diabetes

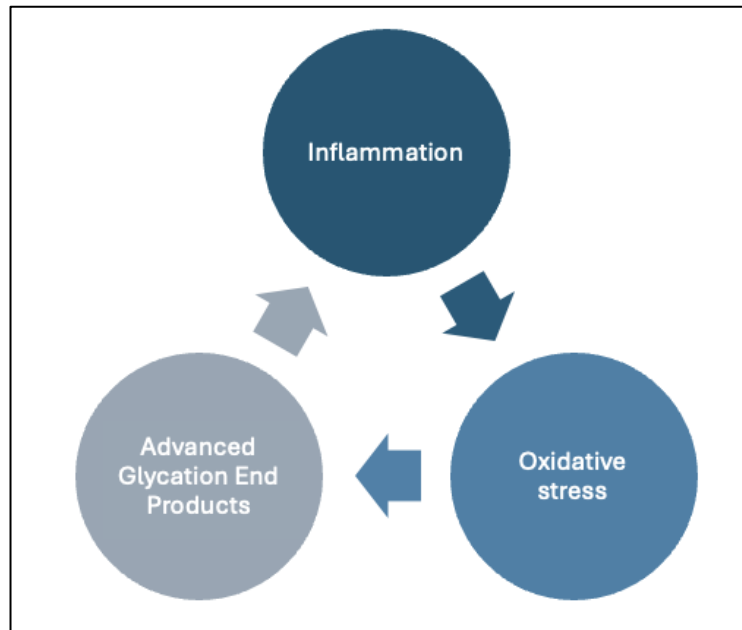


Figure 2: Pathophysiological processes implicated in diabetes related complications

These complications occur with the convergence of multiple risk factors and concurrent ongoing pathological processes, leading to disease manifestation. Diabetes creates a chronic, low level pro-inflammatory state within the body and is mediated by the ongoing insulin resistance.

1.2.1 Pro-inflammatory state

Cytokines are proteins which play an important role in cell signaling. Inflammatory cytokines including tumour necrosis factor- α (TNF- α), transforming growth factor- β 1 (TGF- β 1), Interleukins-1,6,16,17A which promote inflammatory cell infiltration are upregulated [12, 17]. These cytokines drive monocyte recruitment and macrophage activation which further perpetuates the inflammatory state. Patients with diabetes often have an elevated C-reactive protein, a marker of inflammation, at baseline [12, 17].

TNF- α is implicated in cardiomyocyte apoptosis and release of adhesion molecules which damage the retinal tissue. It is also associated with secretion of matrix metalloproteinases which damages the glomerular basement membrane in the kidney. TNF- α and TGF- β 1 expression is also linked to nerve conduction velocity reduction as well as axonal nerve degeneration [17].

Interleukin-6 and 17-A are overexpressed in patients with diabetic nephropathy and is related to glomerular injury and increased autophagy which drives the interstitial fibrosis in the kidney. It is implicated in retinopathy through retinal neuron and endothelial cell apoptosis which damages the smaller capillaries [17]. IL-6 expression is also correlated with muscle atrophy and loss of muscle strength [12].

Chronic inflammation leads to an increase in the vascular endothelial permeability and oxidization of low-density lipoprotein. This can drive the presence of foam cells (macrophages with excess lipids) which can expedite the process of atherosclerosis in the blood vessels leading to macro and microangiopathic disease [17, 18].

The pro-inflammatory state also upregulates protein catabolism through stimulation of the ubiquitin-proteasome system and downregulates the phosphoinositide 3-kinase/protein kinase B /mammalian target of rapamycin (PI3K-Akt mTOR) pathway which downregulates protein anabolism in skeletal muscle [12].

1.2.2 Oxidative Stress

Oxidative stress is the result of an imbalance between production of reactive oxygen species (ROS) or free radicals and the body's ability to neutralise them [19, 20]. ROS are unstable, unpaired electrons prone to react with proteins, lipids and DNA causing damage and impair healing which also contributes to the aging process [19].

Hyperglycemic states increase the production of ROS which in turn leads to DNA and mitochondrial dysfunction, resulting in a reduction in the ability to repair and leads to a deterioration in overall metabolic health [12]. NLR family pyrin domain containing 3 (NLRP3), a protein involved in innate immunity is activated by ROS and leads to a highly inflammatory programmed cell death (pyroptosis) of Schwann cells and is highly correlated to hyperalgesia and diabetic polyneuropathy [17].

ROS accelerates muscle damage through activation of the ubiquitin-proteasome pathway and degradation of muscle proteins [12]. Oxidative stress also inhibits the PI3K-Akt mTOR pathway which promotes atrophy of the muscles by reducing protein anabolism.

1.2.3 Advanced Glycation End-Products (AGE's)

AGE's are a group of heterogenous molecules including lipids and proteins which are glycated when exposed to glucose or glucose derivatives [12, 21]. The formation of these is irreversible and they tend to accumulate in the neuromusculoskeletal tissues, adversely impacting their biochemical properties [12, 21].

AGE's usually accumulate with age however, in patients with diabetes, this process is significantly expedited. They increase the cross-linkage of proteins within tissues leads to muscle fibrosis and atrophy, myocardial and vascular stiffness which impairs overall contractility [12, 21]. AGE's also induce further inflammation which increases overall oxidative stress and ROS's in the body.

1.2.4 Multiorgan adverse impact of diabetes

The culmination of these simultaneously occurring processes impacts all organs of the body in patients with diabetes which globally has a dire impact on multiorgan dysfunction. It affects the eyes (retinopathy), kidneys (nephropathy), nerves (neuropathy), blood vessels (cardiovascular, cerebrovascular and peripheral arterial disease).

Damage to the blood vessels, known as macroangiopathic disease including cardiovascular disease. Patients with diabetes are at a 72% higher risk of a myocardial infarction, 84% of heart failure and a 52% higher risk of stroke and is associated with dementia [10, 22, 23]. This occurs as a result of metabolic changes including hyperglycemia, hypertension and hypercholesterolaemia which also affects the major blood vessels.

Diabetes is also responsible for changes to the overall microcirculation in the body, called microangiopathic disease and contributes to the development of peripheral arterial disease (PAD). The prevalence of PAD increases up to 28% in patients with diabetes compared to their non-diabetic counterparts who have a reported prevalence of 10-26% [24].

The chronic damage to blood vessels culminates in damage to the vessels in the eye leading to retinopathy and is the most common cause of visual impairment in the western world [23]. Diabetes also effects the kidneys leading to nephropathy which may result in the need for renal replacement in the form of dialysis or renal transplantation [22, 23]. Furthermore,

diabetes leads to damage of the peripheral nerves leading to neuropathy which can result in reduced sensation, pain and muscle atrophy [12, 23].

1.3 Diabetic Foot Disease

Diabetic foot disease is defined as disease in the foot of a patient with an existing or previous diagnosis of diabetes mellitus which includes one or more of the following: peripheral neuropathy, peripheral arterial disease, ulcer, infection, neuro-osteoarthropathy, gangrene or amputation [25].

Diabetic Foot Ulcer (DFU) is a breach in the skin in the foot involving the epidermis and/or part of the dermis with the presence of neuropathy and/or peripheral arterial disease (PAD). Almost one in three diabetic patients will develop an ulcer in their lifetime [25]. The lifetime incidence of DFU is almost 34% [24].

The post-insulin era heralded a time where patients with diabetes were living for longer but were developing complications which were not noted in the patient cohort before. Thomas Willis (1621-1675) first observed the formation of ulcers and gangrene in patients with diabetes [4]. A century later, Thomas Hodgkin (1798-1866) made the connection between diabetes and gangrene purporting that diabetes may be contributing to nerve and vessel damage in the foot [4]. In 2024, up to 50% will die within 5 years of developing a diabetic foot ulcer [26]. DFU's can progress to infection, gangrene and tissue loss necessitating amputations.

1.3.1 Anatomy of the foot

The foot is comprised of 28 bones, 33 joints and 112 ligaments which are surrounded by skeletal muscle (Figures 4-6) [27]. These work in synergy to provide stability and allow mobility of the foot.

1.3.2 Intrinsic and extrinsic muscles of the foot

The extrinsic muscles aid in plantarflexion, dorsiflexion, inversion and eversion movements in the foot (Tables 1-2). The intrinsic muscles, which originate and insert within the foot, provide support by stabilising the arches of the foot and balance. They are also thought to

have a role in proprioception [28]. They are further divided into a plantar group comprising of 4 layers and, a dorsal group (Table 2) [29]. Interestingly, intrinsic muscles are almost non-existent in quadrupeds [29]. It is thought that the intrinsic muscles of the foot also offer energy storage and resistance to tensile loads during the various phases of gait [28].

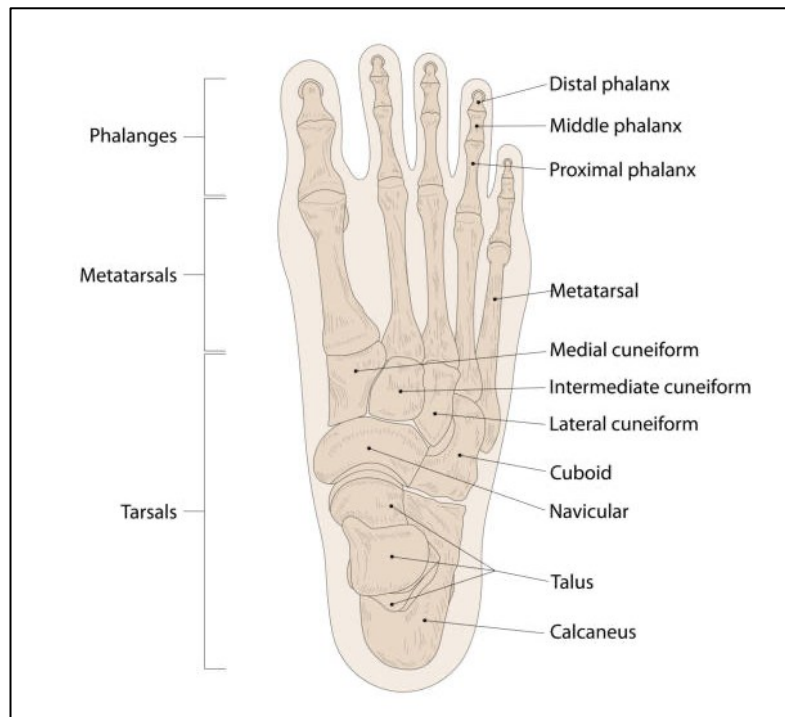


Figure 3: Osteal anatomy of the foot

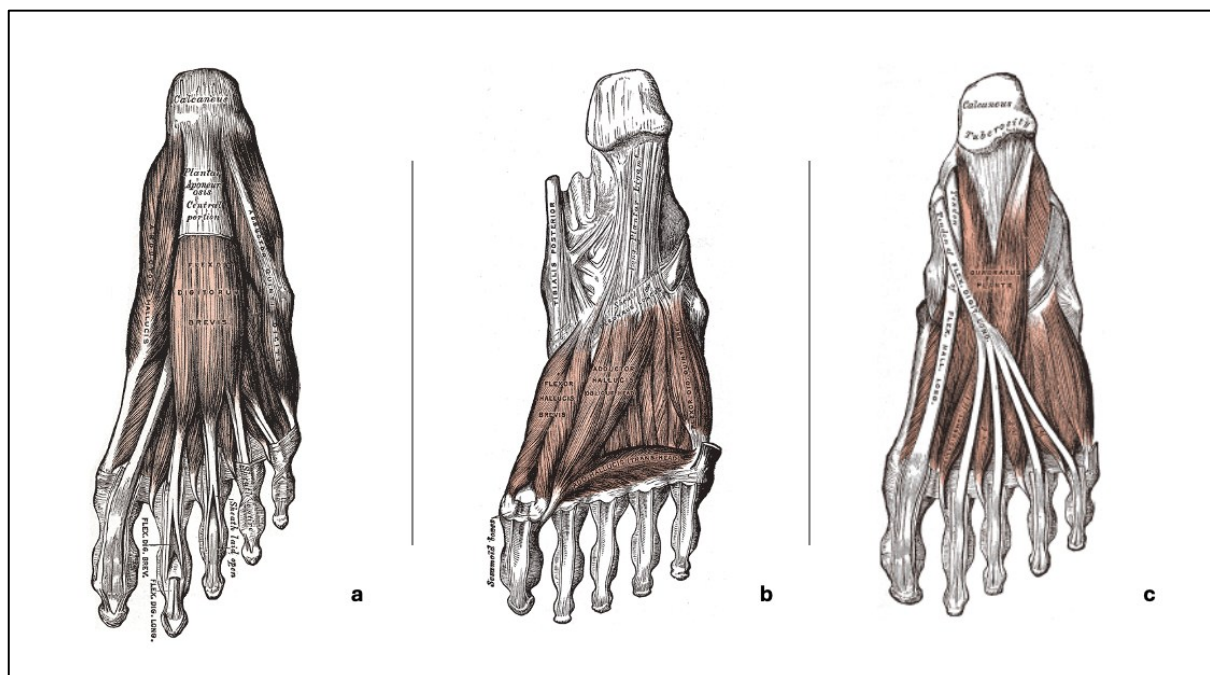
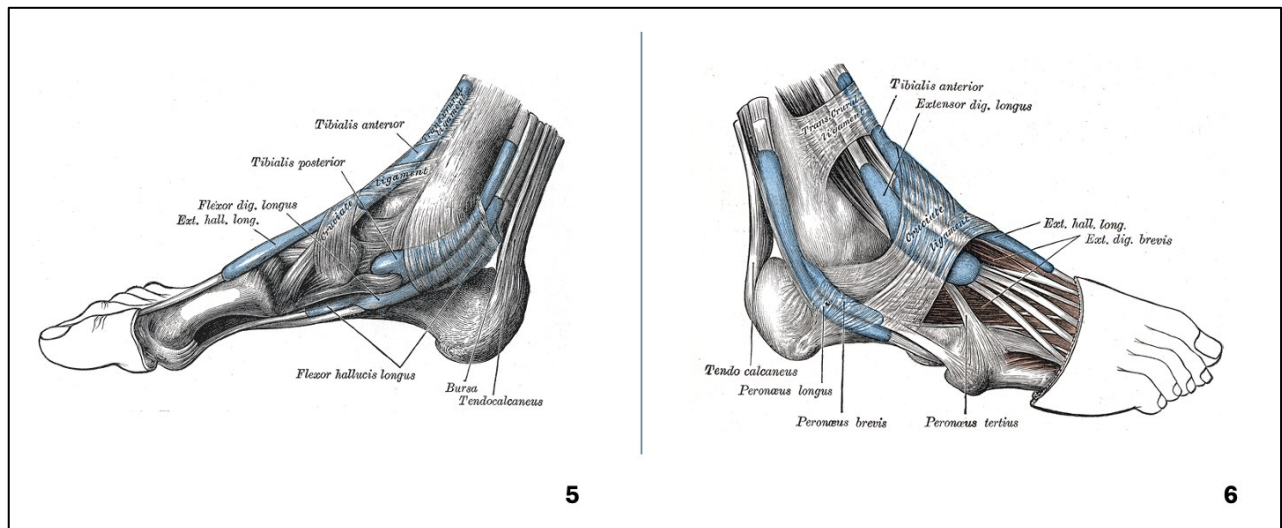


Figure 4: a – Plantar view of the 1st layer of intrinsic muscles (flexor digitorum brevis, abductor hallucis, abductor digiti minimi); b – Plantar view of the 2nd layer of intrinsic muscles (adductor hallucis [oblique and transverse head], flexor hallucis brevis)



	Muscle	Function
Anterior Compartment	Tibialis Anterior	Ankle dorsiflexion + Foot inversion
	Extensor Hallucis Longus	Hallux extension
	Extensor Digitorum Longus	Extension of 2nd-5th digits
	Peroneus Tertius	Ankle plantarflexion + Foot eversion
Lateral Compartment	Peroneus Longus	Ankle plantarflexion + Foot eversion
	Peroneus Brevis	
Posterior Compartment	Deep Muscles	
	Tibialis Posterior	Foot inversion + plantarflexion + adduction
	Flexor Hallucis Longus	Hallux Flexion
	Flexor Digitorum Longus	Flexion of 2nd-5th digits

	Superficial Muscles	
	Gastrocnemius	Ankle plantarflexion
	Soleus	
	Plantaris	

Table 1: Extrinsic muscles of the foot and their functions. Adapted from Ahmad et al (2024)

	Muscle	Function
Plantar Group		
1st Layer (superficial)	Abductor Hallucis	Hallux flexion + abduction
	Abductor Digiti Minimi	5th Digit abduction
	Flexor Digitorum Brevis	2nd-5th Digit flexion
2nd Layer	Lumbricals	Flexion of the metatarsophalangeal joints + Extension at the interphalangeal joints
	Quadratus plantae	2nd-5th Digit flexion
3rd Layer	Adductor Hallucis	Hallux adduction at the metatarsophalangeal joint
	Flexor Hallucis Brevis	Hallux flexion at the metatarsophangeal joint
	Flexor Digiti Minimi Brevis	5th Digit flexion at the metatarsophalangeal joint
4th Layer (deep)	Plantar Interossei	3rd-5th Digit adduction
	Dorsal Interossei	2nd-5th Digit adduction
Dorsal Group		
	Extensor Hallucis Brevis	Hallux extension
	Extensor Digitorum Brevis	2nd-4th Digit extension

Table 2: Intrinsic muscles of the foot and their functions. Adapted from Ahmad et al (2024)

1.3.3 Plantar fat pad

The plantar fat pads are located in the sole of the foot have robust fibrous septations which hold the adipocytes firmly together and have incorporated elastin which makes it ideal for shock absorption, insulation and weight bearing. There are two main fat pads of relevance in the foot: the calcaneal fat pad and the forefoot fat pad at the base of the metatarsals, i.e. the sub metatarsal head fat pad. Plantar fat pads vary in their composition to adipose tissue compared with other areas of the body which demonstrate fewer septae and loosely held fibrous tissues - not ideal for weight bearing and shock absorption [30].

Sensory nerves innervate the fat pad with twice as many supplying the metatarsal area than the rest of the foot and the Vater-Pacini corpuscle, a mechanical baroreceptor, found in the plantar fat pad [30]. In individuals with diabetes, plantar fat has been shown to migrate distally due to the hyperextension of the metatarsophalangeal joints. This reduces the area between the metatarsal heads and skin leading to increased plantar pressure and susceptibility to morphological changes such claw toe and hammer toe deformities [31]. Changes to the plantar fat pads, such as changes in their thickness or reduction in the elasticity, can occur over time in patients with ongoing inflammatory processes such as diabetes leading to altered foot biomechanics, displacement of where foot peak pressures are applied and leading to skin breakdown and ulceration [31, 32].

1.3.4 Pathophysiology of Diabetic Foot Disease

The development of DFD is complex and multifactorial. As discussed earlier, insulin resistance leads to hyperglycemia which creates a state of chronic, low grade inflammation with free radicals and circulating AGE's all of which damage the tissues and impair healing. These culminate in DFD (Figure 7).

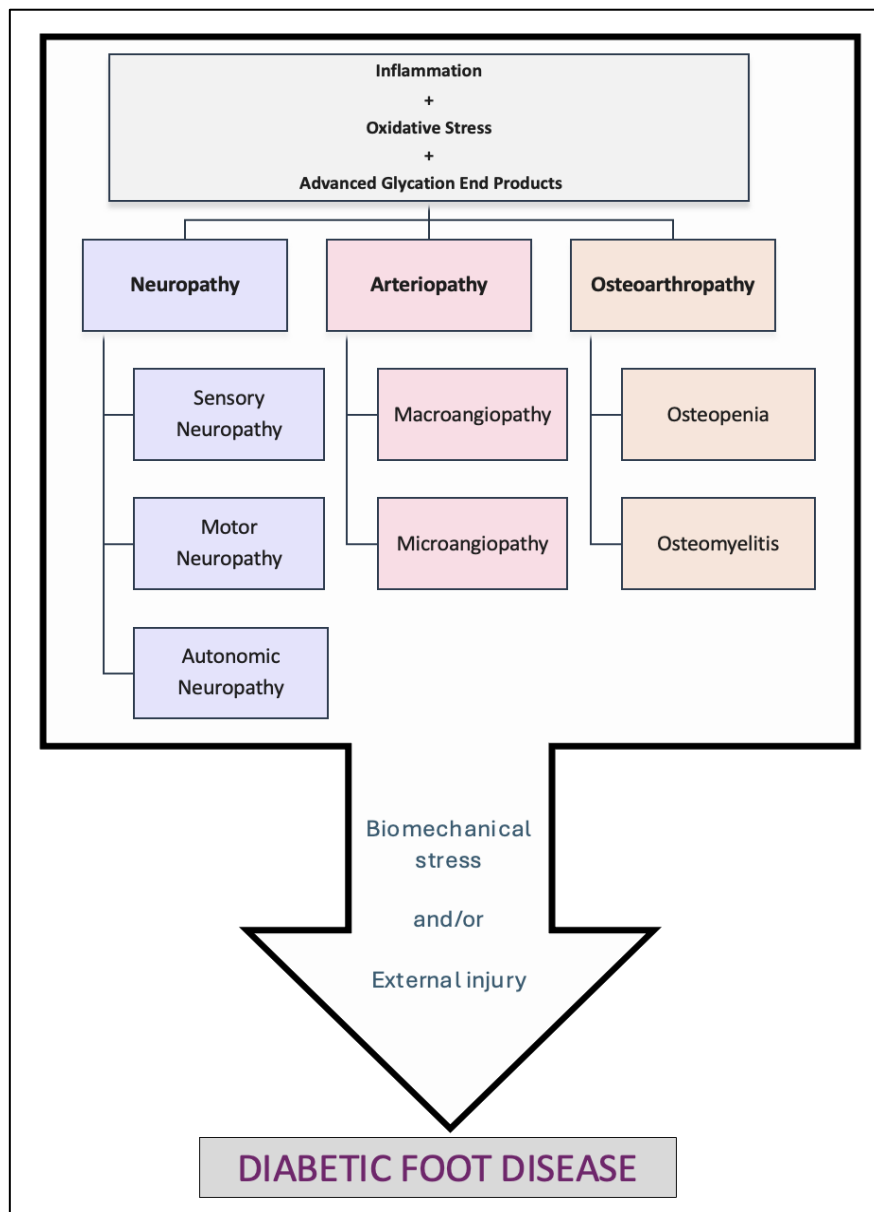


Figure 7: Development of diabetic foot disease

Diabetic peripheral neuropathy

Sensory neuropathy involves dysfunction of the peripheral nerves which leads to altered perception of pain, pressure and temperature. It may initially present with a deep seated, burning sensation but can eventually lead to complete loss of sensation and follows a “glove and stocking” pattern of distribution [33].

Motor neuropathy damages the innervation to skeletal muscle which leads to muscle fibrosis, atrophy and structural changes in the foot whereby the skeletal muscle is gradually replaced by intramuscular fat which changes the overall biomechanics of the foot including where pressure is applied during various gait stages [12].

Autonomic neuropathy can involve the entirety of the autonomic nervous systems (sympathetic and parasympathetic) and impacts the vasomotor and visceromotor fibers innervating organs including cardiovascular, gastrointestinal and genitourinary systems [34]. In relation to DFD, autonomic neuropathy affects the sudomotor system leading to dry skin and anhidrosis. Neuropathic patients can incur minor foot injuries and poor footwear resulting in friction which can lead to breaks in the skin which may go unnoticed.

Neuropathy also leads to foot deformities such as claw/hammer toes which changes the biomechanical loading of the foot and leading to higher mechanical stress in certain areas which can lead to callus formation and eventual ulceration.

Arteriopathy

Macroangiopathic disease is accelerated in patients with diabetes and affects all major blood vessels which leads to arterial stiffening and impacts the overall elasticity of the blood vessels. This may culminate in arterial stenosis or occlusions in the lower limb vessels. This may be asymptomatic or symptomatic and can present with intermittent claudication or chronic limb threatening ischaemia (characterised by rest pain, night pain and tissue loss). In the presence of neuropathy, this typical presentation may be delayed which makes the diagnosis of DFD early imperative. PAD affects 113 million individuals globally[35]. 50% of patients with DFU have PAD [24].

A good blood supply is required for effective wound healing and reducing the risk of infection. Microangiopathic disease affects the small capillaries and arterioles in the foot, adversely affecting the perfusion to the distal peripheries in the foot which further perpetuates poor wound healing through microinfarction in the muscle and nerves[35]. Macro- and Microangiopathic disease is a significant risk factor to poor outcomes and development of the advanced stages of DFD which include gangrene and necrosis.

Osteoarthropathy

The combination of neuropathy and arteriopathy also leads to osteopenia and osteoporosis in the bones of the foot. Charcot Foot, also known as Charcot neuro-osteoarthropathy is an inflammatory process in patients with diabetes and peripheral polyneuropathy which results

in injury to the bones (microfractures and dislocations), joints and soft tissues of the foot and ankle, leading to deformities of the foot [36]. Charcot foot was first described by Jean-Martin Charcot (1825-1893) in 1883. The most common aetiology at that time was syphilis related neuropathy however, the most common cause for Charcot foot now is diabetic polyneuropathy [4]. DFD can also culminate in osteomyelitis which often requires long courses of antibiotics and/or amputation.

Patients with diabetes often have a level impaired immunity and this may be due to a number of reasons. One theory proposed for this neuroimmune impairment is the immunocompromised district theory which purports that there are obstacles in the normal pathways of immunocompetent cells or neuropeptide signal interference as a result of neuropathy and microangiopathy leading to localised areas of immunocompromising [37]. Whilst this has not been confirmed, it is certain that the chronic inflammation and diabetes related complications impair the ability to heal and fight pathogens which makes indolent and overt infections common.

1.3.5 Risk assessment

Risk assessment is recommended in the UK National Institute for Health and Care Excellence (NICE) guidelines published in 2015 on the prevention and management of diabetic foot problems. This involves clinical examination and checking for neuropathy using a 10g monofilament, limb ischaemia with an ankle-brachial pressure index (ABPI), ulceration, callus, infection and/or inflammation, deformity, gangrene and Charcot arthropathy [38]. This then allows stratification of patients into low, moderate, high risk and those with active problems as shown below in Table 3:

Risk	Features	Frequency of risk assessment
Low risk	No risk factors + callus	Annually
Moderate risk	Deformity OR Neuropathy OR Peripheral Arterial Disease (PAD)	3-6 months

High risk	previous ulceration OR previous amputation OR on renal replacement therapy OR neuropathy + PAD together OR neuropathy + callus +/- deformity OR PAD + callus +/- deformity	1-8 weeks + referral to foot protection service
Active DF problem	Ulceration OR infection OR chronic limb-threatening ischaemia OR gangrene OR suspicion of an acute Charcot arthropathy OR an unexplained hot, swollen foot with a change in colour +/- pain	1-2 weeks + referral to foot protection service

Table 3: Risk stratification of developing DFU or requiring an amputation (Adapted from UK NICE Guidelines on Diabetic Foot Problems: diagnosis and management (2015) (37))

The risk stratification system determines the frequency at which patients have their foot assessment i.e. annually for low risk patients, 3-6 months for moderate risk and 1-8 weeks for high risk patients [38].

Moderate and high-risk patients are referred to a foot protection service where they ought to be seen within 2-8 weeks depending on the risk category. Further assessment of the vasculature and biomechanical status is also recommended [38].

The 2023 IWGDF guidelines have a similar risk stratification system which screens the patients for symptoms and potential underlying causes [39]. This includes examination of the foot looking for ulcers, testing for loss of protective sensation including pressure (using a 10g monofilament), vibration (using a 128 hertz tuning fork), tactile sensation and the

vasculature (symptoms and presence or absence of pedal pulses) [39]. This is summarised in Table 4 below:

Category	Ulcer risk	Characteristics	Frequency of risk assessment
0	Very low	No loss of protective sensation (LOPS) + No signs of peripheral arterial disease (PAD)	Annually
1	Low	LOPS OR PAD	6-12 months
2	Moderate	LOPS OR PAD OR LOPS + foot deformity OR PAD + foot deformity	3-6 months
3	High	LOPS OR PAD + ≥ 1 of the following: - History of foot ulcer - History of minor or major lower limb amputation - End-stage renal disease	1-3 months

Table 4: IWGDF Risk stratification and screening (Adapted from IWGDF Guidelines on prevention and management of Diabetic Foot Disease (2023))

1.3.6 Diabetic foot ulcer diagnosis and classification

DFU's can be broadly categorised into neuropathic, neuroischaemic or ischaemic ulcers. The majority of ulcers are neuropathic or neuro-ischaemic. All current classification systems offer aid in identifying the type of ulcer a patient may present with. The first DFU classification systems were the Meggitt-Wagner classification in 1976 followed by the University of Texas wound classification system (UTWCS) in 1996 [4].

The International Working Group of the Diabetic Foot (IWGDF) 2023 guidelines recommend the SINBAD [Site, Ischaemia, Neuropathy, Bacterial Infection, Depth] system for patients

with DFU when communicating between healthcare individuals[40]. The SINBAD classification system (Table 5) grades 5 areas from a scale of 0-1 with a maximum score of 6. It does not require specialist equipment and has been validated by 12 studies [40].

Category	Details	Score
Site	Forefoot	0
	Midfoot + Hindfoot	1
Ischaemia	At least one palpable pedal pulse	0
	Reduced pedal flow clinically	1
Neuropathy	Protective sensation intact	0
	Protective sensation lost	1
Ulcer Area	< 1cm ²	0
	≥1cm ²	1
Depth	Confined to skin and subcutaneous tissue	0
	Extending to muscle, tendon or deeper	1

Table 5: SINBAD Classification system (Adapted from the IWGDF 2023 Classification Guidelines (37))

The 2023 IWGDF guidelines recommend the Wifi [Wound, Ischaemia and foot infection] system where the appropriate resources and expertise is available [40]. The Wifi combines scores for wound, ischaemia and foot infection and offers the benefit of allowing decisions regarding revascularisation to be made (Tables 6-8). The Wifi grading provides a 1 year risk

for amputation and 1 year benefit for revascularisation which can be graded as very low, low, moderate or high [40].

Other classification systems include the Diabetic Foot Risk Assessment (DIAFORA) score which evaluates neuropathy, foot deformity, arteriopathy, previous foot ulcer or amputation, presence of multiple ulcers, infection and osteal involvement but had not been externally validated at the time of guideline publication [40]. In 2025, external validation results were found to be similar to the initial study [41].

In patients with diabetes and an infected foot ulcer, the Infectious Disease Society (IDSA)/IWGDF system is recommended [40]. The IDSA/IWGDF is partially incorporated in the Wifi score but can function in isolation in scenarios where infection is being assessed and there is paucity of equipment [40].

Diabetic Foot Ulcer	Gangrene	Grade
No ulcer	No gangrene	0
Clinical description: minor tissue loss. Salvageable with simple digital amputation (1 or 2 digits) or skin coverage.		
Small, shallow ulcer(s) on distal leg or foot; no exposed bone, unless limited to distal phalanx	No gangrene	1
Clinical description: minor tissue loss. Salvageable with simple digital amputation (1 or 2 digits) or skin coverage.		
Deeper ulcer with exposed bone, joint or Tendon not involving the heel; shallow heel ulcer, without calcaneal involvement	Gangrenous changes limited to digits	2
Clinical description: major tissue loss salvageable with ≥ 3 digital amputations or standard Transmetatarsal amputation $\pm$ skin coverage.		
Extensive, deep ulcer involving forefoot and/or midfoot; deep, full thickness heel ulcer $\pm$ calcaneal involvement	Extensive gangrene involving forefoot and /or midfoot; full thickness heel necrosis $\pm$ calcaneal involvement	3

Clinical description: extensive tissue loss salvageable only with a complex foot reconstruction or non-traditional TMA (Chopart or Lisfranc); flap coverage or complex wound management needed for large soft tissue defect.

Table 6: WiFi Classification system – Wound description; TMA: Transmetatarsal Amputation (Adapted from the IWGDF 2023 Classification Guidelines (37))

Ankle-Brachial Index	Ankle systolic pressure (mmHg)	TP, TcPO ₂ (mmHg)	Grade
≥ 0.80	> 100	≥ 60	0
0.6 – 0.79	70 – 100	40 – 59	1
0.4 – 0.59	50 – 70	30 – 39	2
≤ 0.39	< 50	<30	3

Table 7: WiFi Classification system – Ischaemia; TcPO₂: Transcutaneous Partial Oxygen Pressure; TP: Toe Pressure; mmHg – millimeters of mercury (Adapted from the IWGDF 2023 Classification Guidelines (37))

Clinical manifestations – Foot Infection	Grade
No symptoms or signs of infection Infection present, as defined by the presence of at least 2 of the following: - Local swelling or induration - Erythema > 0.5 to ≤ 2 cm around the ulcer - Local tenderness or pain - Local warmth - Purulent discharge (thick, opaque to white, or sanguineous secretion)	0
Local infection involving only the skin and the subcutaneous tissue (without involvement of deeper tissues + no systemic signs as described below). Other causes of an inflammatory response of the skin excluded (e.g., trauma, gout, acute Charcot neuro-osteoarthropathy, fracture, thrombosis, venous stasis)	1

Local infection (as described above) with erythema >2 cm, or involving structures deeper than skin and subcutaneous tissues (e.g., abscess, osteomyelitis, septic arthritis, fasciitis), and No systemic inflammatory response signs (as described below)	2
Local infection (as described above) with the signs of SIRS, as manifested by two or more of the following: - Temperature > 38°C or < 36°C - Heart rate > 90 beats/min - Respiratory rate > 20 breaths/min or PaCO ₂ < 32 mm Hg - White blood cell count > 12,000 or < 4000 cu/mm or 10% immature (band) forms	3

Table 8: WIFi Classification system – Infection; ABI: Ankle-Brachial Index; PaCO₂: Partial Pressure of Carbon Dioxide; SIRS: Systemic Inflammatory Response Syndrome; TcPO₂: Transcutaneous Partial Oxygen Pressure; TP: Toe Pressure (Adapted from the IWGDF 2023 Classification Guidelines (37))

The current UK National Institute for Health and Care Excellence (NICE) guidelines published in 2015 on the prevention and management of diabetic foot problems also suggest using the SINBAD classification or the UTWCS system to document patients with ulcers [38].

1.4 Management of Diabetic Foot Disease

The management of DFD requires a multidisciplinary input at primary, secondary and tertiary levels and warrants input from a range of specialists including endocrinology, podiatry, orthotics, orthopaedics, microbiology, radiology and vascular surgeons.

The concept of methodical wound examination and considering factors such as the size, depth, location and signs of infection as well as the role of debridement and clean dressings to expedite healing was described by Al-Razi (854-925 A.D.) in the Kitab Al-Hawi [4]. The underlying principles of managing DFD has not changed significantly over the last few millennia. Sumerian clay tablets from 2100 B.C. record the three ‘healing gestures’ which formed the cornerstone of wound management at that time i.e. washing the wound, applying a dressing/plaster and careful bandaging [4].

Modern day practices focus on mitigating risk factors such as:

1. Improving the vasculature to restore adequate tissue perfusion: This may be done through endovascular options including angioplasty and/or stenting or open intervention in the form of a bypass using autologous vein/ prosthetic graft. Medically, this may also be improved with vasodilators such as a Prostaglandin (Iloprost).
2. Offloading areas of pressure: The IWGDF guidelines recommend a non-removable knee-high offloading device such as a total contact cast. Where this is not appropriate, a removable knee or ankle length offloading device can be utilised. Other more invasive options also include digital flexor tenotomy, Achilles tendon lengthening, metatarsal head resection, metatarsal osteotomy[39].
3. Treating infection: This may be done with antibiotics or debridement of the wound. If a diabetic foot infection is suspected, the UK NICE guidelines recommend taking a soft tissue or bone sample from the wound for microbiology, excluding osteomyelitis and commencing antibiotics [38]. If risk of diabetic foot sepsis is evident or highly suspected, urgent referral to an appropriate surgical centre for urgent debridement, drainage of foot sepsis and or amputation is warranted.
4. Regular, meticulous wound care: The UK NICE guidelines from 2015 recommend one or more of the four options in managing diabetic foot ulcers: offloading, control of foot infection, wound debridement and wound dressings [38]. The IWGDF 2023 guidelines do not recommend certain types of wound dressings such including topical antiseptic or antimicrobial, honey, collagen or alginate dressings and recommend a sterile environment for sharp debridement [42].

Consideration to hyperbaric oxygen and topical oxygen is suggested as an adjunct in neuro-ischaemic or ischaemic DFU's where standard care has failed and if resources are available in the IWGDF guidelines however, this is not recommended in the UK NICE guidelines as a treatment option [38, 42].

Similarly, NICE recommends considering dermal/skin substitutes which is not the case in the 2023 IWGDF guidelines [39]. The UK NICE guidelines from 2015 also do not recommend growth factors or autologous platelet-rich plasma gel and regenerative wound matrices. Conversely, the IWGDF guidelines suggest considering the use of multi-layered patches of

autologous leucocytes, platelets and fibrin, placental membrane allografts in non-infected ulcers which fail to heal after 4-6 weeks of optimal treatment [39].

It also imperative to remember that DFU extends beyond the lower limb and a holistic approach is necessary. This includes assessment and optimisation of glycaemic control and cardiovascular risk factors, treating malnutrition and addressing the mental and/or social concerns and challenges in the lives of the cohort of patients presenting with DFD [39]. These are often pertinent contributing factors which play a significant role prior to the manifestation of disease which is more often than not a culmination of multiple decisions and influences over a lifetime.

1.5 Amputations

Lower Limb Amputations (LLA) are a serious complication of DFD. DFU's precede up to 85% of all non-trauma related amputations [24]. An amputation is defined as 'resection of a segment of a limb through a bone or joint' [25]. This is further divided into minor amputation (below the ankle) or a major amputation (above the ankle). Amputations are further characterised in the table below:

Amputation type	Name	Anatomical level
Minor Amputation	Toe amputation	Partial or complete removal of one or more toes
	Metatarsal-phalangeal disarticulation	Removal of one or more toes at the metatarsal – phalangeal joint
	Transmetatarsal amputation	Through one or more metatarsals
	Tarso-metatarsal disarticulation	Through one or more tarso-metatarsal joints
	Midtarsal disarticulation	Through the tarsal bones or joints
	Ankle disarticulation	Through the ankle joint
	Transtibial / Below Knee Amputation	Through the tibia and fibula

Major Amputation	Knee disarticulation / Through-knee amputation	At the knee joint
	Transfemoral/ Above Knee Amputation	Through the femur
	Hip disarticulation/Transpelvic	At the hip joint with or without part of the pelvis

Table 9: Summary of anatomical amputation levels - adapted from Van Netten 2024 (19)

1.5.1 Impact of amputations

Amputations have a profound impact on the personal, physical, social, psychological and occupational lives of those involved. The burden of undergoing a LLA takes a toll on the activities of daily living leading to a reduction in work as patients have to attend regular hospital appointments to wound clinics. The resulting absenteeism and the costs associated with travel collectively has a financial impact on the individuals. The domino effect from amputations also directly effects family members, caretakers/care providers and those close to the individual who have to adapt to the role of caregivers. The impact of DFD echoes across multiple hospital specialities including primary care, secondary and tertiary care.

1.5.2 Diabetes and amputations

S [43]. Poorly managed diabetes contributes to over 9,000 LLA's annually in the UK [44]. 80% of all amputations performed in the UK on a weekly basis are a result of ulceration in the preceding weeks and months which could be prevented [16, 22].

16,693 amputations were undertaken in the UK between 2007 and 2010 in diabetic patients [44]. This number jumped to 27,465 between 2015 and 2018 [43]. Between 2017-2020, 171,759 hospital admissions, 21,738 minor amputations and 7,957 major amputations were directly attributed to DFD [45]. In the UK, white male patients over the age of 65 from lower socioeconomic areas were at the highest risk of amputation [45].

In the United States of America, an estimated 2 million individuals were living with LLA's in 2024 [46]. This number was projected to increase by 145% by 2060 with the majority as a result of diabetes (67%) and vascular disease (36%) [46].

1.5.3 Morbidity and mortality associated with amputations

The morbidity and mortality associated with LLA's cannot be ignored. Patients who undergo an amputation have a 16.5% emergency 30-day hospital readmission rate and have a 7.5% mortality rate [47].

1.5.4 Amputations and litigation

The rise in DFD and amputations have led to growing litigation pertaining to medical negligence resulting in amputation in recent years [48]. A review of ninety-two claims related to diabetes identified recurring themes related to preventative care, pathways and management of DFD, management of peripheral arterial disease, biomechanics and offloading, emergency department discharges, education and psychological support [48].

1.6 Cost of diabetes and diabetes related complications

Health expenditure related to diabetes and its complications has grown considerably from \$232 billion in 2007 to over a trillion US dollars in 2024, a 338% increase over the last 17 years [10]. Diabetes and its complications cost the UK National Health Service almost £10.7 billion annually in 2021, consuming almost 10% of the NHS budget [26, 49]. Indirect costs from absenteeism and early mortality are estimated at over £3.3 billion [50]. This number is expected to grow up to £20 billion by 2036 and could consume almost 17% of the overall NHS annual budget [49].

In 2014/2015, DFD related health expenditure comprised of 1% of all NHS spending [26, 45]. Diabetic foot ulceration costs over £147 million whilst amputations cost almost £467 million every year to the NHS [50]. According to the NHS payment scheme for 2024/2025, the cost for amputation coded procedures is as follows [51]:

- A Single, Amputation Stump or Partial Foot Amputation Procedure, for Diabetes or Arterial Disease, with Other Open Blood Vessel Procedure is between £14,328-£21,173
- Multiple, amputation stump or partial foot amputation procedures, for Diabetes or Arterial Disease is between £14,328-£21,173
- Amputation of single limb is between £10,501-£24,981

- Multiple limb amputation procedures £18,730-£27,202

These are not insignificant numbers and fail to take into the account the costs accrued from daily in-hospital stays, rehabilitation, providing orthotic footwear or the multiple hospital follow-up appointments after.

1.7 Risk prediction

A risk prediction marker is any measure (quantitative or qualitative) used to predict an individual's risk for a particular event [52]. Risk prediction models (RPM) combine information from several markers. Statistical models for risk prediction include logistic regression, Cox proportional hazard models and classification trees. RPM's may be diagnostic i.e. where the risk of pre-existing disease is calculated or prognostic where the risk of a particular outcome in a defined disease is measured [53].

The accuracy of RPM's depends on two factors - the model's discrimination and calibration [52]. RPM discrimination is evaluated using the receiver-operating characteristic (ROC) curve and or Harrell's C-statistic. ROC curves provide a visual representation of the trade-off between the true positive rate and false positive rates and can useful when dealing with imbalanced datasets as it is insensitive to class distribution. Class distribution is the proportion or frequency of observations assigned to each distinct class label within a dataset.

Calibration allows the model's probability to predict an event for an individual with specific predictor values similar to the proportion of all individuals in the population who experience the same event [52]. Calibration is evaluated using the Hosmer-Lemeshow test which assesses the goodness-of-fit for logistic regression models and measures the predicted probabilities from the logistic regression match the actual observed outcomes across different groups.

Another aspect of RPM's is model validation which may be internal validation or external validation. Internal validation involves using the same data used to develop it whilst external validation is when the data used is different to the initial set used to create the

model [53]. The latter offers robust testing of a given RPM and is required before being incorporated in a clinical setting.

1.7.1 Classification and risk prediction models

The rationale for classification systems is multipurpose. They may be used to introduce uniformity when communicating between different specialities in a healthcare environment, in standard setting exercises and auditing. They also allow the evaluation of individual risk and stratification of ulcer development, progression of disease or complications such as a LLA.

Systematic reviews by Kaka et al in 2023 and Monteiro-Soares in 2020 and 2024 synthesise findings from several prediction models [40, 54, 55].

Current models suggested to date have attempted to predict:

1. DFU onset

- a. Diabetic Foot Risk Assessment (DIAFORA)
- b. Martins-Mendes classification
- c. Prediction of Diabetic Foot UlcerationS (PODUS 2020)
- d. Scottish Intercollegiate Guidelines Network (SIGN)/Scottish Care Information – Diabetes Collaboration

2. DFU healing

- a. Boyko classification
- b. Diabetic ulcer severity sore (DUSS)
- c. Margolis et al

3. Lower Limb Amputation

- a. Eurodiale study
- b. Lipsky et al
- c. Martins-Mendes classification
- d. Pickwell et al

- e. Scottish Intercollegiate Guidelines Network (SIGN)/Scottish Care Information – Diabetes Collaboration
- f. Van Acker-Peter
- g. Wound depth, ischaemia, and foot infection

1.7.2 Predicting the risk of lower limb amputation

The 2023 IWGDF do not recommend any classification or scoring systems to offer individual patients with DFD/DFU with an outcome prognosis or predicting risk [40]. However, there are risk prediction models for predicting amputation risk which have been suggested over the preceding decade.

Eurodiale Study

The Eurodiale study was a multicentre study across 14 European diabetic foot centres. The prospective cohort study from 2010 comprised of 1232 patients who were followed up for a year and identified baseline characteristics independently associated with minor amputation [55, 56]. The multivariable predictive model predicts minor LLA and suggested the following formula:

$$(4 \text{ points for males}) + (18 \text{ points for deep ulcer}) + (5 \text{ points for infection}) + (6 \text{ points for ischaemia})$$

However, this has not been externally validated and there was paucity in the details of the model's interpretability [55].

Lipsky et al

This predictive model was suggested in 2011 and is based on a multicentre retrospective cohort of patients from the United States of America [40, 55]. This utilised data from inpatients with DFU's. The final model and score formula is as follows:

$$0.1372 \times (\text{chronic renal disease or creatinine} > 3 \text{ mg/dL}) + 0.1988 \times (\text{male sex}) + 0.2830 \times (\text{temperature} < 96 \text{ F or} > 100.5 \text{ F}) + 0.5477 \times (\text{age} \geq 50 \text{ years}) + 0.5168 \times (\text{infected ulcer vs$$

$$\text{cellulitis}) + 0.5020 \times (\text{LEA history}) + 0.6203 \times (\text{albumin} < 2.8 \text{ g/dL}) + 0.7485 \times (\text{PAD history}) + \\ 0.9596 \times (\text{white blood cell count} \geq 11) + 1.3845 \times (\text{surgical site vs cellulitis}) + 1.6418 \times \\ (\text{transferred from other acute care facility})$$

Each term within the brackets received a binary score of zero or one. Internal validation of this demonstrated good discrimination and calibration [55].

Martins-Mendes classification

The Martins-Mendes classification models from 2014 for amputation risk over a three-year period has two versions. The original, which evaluates the risk of DFU or amputation includes 3 variables i.e. PAD, Previous DFU or diabetes related complications (retinopathy, nephropathy, neuropathy, cerebrovascular disease, cardiovascular disease or metabolic abnormalities including ketoacidosis).

The simplified model, which focuses only on amputation has two variables i.e. prior DFU and diabetes complications [54]. The simplified model for predicting amputations uses the following formula:

$$-5.35 + 0.61 \times \text{number of complications} + 1.91 \times \text{previous DFU}$$

Both versions have been externally validated with good discrimination and prognostic value [54].

Pickwell et al

This model from 2015 utilised data from the Eurodiale study to predict total amputation and minor amputation (excluding 2nd-5th digits) and included 575 patients of whom 159 underwent an amputation [57].

The following parameters were used: sex, PAD, pain/tenderness on palpation, ulcer size, ulcer depth and oedema around the wound. Only 50% of the patients in the highest tertile and this was internally validated [55, 57].

Scottish Intercollegiate Guidelines Network (SIGN)/Scottish Care Information – Diabetes Collaboration

Initially created to predict DFU onset, it includes factors such as the presence or absence of foot deformities, pedal pulses, sensation with a 10g monofilament, previous history of DFU and ability to self-care. This has been externally validated but had one of the lowest ROC values for predicting LLA's [55].

Van Acker-Peter

The Van Acker-Peter RPM builds on the UTWCS system and provides a chromatic code to stratify risk in patients with pre-existing DFU.

A two-dimensional matrix grades depth and extent of the wound on the horizontal axis i.e. superficial dermis/epidermis, minor/major soft tissue, periosteitis, extensive osteomyelitis. The vertical axis includes pathological features including loss of sensation, bony deformities, Charcot foot, ischaemic foot, mixed insensitive and vascular [55]. This has been internally and externally validated and found to have similar accuracy compared to other classification systems in predicting LLA's[55].

Wound depth, ischaemia, and foot infection

This is a vascular based classification system focusing on the threatened limb. It was created in 2014 with the aim of predicting the likelihood of revascularisation being needed for wound healing and limb salvage and stratifies patients according to their 1 year risk of major LLA [58]. The scoring system incorporates the Wifi system and the IDSA/IWGDF for wound infection. It has been externally validated by multiple studies and found to be useful in predicting LLA, LLA-free survival, independent living status and mortality [55].

1.8 Imaging

Imaging can provide useful information in patient assessment, surveillance and management. The IWGDF 2023 guidelines recommend MRI as an imaging modality when there is suspicion of diabetes related osteomyelitis despite clinical, x-ray and biochemical markers [59]. This is not dissimilar to the UK NICE guidelines on diabetic foot management

which suggest x-rays as an initial modality to exclude osteomyelitis which may be escalated to an MRI if there is ambiguity regarding the presence of osteomyelitis [38].

1.8.1 Magnetic Resonance Imaging (MRI)

Hydrogen nuclei (^1H) are present in abundance in the body's tissue which is composed of water and fat. Each nuclei (which has positively charged protons) possesses a property called 'spin' so that they are rotating about an axis at a constant rate which then creates a magnetic field called a 'magnetic moment' [60].

When the tissue is placed inside a strong, external magnetic field ($\mathbf{B}_0$), the magnetic moments align parallel (low energy) or antiparallel (high energy) to the field. Protons rotate perpendicular to or 'precess' about the magnetic field instead of aligning perfectly. The frequency of this precession is called Larmor frequency/equation:

$$\omega_0 = \gamma B_0$$

- ω_0 = angular precession [Larmor] frequency (in megahertz)
- γ = gyromagnetic ratio (~ 42.58 Megahertz/T) i.e. constant for hydrogen
- $\mathbf{B}_0$ = strength of the magnetic field (in Tesla)

Radiofrequency (RF) excitation

Radiofrequency excitation at Larmor frequency causes protons to absorb energy and flip into a higher energy state and tipping the net magnetisation vector in the transverse plane [60, 61]. Once the radiofrequency excitation stops, the protons relax into the following planes:

- Longitudinally [T1 Relaxation]
 - Magnetisation recovery along the $\mathbf{B}_0$ direction and energy exchange with surrounding molecules
 - Short T1 appear bright on T1-weighted images i.e. fat
- Transversely [T2 Relaxation]
 - Magnetisation decay due to loss of phase coherence. This is called dephasing and is caused by interactions between nearby 'spins'

- Long T2 appear bright on T2-weighted images i.e. fluid

Free Induction Decay

Protons emit radiofrequency signals as they relax. A receiver coil detects the free induction decay.

Spatial encoding

MRI uses magnetic field gradient which is the change in the strength of the magnetic field over a specific distance, to locate precisely the origin of the signal within the object being scanned [60, 61].

Fourier Transform

Converts signals from a time or spatial domain into a frequency domain to produce an image [60, 62].

Repetition Time (TR) and Echo Time (TE)

Repetition Time, measured in milliseconds, is the time between successive RF excitation applied to the same tissue and has an impact on the image contrast. Echo Time, also measured in milliseconds, is the time between RF excitation applied and the peak of the signal received by the scanner [60]. These concepts become relevant in the images rendered. T1-weighted images have a short TR and TE, T2-weighted images have long TR and TE whilst Proton density images have a long TR but short TE [60].

1.8.2 Imaging the Foot and Ankle

The foot and ankle is a difficult area to image and ideally should be done with the patient supine and the foot at a right angle to the limb, ensuring that images are perpendicular to the long axis of the foot and that the metatarsals are seen as five circles of bone cut transversely [62]. MRI in patients with diabetes can demonstrate neuropathy related changes as well including muscle myopathy/atrophy which is indicated by high T1-weighted signals, bony fragmentation and resorption, changes in the marrow signals indicative of osteomyelitis (low T1-weighted signal) [62].

1.8.3 MRI in Diabetic Foot Disease

It is an invaluable diagnostic tool as it provides detailed, multiplanar overview of the soft tissues (sarcopenia, sinus tracts, collections or abscesses in the foot), bones (osteomyelitis, bone marrow oedema indicative of infection) and structure of the foot. In patients with diabetes, this becomes clinically relevant as it may dictate the route and duration of antimicrobials, warrant surgical debridement, biopsies or amputation. It is non-invasive, does not expose patients to radiation and is not operator dependant.

However, it is not infallible. It lacks portability and requires space and initial financial investment. MRI use requires trained technical experts who need to be able to operate the machines according to set protocols. MRI scans are contraindicated in patients with metal implants including certain cardiac pacemakers or defibrillators, aneurysm clips, neurostimulators, ocular metal fragments or shrapnel. Images may also be obscured by movement artefact.

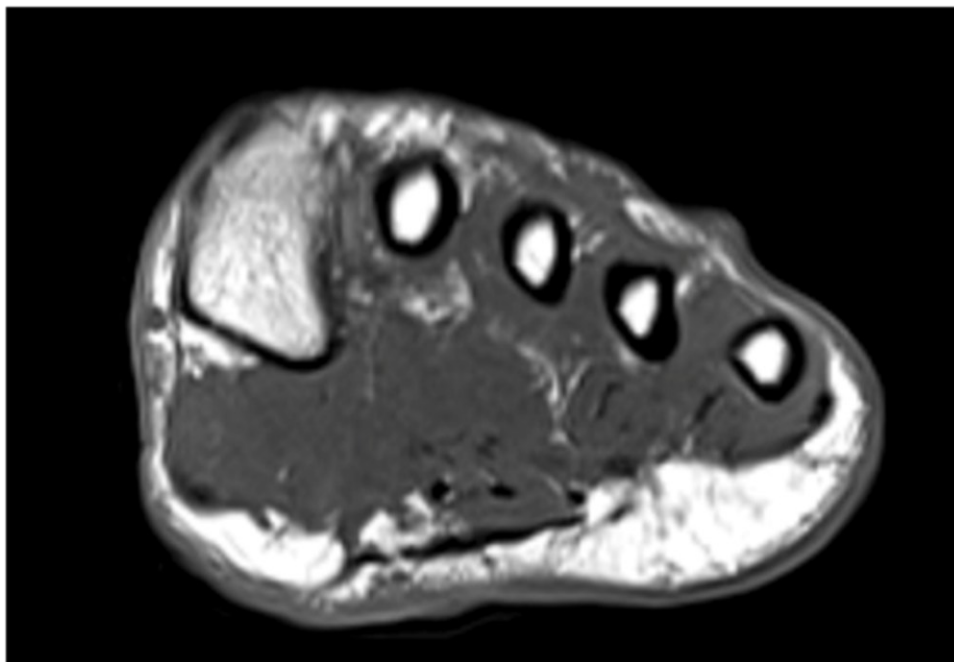


Figure 8a: Coronal view of foot at the base of the 1st Metatarsal on MRI (normal foot)

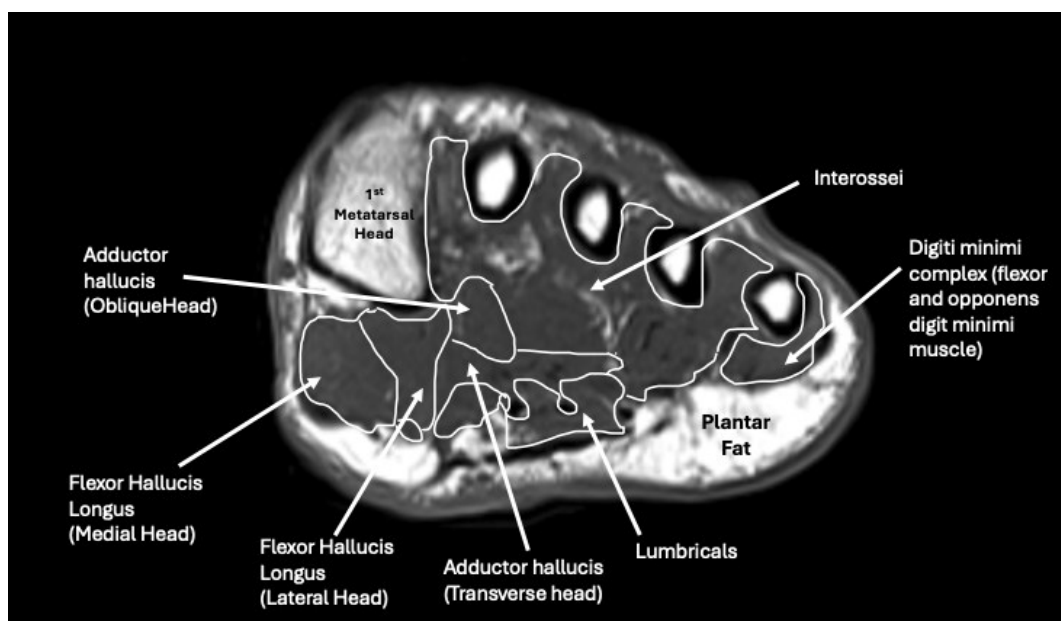


Figure 9b: Coronal view of foot at the base of the 1st Metatarsal on MRI (normal foot – annotated)

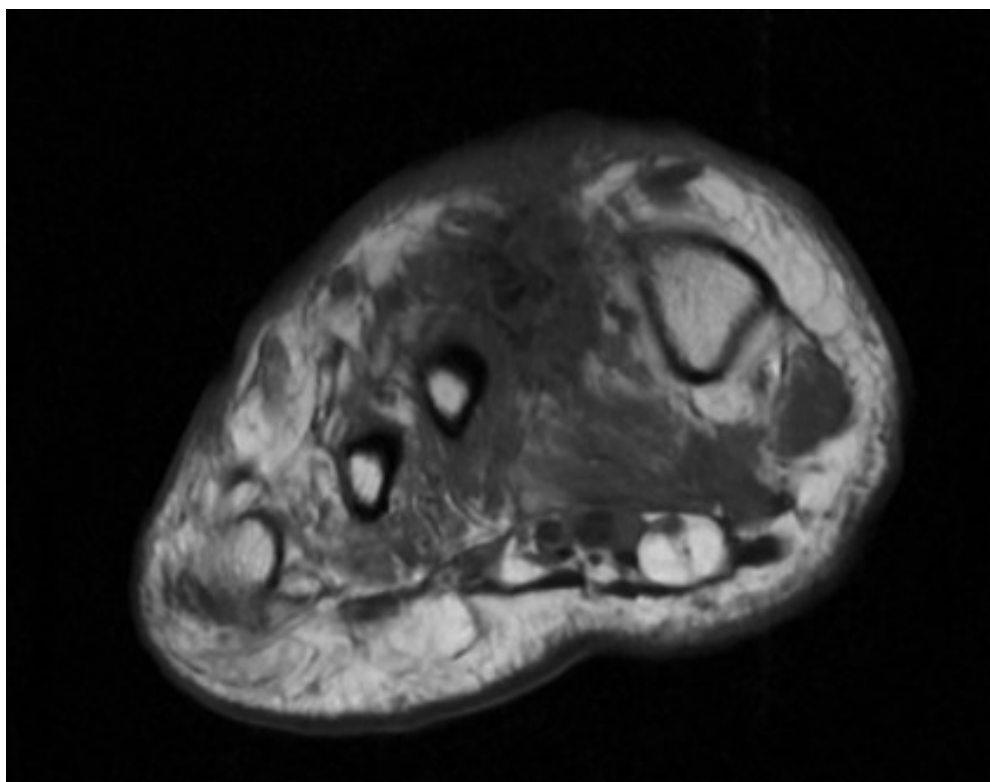


Figure 10: Coronal view of foot following amputation of the 2nd toe

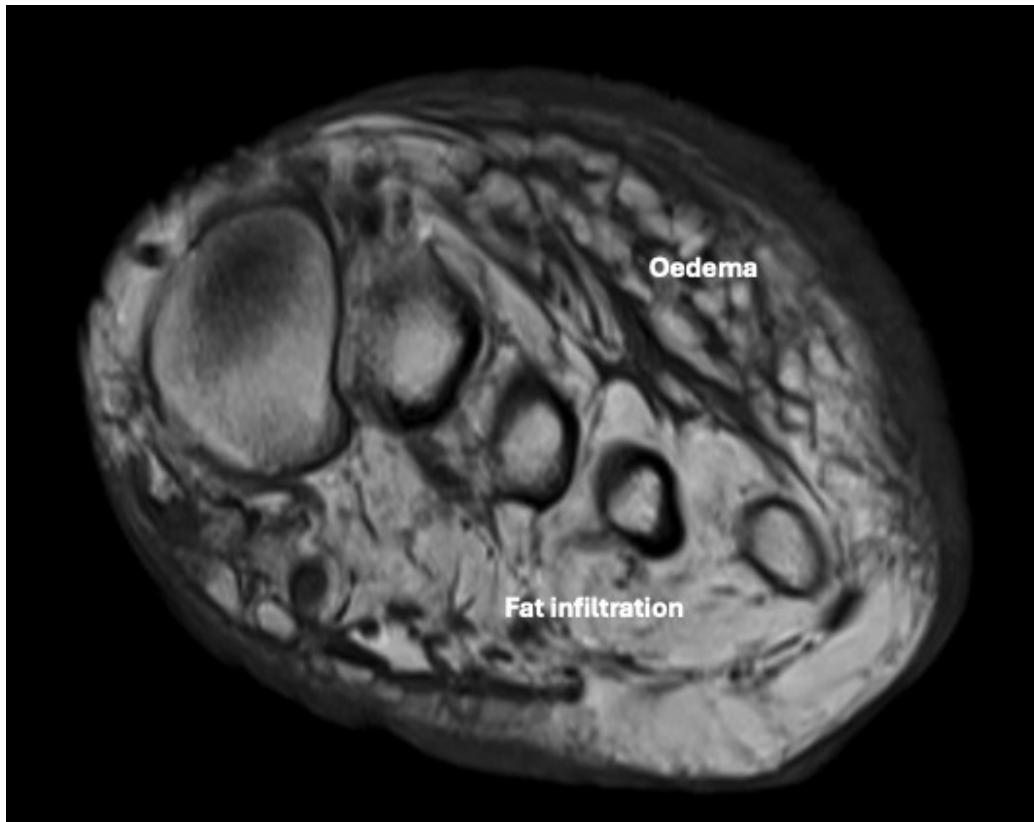


Figure 11: Coronal view of the foot demonstrating extensive oedema and fat infiltration in the muscle

1.8.4 Cost of MRIs

MRI's cost more than x-rays, but the overall associated costs are still lower than other advanced imaging techniques for the information it provides which includes information about the quality of the muscle and changes in the bone which may not be immediately evident on other imaging modalities such as computed tomography [59]. The NHS payment scheme for 2024/2025 cites the tariffs for MRI with reporting [51]:

Scan type	Cost
MRI scan of One Area, without Contrast, 19 years and over	£122
MRI scan of One Area, without Contrast, between 6 and 18 years	£205
MRI scan of One Area, without Contrast, 5 years and under	£292
MRI scan of One Area, with Post-Contrast Only, 19 years and over	£178

MRI scan requiring extensive patient repositioning	£247
--	------

Table 10: MRI costs

Whilst this may seem costly for a single investigation, it is on balance offset by the multiple late costs of DFD including wound clinics, transport, long inpatient stays, operation related costs and rehabilitation costs.

1.9 Sarcopenia

The term sarcopenia arises from the Greek terms σάρξ ‘sarx’ meaning flesh and πενία ‘penia’ meaning poverty or lack of. Sarcopenia is the age-related, progressive and global loss of skeletal muscle mass and function and is associated with distinct adverse clinical outcomes including disability, falls and mortality[63]. The term was introduced by Rosenberg in 1988 before being superseded by Baumgartner’s definition a decade later, which offered a method of measuring sarcopenia using height-adjusted appendicular lean mass [64]. Sarcopenia was also officiated as a muscular disease in the International Classification of Diseases (ICD) 10th edition in 2019 [65].

Current definitions of sarcopenia are the culmination of a group of collaborative consensus from the European Working Group on Sarcopenia in older people (EWGSOP2) 2018, The Asian Working Group for Sarcopenia (AWGS) 2019, the International Working Group on Sarcopenia (IWGS), Foundation for the National Institutes of Health (FNIH) Sarcopenia Project and Australian and New Zealand Society for Sarcopenia and Frailty Research (ANZSSFR) [63, 66]. Thus, sarcopenia is defined as the presence of the following criteria:

1. Low muscle strength
2. Low muscle quantity and/or quality
3. Low physical performance

The EWGSOP2 2018 operational definition confirms sarcopenia with the presence of two of these criteria, whilst the presence of all three is indicative of severe sarcopenia [63].

1.9.1 Impact of Sarcopenia

The overall prevalence is reported at 10% in all patients over the age of sixty [67]. There is also variability in community dwelling patients (1-29%) and those in residential care (14-33%) [68]. It is estimated to be directly attributable to medical costs of up to \$18.5 billion in

the US in 2000 and up to £2 billion in the UK has continued to grow with an overall aging and co-morbid global population [68, 69].

Whilst previously thought to be primarily present in the elderly, it has become apparent in the last decade that the process of sarcopenia can start as early as the third decade and can be accelerated by a number of factors including ethnicity, geography, nutrition (low protein intake), Vitamin D deficiency, poor lifestyle choices i.e. unhealthy diet, smoking alcohol, lack of physical activity, obesity and chronic inflammatory conditions such as diabetes [12, 68-73]. Men appear to have a higher rate of strength decline compared to women as has been reported consistently across multiple studies[68]. There is, however, discrepancy in terms of ethnicity and sarcopenia. Sarcopenia has a high prevalence in Asian populations [68, 74]. Black and Hispanic populations were also found to have higher lean muscle mass but lost more lower limb lean muscle mass compared to Caucasian populations [68].

As the concept of sarcopenia is understood better, it has also come to light that it can be acute or chronic. Acute sarcopenia (less than six months) is related to an acute illness or injury whilst chronic sarcopenia lasts over six months and is associated with chronic and progressive conditions [63]. This highlights the dynamic nature of sarcopenia and the need for periodic screening and assessment. Sarcopenia is associated with a higher risk of falls, fractures, hospital stays, functional decline and mortality – all of which collectively add the already existing burden within the healthcare systems globally [63, 68, 75].

1.9.2 Diagnosing Sarcopenia

Current classifications of sarcopenia are based on a range of anthropomorphic features and physical tests which assess the three domains i.e. muscle mass, muscle strength and physical performance. Initial screening is done using the SARC-F or the updated version SARC-CalF questionnaire which can be easily administered in the community [63, 67]. It assesses strength, assistance with walking, rising from a chair, climbing stairs and falls along with the calf circumference. The latter version has a 57.4%-65.5% sensitivity and is currently recommended by the EWGSOP2 classification [63, 67].

Another screening tool is the Mini Sarcopenia Risk Assessment questionnaire which has a short and long format. The long format evaluates age, recent hospital admission, physical activity, frequency of meals, dairy intake, protein intake and unintentional weight loss [67]. This has a slightly higher diagnostic accuracy of 85% compared to SARC-CalF which is 70% with a p-value <0.001 [67]. These screening tools can then allow for further assessment of the three domains of sarcopenia which have been adapted and summarised in Table 11 from the review by Voulgaridou et al [67].

Muscle mass

Muscle mass can be assessed with Dual-energy X-ray absorptiometry (DXA) which allows measurement of lean mass, bone mass and mineral density of the bone but fails to provide information about the quality of the muscle mass and results can vary based on hydration and the equipment being used [67]. Bioimpedence analysis estimates muscle mass and muscle fat through electrical conduction through tissues but does not provide direct information about the quality of the muscle [63, 67, 76]. The gold standard for evaluating body composition is CT and/or MRI [71, 77].

Muscle strength

A hand dynamometer is used to measure the handgrip strength and is the current gold standard method of measuring strength [63, 66, 76]. It is used as a proxy for global strength in the body.

Physical strength

Several tests have been suggested and incorporated into the existing diagnostic criteria of sarcopenia. These include gait speed, 4 meter- and 6 meter- gait speed, time up and go (TUG), stair climb power test (SCTP), five-time stand-to-sitting test (5TSST). These tests are essentially indicative of extremity strength and balance which, in turn is reflective of functional ability [63, 66, 67, 76].

International Working Group on Sarcopenia (IWGS)	American Foundation for the National	Asian Working Group for Sarcopenia (AWGS2)	European Working Group on Sarcopenia in Older People	Sarcopenia Definition and Outcomes Consortium
---	--	---	---	--

	Institutes of Health (FNIH)	(EWGSOP2)	(SDOC)
Muscle mass	DXA (ALM/height ²): <7.23 kg/m ² in men <5.67 kg/m ² in women	DXA (ALM/BMI): <0.789 kg/BMI in men <0.512 kg/BMI in women DXA (ASM): <7.0 kg/m ² in men <5.4 kg/m ² in women OR Bioimpedance Analysis(ASM): ≤7.0 kg*m ⁻² in men <5.7 kg*m ⁻² in women	DXA (ALM/height ²): <7.00 kg/m ² in men <6.00 kg/m ² in women OR BIA OR CT OR MRI
Muscle strength	-	Grip strength: <26 kg in men <16 kg in women	Grip strength: <28 kg in men <18 kg in women Grip strength: <27 kg in men <16 kg in women OR Chair stand test >15seconds
Physical performance	4MGS: <1.0 m/s OR Standing up from a chair (> or < than 15 seconds)	4MGS, 6MGS: <0.8 m/s 6MGS: <1.0 m/s OR 5TSST: >12 s OR SPPB: ≤9	Gait speed: <0.8 m/s OR SPPB: ≤8 OR TUG: ≥20 seconds OR 400 m walk: >6 min

Table 11: Summary of diagnostic criteria for sarcopenia. Adapted from Voulgaridou et al (65). 5TSST – 5 time stand-to-sit test; 6MGS – 6 m gait speed; ALM – Appendicular lean mass; ASM – Appendicular skeletal mass; BIA – Bioimpedance Analysis; BMI – Body mass index; CT – Computed tomography; DXA – Dual-energy X-ray absorptiometry; Kg – Kilogram; m – Meter; MRI – Magnetic resonance imaging; s – seconds; SCTP – Stair climb power test; SPPB – Short Physical Performance Battery; TUG – Time Up and Go

1.9.3 Diabetes and Sarcopenia

80% of all glucose clearance in the body is in skeletal muscle via the glucose transporter 4 (GLUT4) protein [12]. This is primarily in skeletal and cardiac muscle and is influenced by insulin. Aging contributes to shrinkage of myofibers as well as a shift from type II myofibers (fast-twitch and rely on anaerobic metabolism) to type I myofibers (slow-twitch and rely on aerobic metabolism) [12].

In contrast to young, healthy individuals, the oxidative capacity per muscle unit is halved in older people and with increasing oxidative stress, ROS, AGEs and inflammation, the process of sarcopenia is accelerated [12]. This process becomes cyclical as intramuscular fat accumulation increases adipokine secretion and low-grade inflammation, further perpetuating the insulin signalling and mitochondrial dysfunction in skeletal muscle, which further accelerates muscle atrophy [12] .

Elderly patients with diabetes accrue three times the intramuscular fat stores compared to their healthy counterparts [12]. There is an established bidirectional relationship between diabetes and sarcopenia [69].

1.9.4 Sarcopenia and diabetes related complications

The Health, Aging and Body Composition study correlated diabetes with rapid loss of skeletal muscle strength and mass [78]. The prevalence of sarcopenia may be as high as 15% in patients with diabetes [78]. In 2018, Kim et al reported the five year mortality rates in patients who underwent minor LLA with and without sarcopenia (57.7% versus 37%) and major LLA with and without sarcopenia (67.6% versus 33.3%) [78]. A 46 study meta-analysis found the duration of diabetes, In terms of risk, age, BMI, HbA1c, CRP, and diabetic nephropathy were probable risk factors of sarcopenia [74].

The Age, Gene/Environment Susceptibility (AGES) – Reykjavik study found increased intramuscular adipose tissue in the thigh muscles on CT was associated with an 8% higher mortality risk (HR = 1.08, 95% CI: 1.01, 1.16) [79].

Similarly, Cheng et al reported sarcopenia to be independently associated with more extensive DFD clinically including more foot ulcers, higher Wagner grades and higher percentage of amputations (21.4% versus 7.8%) [80].

A three-year prospective study published in 2025 from Australia found patients with low hand grip strength and low psoas muscle index on CT had a higher risk of death – hand grip strength (high versus low), hazard ratio = 0.46, 95% CI:0.22–0.997; psoas muscle index (low versus high), HR= 2.15, 95% CI: 1.17–3.96. The study also found patients had higher odds of wound healing if they had higher hand grip strength [81].

Yang et al also identified sarcopenia (HR 5.051, 95% CI:1.968, 12.961, P= 0.001) and serum creatinine (HR 1.007, 95% CI:1.003, 1.010, P < 0.001) as an independent risk factor for all-cause mortality in patients with DFUs [82].

Sarcopenia has also been identified as a prognostic factor for overall survival in patients with CLTI (HR = 3.22; 95% CI, 1.24–9.11; p= 0.02) [83]. Similar results from another study demonstrated that low muscle density [abdominal/thigh/lower limb muscles] on CT were significant predictors of amputation in patients with CLTI [84].

The impact of sarcopenia also extends to revascularisation cases in patients with CLTI and diabetes. A study with 116 patients undergoing lower limb (infrainguinal) open bypasses were found have higher amputation rates and lower overall survival when they had lower psoas muscle inDICES on CT [85].

The presence of sarcopenia in patients with CLTI is also associated with statistically significant higher in-hospital mortality as well as 30 day- and 6 month mortality [86].

Another study from 2025 with 72 patients with diabetes found that low psoas muscle and skeletal muscle inDICES were associated with low overall survival rates (PMI/SMI: HR 4.120, 95% CI: 1.692–10.032/HR 2.487, 95 % CI: 1.091–5.666) and amputation free survival (PMI/SMI: HR 3.561, 95% CI: 0.938–13.516/HR 3.982, 95% CI: 1.080–14.677) [87].

1.10 Sarcopenia in the foot on magnetic resonance imaging in patients with diabetes

In the normal, non-sarcopenic foot, skeletal muscle is present in a high ratio compared to the surrounding fat (figures 8 and 11). With progressive sarcopenia, fatty infiltration within and around the muscle changes the overall composition of the foot (figures 10, 12-14).

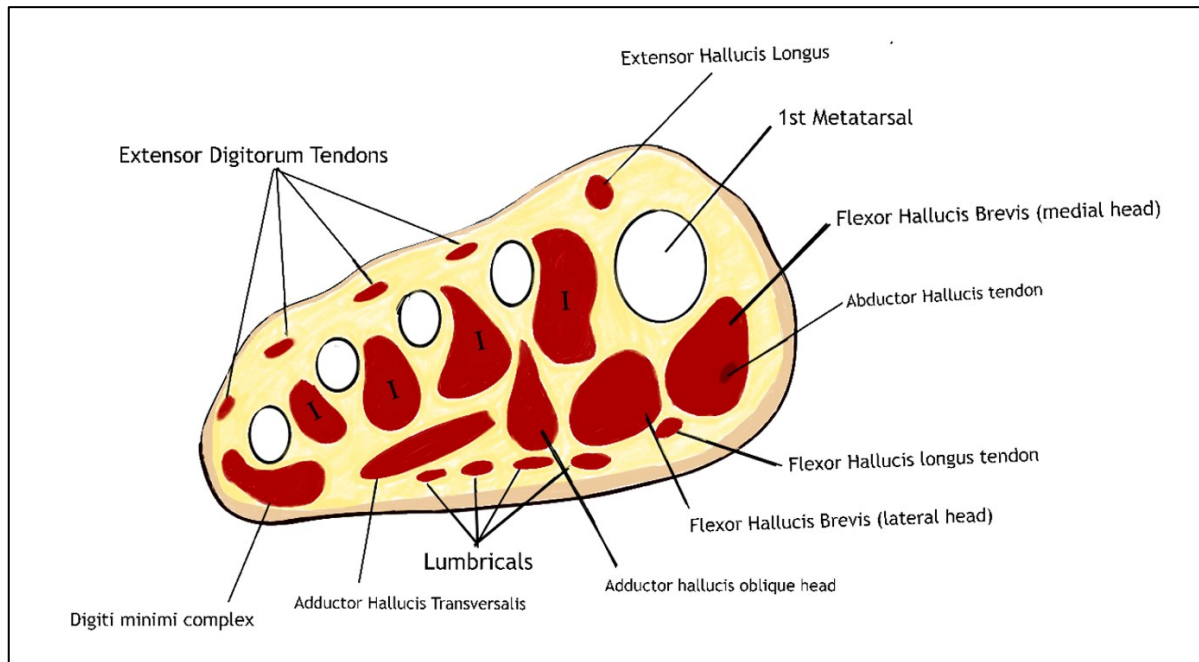


Figure 12: Schematic representation of a coronal view of the normal foot without sarcopenia. Adapted from Ahmad et al (29)

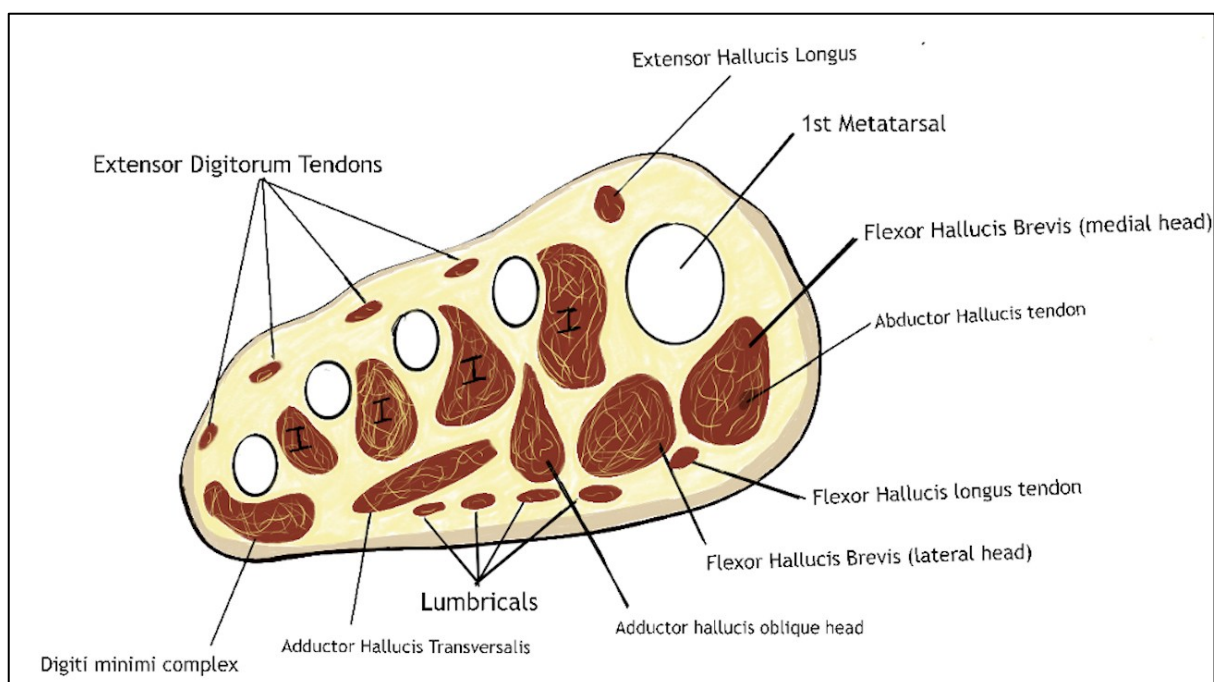


Figure 13: Schematic representation of a coronal view of the foot with mild and early sarcopenia. Adapted from Ahmad et al (29)

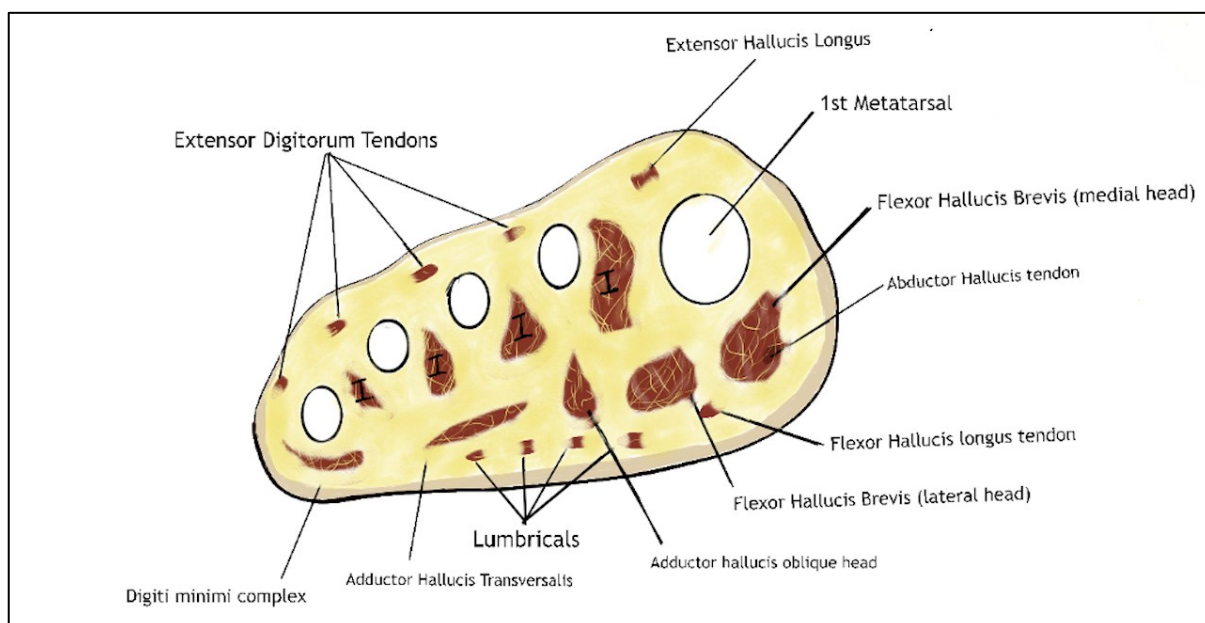


Figure 14: Schematic representation of a coronal view of the foot with moderate sarcopenia. Adapted from Ahmad et al (29)

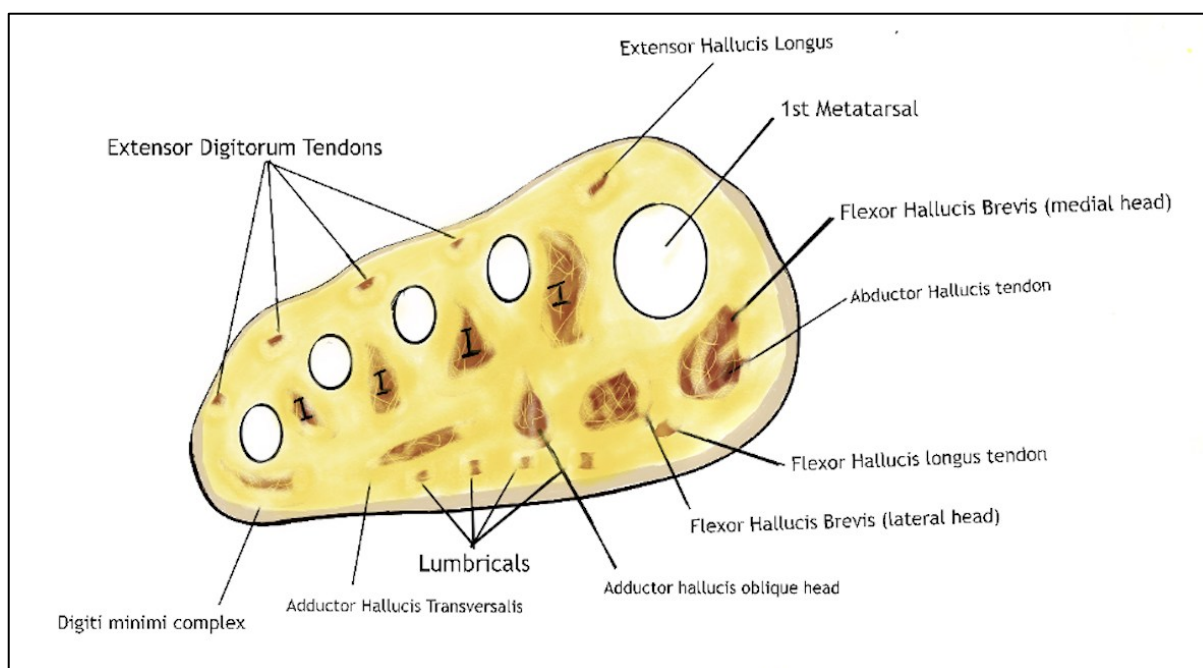


Figure 15: Schematic representation of a coronal view of the foot with severe sarcopenia. Adapted from Ahmad et al (29)

A systematic review from 2024 found seven studies evaluating sarcopenia in the foot and comprised of seven case-control studies and a single cross sectional study [29]. A pooled total of 234 patients (129 males and 110 females) were included of which 68.5% were diabetic. The vast majority of the study evaluated the intrinsic muscles of the foot and found sarcopenia was positively correlated with the extent and duration of diabetic polyneuropathy. Across all studies, reduction in the skeletal muscle volume and increased fatty infiltration was noted in patients with diabetes [88-94]. These findings are summarised in table 12.

Author	Year	Muscles Measured	Method of Measurement	Control	Diabetic (non-neuropathic)	Diabetic (Neuropathy)
Andersen [88]	2004	Intrinsic muscles of the foot	Cross sectional area (CSA) - Stereological total volume	168+/-42 cm ³	165+/-34 cm ³	86+/-52 cm ³
Andreassen [89]	2009	Intrinsic muscles of the foot	Cross sectional area (CSA) - Stereological total volume	Annual loss of muscle volume: 0.2%	Annual loss of muscle volume: -1.1% (-4 to -0.2, p<0.001)	Annual loss of muscle volume: -2.0% (-4 to -0.2, p<0.001)

				(-2.5 to 2.4)		
Bus [90]	2002	Intrinsic muscles of the foot	Compositional analysis of the soft tissue (pixel by pixel parametric T 2 colour maps) to measure the cross-sectional area	30.8% [SD +/- 3.9%] (range 24.4–36.8)	-	8.3% [SD+/- 2.9%] (range 3.7–11.8)
Bus [91]	2009	Intrinsic muscles of the foot, EDL, FDL	5-point grading atrophy score	-	2.6 +/- 1.2 (non-deformity group)	3.1 +/- 1.1 (deformity group)
Cheuy [92]	2013	Intrinsic muscles of the foot	Segmentation and volume measurement	-	-	-
Greenman [93]	2005	Intrinsic muscles of the foot	Stereological total volume	0.55 [SD+/- 0.04] p<0.0001	0.44 [SD+/- 0.05] p<0.0001	0.06 [SD+/- 0.06] p<0.0001
Lue [94]	2022	Abductor Hallucis	Visual grading scale - Goutallier Grading (Fatty Infiltration)	None 85.7% Diffuse 11.9% Patchy 2.4%	-	None 22.1% Diffuse 70.6% Patchy 7.4%

Table 12: Summary of sarcopenia on MRI in the foot. Adapted from Ahmad et al (29)

Five primary methods of measuring sarcopenia in the foot on MRI have been reported.

1.10.1 Stereological total volume measurement

Stereology is a branch of morphometry which allows acquisition of three-dimensional data from two-dimensional sections of any given structure by evaluating small sections of an area of interest called “slices” or “slabs” [95]. It draws on Cavalieri’s principle which states that if two three dimensional structures have the same height and width/cross sectional area at every level, then they have the same volume (figure 11). The surface area at different slices is calculated as well as the distance between each slice. The two values are then multiplied to give an estimate of the overall volume. This method was used by Andersen et al and

Andreassen et al who reported a significant loss of volume in patients with diabetes and particularly in patients with diabetic polyneuropathy when compared to controls [89, 96].

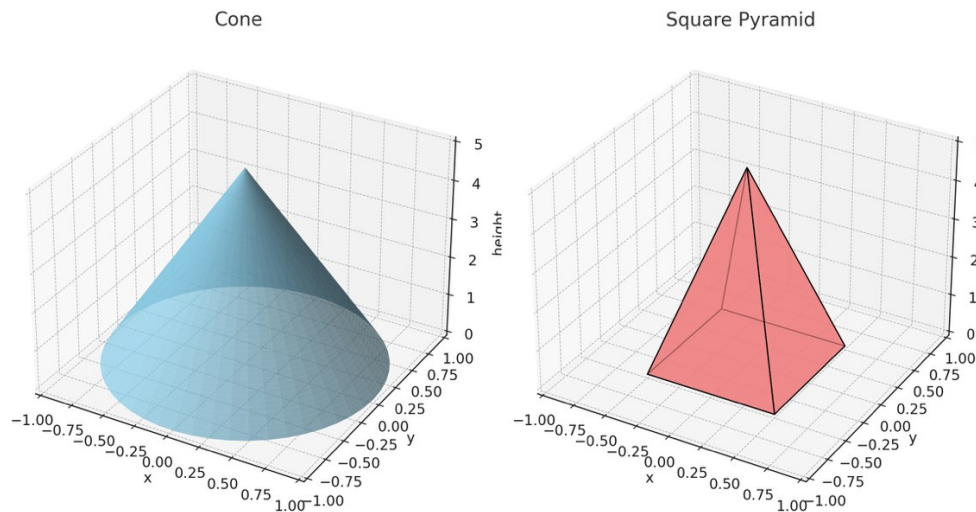


Figure 16: Illustration of Cavalieri's principle

1.10.2 Soft tissue compositional analysis

Extensive inter- and intramuscular fat in sarcopenia can lead to challenges in delineating muscle from fat. This may be overcome by using pixel based T1 colour maps based on echo spins and by creating a threshold cut-off point to help determine the extent of muscle and fat [29]. This is done by establishing a threshold cut-off indicating fat whereby readings below this can imply muscle.

1.10.3 Segmentation based volume measurements

Segmentation of a select group of muscles by rater/s at established anatomical slices from two dimensional renditions can also be utilised. The voxel which is a three-dimensional rendering of a two-dimensional image i.e. a single data point in a three-dimensional grid, calculates the signal intensity in the skeletal muscle and as well as the overall cross-sectional area within that region of interest. This could then be used to create a histogram of the signal intensities and help establish threshold cut-offs to differentiate muscle from fat [29]. In 2013, Cheuy et al attempted to use semi-automated process for MRI based volumetric analysis of the intrinsic foot muscles in 19 subjects with diabetes and associated neuropathy

[92]. This involved manual segmentation of MRI scans from 19 subjects at the forefoot, midfoot and hindfoot by two raters (both engineers trained in anatomy) which was then used to calculate the volume. The study demonstrated intraclass correlation co-efficient (ICC) values of over 0.95 which implies that it may be a possible valid method for MRI based volumetric analysis.

1.10.4 Five-point atrophy scale

This five point scale was proposed by Bus et al as a visual grading system [97]. The study graded sarcopenia as zero (no sarcopenia) and one to four indicating mild, moderate, severe and complete loss of muscle, respectively, as evaluated by two experienced radiologists. Measurements were taken at metatarsophalangeal joint, toe angle and also reviewed the plantar fat pads in twenty-eight subjects of which twenty-three had established diabetic polyneuropathy. The ICC intraobserver agreement for intrinsic muscle atrophy was 0.94 [97]. This method was subsequently used by the same group in other studies but has yet to be validated externally [91, 98].

1.10.5 Goutallier Grading

This four point grading scale was suggested by Goutallier et al but was initially intended for grading fatty muscle degeneration in rotator cuff injuries on CT. The grading system ranged from zero (normal muscle) to grade four where more than fifty percent of the muscle was infiltrated by fat [99]. There is significant variation in the grading reliability across specialities i.e. radiologists versus orthopaedic surgeons but not due to experience or reporting frequency [29].

1.11 Artificial Intelligence

The Oxford English dictionary defines artificial intelligence (AI) as the “theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages” [100]. The AI market in health care is expected to be worth \$45.2 billion by 2026, with patient data and risk analysis capturing a significant share [101].

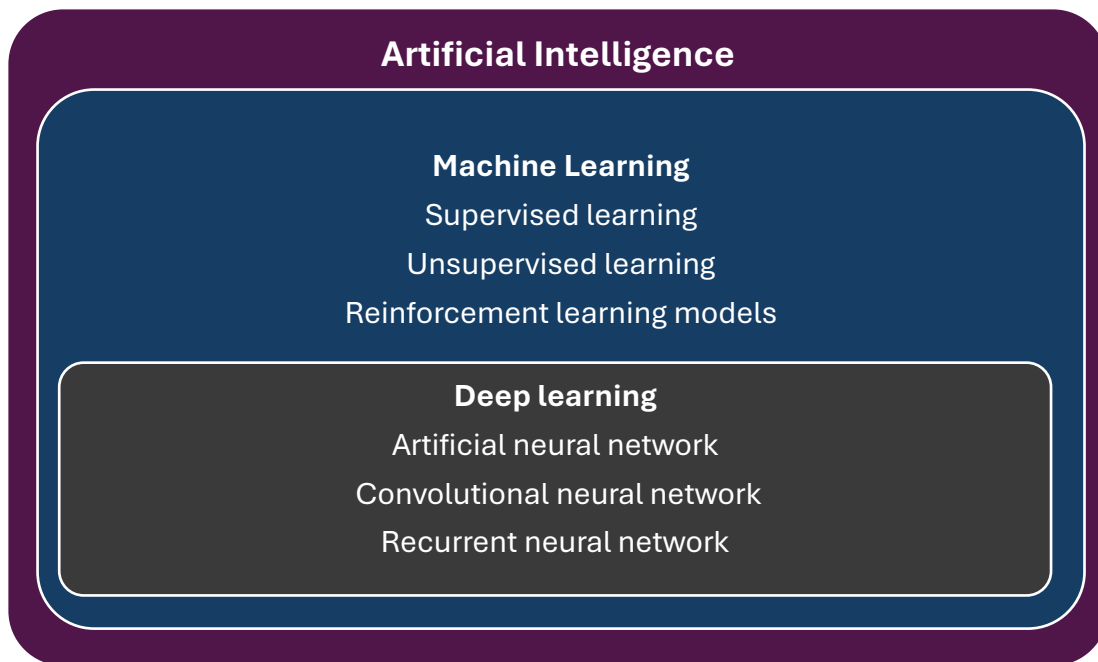


Figure 17: Subtypes of artificial intelligence

1.11.1 Machine Learning Models

The terms Machine Learning (ML) and Deep Learning (DL) are nested within the broader term of artificial intelligence. ML is a subset of AI which uses statistical analysis to execute certain tasks. This process includes supervised learning models, unsupervised learning models and reinforcement learning models [102]. Supervised learning can include classification or regression tasks. Unsupervised learning models use unlabelled datasets to identify similarities and cluster datapoints [102]. Reinforcement learning models use a reward or penalty mechanism to encourage positive decision-making based on critical feedback and are based on game theory but are not commonly used in the realm of medicine.

1.11.2 Deep Learning Models

Deep Learning (DL) draws on the architecture of human body's central nervous system. These neural networks contain structures similar to the network of individual neurons in our brains. The individual decision unit (neuron) is called a "perceptron." A dense architecture of thousands of perceptrons allows information - the 'input' - to be filtered through multiple hidden computational layers to extract complex information, formulate decisions and give an output. Each connection within the hidden layer has a weighting which is adjusted as the

model is trained. There are several subtypes of DL models. This includes artificial neural networks, convolutional neural networks, recurrent and recursive neural networks [102].

Artificial Neural Networks (ANN) are the basic DL models for general regression or classification tasks and work using the input, hidden layer and output format. These are useful for general purpose DL model implementation.

Recurrent neural networks are for sequential data where time and order matter and are used for time series and speech [103]. Recursive neural networks work on hierarchy and combine smaller parts into an overall larger representation through parsing which is the conversion of raw input into a ML format [102]. These are primarily used in natural language processing and sentiment analyses.

Convolutional neural networks (CNN) are designed for spatial data and are ideal for two- or three-dimensional images as they automatically detect patterns including edges and shapes. They are the ideal choice for object recognition and image detection [102, 104].

1.11.3 Radiomics and AI

Radiomics is the conversion of digital medical images into minable, high-quality datapoints for further evaluation [105]. There has been an exponential rise in the number of studies combining radiomics with clinical, multidimensional data [102]. Implementation of AI in vascular surgery continues to evolve with improved imaging modalities, lower costs, availability of data, improved data storage and processing capabilities, as well as the availability of low coding platforms whereby minimal coding skills are required as well as coding repositories such as GitHub.

Three-dimensional imaging has been used in vascular surgery in the surveillance, grading, and prognostication [106]. In the realm of diabetes and its related areas, ML models have been applied to blood glucose management, predicting adverse glycaemic events and diabetic complications and as a decision support system for monitoring lifestyle changes [29, 106, 107]. The use of ANNs have been applied in the classification of peak plantar pressures in high-risk regions of the foot [13]. DL models have also been utilised for plantar

image analysis for detecting gait abnormalities [108]. The number of studies incorporating deep learning models and DFD has risen in the preceding decade but are primarily based on two dimensional imaging such as for ulcer detection, wound classification and analysis of plantar thermography images [102].

A systematic review from 2024 identified a single study using three-dimensional imaging – CT angiograms applied to an ANN to predict DFU prognosis i.e DFU resolution or amputation [102, 109]. The study combined CT reports along with patient metadata (age, sex, BMI, duration of diabetes, duration of DFU and limb symptoms to predict the outcome. The group reported accuracy of 91.6% in the training set and 88.9% in the validation set. The receiver operator characteristics area under the curve was 0.955 [109]. This, however, has not been externally validated and has not been replicated.

The availability of open-source coding and software platforms for image analysis has allowed for expansion in the area of radiomics based research. Semantic segmentation is the process of using raw biomedical data to create spatially structured information [110]. Several years ago, this would have required heavy input from engineers and computational teams. In 2020, Isensee et al created nnU-Net, a deep learning based empirical segmentation model which was self-configuring and therefore provided an out-of-box solution including the preprocessing, network architecture, training and post-processing for a given task [110]. The code adapts to CT, MRI and histological images and is capable of handling a vast variety of target structures and image properties having been initially trained on 23 different datasets.

1.11.4 Evaluating the performance of convolutional neural networks

Several concepts and terms are relevant when measuring segmentations and gauging the accuracy of the CNN. These are discussed below:

Ground truth

This is the verified or true data and is representative of the accurate data upon which ML models are trained [111]. This may be done by labelling the data if numerical or

alphabetical. In the case of two- or three- dimensional images, the region of interest may be labelled or segmented by a human. This sets the rules for the boundaries or regions within an image which the ML model is taught to use. Ground truth provides the benchmark for what the accurate predictions are supposed to be. An example of this is segmenting various muscles in the lower limb by a human. This then teaches the ML model to identify those areas based on the manual areas segmented in the training phase. During the validation and testing phase, it is the ground truth against which the performance of the ML model is evaluated.

This is not without its challenges. Ground truth is subject to human interpretation and judgement which is subjective and may vary between one person and another. This can lead to inconsistencies which may have an impact on the overall behaviour of how a model performs [111]. Another aspect to consider is the dataset being used as it may be skewed and or biased and might not always be applicable to real-world application when exposed to a new testing set.

Cross-entropy

Cross-entropy (CE) measures the difference between two probability distributions and the amount of loss between the actual label/segmentation and that predicted by the ML model. In the realm of ML, CE measures how well the probabilities predicted actually match the class labels. A high CE indicate the predictions are close to the true distribution whilst the converse is true for low CE.

An example of this is if a model states the intrinsic muscles = 0.9 when it is true. This implies the CE loss is low. If the ML model were to state the intrinsic muscles = 0.1 when the region of interest is still the intrinsic muscle, the CE loss is considered to be high. CE loss can be improved by using more training data, using clear images and optimising the parameters of the ML model.

DICE-Sørensen coefficient/DICE score

This is a statistic used to gauge the similarity between two samples. In recent years, it has gained popularity as a performance metric in medical image segmentation. It was first

proposed by Zijdenbos et al in 1994 for white matter lesion segmentations in the brain [112]. In image based CNNs where images are segmented, the DICE score is used to compare the ground truth with the model's predicted areas [112].

This is calculated by multiplying the area of overlap (between the ground truth and predicted area) and dividing this by the total area (figure 17). It also captures the true positive, false positive and false negative predictions using the following formula:

$$\text{Dice Score} = \frac{2 \times \text{True Positive}}{(2 \times \text{True Positive}) + (2 \times \text{False Positive}) + (\text{False Negative})}$$

The score ranges between zero and one. A score of one implies a perfect overlap between the ground truth and the model's performance whilst a score of zero indicates no overlap.

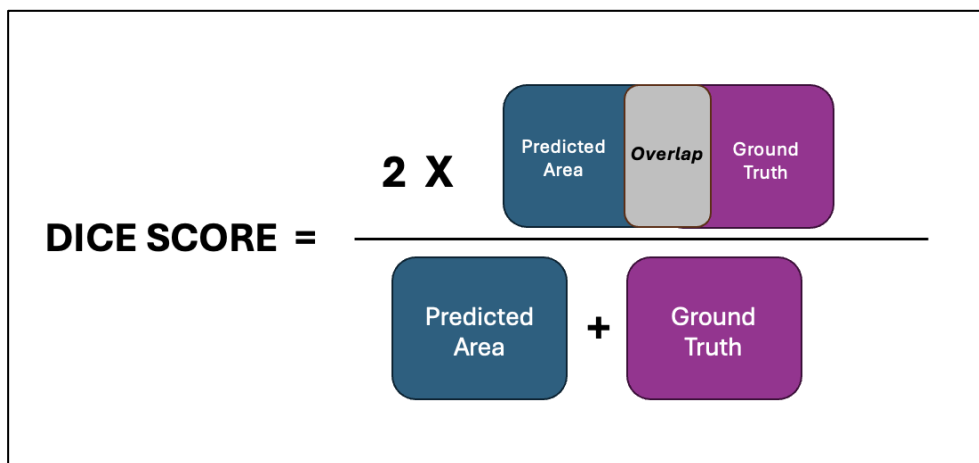


Figure 18: Calculating the DICE score for segmentations

Jaccard Index/Intersection over Union (IoU)

This is another statistical measure to compare diversity and similarity between two sets but is used less frequently in medical imaging where it can be used for granular evaluation by comparing pixel sets [112].

This is calculated by dividing the number of elements common to two sets called the intersection (such as the ground truth and the predicted area) by the number of unique elements in each set called the union (figure 18).

The score ranges between zero and one. A score of one implies the two sets are identical whilst a score of 0 means the sets do not have any common elements.

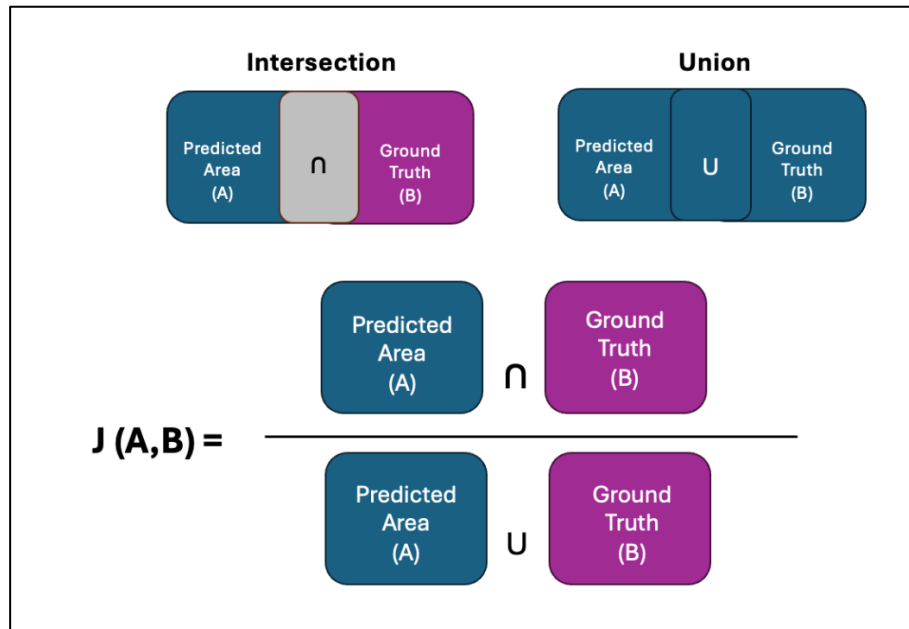


Figure 19: Calculating the Jaccard index

Kim et al used this methodology in 2024 for individual muscle segmentation in seventy-two whole thigh CT scans from patients with hip fractures to assess for sarcopenia [113]. The parameters evaluated included DICE score, average symmetric surface distance, volume correlation, relative absolute volume difference and reported a DICE score of up to 0.94. [113]. The group also reviewed the correlation between sarcopenia and individual thigh muscle volumes was and found a correlation between sarcopenia and individual thigh muscles in the male cohort.

1.11.5 Challenges in the use of artificial intelligence for research

Challenges in the area of AI related research have yet to be fully resolved. These include small datasets, lack of clear provision of coding (although this is now being overcome by coding repositories), the black box nature of neural networks, security/copyright- and ownership-related matters with the advent of General Data Protection Regulation.

The research community has been cognizant of these and have incorporated these in the 2020 consensus statement – ‘Guidelines for clinical trial protocols for interventions involving AI: the SPIRIT-AI extension’ [114]. The consensus statement asks researcher to clarify the clinical context and intended use, procedure for acquiring and selection of data, output from AI use, identifying performance errors, access to the code or restrictions to access and/or reuse [114].

1.12 Research priority and rationale for this study

Reducing the number of amputations has been highlighted by several national and international bodies. The Diabetes UK Research Steering Group 6 identified amputations as a top priority in addressing diabetes related complications.

In 2021, the Vascular Society of Great Britain and Ireland amputation research special interest group in association with the James Lind Alliance priority setting partnership conducted two Delphi processes involving a multidisciplinary clinician survey and patient survey. This resulted in responses from 481 healthcare professionals and 373 vascular patients and carers [115]. The primary clinician priority question was “How can we reduce the rates of major lower limb amputation?” whilst the patient/carer priority was “In a person who has undergone amputation, how do you reduce the chances of amputation in the other limb?” [115]. A Canadian Delphi survey with 31 respondents also sought to identify priorities and strategies to help build capacity for research related to amputations [116]. This identified building a national dataset to obtain data to help improve strategies related to measuring outcomes in patients undergoing amputations [116].

The following year, the Vascular Society of Great Britain and Ireland diabetic foot research specialist interest group joined forces with the James Lind Alliance to identify areas for research priority in diabetic foot disease [117]. Amongst the 12 ranked priorities, the areas highlighted included further research into the most effective way of preventing DFUs, the most effective way of preventing further amputation after toe amputation for DFD, improving outcomes in diabetic patients with foot infection, preventing recurrence, improving risk assessment, if the annual foot check was worthwhile and which factors

affected healing times in DFD [117]. These highlight the clear public and professional need to proactively anticipate DFD in high-risk patients.

Early detection and risk stratification is imperative to identify high risk patients through identification of potential surrogate markers which may help predict the future likelihood risk of amputation. Specific changes can be detected on MRI prior to the development of clinical symptoms. These include changes to the quantity and quality of the muscle (sarcopenia), changes in the architecture of the foot and the presence of quiescent infections in the bone. The direct impact of this project would allow detection of patients at risk of developing diabetic foot disease prior to the development of clinical symptomatology by exploring the interaction between patient factors and existing findings on magnetic resonance imaging (MRI).

MRI's of the foot are recommended in the current UK NICE and IWGDF guidelines to exclude certain pathologies discussed earlier however, ML and DL models have not been applied to these to evaluate the risk of diabetes related complications and is an area worth exploring.

This work has the potential to benefit asymptomatic and the minimally symptomatic population of patients with diabetes as early detection would allow active surveillance, risk factor mitigation, medical optimisation and lifestyle changes. Early sarcopenia is reversible with dietary modification (higher protein intake, better glycaemic control, resistance training) [12]. It would also allow monitoring in patients with diabetes who have had minor amputations to reduce the risk of major limb amputation or disease development in the contralateral lower limb. Furthermore, this may help to directly alleviate the multispecialty burden on secondary and tertiary healthcare resources required for managing diabetic foot disease.

1.13 Aims of this thesis

The aims of the project are classified into the primary, overarching aim and the secondary aims which are detailed below:

Primary aim:

To use deep learning models to identify and measure surrogate marker/s which may predict limb loss in the form of minor or major amputation in patients with diabetic foot disease. This study also aims to identify the correlation between patient demographics, risk factors and their MRI findings to establish the risk of limb loss.

Secondary aims:

Sarcopenia

- To establish the severity of sarcopenia in patients with and without diabetes.
- To explore measuring tools to define and measure the extent of sarcopenia in the foot including area and severity (i.e. extent of fat infiltration).
- To validate methods to measure the extent of sarcopenia in the lower limb.
- To measure the accuracy of ML models in measuring sarcopenia when compared to current measurement methods including visual scales and fat to muscle ratio measurement.

Metadata

- Correlate the imaging findings with patient metadata.
- To evaluate the correlation between sarcopenia and patient factors including control of diabetes, co-morbidities, social history and biochemistry.
- To validate the deep learning model and establish the inter-rater reliability with human radiology measurements.
- To create a risk prediction model and establish the validity of the risk prediction model to aid risk stratification of patients at risk of limb loss.
- To utilise deep learning models to attempt automation of this tool.

Chapter 2: Methodology

2.1 Introduction

This chapter aims to discuss the preamble to work undertaken as part of the thesis. As discussed in the preceding chapter, there is paucity of data with regards to the use of surrogate markers on MRI to predict outcomes in patients with DFD.

2.2 Patient and Public Involvement (PPI)

A research project will be better designed and more impactful when it is shaped by the population to whom it is relevant. The Health Research Authority (HRA) define Patient and Public Involvement (PPI) as research done ‘with’ or ‘by’ the public, not ‘to’, ‘for’ or ‘about’ them [118]. PPI forms the foundational layer of a research question. It integrates the perspective of the patients and members of the public and is a gauge of the vox populi with the intent to improve the relevance and applicability of the outcomes.

The National Institute for Health and Care Research (NIHR), Chief Scientist Office Scotland, Health and Care Research Wales and the Public Health Agency Northern Ireland introduced the UK standards for public involvement to provide a framework for effective public involvement [119]. This is comprised of six domains and are outlined below:

- Communications – Using plain language for well times and relevant communications.
- Governance – Involving the public in research management, regulation, leadership and decision making.
- Working together - Work together in a way that values all contributions, and that builds and sustains mutually respectful and productive relationships.
- Inclusive opportunities – Establishing public involvement partnerships which are accessible and include a range of people and groups, as informed by community and research needs.
- Support and learning - Offer and promote support and learning opportunities that build confidence and skills for public involvement in research.
- Impact - Seek improvement by identifying and sharing the difference that public involvement makes to research.

The HRA further provides four key principles for meaningful PPI in health and social care research. These include involving the right people, involving enough people, involving those people enough and describing how it helps [118].

PPI may be approached in a number of ways based on the nature of the project and the setting. This may include questionnaires, structured or unstructured interviews, engaging with key stakeholders through community platforms and patient groups.

PPI for this project was completed using three complementary methods:

- Administration of structured questionnaires.
- Facilitation of patient group discussions.
- Direct engagement with patients attending the weekly Diabetic Foot clinics.

The use of questionnaires enabled the collection of quantifiable data on patient priorities whilst group discussions offered an open forum for in-depth discussion which shed light on the shared patient perspectives and facilitated dialogue. Furthermore, direct engagement with patients in the clinical setting allowed context-specific insights and themes to emerge which might not have necessarily become apparent in patients who had not been affected by DFD. These approaches collectively ensured that the perspectives of patients were integrated into the research process in a manner which helped to improve the relevance and acceptability of the project.

2.2.1 Survey questionnaire

An initial seven item questionnaire with further space for comments was circulated in the vascular outpatient clinics. This was undertaken over a four-week period in March 2023. This included patients with diabetes with no previous LLAs, patient with diabetes with LLAs, as well as non-diabetic patients. Twenty-six respondents completed the questionnaire. The responses are summarised in Table 13.

Question	Yes	No	N/A
----------	-----	----	-----

Would you be willing to have a scan of your legs at regular intervals if it helps to identify that you may at a higher risk of developing foot problems related to your diabetes?	57%	35%	8%
Would you want to know if you are at a higher risk of developing foot problems and amputation?	65%	23%	11.5%
Would this information change your perception of your diabetes and how well controlled your blood sugars are?	62%	34.6%	8%
Would you be happy for artificial intelligence or a smart computer to monitor and detect these features on the scan of your legs?	62%	27%	11.5%
Would you trust a smart computer to read your scans to detect high risk features which may predict your risk of developing foot problems and amputation?	62%	23%	15%
Would knowing your risk of future amputation change your actions towards managing your diabetes?	73%	15%	11.5%
If a smart computer is able to predict your risk, your management of diabetes may change, for example closer monitoring of your blood sugars, monitoring and/or changes to your medications to optimise your blood sugar levels, more regular foot checks. What are your thoughts on this?	see discussion		

Table 13: Responses from the PPI survey questionnaire

Almost two-thirds wanted to know if they would be at a higher risk of developing foot problems and that it would change their perspective on how their diabetes was managed. 62% were happy to have the involvement of AI and smart computers to detect high risk features on a scan. 73% of all respondents agreed that they would alter their actions in terms of managing their diabetes if they knew their future risk of amputation.

The final question provided space for free text to capture qualitative responses. These provided fairly polarised views but the majority of these were positive including statements such as “fine by me”, “good”, “that’s fine”, “I’m all for it”, “agreeable”, “I think this is great” and “I have to protect what I have left”. However, there was some reservation from some respondents including “I would approach such results with caution” and “it’ll never follow the choices of a real doctor”. It is worth noting that respondents completing the survey

were mostly all over the age of fifty and repeating the same questionnaire now and perhaps in a younger cohort might yield different sentiments.

2.2.2 Open forum discussion

A two-hour discussion session with twenty-one members of the Hammersmith and Fulham Diabetes Group (part of the Know Diabetes Group) was also undertaken at Charing Cross Hospital. The group consisted of members over the age of forty with recently diagnosed and with long standing diabetes. All individuals were actively engaged in managing and optimising their diabetes. The group also included patients who had DFD and subsequent amputations.

The session included an introductory presentation of the problem (DFD and LLAs) and what the research question aimed to address. This was followed by open floor discussion around the various aspects of the project. This provided space for members to ask questions about how it would be integrated into their current management in future, the implications it would have and the relevance of scanning which helped to shape the overarching aims of this thesis.

Patients highlighted issues around variability in accessing resources based on their geographical location including foot checks, local groups for diabetic patients and access to primary care which they felt had an impact on their overall care.

Interestingly, the session highlighted the lack of information about the disease trajectory pertaining to DFD and the fatal consequences of ulceration. It was interesting to see how a majority of the members were alerted to the gravity of the problem when shown visual references of the disease. We discussed how education around this could be improved in the community and amongst the general public and several ideas were proposed by the members including a visual timeline of the disease trajectory, accessible information and easy to understand format for sharing information.

The group unanimously agreed that they would attend for serial MRI scans in the future if it could help flag up their future risk of LLAs and place them in low, medium or high-risk

categories. During the session, we also extensively discussed the use of artificial intelligence in the interpretation of scans and predicting risk. The group had varying opinions on this and was reflective of the survey questionnaire in that the views were polarised across the group. Certain members who had previously been involved in research, accounting and mathematics were interested in knowing how this pilot project would work, how many scans this project would be using and validity. The group differed in their opinion on whether AI should be used on its own for diagnosis, with the majority wanting to have a level of confirmation and certainty in how accurate the AI model would be which I discussed would be a priority and incorporated in the fabric of this project. The majority of the group was in favour of the use of AI, especially when it could be confirmed by radiologists. They highlighted it could potentially reduce human errors and help with doctors who would otherwise have a lot of scans to review but would be very binary in its diagnosis and may not always factor in human aspects which I thought was interesting as this was perceived as a positive and a negative point.

The group also highlighted the potential issues around data protection and malware which I assured would be addressed before any new technology would be used in a clinical setting as it undergoes robust review prior to implementation. The session provided extremely useful insight into how patients and members of the public perceive the project and has allowed implementation of the feedback in the evolution of how the project was shaped, in particular the role of verifying the accuracy of AI in patient care which was a prevalent theme.

2.2.3 Diabetic Foot clinics

Imperial College Healthcare NHS Trust has daily diabetic foot clinics. This includes multidisciplinary DFD clinics which is comprised of Endocrinologists, Orthopaedic Surgeons, Vascular Surgeons, Microbiologists and Podiatrists. This allows for a one-stop clinic for patients with ongoing DFUs and Charcot Foot to attend where they can be managed regularly. There are also weekly Vascular wound clinics where patients who have undergone DFD related LLAs can be followed up on a regular basis following operative intervention.

The vast majority of patients attending had pre-existing or new ulceration or had undergone an amputation and were keen to engage with the service. Working within the team and interacting with the patients on a weekly basis allowed me to fully understand the gravity of the disease, the long-term implications of DFD and LLAs and the time-consuming effort required by patients and the parent teams in achieving wound healing and ambulation.

2.2.4 Emerging themes from PPI

During the PPI process, it became apparent that there was a lack of understanding of the gravity of DFD and patients wanted more information when informed about it. There was invested interest in identifying methods to mitigate the risk of DFU, DFD and LLA in patients with and without diabetes and more so in patients who had undergone LLA (whether minor or major, ipsi- or contralateral). Patients were keen to engage with services which would help to reduce their risk of LLA and were cognizant of improving the factors which would contribute to it i.e. better glycaemic control but also had varying perceptions of how their willingness to engage would be accepted in an already strained primary care system.

The major area of polarity in opinion was around the acceptability of AI as a sole method of identifying risk and this was evident across all formats of PPI undertaken. Patients appeared wary of an unchecked AI system for radiology. It was an aspect that became essential as the project developed – that the reliability of any system used was robust and replicable. It is interesting to note that the apprehension towards AI seems to stem from older cohorts of patients and would be worth repeating in a younger group of patients to identify for any variability in the responses.

2.3 Power calculation and sample size

Power calculation is a statistical method used to determine the sample size required within the study to improve the chances of detecting a true effect, should it exist. Statistical power analysis was largely pioneered by American psychologist and statistician, Jacob Cohen for his work in power and effect sizes [120]. The terms Cohen's kappa (used in measuring inter-rater reliability), Cohen's d (difference between two means divided by the standard deviation and used in measuring sample sizes) and Cohen's h (measure of distance between

two probabilities) which have made their way into statistical vernacular can be attributed to his lifelong work.

Power calculations were based on the prevalence of diabetic foot disease ulceration and amputation codes. The rationale for this was grounded in the paucity of numbers indicating the true prevalence of sarcopenia which remains underreported and undetermined. However, sarcopenia is present to some extent in patients presenting with DFD and therefore the sample size was based on the prevalence of DFD. The national diabetes foot care audit interval review between July 2014 - March 2021 reported 108,450 registered ulcers with approximately 20,000 - 25,000 new cases registered annually. The true incidence may well be much higher. Hospital Episode Statistics recorded 96,492 admissions with diabetes and foot ulcer or amputation codes between 2014–2015 [16].

The sample size was calculated using Raosoft software [121]. Calculations were based on a 5% margin of error and a 95% confidence interval using the formula below:

$$n = \frac{(Z \text{ squared}) \times (p) \times (1 - p)}{E \text{ squared}}$$

- **n** = sample size
- **Z** = Z-score/critical value for the confidence level
- **p** = estimated population proportion
- **E** = margin of error

Based on an estimated reported prevalence between 96,492 and 108,450, a minimum sample size of 384 was required.

2.4 External peer review

External peer review was obtained by the Imperial College Peer review office. The proposed study plan including the rationale, outcomes and methodology was assessed by two independent peer reviewers prior to commencement of the study.

2.5 Sponsorship and intellectual property

Study protocols and all revisions were reviewed, discussed and assessed by the Research and Governance Integrity Team (RGIT) at Imperial College London (Reference number 23CX8499). Sponsorship approval, indemnity and insurance was granted on the 6th of October 2023.

Imperial College London agreed to hold the negligent harm and non-negligent harm insurance policies which were applicable to this study. Delegated responsibilities were assigned to the NHS trust i.e. Imperial College NHS Trust, taking part in this study. Capacity and capability was also confirmed by Imperial College Healthcare NHS Trust to deliver the study.

The project was also extensively discussed with the Imperial College London's Enterprise department prior to commencing the project to ensure all data and products/software is managed in line with the college's policies.

2.6 Ethics approval

Ethics approval via the Integrated Research Application System (IRAS Project ID: 325124) was applied for. The HRA and Health and Care Research Wales approved the study on the 6th of November 2024 (HRA Reference 23/HRA/4210).

The study was conducted in accordance with the recommendations for physicians involved in research on human subjects adopted by the 18th World Medical Assembly, Helsinki 1964 and later revisions [122].

2.7 Consent

Given the retrospective nature of the study and use of pseudonymised dataset collected by the direct care team, consent was not required in line NIHR/HRA guidelines and was compliant with the General Data Protection and Regulation (GDPR) laws.

2.8 Inclusion criteria

- Patients aged 18 and above.

- Patients with and without diabetes who have undergone an MRI scan of the foot between 2000-2022.

2.9 Exclusion criteria

- Patients under the age of 18.
- Patients with pre-existing or concurrent neuromuscular disorders as they would have pre-existing sarcopenia.
- Patients with advanced oncological conditions as they would have accelerated, pre-existing sarcopenia.
- Patients with metal implants/foreign bodies in the foot as this would add significant noise and artefact for scans to be interpreted and segmented appropriately.
- Patients with prolonged periods of immobility or who may be bedbound as they would have pre-existing and advanced sarcopenia.

2.10 Site Selection

Patients were identified across a single hospital trust in central London (Imperial College NHS Trust). The trust comprises of three hospital sites – St. Mary’s Hospital, Charing Cross Hospital and Hammersmith Hospital.

2.11 Patient identification

Initial patient identification was centred around the availability of relevant MRI scans. All MRI scans of the foot done at Imperial College Healthcare NHS Trust (ICHT) over the preceding two decades were identified via the patient archiving and communication system (PACS). A filter was added to include all MRIs of foot between 2000-2022. MRI scans and their corresponding request and reports were checked and verified to check if the inclusion and exclusion criteria were met.

2.11.1 MRI identification and storage

Scans which met the criteria were saved to a local drive PACS folder to generate a list of patients. The list generated comprised of the patient research case identification (ID) code, the patient name and hospital identifier, the date of the scan and the laterality.

The generated list was sent to the Imaging Research department within the Department of Imaging at ICHT who removed any patient identifiable details and allocated the research case identification code to the scans. These scans were provided in Digital Imaging and Communications in Medicine (DICOM) format for each case ID on a secure cloud drive.

Access to eXtensible Neuroimaging Archive Toolkit (XNAT) was applied for via Imperial College London. XNAT is an open-source software platform for managing, storing, organising and sharing medical imaging data and associated metadata. It is capable of storing diverse data types including imaging and non-imaging data. XNAT allows images to be segmented, annotated and code plug-ins to be added. Furthermore, it provides a platform for storing images and processing, thus allowing for a curated imaging repository.

Approval was provided by the Imaging Research Department, and a separate directory was created for this project within XNAT. The pseudonymised MRI images of the foot were downloaded from the secure cloud drive and uploaded to XNAT in batches in DICOM format within the new directory created for this project (figure 19).

The screenshot displays the XNAT web interface. At the top, a navigation bar includes links for 'Browse', 'New', 'Upload', 'Tools', and 'Help'. The user is logged in as 'manal_ahmad' with an auto-logout timer of 1:59:45. The main content area is titled 'Sarcopenia and Diabetes of the lower limb'. On the left, the 'Details' tab shows project information: ID: SAD-LL, Description: Predicting limb loss in patients with diabetic foot disease on magnetic resonance imaging using deep learning models, PI: Davies, Alun, and Investigators: Rockall, Andrea. On the right, an 'Actions' menu lists options like 'Add', 'Add to Favorites', 'Download XML', 'Download Images', 'Manage Files', 'Upload Images', 'View Prearchive', and 'Scan Type Cleanup'. Below this, the 'Subjects' tab is active, showing a table of subjects. The table has columns for 'Subject', 'MR Sessions', and 'Creator'. It lists 16 subjects (SD1 to SD116) with their respective MR session counts and the creator 'manal_ahmad'. Navigation controls at the bottom of the table indicate '1 of 3 Pgs (418 Rows)'.

Subject	MR Sessions	Creator
SD1	4	manal_ahmad
SD10	1	manal_ahmad
SD100	1	manal_ahmad
SD101	1	manal_ahmad
SD102	4	manal_ahmad
SD103	4	manal_ahmad
SD104	1	manal_ahmad
SD105	2	manal_ahmad
SD106	1	manal_ahmad
SD107	2	manal_ahmad
SD108	1	manal_ahmad
SD109	1	manal_ahmad
SD11	1	manal_ahmad
SD110	6	manal_ahmad
SD111	1	manal_ahmad
SD112	2	manal_ahmad
SD113	1	manal_ahmad
SD114	1	manal_ahmad
SD115	1	manal_ahmad
SD116	1	manal_ahmad

Figure 20: XNAT Directory for the project with pseudonymised scan folders

2.11.2 Patient metadata

Patient metadata was sought via the ICHT's electronic patient care record system, Cerner, which combines patient notes, reports, blood test results and imaging information across the three hospital sites. The original hospital identifier was kept securely to identify the records, and metadata was recorded on an ICHT/ICL approved web-based application - research electronic data capture (REDCap) which was applied for and authorised prior to data collection.

Baseline patient characteristics and factors

As discussed in Chapter one, certain factors are associated with higher risk of developing DFU/DFD. These can be broadly categorised into patient related, medical, social, biochemical, imaging, arterial intervention and outcomes (table 14). Salient patient related factors such as increasing age, sex, ethnicity, presence of an infected ulcer, smoking and

associated lung disease as well as obesity are associated with poor outcomes. The presence and severity of other concurrent diabetes related complications including cardiovascular, cerebrovascular and peripheral arterial disease is a poor prognostic indicator. Furthermore, presence of direct diabetes related complications – retinopathy, neuropathy, nephropathy is indicative of a disease process that is progressing rapidly.

Biochemistry

Glycated haemoglobin (HbA_{1c}) indicates glycaemic control over a preceding period of 120 days. There is a complex interplay between Haemoglobin (Hb) levels and diabetes. Anaemia is both a consequence and a culprit in diabetes related macro- and microvascular complications. Erythrocytes in patients with diabetes have increased aggregation and decreased deformability [123]. This also contributes to the rapid development of diabetic retinal ischaemia and chronic kidney disease [123, 124]. Furthermore, anaemia can increase HbA_{1c} leading to poorer glycaemic control but also lead to false low HbA_{1c} readings thus making it a double edged sword [125]. White cell count (WCC) and C-reactive protein (CRP) are indicators of an ongoing inflammatory process as well as active or indolent infection [12, 39, 59].

A similarly complex dynamic applies to Albumin levels in diabetes. The presence of AGEs and ROS leads to structural changes in albumin which undergoes increased glycation [126]. In a UK Biobank study evaluating 398,146 patients without diabetes and 30,952 patients with diabetes, it was found that higher serum albumin levels were significantly lowered the risk of diabetes related microvascular complications and a 10gram/litre increase in serum albumin almost halved the hazard ratios for nephropathy and reduced the hazard ratios by a third for neuropathy and retinopathy [127].

MRI reports

MRIs of the foot are often done with to exclude certain DFD related features which include collections, presence and the extent of osteomyelitis (OM), presence of previous OM on scans if the patient has multiple scans, presence of sarcopenia and/or diabetic myopathy. The presence or absence of these was also recorded for each scan completed.

MRI scans

For each scan segmented, the muscle signal and fat signal intensity was also recorded. This will be discussed in greater detail in subsequent chapters.

Arterial intervention

As discussed extensively in Chapter one – PAD is a major contributing factor in the development of DFD. The intersocietal IWGDF/ESVS/SVS guidelines on peripheral artery disease in people with diabetes and a foot ulcer emphasise the necessity of excluding PAD in patients with non-healing ulcers or where there may be a suspicion of concurrent PAD [24]. The guidelines also recommend that a person with diabetes, PAD and DFU or gangrene and has adequate single segment saphenous vein and has the indications for infrainguinal revascularisation should have open intervention preferentially over endovascular intervention provided that they are physiologically fit for it [24]. Best practice dictates that the decision for the type of intervention should be in a centre with vascular expertise and the decision should be based on “the risk to and preference of the individual, limb threat severity, anatomical distribution of PAD, and the availability of autogenous vein” [24]. Information pertaining to the presence of arterial disease, the extent was recorded alongside the type and number of interventions.

Outcomes

Patient outcomes including the type of amputation – minor/major/no amputation (table 9), whether the patient had a previous history of amputations, the time from onset of disease to the outcome as well as the time to development of contralateral disease.

Domain	Metadata collected
Patient	- Alive/ Deceased - Age - Sex - Ethnicity - Weight (in Kg) - Infected Ulcer
Medical	- Cardiovascular Disease - Cerebrovascular Disease - Peripheral arterial disease - Diabetes (T1/T2)

	- Chronic Kidney Disease/Renal Replacement - Retinopathy - Neuropathy - Lung Disease - Other relevant past medical history
Social	- Smoking
Biochemical	- Glycated Haemoglobin (HbA_{1c}) - White Cell Count (WCC) - C-Reactive Protein (CRP) - Haemoglobin (Hb) - Albumin
Imaging	- Presence of osteomyelitis (OM) - Previous OM on MRI (if more than one MRI) - Extent of OM - Presence of Charcot foot - Presence of sarcopenia/diabetic myopathy - Presence of ulcer
Arterial Intervention	- Arterial disease - Extent of arterial disease - Intervention to improve vasculature – endovascular/open/hybrid - Number of interventions
Outcome	- Outcome - Previous amputation - Time between onset and outcome - Time between onset and development of contralateral disease [if applicable]

Table 14: Categories and details of patient metadata collected

2.12 Data storage

The project was approved and registered with the Imperial College London's Data Asset Registration Tool (DART) team to fulfil the legal Data Protection Impact Assessments (DPIA) requirement, defined within Article 35 of the UK GDPR. DART registration also intends to ensure that all personal data is recorded and managed in an effective manner and in line with existing protocols.

A transparency notice and information pertaining to the study was made available to the public and potential participants on the Imperial College London's website under the vascular research section to provide details pertaining to the study as well as providing a point of contact for individuals interested in the study with the option for participants to withdraw up until the time of data analysis.

It was agreed and approved that the data and all appropriate documentation will be stored for a minimum of 10 years after the completion of the study.

2.13 Data cleaning

2.13.1 Metadata cleaning

Following data collection, REDCap data was transferred to a comma-separated values (CSV) file where it was inspected. Duplicates were identified and removed. Inconsistent values e.g. weight in kilograms/pounds, etc was formatted to the same unit and value to ensure text and numerical values were consistent across the rows and columns. All dates and timelines were parsed to ensure uniformity i.e. described in months instead of years, weeks, dates, etc.

Text data was formatted to ensure consistency e.g. atrial fibrillation, AF, AFib, etc were all converted to AF or Male, male, males, M were all changed to the same data label 'Male'. Extra spaces and symbols were removed. Where missing data was significant across a patient case ID, the rows were removed.

Following this initial data cleaning, each row and column was inspected visually using Statistical Package for Social Sciences (SPSS) version 29.0.2.0 and R version 4.5.0. Visual distribution of the data for variables was created using histograms and boxplots to evaluate for skewed data [128].

The mean, median and mode was calculated numerical row and inspected for outlier values. These values were verified again against the patient records again and extensive outliers were removed. Means with standard deviations were used to describe numerical/continuous variables which had a normal distribution whilst skewed values/variables were summarised using the median value and interquartile ranges. Categorical values e.g. male/female or smoker/non-smoker were summarised using frequencies and overall percentages.

2.13.2 Segmentations and cleaning the imaging data

All MRI scans were completed using a 1.5 Tesla, Signa Artist MRI machine from General Electric (GE)© Medical System. This is a commonly used wide bore MRI machine for multiple purposes.

The images used were 512 x 512 millimetres. This essentially means a grid of 512 pixels vertically and 512 pixels horizontally such that each slice has 262,144 pixels. This provides a larger data set to work with and has better resolution i.e. a high-resolution matrix. Each slice of the MRI scan was 3 millimetres in thickness. When the pixels are multiplied by the thickness, it adds a third dimension to the image called a voxel (figure 20).

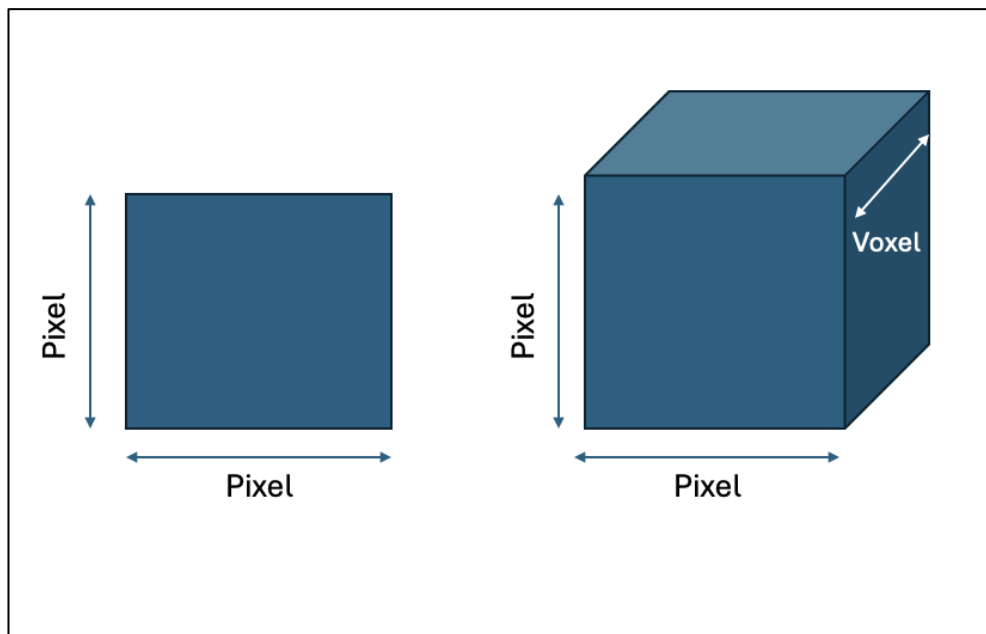


Figure 21: Schematic representation of a pixel versus a voxel

All images were uncompressed and lossless i.e. no reduction to the file size was made in order to save the image data. All images were transferred exactly as they were acquired without any change to the resolution. ICL/ICHT mandates uncompressed/lossless images for storage in PACS. MRI scans followed a uniform protocol of imaging from the digits to just distal to the calcaneum. T1- and T2- weighted images in coronal, axial and sagittal views were available.

Scans were inspected to ensure the relevant anatomical region of interest was available and visible. Incomplete scans were flagged for removal. Included images had to be without artefact from metal, foreign objects or movement or overlap (images partially covering the same anatomy multiple times) which would distort the images. MRI scans which did not have T1-weighted coronal views were also removed as it was agreed that all images would have the same orientation to best visualise the intrinsic muscles of the foot. Patients and scans were also removed in the event that they had reports but no images available.

2.14 Dataset

1027 scans across 427 patients were initially identified via PACS. These were reviewed to ensure they met the inclusion and exclusion criteria. 203 scans were excluded due to significant artefact (n=63), incomplete imaging (n=41), overlap (n=9), unavailability of T1-weighted coronal views (n=9). Other (n=169) were also excluded as the scans were not suitable for segmentations due to lack of clear anatomy, significant pre-existing charcot and destruction of anatomical landmarks and metal implants. Forty-one patients were also subsequently excluded.

The final dataset consisted of 824 reports with scans and 386 patients as reported in the Consolidated Standards of Reporting Trials (CONSORT) flowchart below (figure 21).

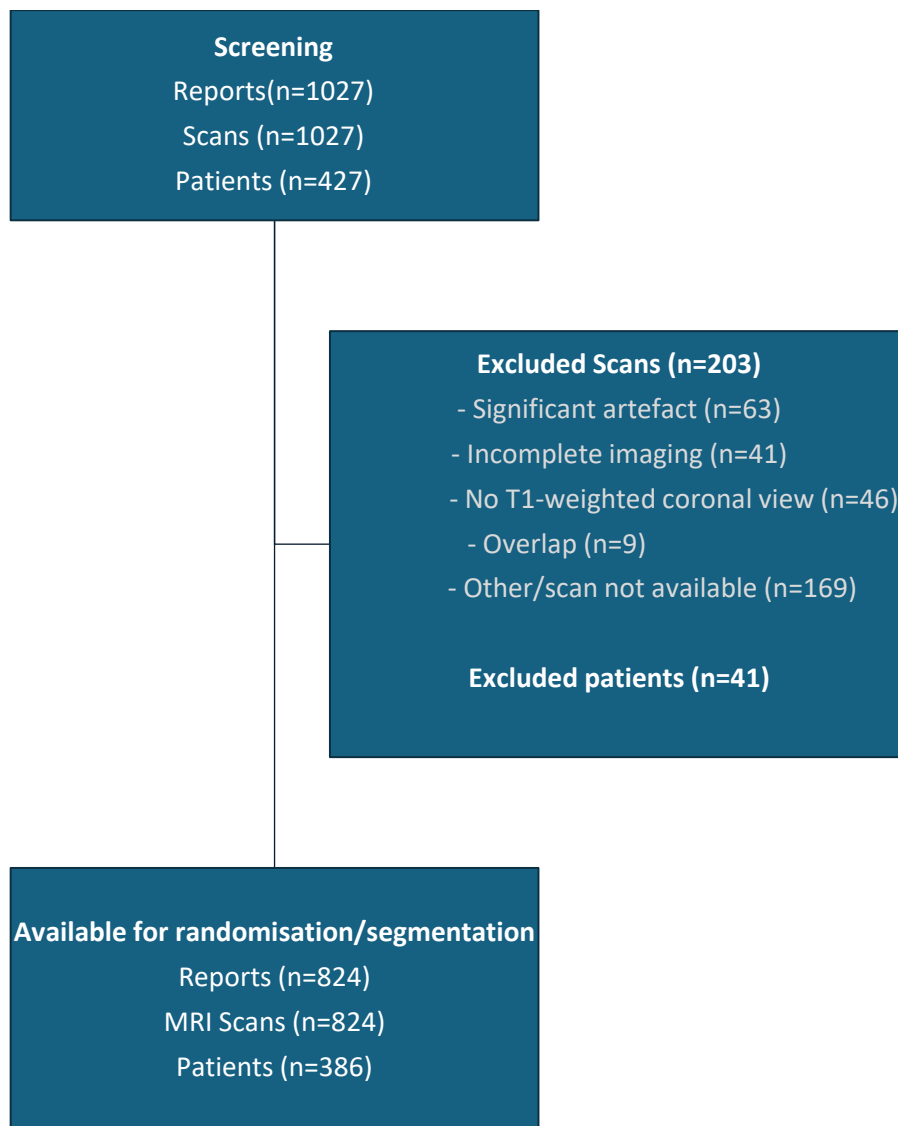


Figure 22: CONSORT diagram of final dataset

Chapter 3: Risk prediction using MRI reports

3.1 Introduction

As discussed in chapter one, Zhang et al previously used CT angiography reports to predict the Wagner score as the DFU outcome [102, 109]. Beyond this paper, the use of scan reports for prediction purposes is scarce. MRI reports provide a plethora of information about a patient with pre-existing or suspicion of DFD. Initial screening for the dataset was completed by reviewing all MRI foot scan reports via PACS.

Several models may be used for general classification and prediction purposes in the world of data science of which the three most relevant and accurate models for classification purposes in this scenario were selected – Random Forest (RaF), Extreme Gradient Boosting (XGB) and Multilayer perceptron neural network (MLP).

These are forms of ensemble learning; a ML method which combines predictions from multiple models/learners to improve the performance of the model[129].

3.1.1 Random Forest

Random Forest (RaF) is a type of supervised ML model which can be used for improving prediction accuracy. It is commonly used for classification and regression purposes and builds on the concept of a Decision Tree (DT).

Decision Trees

The idea of a DT is not new to the world of medicine and replicates human decision making. It consists of a hierarchical structure with node branches, internal nodes and leaf nodes as demonstrated in figure 22 [130]. The root node can be interpreted as the problem, internal nodes as influencing factors or predictors and leaf nodes as the decisions or outcomes.

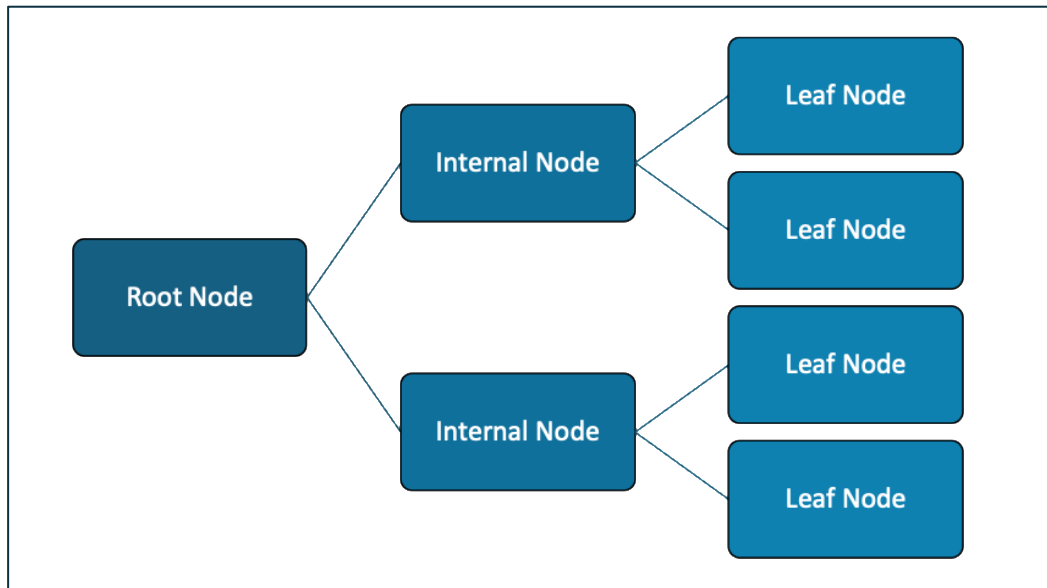


Figure 23: Structure of a decision tree. Adapted from IBM [130].

In its simplest form, it is the basic structure of an algorithm. DTs learning aims to identify the optimal split points within the tree in a recurring top-down pattern until all records are classified under specific class labels or categories. A DT then learns how a previous set of questions were answered or factored to decide the outcome. An advantage of using decision trees is in their ability to integrate categorical or numerical data as variables and can model problems with multiple outcomes [131].

Overfitting

A commonly encountered problem in ML which will be discussed at multiple points during this thesis is overfitting. This is whereby the ML model learns the training data including the noise (random or erroneous data which deviates from the actual pattern), fluctuations and corresponds exactly to the training set. It can therefore perform very well on that particular dataset because the model has learned by rote rather than understanding why each decision was made. When applied to a validation or test set, the ML falters.

Random Forest Structure

RaF, as the name suggests, is when multiple decision trees are combined to reach a decision. First proposed by Leo Breiman in 2001 before being trademarked in 2006 by Leo Breiman and Adele Cutler [130, 132]. Breiman defined RaF as a “classifier consisting of a collection of tree-structured classifiers ($h(x, \Theta_k)$, $k= 1, . .$) where the Θ_k are independent

identically distributed random vectors and each tree casts a unit vote for the most popular class at input x " [132].

Given a dataset, multiple DTs reach a decision/prediction, and the overall output is based on the cumulative majority, or aggregate decision of the multiple DTs (figure 23).

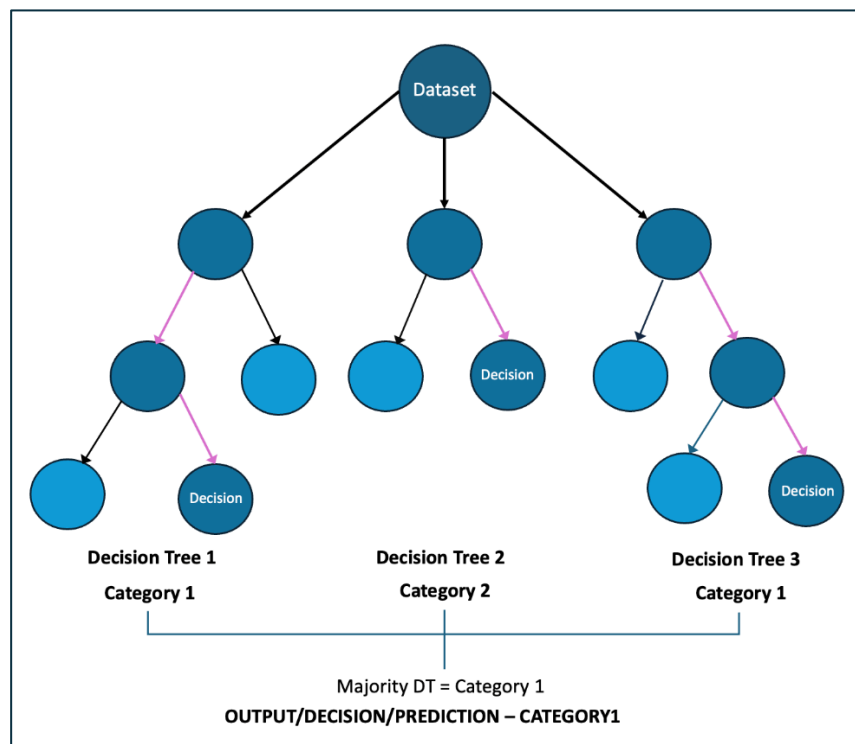


Figure 24: Schematic representation of Random Forest. The purple arrows demonstrate the pathway the DT uses to reach the decision. The dark blue circles demonstrate the options chosen at each point in the decision-making process. The light blue circles indicate the decision which was not chosen. Adapted from IBM [130].

This is based on Breiman's early work who introduced the concept of Bootstrap Aggregating or 'Bagging'. Bootstrapping is a resampling technique which involves recursive or recurrent resampling/redrawing of random samples from a dataset to create multiple, small datasets [133]. The predictions from bootstrapping or each small dataset is then aggregated through a majority vote for classification and averaging for regression to improve accuracy [134]. Bagging overcomes the problem of overfitting and can improve the model performance and accuracy. Bagging is also useful as it can factor in bias and variance, and works well with small datasets.

3.1.2 Extreme Gradient Boosting

XGB is a DT ensemble learning algorithm similar to RaF which shares a similar structure as RaF however, instead of assessing trees in parallel, it assesses each tree sequentially (figure 24) [135].

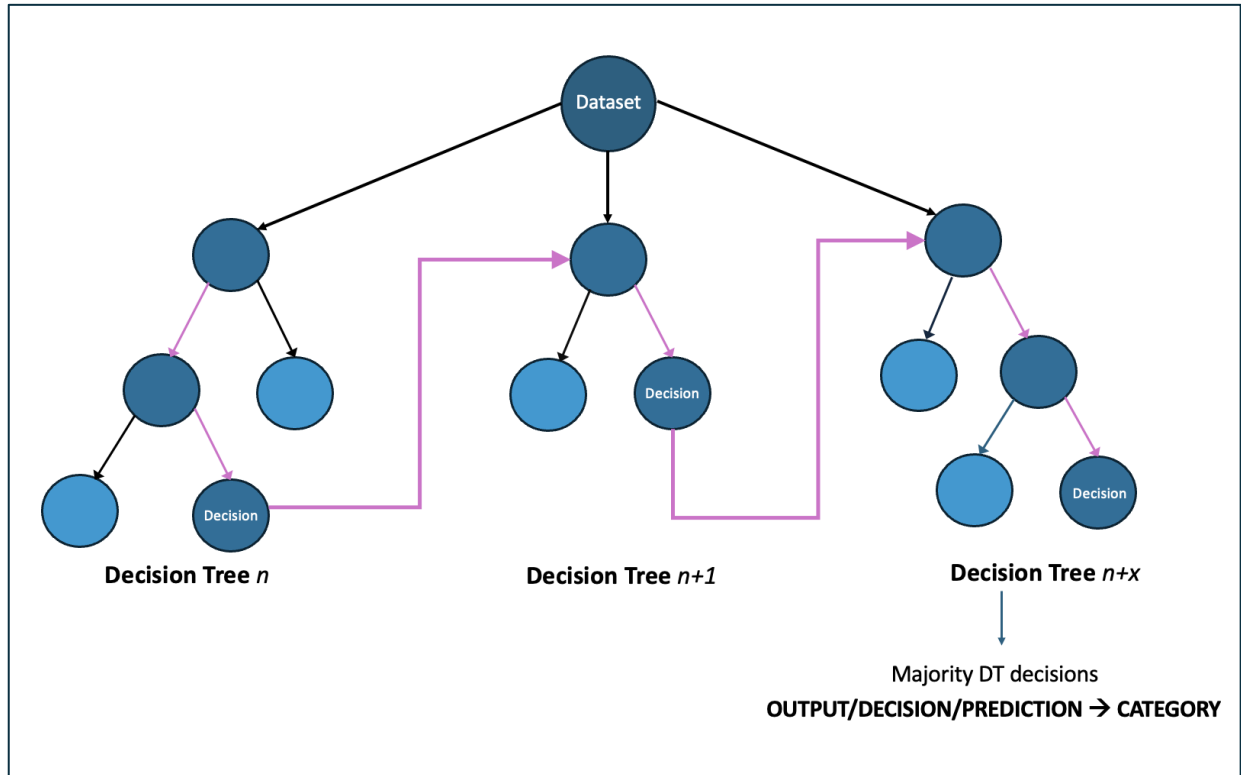


Figure 25: Schematic representation of Extreme Gradient Boosting structure. The purple arrows demonstrate the pathway the DT uses to reach the decision. The dark blue circles demonstrate the options chosen at each point in the decision-making process. The light blue circles indicate the decision which was not chosen. Adapted from Kavlakoglu [136].

XGB is also known as a gradient boosting decision tree. XGB uses boosting whereby multiple weak classifiers are combined to create strong classifiers. Every sequential model is trained to address the error in decision making of its predecessor [137]. It does this by assigning weights so that correctly classified decisions are attributed lower weights and vice versa [136, 137]. Boosting aims to reduce bias rather than overfitting, but this can be fine-tuned. It further allows scalability and has higher predictive performance [129]. Furthermore, XGB allows weights to be handled within the performance model and tuning of the parameters or hyperparameters.

3.1.3 Multilayer perceptron model

Neural networks were previously discussed in Chapter one. MLPs are a type of supervised, feed forward neural network which consists of three components – an input layer, a hidden layer and an output layer as shown below in figure 25 [138].

X indicates the input data features; H is the number of nodes in the hidden layer and Y is the output. Each line connecting the hidden layer has a weighting which cannot be changed and is determined by the neural network. Feedforward means that the input data goes through multiple convolutional layers/filters without looping back. The network compares its prediction to the actual label and calculates the error which is the cross-entropy loss (discussed in chapter one). It then backpropagates the gradient of loss with each weighting and updates the weighting for each pass [139]. MLPs can be used for various types of data including text, numerical data and images. MLPs have their advantages but often require more data than RaF, DTs and have a ‘black box’ methodology whereby the process of interpretation within the hidden layers is not shown.

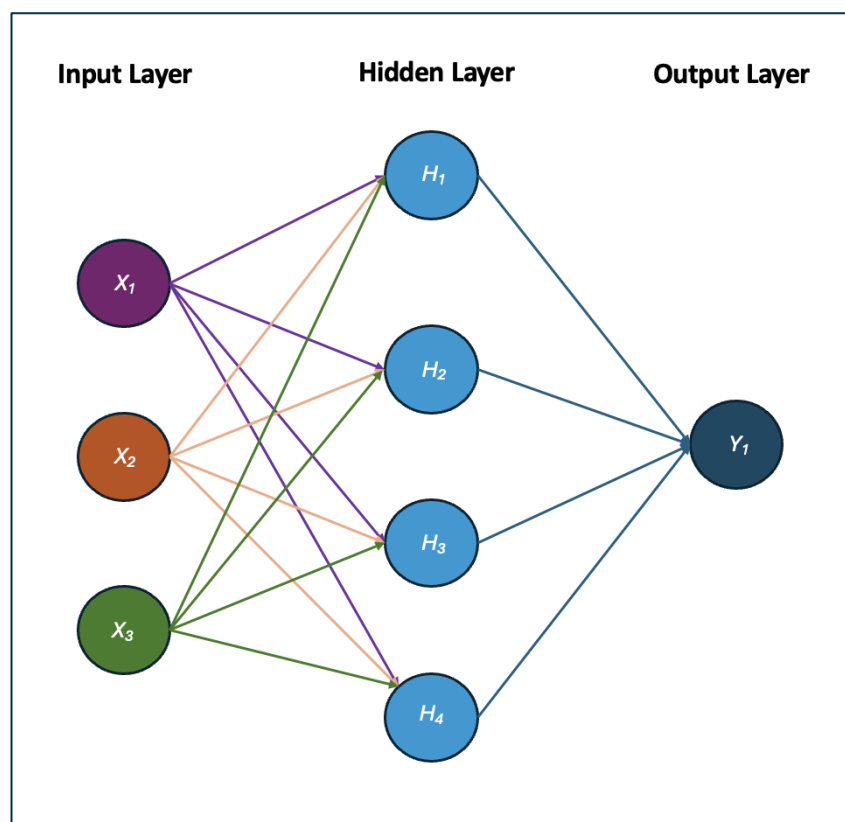


Figure 26: Structure of a multilayer perceptron model. Adapted from Krizhevsky et al [140].

3.2 Aim

The aim of this chapter is to identify potential applications of AI models applied to MRI reports as a prognosticator in highlighting potential predictors for limb loss in patients with DFD.

3.3 Methodology

Five MRI reported features were selected as potential predictors given the reporting format for MRIs of the foot:

- Presence or absence of osteomyelitis along with the extent.
- Presence or absence of diabetic myopathy and/or sarcopenia.
- Presence or absence of ulcer.
- Presence or absence of collections.
- Charcot foot, if stated in the MRI report.

The outcomes selected were no amputation, digit amputation, multiple digit amputations, transmetatarsal, below knee amputation and above knee amputation. These were then categorised into three groups or classes - no amputation, minor amputation and major amputation, with each group representing potential degrees of severity. All outcomes were changed to categories for data analysis such that 0=no amputation, 1=minor amputation and 2=major amputation.

A significant cohort of patients had multiple scans spanning several years with varying laterality. This posed a problem in that each scan would count as a single, discrete episode with the same outcome therefore representing a higher number of the overall number of each outcome. To overcome this and add clarity to the analysis, the last scan prior to the final outcome was selected for each patient to minimise temporal bias.

Following data cleaning, a data spreadsheet was created using Microsoft Excel consisting of five columns stating the absence or presence of each feature and a sixth column stating the outcome. A total of 419 records were included.

3.4 Summary of MRI reports

3.4.1 Feature presence

Table 15 succinctly summarise the findings from the MRI reports below. Out of the 419 patient records, 247 patients had Osteomyelitis (59.1%) and 369 patients had Diabetic Myopathy/Sarcopenia reported (88.3%).

Of the cohort who did have Osteomyelitis, 156 (63%) had single digit involvement, 37 (15%) had multiple digit involvement, 20 (8%) had midfoot involvement and 34 (14%) had hindfoot involvement (figure 26).

A majority of patients also had a reported ulcer (n=295; 70.6%). This is not surprising given the presence of DFUs or suspicion of osteomyelitis would have prompted the MRI scans for a majority of the patients. Charcot foot (n=66; 15.8%) and collections (n=31; 7.4%) were less common.

Feature	Yes – n (%)	No – n (%)
Osteomyelitis	247 (59.1)	171 (40.9)
Collection	31 (7.4)	387 (92.6)
Ulcer	295 (70.6)	123 (29.4)
Diabetic Myopathy/Sarcopenia (DMS)	369 (88.3%)	49 (11.7)
Charcot Foot	66 (15.8%)	352 (84.2)

Table 15: Summary of MRI reported features

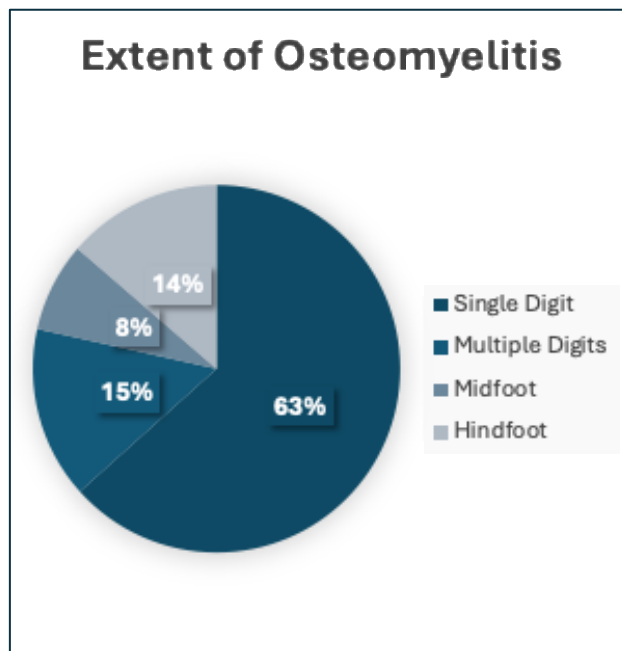


Figure 27: Anatomical distribution of Osteomyelitis as per MRI reports

3.4.2 Summary of outcomes

284 patients did not have limb loss, 94 patients underwent minor limb loss, and 40 patients had major limb loss. This is visualised in figure 31 below.

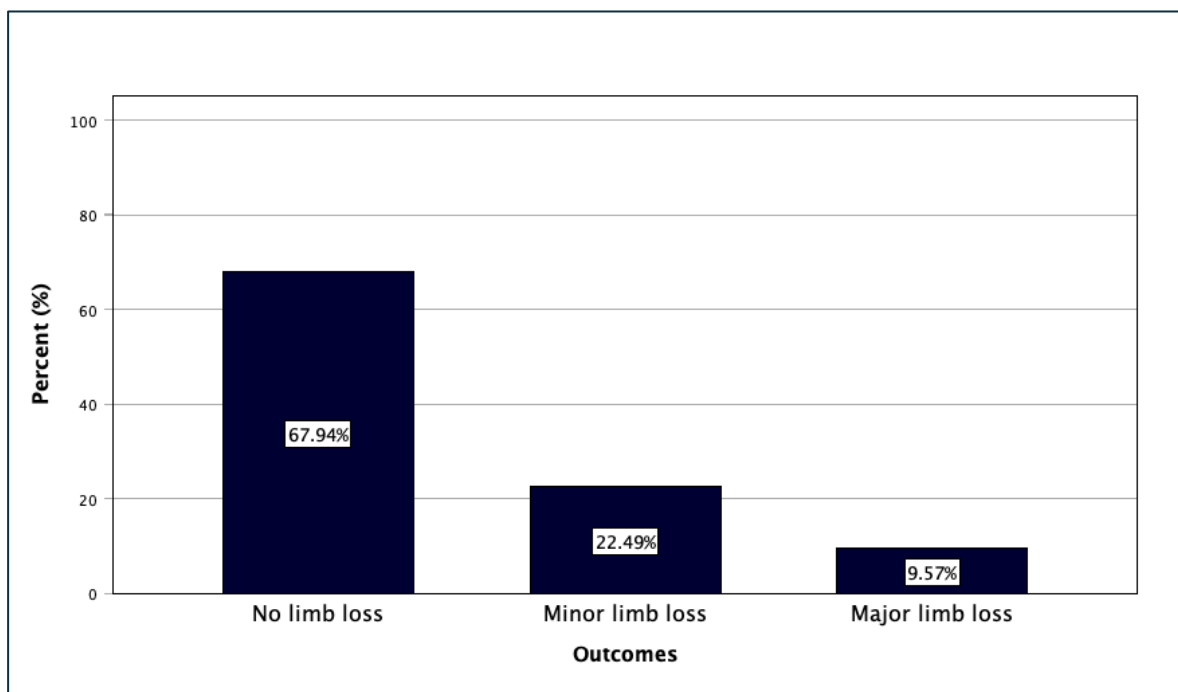


Figure 28: Bar chart demonstrating the corresponding outcomes for all included cases.

3.4.3 Kruskal-Wallis Test

Prior to applying any ML models, a basic non-parametric test was completed as part of the initial data analysis. The data types available include multiple categories of nominal data (MRI features) and ordinal data (outcomes in increasing severity). The skewed visual representation of the data implies a non-parametric test was required to compare the six independent groups therefore the Kruskal-Wallis Test was used.

Method

Using R 4.5.0, the following libraries were downloaded:

- **readxl**: Allows Microsoft excel files to be read in R.
- **dplyr**: Part of the tidyverse collection in R. This allows data manipulation and provides the grammar and syntax with data and tibbles (consistent and readable data frames).
- **ggplot2**: For data visualisation and graphics.

The complete, clean and reproducible R code is available in Appendix 1.

Each variable was identified and the outcome column established. A function to loop through the variables was set so that the test could be run for each variable/feature rather than having to change the code every time for a new column.

If the p-values were significant, further Wilcoxon tests were run which compares two dependent samples with Bonferroni correction for multiple testing to reduce the risk of false positive results (Type 1 error).

Results

The results from the Kruskal-Wallis test are shown in table 16 and figure 32. This demonstrates a strong and statistically significant difference in the outcomes in patients with osteomyelitis and ulcers. Collections also have a moderate impact on the outcome but this is not reach statistical significance. Diabetic myopathy/sarcopenia and Charcot Foot did not have an impact on the outcome.

Outcome by feature	Kruskal-Wallis chi-squared value	p-value
Osteomyelitis	20.6	<0.001
Ulcer	12.689	<0.001

Collection	7.67	0.0057
Charcot	2.2857	0.1306
DMS	0.03	0.8625

Table 16: Kruskal-Wallis results for MRI reported features on outcomes.

3.5 Logistic Regression

It is essential to remember that Artificial Intelligence and all that falls under its vast umbrella is grounded in the principles of mathematics and, before advanced statistics was rebranded as AI, there were statistical models and tests which were sometimes arduous, complex and time consuming.

Logistic regression is a standard method of statistical analysis which predicts the probability or likelihood of an outcome as a linear combination based from one or more variables or predictors. Prior to exploring ML models, a basic Logistic Regression analysis was completed.

Method

A desktop graphical user interface – Anaconda Navigator was downloaded on a MacBook Air using an M1 chip and running macOS Sequoia (version 15.3). Anaconda allows Conda, an open-source package manager and various environments including cloud-based environments to be installed and run (figure 33).

Using the MacBook Terminal application, Python version 3.11.7 was downloaded. Python is a multipurpose, open-source programming language which works in multiple environments[141].

Jupyter is an open-source software which provides a platform capable of handling multiple programming languages and software and provides a notebook type format with integrated desktop use capabilities via Anaconda Navigator (figure 34) [142]. The relevant Python libraries including ‘NumPy’ for dealing with statistics, ‘PyTorch’ – an ML framework, ‘Pandas’ for working datasets, ‘matplotlib’ and ‘seaborn’ for static graph plots and ‘Graphviz’ for graph visualisation were downloaded in a Jupyter notebook (Appendix 2). Other recommended libraries for optimised use including ‘Joblib’ and ‘ipykernel’ for streamlining

and allowing pipelining for bigger datasets were also included. These libraries will also be used for multiple other data analyses throughout the thesis.

The scientific toolkit for machine learning called Sci-kit learn was downloaded in the Jupyter notebook. Sci-kit is an open-source ML library using Python with a plethora of tools for data analysis including classification, regression, clustering and preprocessing amongst many others [143].

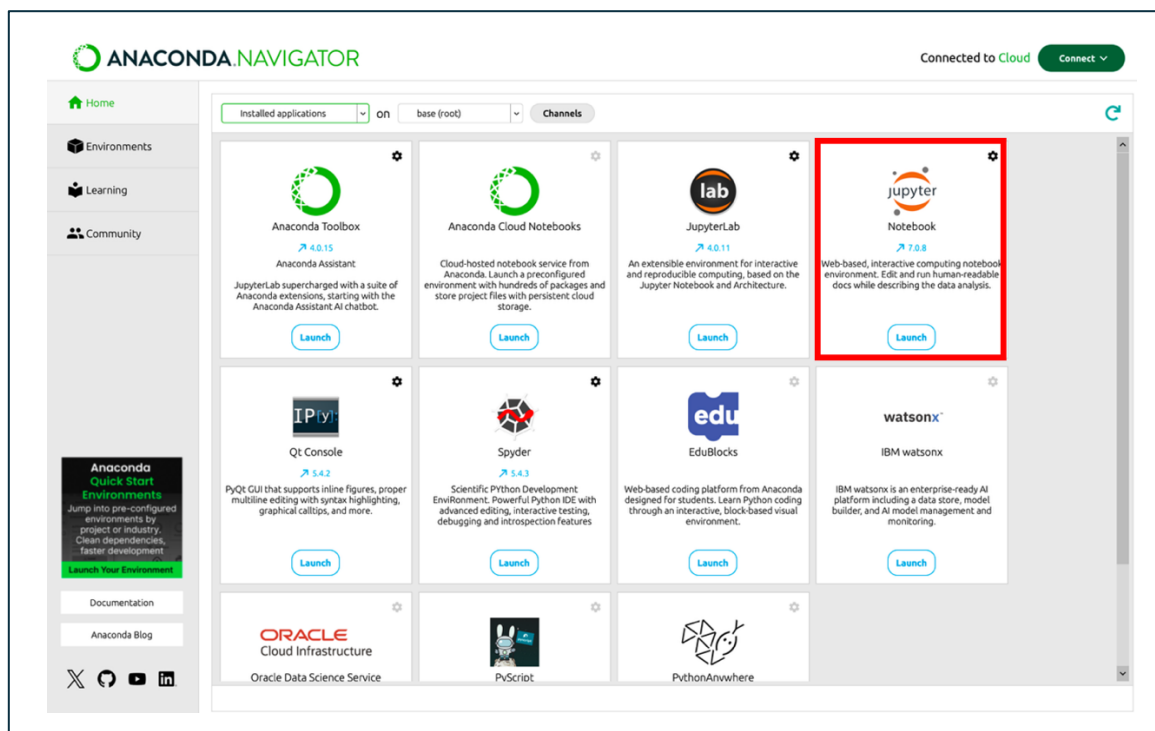


Figure 29: Anaconda Navigator desktop main page with Jupyter notebook highlighted in red.

Once Sci-kit learn was downloaded, relevant libraries were added to run logistic regression including 'LabelEncoder', 'Logistic Regression', 'Classification_report' and 'train_test_split'.

Each Yes/No response for each MRI feature was relabelled as a binary outcome i.e. 0 and 1 and split into a ratio for training and testing ratio (80:20). The random state was set at the commonly used value of 42. The random state is the equivalent of the 'set seed' function in R. Establishing the 'seed' for the random number generator is an important step to ensure reproducibility of the code including adjustments for weighting.

Results

Figure 35 and table 17 demonstrate the co-efficient which indicates whether the direction of the feature's association is positive or negative and what degree. The absolute co-efficient provides the magnitude of the association.

Collections and Osteomyelitis had a negative association on the outcome. Presence of ulcers also had a weak negative association on the outcome. Conversely, DMS and Charcot Foot had a weak association with the outcomes. This is further visualised in figure 34.

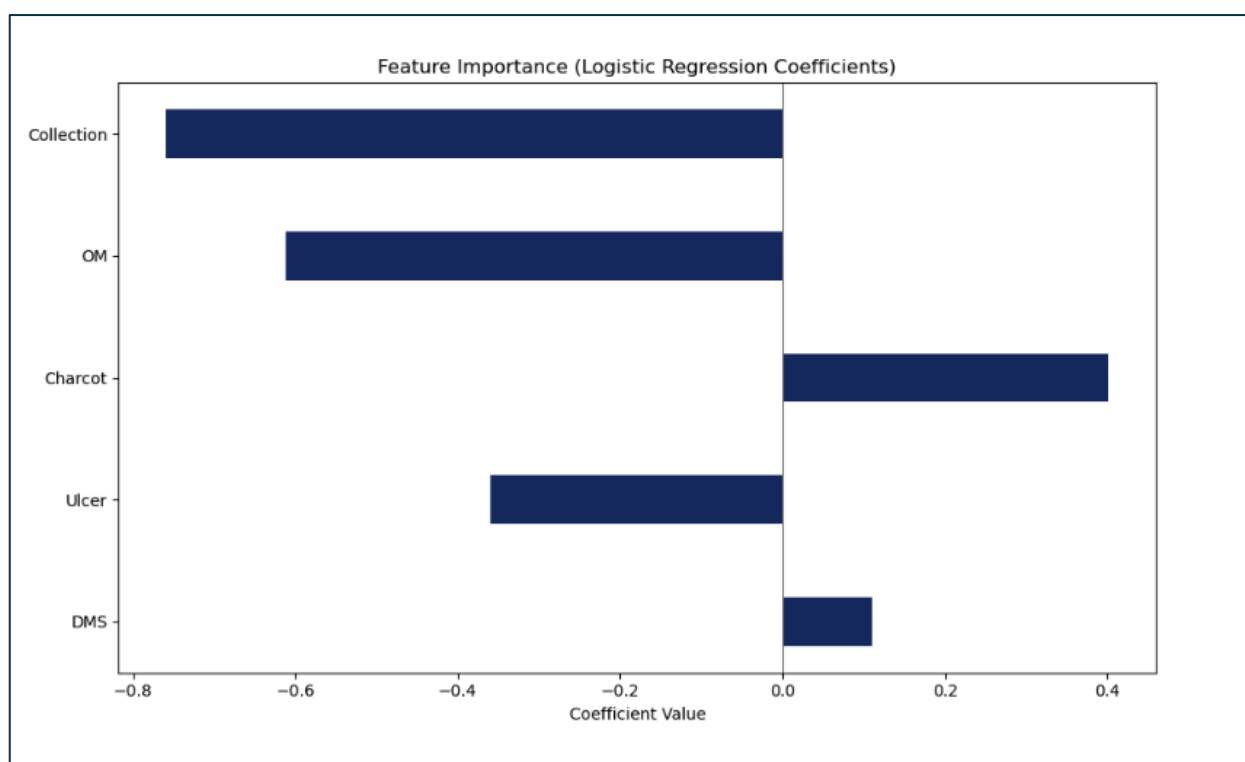


Figure 30: Visual representation of logistic regression coefficients

Feature	Co-efficient	Absolute Co-efficient
Collection	-0.7592942635304410	0.7592942635304410
OM	-0.6123011553866800	0.6123011553866800
Charcot	0.4016547613335740	0.4016547613335740
Ulcer	-0.3601832358863090	0.3601832358863090
DMS	0.11027586058719600	0.11027586058719600

Table 17: Logistic Regression for MRI reports

3.6 Random Forest

The background for RaF was discussed at the start of this chapter and will therefore not be discussed here.

3.6.1 Method

Using R 4.5.0, the following libraries were downloaded: '**readxl**', '**dplyr**' which were previously discussed earlier along with:

- **caret**: An acronym for Classification and REgression Training.
- **randomForest**: Implements Breiman's RaF algorithm in R.

The set seed value was set to the standard value of 42 and the Excel spreadsheet was uploaded to R. All columns which were not labelled as "Outcome" were set as predictors. Predicting variables were converted to numerical value. This is called encoding. A classification RaF was used.

The data was split into an 80:20 set with 80% for training and 20% for testing. The maximum number of DT's for running RaF was set to 500 and the number of variables or predictors set at each split was set to a maximum of 15. The output coded for including a confusion matrix and a multiclass receiver operating characteristics-area under the curve (ROC-AUC).

A confusion matrix is a tabular summary which provides a visual representation of a given model's performance [144]. It provides a visual summary of the number of true positive, true negative, false positive and false negative values. This concept will be discussed again in the next chapter.

		Actual Outcome	
		Positive	Negative
Predicted Outcome	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 31: Confusion Matrix

An ROC-AUC is a performance metric for classification models which indicates a model's performance. ROC curves describe the relationship between the sensitivity and specificity of a test [145]. The x-axis represents the false positive rate whilst the true positive rate is on the y-axis. Interestingly, the origin of ROC is grounded in signal detection theory during World War II to help improve the accuracy of radar detection and helping radar operators distinguish between actual enemy aircrafts and non-sensical 'noise' present in the environment [145]. The ideal model has an AUC of 1.0 which implies a 100% probability of the model correctly predicting an outcome every time.

The code for RaF also reports the out-of-bag error which is a method of internal cross-validation. This is a technique to evaluate the accuracy of the model, estimates the error and allows for tuning of parameters [146]. Another concept which is included in the R code is the mtry which is the number of variables or predictors which are drawn for DT splits [146]. The complete, clean and reproducible R code is available in Appendix 3.

3.6.2 Results

3.6.2.1 Model accuracy

The model opted to try two variables at every split. The out-of-bag error was 34.23% which means this was the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 18) showed the model predicted class 0 (no amputation

group) 220 times and incorrectly predicted it 8 times with a 3.5% error rate. The model predicted class 1 (minor amputation group) poorly, classifying it correctly once with an almost 99% error rate. Class 2 (major amputation group) was classified incorrectly every time.

	Class 0	Class 1	Class 2	Class error
Class 0	220	8	2	0.03508772
Class 1	75	1	0	0.98684211
Class 2	31	1	0	1.00000000

Table 18: Random Forest training set confusion matrix (class 0 = no amputation, class 1=minor amputation, class 2= major amputation)

The RaF test set confusion matrix demonstrates the model was best at predicting class 0 but falters for the other two classes (table 19). The model's accuracy was 67% [95% CI 0.5581-0.7706; p-value 0.6437].

	Class 0	Class 1	Class 2
Class 0	55	18	8
Class 1	1	0	0
Class 2	1	0	0

Table 19: Random Forest test set confusion matrix (class 0 = no amputation, class 1=minor amputation, class 2= major amputation)

3.6.2.2 Class imbalance

This was due to evident class imbalance. Therefore, this was rebalanced by running 'sampsize' within the randomForest library and changing the 'ntree' number to 500. A new test set was created within the dataset and RaF ran again. Class weights define the weight

or importance of each class. Inverse class frequency is therefore the weights which are inversely proportional to the frequency of each class and is used in datasets with class imbalance to add more weight to the less represented classes in a dataset [147]. This resulted in a lower accuracy rate (from 67% to 9.76%) therefore the class imbalance coding was removed and code changed back to its original format.

3.6.2.3 Performance table

The performance table below shows sensitivity was highest for class 0 and when the model did predict class 0, it was correct 67.9% of the time. The balanced accuracy score is used in imbalanced datasets and is defined as the average recalls for each class or category whereby zero is the worst value and one is optimal. For class 0, this was 0.49 which implies a moderate recall but there is evidence of overpredicting class 0. The model seems to falter significantly at predicting class 1 and 2.

	Class 0	Class 1	Class 2
Sensitivity	0.9821	0.0000	0.0000
Specificity	0.0000	0.9844	1.0000
Positive Predictive Value (PPV)	0.6790	0.0000	0.0000
Negative predictive value (NPV)	0.0000	0.7778	0.90244
Prevalence	0.6829	0.2195	0.09756
Detection rate	0.6707	0.0000	0.0000
Detection Prevalence	0.9878	0.0122	0.0000
Balance Accuracy	0.4911	0.4922	0.5000

Table 20: Random Forest performance table for no amputation (class 0), minor amputation (class 1) and major amputation (class 2).

3.6.2.4 Feature importance

Feature importance was coded for in order to highlight features which may have been most predictive of the outcome (table 21). This is best interpreted by reading the mean decrease accuracy and mean decrease Gini.

Mean decrease accuracy quantifies the change in accuracy in of the model when the variables or predictors are randomly permuted [148]. A positive value highlights an important feature whilst a value of zero or below implies irrelevance.

The Gini impurity index is a measure of how “pure” a node is within a DT with zero implying a pure node because all the samples from a dataset in that node are from the same class or category [148] . Mean decrease Gini measures the average decrease in Gini impurity across all the DTs i.e. to what extent did a particular variable or predictor reduce the Gini impurity across a DT split [148, 149]. A high value implies the variable or predictor effectively splits the DT into purer nodes whilst a low value means that the variable is not used or does not help to discriminate the classes/outcomes.

Variable	Class 0	Class 1	Class 2	Mean decrease accuracy	Mean decrease Gini
OM	11.02432	-5.56368	-6.595163	9.315794	5.636243
Collection	6.522906	13.28593	-5.005574	11.585022	3.78352
Ulcer	10.73541	-4.48725	-4.329975	9.501076	3.233866
Charcot	-5.04429	2.125242	-1.194432	-3.212387	3.798439
DMS	-1.10817	-0.55389	-2.975733	-1.545798	1.631404

Table 21: Feature importance (Random Forest)

In the table above, the most influential predictors from the MRI reports were osteomyelitis, collection and ulcers. Osteomyelitis had an impact for class 0 outcomes, whilst collections

were predictive for class 1 outcomes. Charcot and DMS added noise to the model. This is also visualised in the figure below (figure 37).

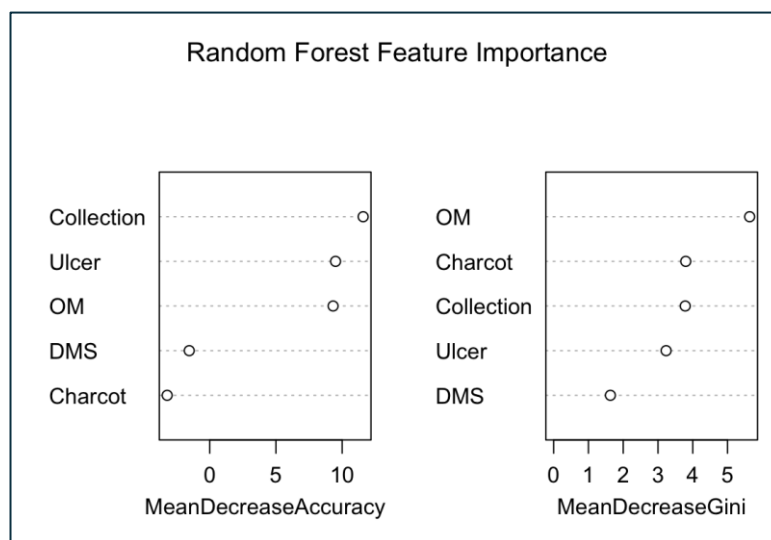


Figure 32: Random Forest mean decrease in accuracy and mean decrease in Gini

3.6.2.5 ROC-AUC for Random Forest

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class. For class 0, the ROC-AUC was 0.397, For class 1, the ROC-AUC was 0.481 and for class 2, the ROC-AUC was 0.42. All of these indicate poor performance. The micro-AUC was 0.742. The micro-averaging AUC adds equal weight to all predictions equally and should be interpreted with some caution as it may be influenced by class imbalance in a dataset.

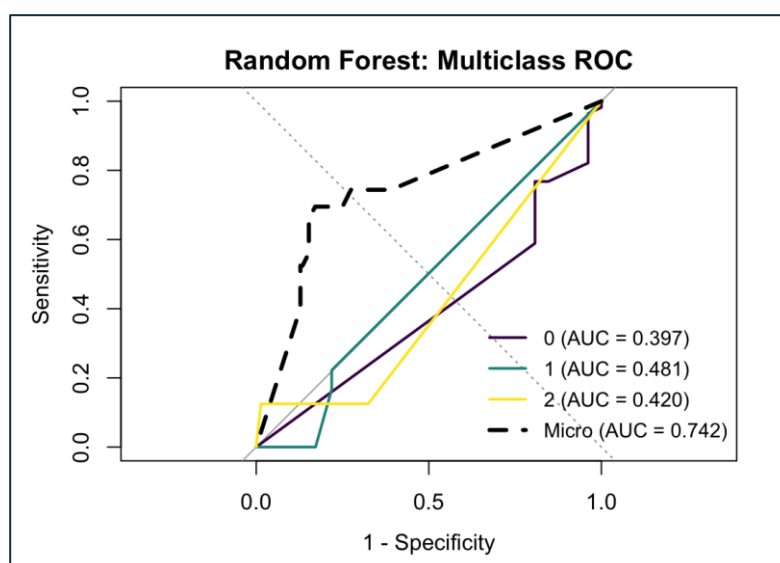


Figure 33: Random Forest ROC-AUC

3.7 Extreme Gradient Boosting

The background for XGB has already been discussed at the start of this chapter and will therefore not be discussed here.

3.7.1 Method

Using R 4.5.0, the following libraries were downloaded: `'readxl'`, `'dplyr'`, `'caret'`, `'ggplot2'`, which were previously discussed earlier along with:

- **lattice:** Creates plots and graphs.
- **Matrix:** Used to create matrices in a table format where all elements in the data structure are the same i.e. either all numbers or all characters. This is different from a data frame which contains numerical values and/or characters.
- **xgboost:** Runs XGB in R.

The complete, clean code which can be reproduced is included in appendix 4. The set seed value was set to the standard value of 42 and the Excel spreadsheet was uploaded to R. All columns which were not labelled as “Outcome” were set as predictors and a filter was used to remove rows where the outcome was not one of the 3 classes. Predicting variables were converted to numerical value after using the `'levels ()'` function which extracted the category labels within the outcome column.

The `num_class` counts the number of each class in the column labelled “outcome”. `'y_all'` and `'as.integer'` converts the categories – no amputation, minor amputation and major amputation to 0,1,2 internally. The same is repeated with the remaining columns to turn it into a matrix.

The data was split into an 80:20 set with 80% for training and 20% for testing. The weights were reviewed for imbalance `'class_wts'` which essentially finds the most prevalent class and the number in that and then divides it by each class to give greater weighting to smaller classes i.e. minor and major amputation groups. This means that during the training of the model, incorrectly classifying to a smaller class is weighted as a bigger error whilst incorrectly classifying to a bigger class (no amputation group) is weighted as a smaller error. The code consisted of setting parameters for binary cases (objective = “binary:logistic”) and

multiclass (objective = “multi:softprob”). The initial hyperparameters for XGB are available in table 22. The model was also asked to store the best iteration prior to early stopping.

Hyperparameter	Code in R	Value set
Learning rate	Eta	0.15
Depth of the DTs	max_depth	6
Number of weights in each leaf node, also called child node	min_child_weight	3
Sample percentage rows per tree	subsample	0.8
Sample percentage columns per tree	colsample_bytree	0.8
L1 regularisation or lasso penalty [shrinks certain coefficients to zero through feature selection and irrelevant predictors are removed]	alpha	0.5
L2 regularisation or ridge penalty or ridge regression [shrinks the regression coefficient towards zero so that all variables are included but their individual influences are reduced]	lambda	1.0
Maximum boosting rounds	nrounds	1000
To stop the model if no improvement was seen by 30 rounds or iterations	early_stopping_rounds	30

Table 22: XGB Hyperparameters used (Round 1)

3.7.2 Results

3.7.2.1 Model accuracy

This initial model achieved an accuracy rate of 45% for all classes (no amputation, minor amputation and major amputation as classes 0,1 and 2). Multiple combinations of hyperparameters were explored to fine tune the model. After finetuning the model by adjusting the hyperparameter values and running XGB with the different combinations, the accuracy improved to 63.4% (CI 0.52-0.7; p-value <0.001). This required 320 iterations to achieve the accuracy rate stated above. The following hyperparameters were tuned to the following combination of settings which yielded a higher accuracy rate (table 23).

Hyperparameter	Code in R	Value set
Learning rate	Eta	0.05
Depth of the DTs	max_depth	4
Number of weights in each leaf node/child node	min_child_weight	15
Sample percentage rows per tree	subsample	0.8
Sample percentage columns per tree	colsample_bytree	0.6
L1 regularisation or lasso penalty	alpha	20.0
L2 regularisation or ridge penalty or ridge regression	lambda	10.0
Maximum boosting rounds	nrounds	1000
To stop the model if no improvement was seen by 30 rounds or iterations	early_stopping_rounds	30

Table 23: XGB hyperparameters used (improved accuracy)

The confusion matrix below from the test set demonstrates the model was better at correctly predicting class 0 – no amputation group and class 1 – minor amputation group compared to the previous model (RaF). It did however, falter for class two despite multiple combinations and weight adjustments.

	Class 0	Class 1	Class 2
Class 0	34	0	3
Class 1	22	18	5
Class 2	0	0	0

Table 24: XGB confusion matrix for the test set. Class 0 - no amputation, Class I- minor amputation and Class II – major amputation.

3.7.2.2 Feature selection

Table 24 and figure 39 below summarise the order in which the features had overall impact on the outcome class. This is defined by three elements – Gain, Cover and Frequency [150].

Gain is the relative contribution of each feature to the overall ML model. It is calculated by calculating the contribution each feature has to every DT. It is a good feature for predictive power and a higher value implies the importance of that feature in the boosted DT [150]. Cover represents the frequency and the relative observations with which that feature helps in the partition or split in the DT [150]. Frequency is the number of times a feature or predictor appears in the DTs in all the various splits [150].

This demonstrates the absence of osteomyelitis is the predominant driver in the overall outcome class. It offers good predictive power and appears in multiple DTs. This is followed by the presence of Charcot Foot but the cover and gain drop.

Feature	Gain	Cover	Frequency
OM (No)	0.425776	0.360521	0.320388
Charcot (Yes)	0.230133	0.159410	0.203884
OM (Yes)	0.162643	0.203251	0.182039
Ulcer (Yes)	0.084301	0.109509	0.118932
DMS (Yes)	0.053385	0.089741	0.104369
Collection (Yes)	0.043762	0.077567	0.070388

Table 25: XGB feature selection; OM – osteomyelitis, DMS – diabetic myopathy and sarcopenia.

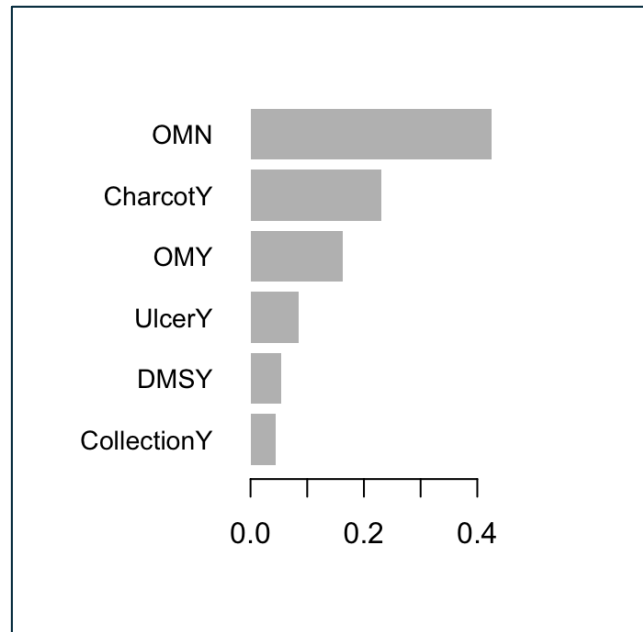


Figure 34: XGB visual chart demonstrating feature importance; OM – osteomyelitis, DMS – diabetic myopathy and sarcopenia.

3.7.2.3 ROC-AUC for Extreme Gradient Boosting.

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class. For class 0 (no amputation), the ROC-AUC was 0.724, For class 1 (minor amputation), the ROC-AUC was 0.779 and for class 2 (major amputation), the ROC-AUC was 0.439.

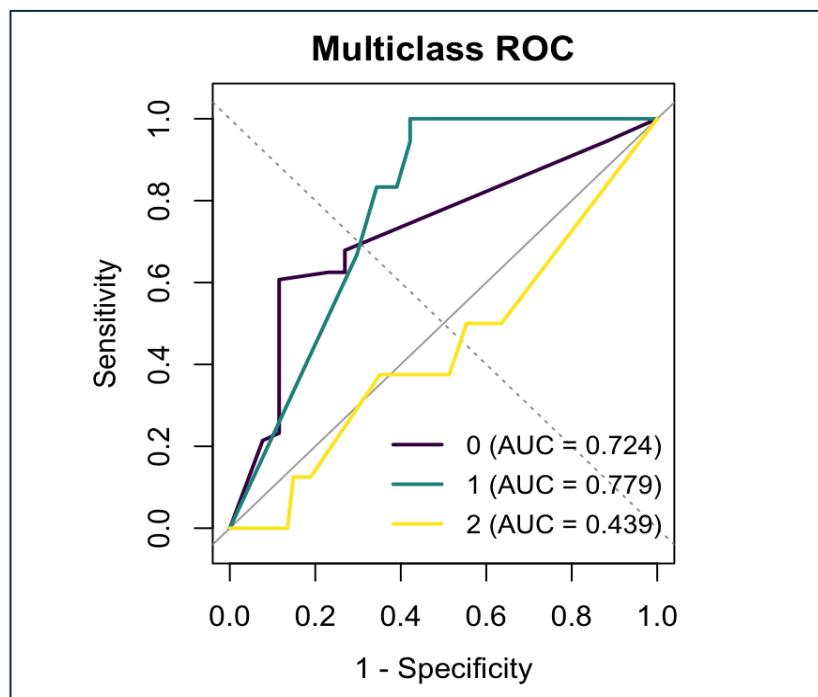


Figure 35: ROC-AUC for Extreme Gradient Boosting

3.8 Multilayer Perceptron

The background for MLPs has already been discussed at the start of this chapter and will therefore not be rediscussed here.

3.8.1 Method

A Python extension was incorporated in Statistical Package for Social Sciences (SPSS) version 29.0.2.0. The dataset was imported from an Excel file and prepared. The dependent variable was set as the 'Outcome' column. The factors/predictors were the remaining columns.

The data training and test set was set at a ratio of 80:20. The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. This defines the learning rate. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output. SoftMax is a function which "converts a vector of raw prediction scores (often called logits) from the MLP into probabilities. These probabilities are then distributed across different classes such that their sum equals 1" [151]. The initial accuracy of the model was low with a 40% error rate. In order to attempt to improve this, the ratio was then set to 80:20 to increase the number available in the test set.

3.8.2 Results

3.8.2.1 Model accuracy

The model 71.6% accurate using the training set but this dropped slightly to 60.6% on the test set. The MLP model performed best on when predicting class 0 and 1 outcomes as shown in the confusion matrix in table 26. The cross-entropy was high i.e. >1 which significantly limits its use or any functionality. Figure 41 below is a visual summary of the MLP process.

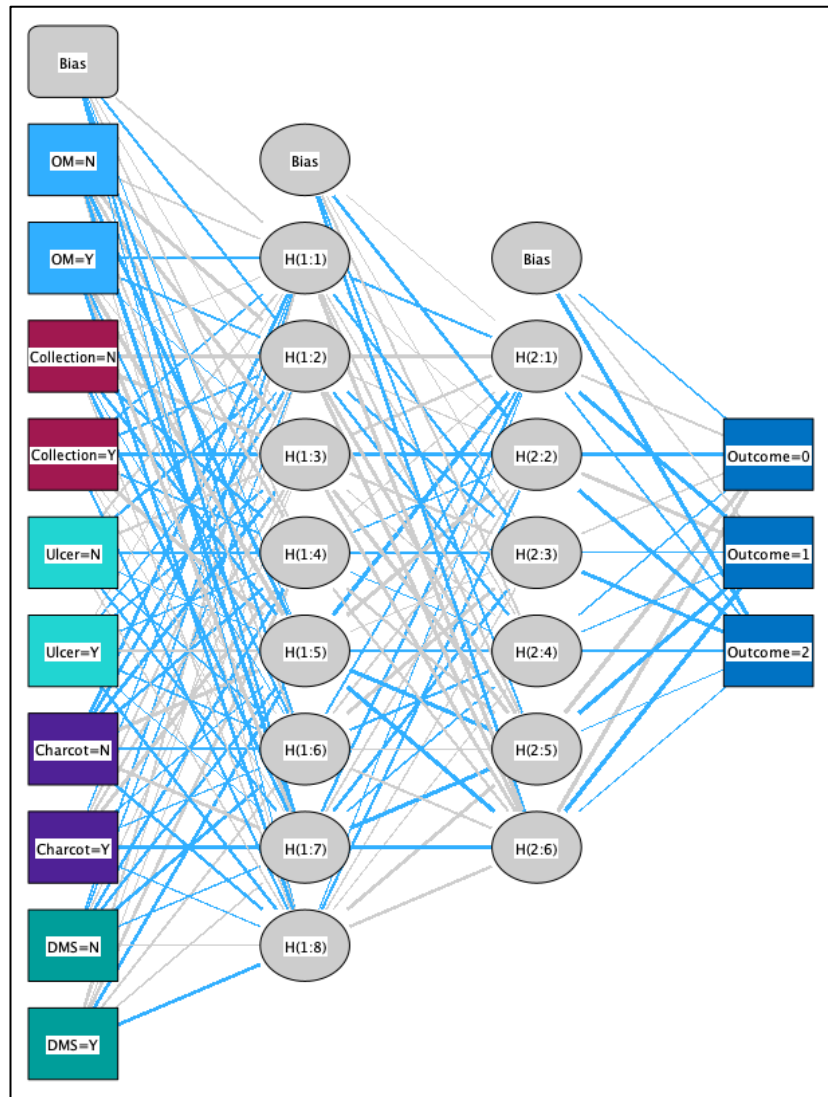


Figure 36: MLP output. This is a visual representation of the MLP. The coloured boxes on the left indicate the predictors. The grey ovals in the centre show the number of hidden layers. The boxes on the right are the outcome classes. The lines in blue and grey show the connections to each hidden layer. Synaptic weights which are indicative of the strength of a connection i.e. greater than or less than zero. It is the equivalent of an action potential between neurons. If it is significant threshold value, it then signals to the next neuron. The blue lines represent a synaptic weight of less than zero. The grey lines represent a synaptic weight greater than zero. Outcome 0 (no amputation), Outcome 1 (minor amputation), Outcome 2 (major amputation).

	Class 0	Class 1	Class 2
Class 0	80	35	17
Class 1	1	0	0

Class 2	1	0	0
----------------	----------	----------	----------

Table 26: Confusion matrix for Multilayer Perceptron Model testing set

3.8.2.3 ROC-AUC for Multilayer Perceptron Model

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class. For class 0 (no amputation), the ROC-AUC was 0.531, For class 1(minor amputation), the ROC-AUC was 0.704 and for class 2 (major amputation), the ROC-AUC was 0.672.

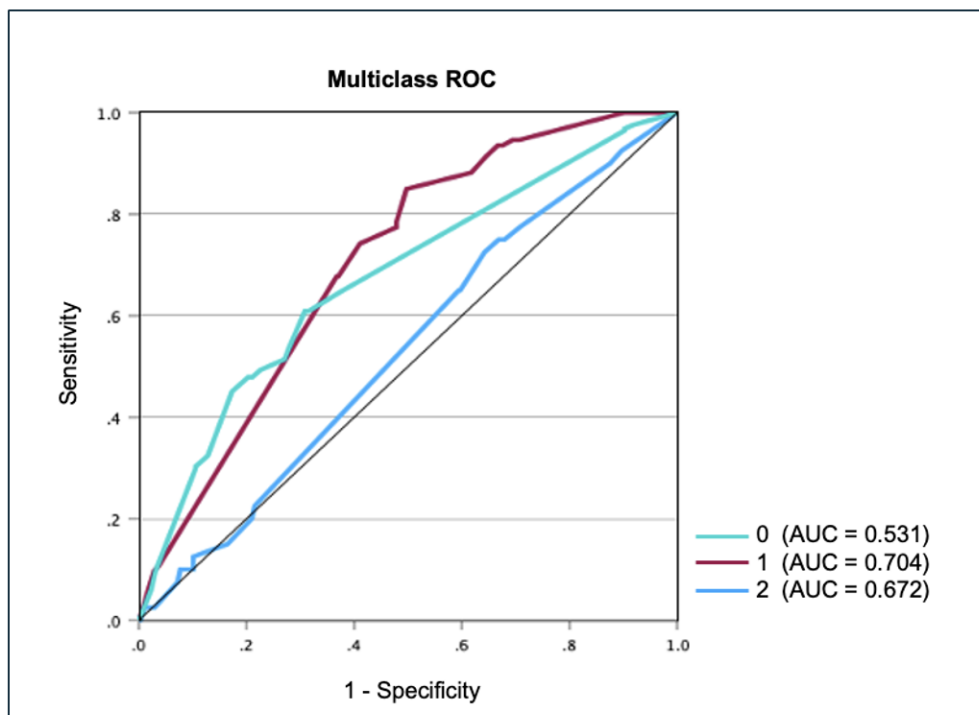


Figure 37: ROC AUC Multilayer Perceptron

3.9 Composite Analysis

Further composite analysis was undertaken to overcome small groups within the individual amputation cohorts and to address the issue with class imbalance. A further outcome added for further analysis was death whether the models could predict death as an outcome based on the MRI reports. RaF, XGB and MLP was run again using the following groups:

1. No amputation versus all amputation (minor and major combined).
2. Major lower limb amputation and death.

3. All amputation (minor and major combined) and death.

3.9.1 Data re-labelling

The data was reviewed and relabelled to create three separate datasets. The following rules were applied to the new datasets:

- All patients who had an amputation (minor or major) and then died were labelled under the death category as this outcome superseded in terms of severity the amputation outcome.
- In the no amputation versus amputation group, patients who had died but did not have an amputation were categorised under the 'no amputation' class.

3.9.2 No amputation versus all amputation group

A total of 418 patients were included. 284 patients were relabelled as Class 0 (no amputation group) and 134 were relabelled as Class 1 (all amputation group).

3.9.2.1 Random Forest (no amputation versus all amputation group)

3.9.2.1.1 Model accuracy

RaF was applied using the methodology discussed earlier. The model opted to try two variables at every split with 500 DTs. The out-of-bag error was 34.52% which was the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 27) showed the model predicted Class 0 (no amputation group) 220 times and incorrectly predicted it 8 times with a 3.5% error rate. Class 0 (all amputation group) was classified incorrectly every time.

	Class 0 (no amputation)	Class 1 (all amputation)	Class error
Class 0 (no amputation)	220	8	0.03508772
Class 1	108	0	1.0000000

(all amputation)			
-------------------------	--	--	--

Table 27: Random Forest training set confusion matrix (class 0 = no amputation, class 1= all amputation)

The RaF test set confusion matrix demonstrates the model was best at predicting class 0 but did not predict the all amputations group (table 28). The model's accuracy was 65% [95% CI 0.5455-0.7597; p-value 0.7265].

	Class 0 (no amputation)	Class 1 (all amputation)
Class 0 (no amputation)	54	26
Class 1 (all amputation)	2	1

Table 28: Random Forest test set confusion matrix (class 0 = no amputation, class 1= all amputation)

3.9.2.1.2 Performance table and feature importance

The performance table below shows sensitivity was highest for class 0 (no amputation group). Class 0 was predicted correctly 67.5% of the time. A single outcome column was generated by R with Class 0 (no amputation group) as the “positive class”. The values did not significantly differ when compared in the non-composite analysis earlier.

This implies ongoing class imbalance which did not change after adjusting for weights. RaF did not predict Class 1 correctly which is reflected by the negative predictive value.

Feature importance did not change between the analysis undertaken prior to the composite analysis with osteomyelitis and collection having the most significant impact in the no amputation group.

	Class 0 (no amputation)
Sensitivity	0.9643
Specificity	0.0000
Positive Predictive Value (PPV)	0.6750
Negative predictive value (NPV)	0.0000
Prevalence	0.6829
Detection rate	0.6585
Detection Prevalence	0.9756
Balance Accuracy	0.4821

Table 29: RaF performance table (no amputation versus all amputation group)

3.9.2.1.3 ROC-AUC for Random Forest

An ROC-AUC was generated to demonstrate the accuracy of the model for both classes (figure 37). For class 0 (no amputation group), the ROC-AUC was 0.664, For class 1 (all amputation group), the ROC-AUC was 0.664 which is an improvement compared to the previous model performance for no amputation versus minor amputation versus major amputation prediction. The micro-AUC was 0.753 which again reflects the class imbalance.

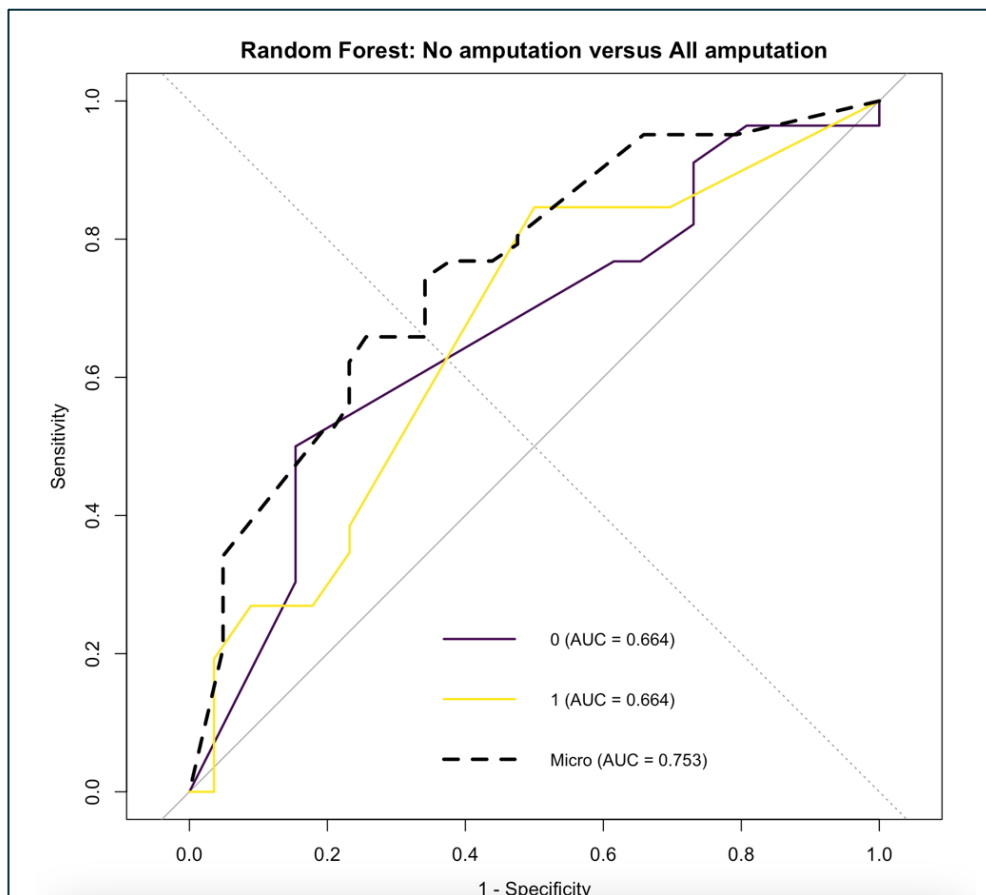


Figure 38: Random Forest ROC-AUC. No amputation versus amputation.

3.9.2.2 Extreme Gradient Boosting (no amputation versus all amputation group)

3.9.2.2.1 Model Accuracy

XGB was applied to the no amputation versus all amputation group dataset. The model demonstrated an accuracy of 68.29% [95% CI 0.57-0.78; p-value 0.55]. This required 97 iterations to achieve the accuracy rate stated. The confusion matrix below from the test set demonstrates the model was good at predicting both classes but was not statistically significant (table 31).

	Class 0 (no amputation)	Class 1 (all amputation)
Class 0 (no amputation)	37	7
Class 1 (all amputation)	19	19

Table 30: Extreme gradient boosting - confusion matrix. No amputation versus all amputation group.

3.9.2.2.2 Feature importance

This demonstrates the absence and presence of osteomyelitis as the predominant drivers in the overall outcome classes. The presence of ulcers also had an influence but to a lesser extent. It offers very good predictive power and appears in multiple DTs. No other features were deemed relevant by the model.

Feature	Gain	Cover	Frequency
OM (No)	0.641213634	0.61462911	0.61290323
OM (Yes)	0.354993038	0.35250677	0.35483871
Ulcer (Yes)	0.2236918	0.3687985	0.3714286

Table 31: Extreme gradient boosting - feature selection. No amputation versus all amputation group. OM- osteomyelitis

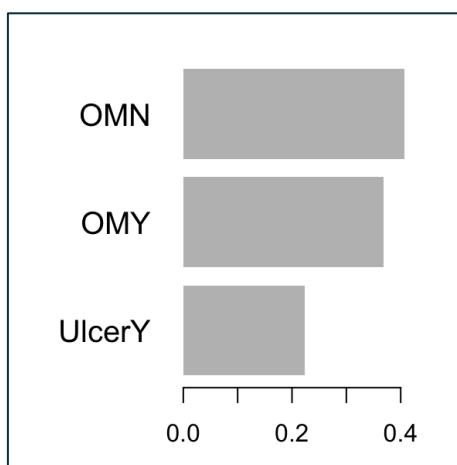


Figure 39: Extreme Gradient Boosting - Feature importance. No amputation versus All amputation group; OM- Osteomyelitis.

3.9.2.2.3 ROC-AUC Extreme Gradient Boosting

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class (figure 39). For class 0 (no amputation group) and class 1 (all amputation group), the ROC-AUC was 0.714.

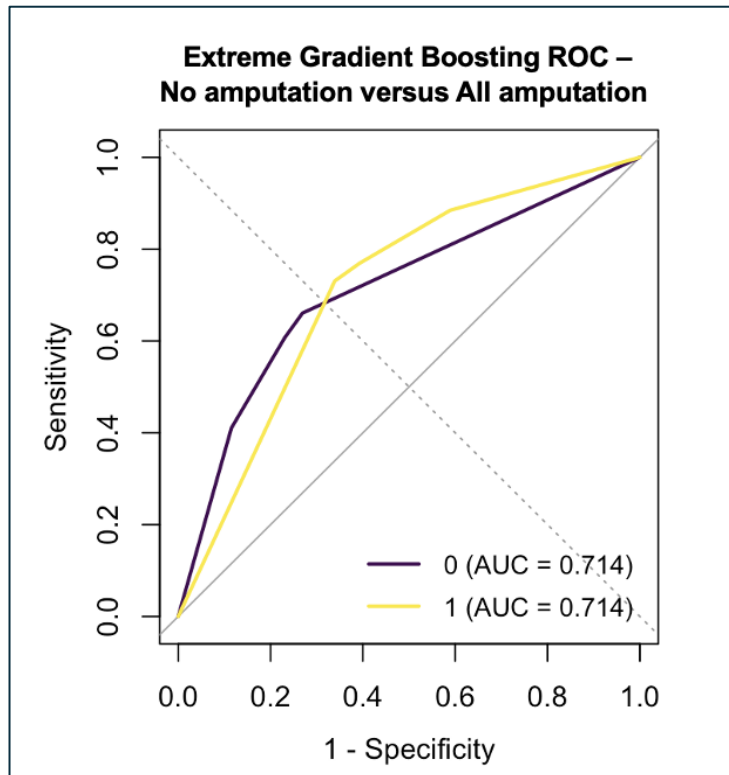


Figure 40: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1(all amputations)

3.9.2.3 Multilayer Perceptron (no amputation versus all amputation group)

As before, the data training and test set was set at a ratio of 80:20. The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output.

3.9.2.3.1 Model accuracy

The model was 70.7% accurate using the training set but this dropped slightly to 61.9% on the test set. The MLP model performed best on when predicting class 0 and 1 outcomes as shown in the confusion matrix in table 33. Two hidden layers were used. There were 7 units in the first hidden layer and 5 units in the second hidden layer. The cross-entropy was still high i.e. >1 which significantly limits its use or any functionality.

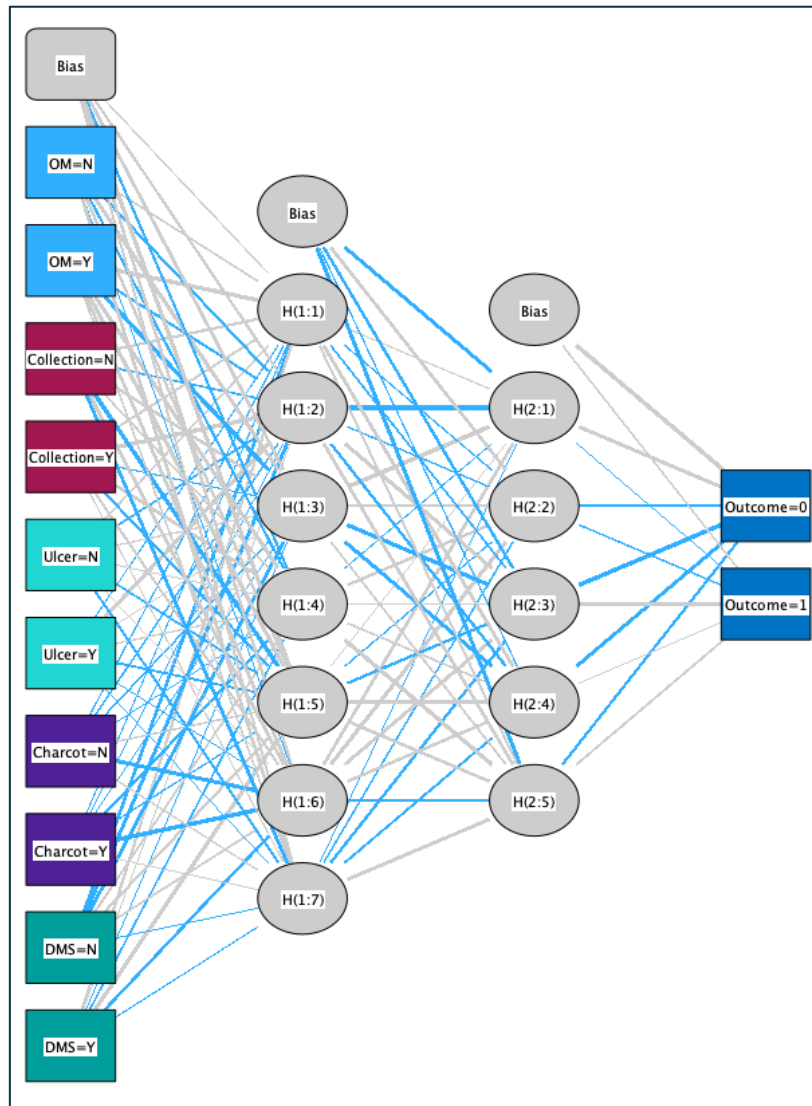


Figure 41: Multilayer perceptron output (no amputation versus all amputation). Class 0 (no amputation), Class 1 (all amputations)

	Class 0 (no amputation)	Class 1 (all amputations)
Class 0 (no amputation)	49	1
Class 1 (all amputations)	31	3

Table 32: Multilayer perceptron confusion matrix (no amputation versus all amputations)

3.9.2.3.2 ROC-AUC for Multilayer Perceptron

The ROC-AUC for both classes 0 and 1 was 0.671 (figure 41). Whilst this was slightly better, class 0 was still overpredicted.

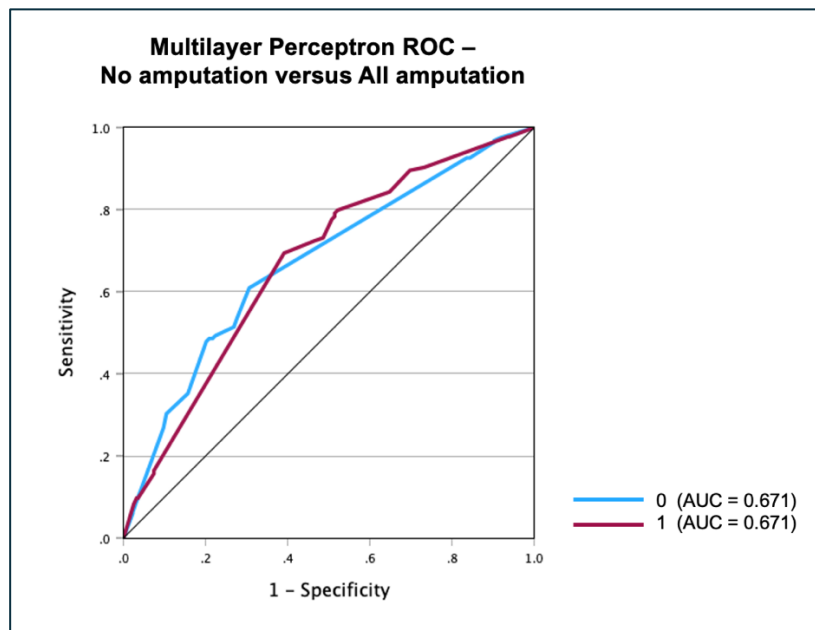


Figure 42: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus all amputations (class1)

3.9.3 Major lower limb amputation versus death

A total of 180 patients were included. 22 patients had a major lower limb amputation and then died. 17 patients underwent a major lower limb amputation but were alive. 163 patients were classed under the death category. The data was relabelled into Class 0 (major LLA) and Class 1 (Death).

3.9.3.1 Random Forest (major lower limb amputation versus death)

3.9.3.1.1 Model accuracy

RaF was applied using the methodology discussed earlier. The model opted to try two variables at every split with 500 DTs. The out-of-bag error was 10.34% i.e. the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 34) shows the model predicted Class 0 (no amputation group) 220 times and incorrectly predicted it 8 times with a 3.5% error rate. Class 0 (all amputation group) was classified incorrectly every time.

	Class 0 (MLLA)	Class 1 (Death)	Class error
Class 0 (MLLA)	0	14	1.000000
Class 1 (Death)	1	130	0.007633

Table 33: Random Forest training set confusion matrix. Major lower limb amputation (MLLA) and death

The RaF test set confusion matrix demonstrates the model was best at predicting class 1 (death) but not major lower limb amputation (table 35). The model's accuracy was 91% [95% CI 0.78-0.98; p-value 0.647].

	Class 0 (MLLA)	Class 1 (Death)
Class 0 (MLLA)	54	26
Class 1 (Death)	2	1

Table 34: Random Forest test set confusion matrix

3.9.3.1.2 Performance table and feature importance

The model heavily leans towards class imbalance and overcalls the death category/class as shown in table 36. Feature importance did not change between the analysis undertaken prior to the composite analysis with osteomyelitis and collection having the most significant impact on the outcomes.

	Class 0 (MLLA)
Sensitivity	0.0000
Specificity	1.0000

Positive Predictive Value (PPV)	Not given
Negative predictive value (NPV)	0.91429
Prevalence	0.08571
Detection rate	0.00000
Detection Prevalence	0.00000
Balance Accuracy	0.5000

Table 35: Random Forest performance table (major lower limb amputation versus death)

3.9.2.1.3 ROC-AUC for Random Forest

An ROC-AUC was generated to demonstrate the accuracy of the model for both classes (figure 42). For class 0 (major lower limb amputation), the ROC-AUC was 0.552, For class 1 (death), the ROC-AUC was 0.522. The micro-AUC was 0.922 which again reflects the class imbalance.

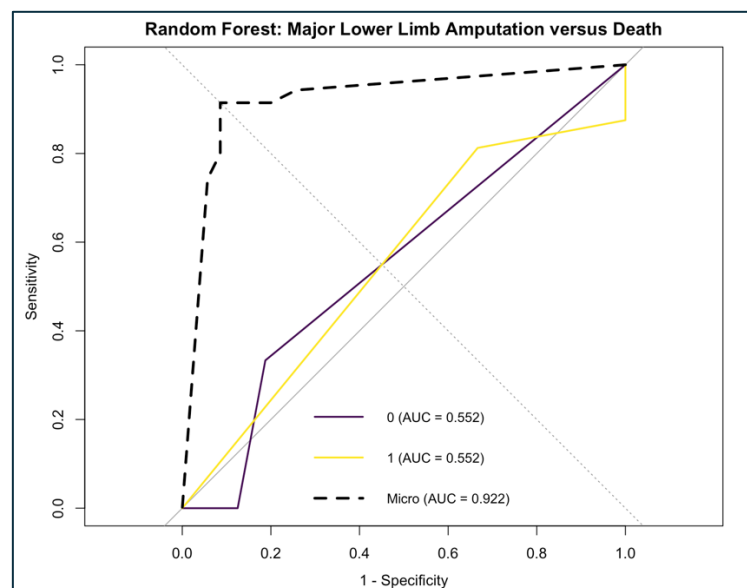


Figure 43: Random Forest ROC-AUC. Major lower limb amputation versus death - Class 0 (major lower limb amputation), Class 1 (death).

3.9.3.2 Extreme Gradient Boosting (major lower limb amputation versus death)

3.9.3.2.1 Model Accuracy

XGB was applied to the major lower limb amputation versus death. The model had an accuracy of 68.57% [95% CI 0.51-0.83; p-value 0.99]. This required 32 iterations to achieve the accuracy rate stated.

The confusion matrix below from the test set demonstrates the model was good at predicting both classes but was not statistically significant.

	Class 0 (MLLA)	Class 1 (Death)
Class 0 (MLLA)	1	9
Class 1 (Death)	2	23

Table 36: Extreme gradient boosting confusion matrix of the test set. Major lower limb amputation (MLLA) versus death.

3.9.3.2.2 Feature importance

Table 38 and figure 43 demonstrate the absence and presence of osteomyelitis as the predominant drivers in the overall outcome. The presence of ulcers also had an influence but to a lesser extent. No other features were deemed relevant by the model.

Feature	Gain	Cover	Frequency
OM (No)	0.6283643	0.5370210	0.5348837
OM (Yes)	0.2714944	0.2744158	0.2790698
Ulcer (Yes)	0.1001413	0.1885632	0.1860465

Table 37: Extreme gradient boosting feature selection. Major lower limb amputation versus death. OM- osteomyelitis.

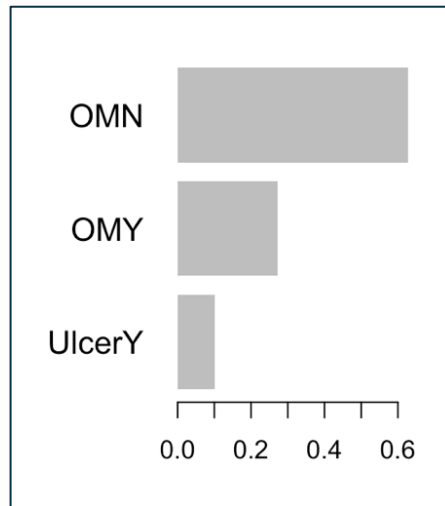


Figure 44: Extreme Gradient Boosting feature importance. Major lower limb amputation versus death; OM- Osteomyelitis.

3.9.3.2.3 ROC-AUC Extreme Gradient Boosting

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class. For class 0 (major lower limb amputation) and class 1 (death), the ROC-AUC was 0.536 (figure 44).

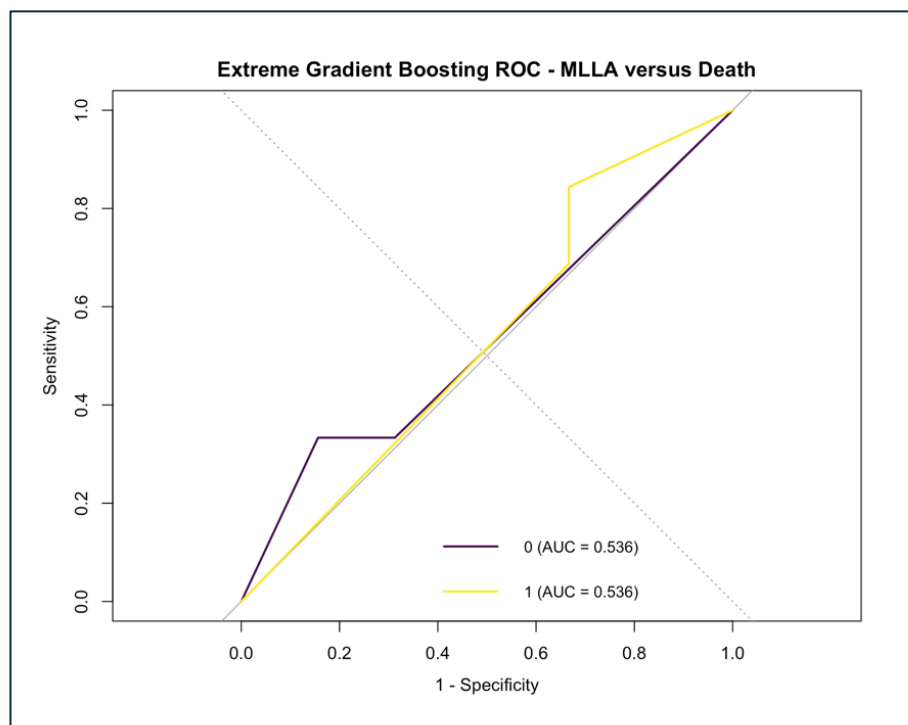


Figure 45: ROC-AUC Extreme Gradient Boosting. Class 0 (major lower limb amputation), Class 1 (death).

3.9.3.3 Multilayer Perceptron (major lower limb amputation versus death)

As before, the data training and test set was set at a ratio of 80:20. The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output.

3.9.3.3.1 Model accuracy

The model was 88.8% accurate using the training set but this increased to 97.3% on the test set. The MLP model performed best on when predicting class 1 (death) as shown in the confusion matrix in table 39 but overpredicted this every time. It is worth noting that there were fewer records and therefore fewer still for the training set which may be a contributing factor. Two hidden layers were used (figure 44). There were 7 units in the first hidden layer and 5 units in the second hidden layer. The cross-entropy was slightly lower but still >1 which limits its use.

	Class 0 (MLLA)	Class 1 (Death)
Class 0 (MLLA)	0	1
Class 1 (Death)	0	36

Table 38: MLP confusion matrix. Major lower limb amputation (MLLA) versus Death.

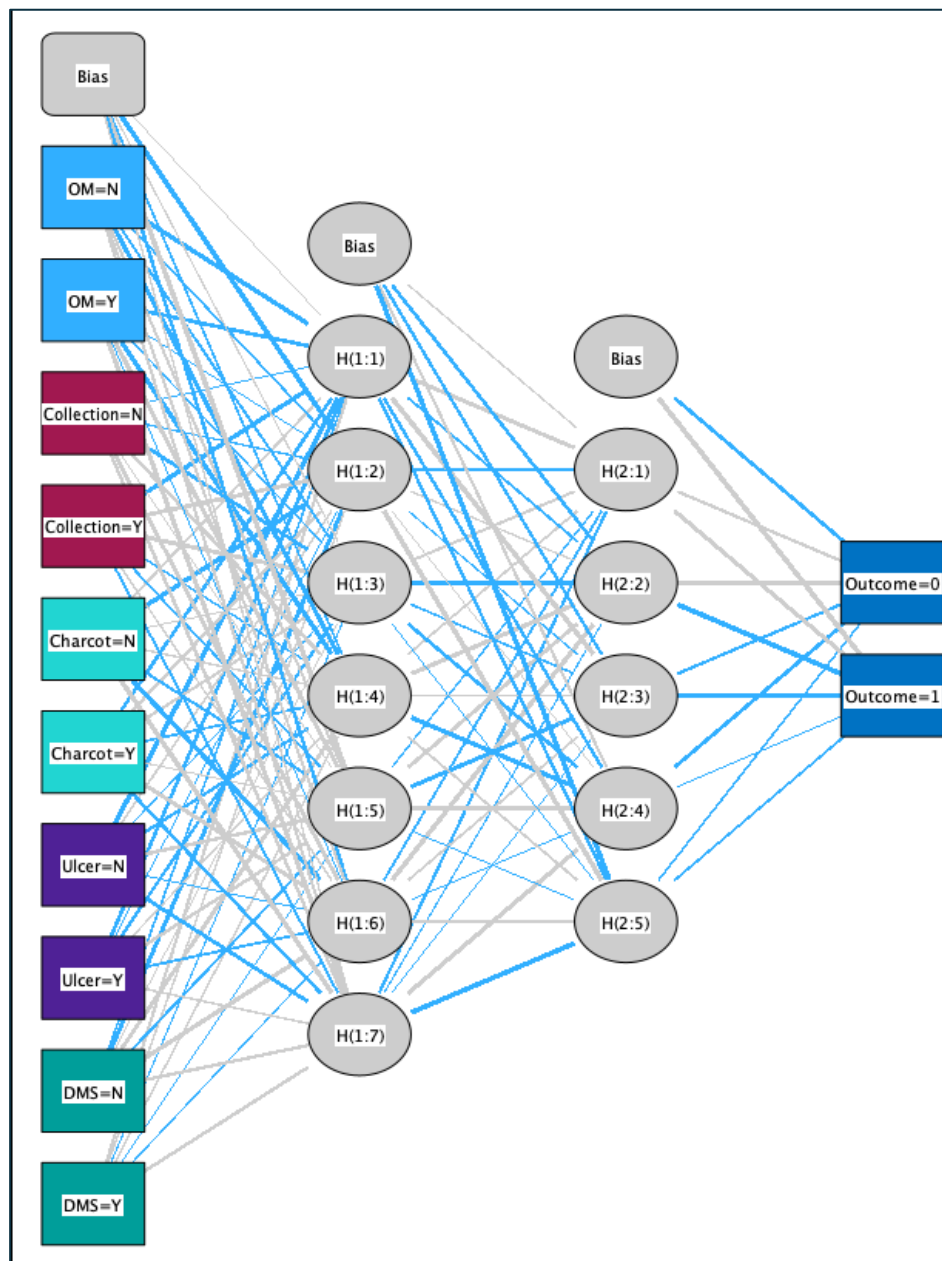


Figure 46: MLP output. Major lower limb amputation versus death

3.9.2.3.2 ROC-AUC for Multilayer Perceptron

The ROC-AUC for both classes 0 and 1 was 0.676 (figure 45).

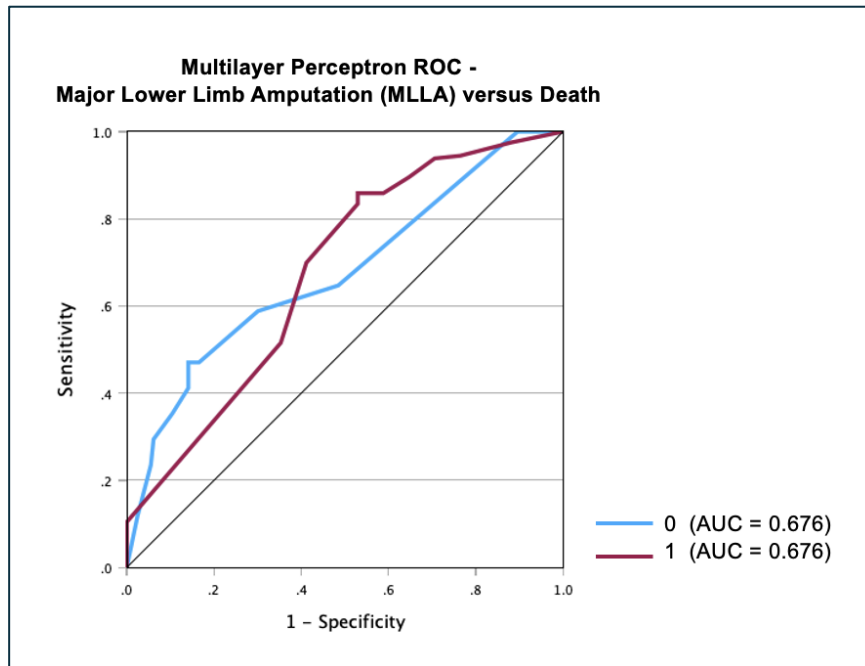


Figure 47:ROC-AUC for MLP. Class 0 (major lower limb amputation), Class 1 (death)

3.9.4 All amputations versus death

A total of 240 patients were included. Class 0 (all amputation group) had 84 patients and Class 1 (Death) had 155 patients.

3.9.4.1 Random Forest (major lower limb amputation versus death)

3.9.4.1.1 Model accuracy

RaF was applied using the methodology discussed earlier. The model opted to try two variables at every split with 500 DTs. The out-of-bag error was 36.46% i.e. the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 40) showed the model predicted Class 0 (all amputations group) 5 times and incorrectly 63 times with a 92.6% error rate. Class 1 (death) was classified correctly 117 times with an error rate of 5.6%.

	Class 0 (All amputations)	Class 1 (Death)	Class error
Class 0 (All amputations)	5	63	0.92647059

Class 1 (Death)	7	117	0.05645161
----------------------------	---	-----	------------

Table 39: Random Forest training set confusion matrix (all amputation versus death)

On the test set, the model had an accuracy of 61.7% [95% CI 0.4638-0.7549, p-value 0.78] as shown by the correctly identified classes in table 41 below.

	Class 0 (All amputations)	Class 1 (Death)
Class 0 (All amputations)	1	3
Class 1 (Death)	15	28

Table 40: Random Forest test set confusion matrix (all amputations versus death)

3.9.4.1.2 Performance table

The performance table below shows a very low sensitivity for detecting class 0 but had a very high specificity such that it correctly identified the true negatives almost 90% of the time and therefore was better at classifying the death group. This is also reflective in the negative predictive value of almost 65% i.e. when it predicts the death group, it was 65% correct.

	Class 0 (all amputation)
Sensitivity	0.06250
Specificity	0.90323
Positive Predictive Value (PPV)	0.2500
Negative predictive value (NPV)	0.34043

Prevalence	0.6829
Detection rate	0.02128
Detection Prevalence	0.08511
Balance Accuracy	0.48286

Table 41: Random Forest performance table (all amputations versus death)

3.9.4.1.3 Feature importance

Interestingly, feature importance did change slightly when comparing all amputations versus death as outcomes (figure 46). Diabetic Myopathy and Sarcopenia were strong predictors for all amputation and death as outcomes whilst osteomyelitis reported on MRI was a strong predicting feature for death but not amputation (table 43).

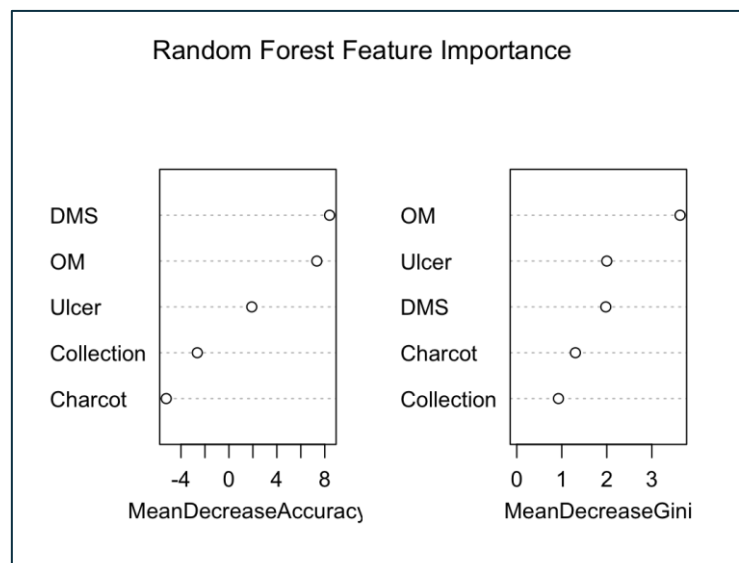


Figure 48: Random Forest feature importance (all amputations versus death)

Variable	Class 0 (All amputations)	Class 1 (Death)	Mean decrease accuracy	Mean decrease Gini
OM	-3.5459734	8.696727	7.338373	3.6265437

Collection	0.9860347	-4.589809	-2.636424	0.9265337
Ulcer	-8.3519702	8.032861	1.910669	1.9996572
Charcot	-3.9837178	-4.104532	-5.226162	1.3025948
DMS	6.8487400	7.441170	8.400551	1.9771219

Table 42: Random Forest feature importance (all amputations versus death)

3.9.4.1.4 ROC-AUC for Random Forest

An ROC-AUC was generated to demonstrate the accuracy of the model for both classes (figure 47). The ROC-AUC was 0.557 for both classes.

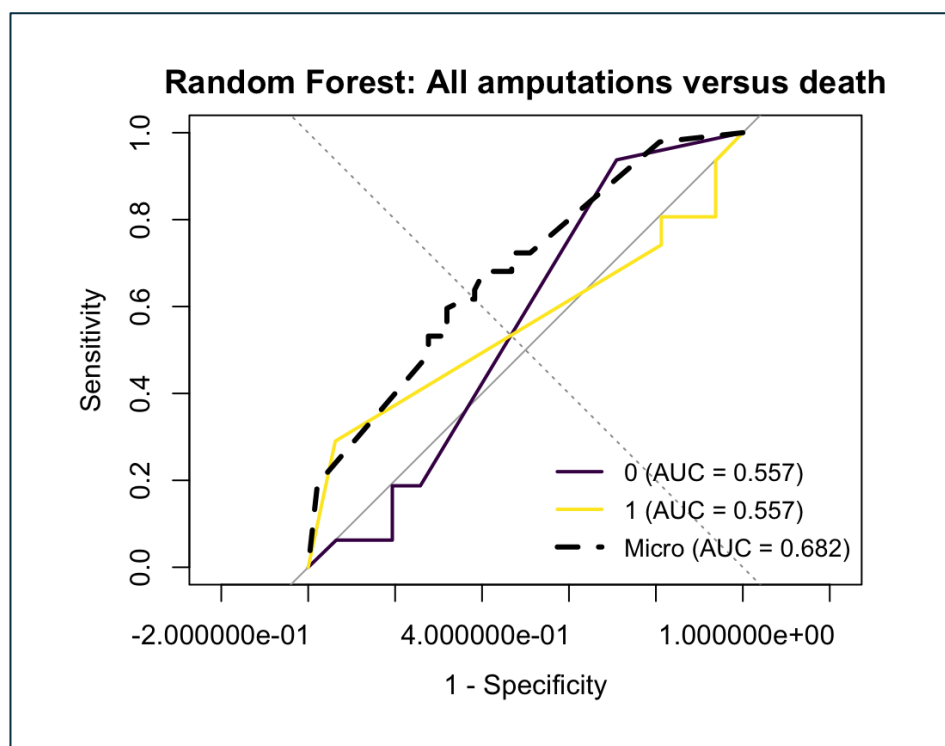


Figure 49: Random Forest ROC-AUC. All amputations versus death.

3.9.4.2 Extreme Gradient Boosting (major lower limb amputation versus death)

3.9.4.2.1 Model Accuracy

XGB was applied to the cohort using the MRI report to compare the outcome of major lower limb amputation versus death.

The model had an accuracy of 57.45% [95% CI 0.42-0.71; p-value 0.9]. This required 18 iterations to achieve the accuracy rate stated. The confusion matrix below from the test set

demonstrates the model was good at predicting both classes but was not statistically significant (table 44). Further analysis demonstrated a high sensitivity for predicting the all amputation group but had low precision/positive predictive value (table 45). However, compared to previous models, this offers a middle ground with the least class imbalance.

	Class 0 (All amputations)	Class 1 (Death)
Class 0 (All amputations)	15	19
Class 1 (Death)	1	12

Table 43: Extreme Gradient Boosting. Test set confusion matrix. All amputations versus death.

	Class 0 (all amputation)
Sensitivity	0.9375
Specificity	0.3871
Positive Predictive Value (PPV)	0.4412
Negative predictive value (NPV)	0.923
Prevalence	0.3404
Detection rate	0.3191
Detection Prevalence	0.7234
Balance Accuracy	0.6623

Table 44: Random Forest performance table. All amputations versus death.

3.9.4.2.2 Feature importance

This demonstrates the absence of osteomyelitis and the presence of diabetic myopathy/sarcopenia as the predominant drivers for the overall outcome classes (table 46 and figure 49). The presence of ulcers also had an influence but to a lesser extent.

Feature	Gain	Cover	Frequency
OM (No)	0.44411051	0.23733756	0.20930233
DMS (Yes)	0.15697081	0.14245894	0.13178295
Ulcer (Yes)	0.14224939	0.18895296	0.20930233
Charcot (Yes)	0.13508340	0.18305358	0.20155039
Collection (Yes)	0.06615906	0.16912543	0.17829457
OM (Yes)	0.05542684	0.07907154	0.06976744

Table 45: Extreme gradient boosting. Feature importance. All amputation versus death; OM- Osteomyelitis, DMS- Diabetic Myopathy and Sarcopenia.

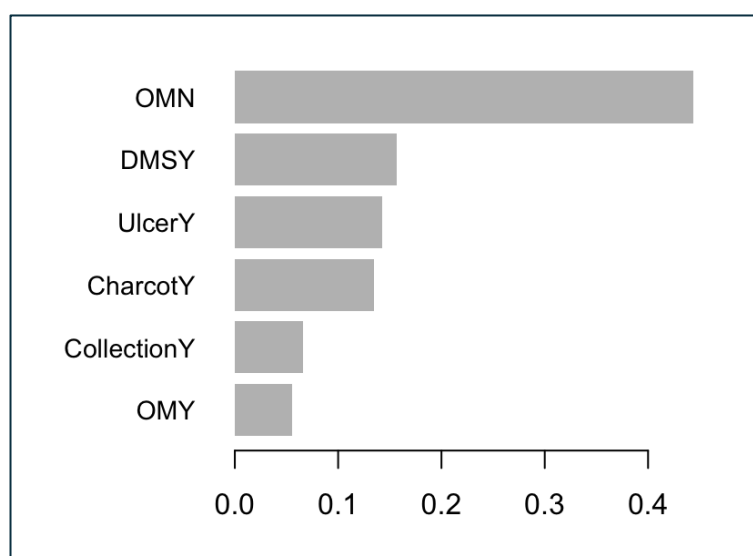


Figure 50: Extreme gradient boosting. Feature importance (visual). All amputation versus death; OM- Osteomyelitis, DMS- Diabetic Myopathy and Sarcopenia.

3.9.4.2.3 ROC-AUC Extreme Gradient Boosting

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class. Both groups i.e. all amputation and death had an AUC of 0.605.

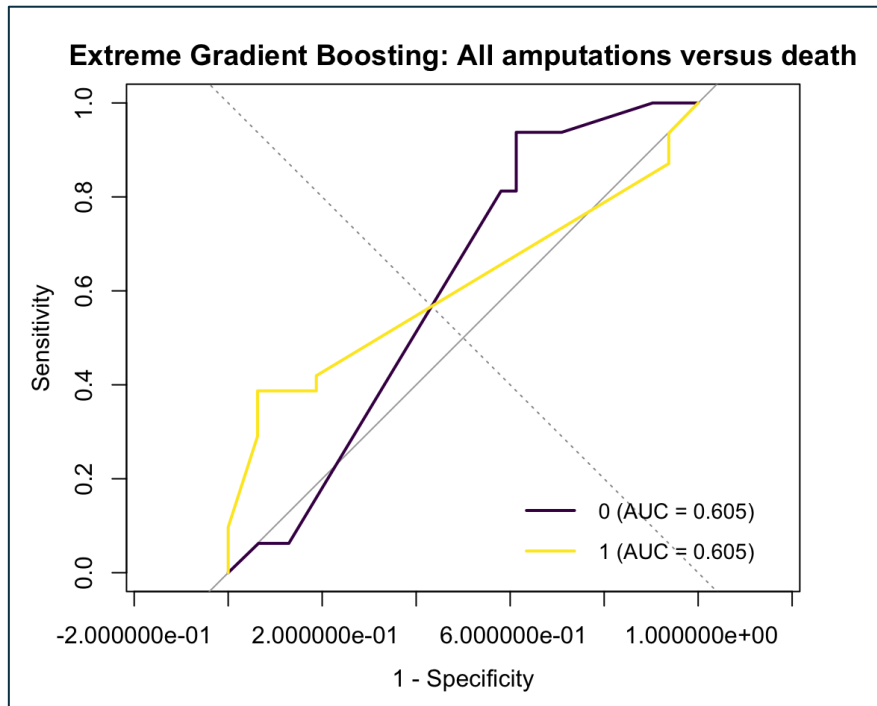


Figure 51: Extreme Gradient Boosting ROC-AUC. Class 0 (all amputations), Class 1 (death).

3.9.4.3 Multilayer Perceptron (major lower limb amputation versus death)

The data set was divided into the training and test set (80:20). The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output.

3.9.3.3.1 Model accuracy

The model was 63.9% accurate using the training set but this increased to 67.1% on the test set. The MLP model again performed best when predicting class 1 (death) as shown in the confusion matrix in table 47. Two hidden layers were used (figure 51). There were 7 units in the first hidden layer and 5 units in the second hidden layer. The cross-entropy was comparable to the major lower limb amputation and death group MLP model - slightly lower but still >1 which limits its use.

	Class 0 (All amputations)	Class 1 (Death)
Class 0 (All amputations)	0	15

Class 1 (Death)	0	31
----------------------------	----------	-----------

Table 46: Multilayer Perceptron. Test set confusion matrix. All amputations versus death.

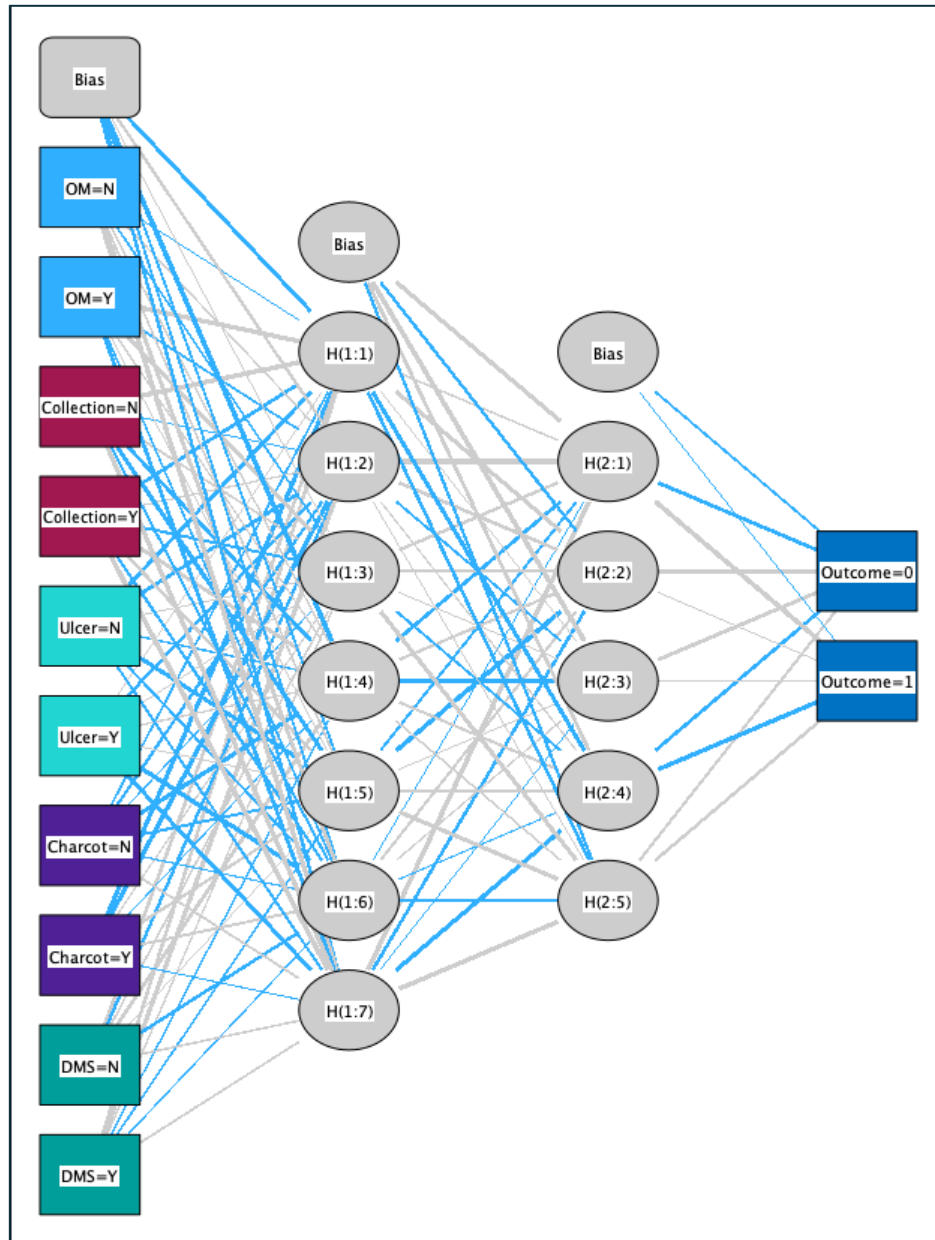


Figure 52: Multilayer perceptron output. Class 0 (all amputations) and Class 1 (death).

3.9.4.3.2 ROC-AUC for Multilayer Perceptron

The ROC-AUC for both classes 0 and 1 was 0.620 (figure 52).

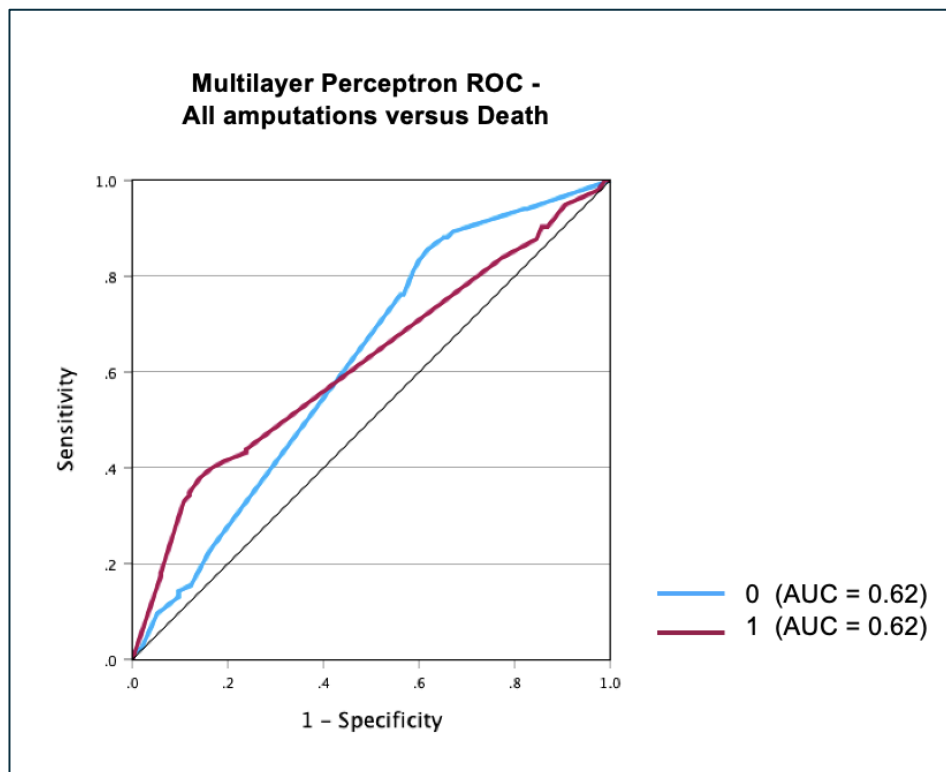


Figure 53: Multilayer perceptron ROC-AUC. Class 0 (all amputations), Class 1(death)

3.9.5 No amputation versus major lower limb amputation and death

A total of 359 patients were included. 176 patients were labelled as Class 0 (no amputation group). 183 patients were included in Class 1 (18 MLLAs and 165 deaths).

3.9.5.1 Random Forest (no amputation versus major lower limb amputation and death)

3.9.5.1.1 Model accuracy

RaF was applied using the methodology discussed earlier. The model opted to try two variables at every split with 500 DTs. The out-of-bag error was 40.28% which was the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 47) showed the model predicted Class 0 (no amputation group) 73 times and incorrectly predicted it 68 times with a 48% error rate. Class 1 (MLLA and death) was classified correctly 99 times and incorrectly 48 times with an error rate of 32.65%.

	Class 0 (No amputation)	Class 1 (MLLA + Death)	Class error
--	----------------------------	---------------------------	-------------

Class 0 (No amputation)	73	68	0.4822695
Class 1 (MLLA + Death)	48	99	0.3265306

Table 47: Random Forest training set confusion matrix (class 0 = no amputation, class 1= MLLA and death).

The RaF test set confusion matrix demonstrates the model was good at predicting both classes. The model's accuracy was 57.8% [95% CI 0.4544-0.6939; p-value 0.1427].

	Class 0 (no amputation)	Class 1 (MLLA + Death)
Class 0 (no amputation)	16	11
Class 1 (MLLA + Death)	19	25

Table 48: Random Forest test set confusion matrix (class 0 = no amputation, class 1= MLLA and death).

3.9.5.1.2 Performance table

The performance table below shows sensitivity was highest for class 0 (no amputation group). Class 0 was predicted correctly 59% of the time. A single outcome column was generated by R with Class 0 (no amputation group) as the "positive class". The model correctly identifies the no amputation group almost 46% of the time and the combined MLLA and death group 69% of the time. Almost 49% of the dataset was in Class 0 (no amputation group) which shows a more balanced group. The detection prevalence of 38% implies that the model predicts the no amputation group for 38% of the patients. The balanced accuracy which is an average of the sensitivity and specificity has a value of 0.5758 which means the model is no better than random guessing i.e. a 50/50 chance.

	Class 0 (no amputation)
--	------------------------------------

Sensitivity	0.4571
Specificity	0.6944
Positive Predictive Value (PPV)	0.5926
Negative predictive value (NPV)	0.5682
Prevalence	0.4930
Detection rate	0.2254
Detection Prevalence	0.3803
Balance Accuracy	0.5758

Table 49: RaF performance table (no amputation versus MLLA and death combined).

3.9.5.1.3 Feature importance

Feature importance did change between the analysis undertaken prior to the composite analysis with Charcot having the most significant impact in the no amputation group but also to the MLLA and death group therefore is not specific to either class. This is followed by followed by ulcer and osteomyelitis. The presence of osteomyelitis strongly pushes towards no amputation but equally to class 1 (MLLA and death). Ulcer has a large positive effect for class 0.

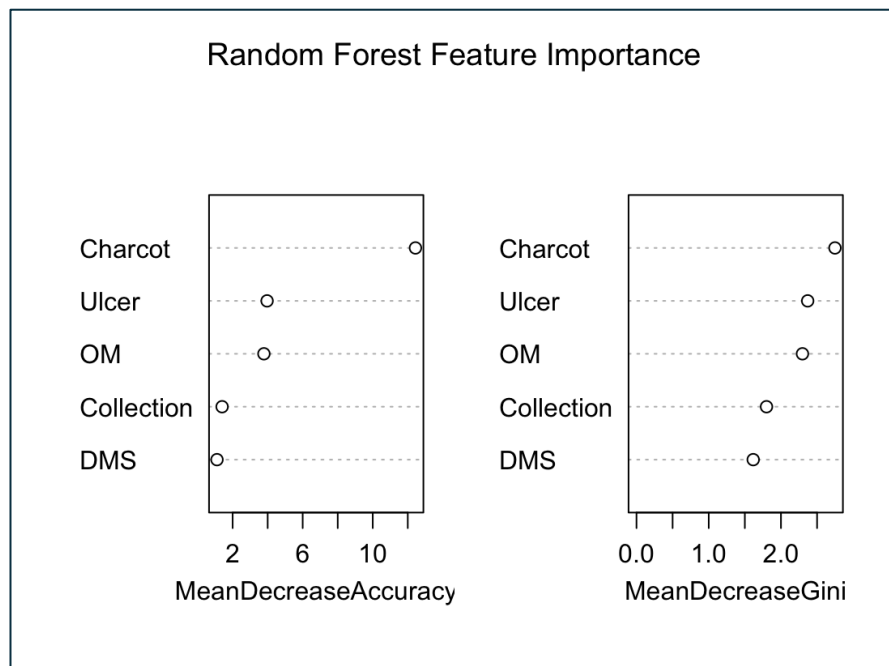


Figure 54: RaF feature importance (no amputation versus MLLA and death combined).

Variable	Class 0 (no amp)	Class 1 (MLLA and death)	Mean decrease accuracy	Mean decrease Gini
OM	16.125631	- 13.092263	3.784615	2.297681
Collection	4.017644	- 1.674431	1.397132	1.799839
Ulcer	12.131811	-9.464579	3.965866	2.369875
Charcot	11.777887	9.724882	12.438827	2.747574
DMS	3.549730	- 1.739463	1.105495	1.615933

Table 50: RaF feature importance – Class 0 (no amputation), Class 1 (MLLA and death combined).

3.9.5.1.4 ROC-AUC for Random Forest

An ROC-AUC was generated to demonstrate the accuracy of the model for both classes (figure 54). For class 0 (no amputation group), the ROC-AUC was 0.613, For class 1 (all amputation group), the ROC-AUC was 0.613. The micro-AUC was 0.608.

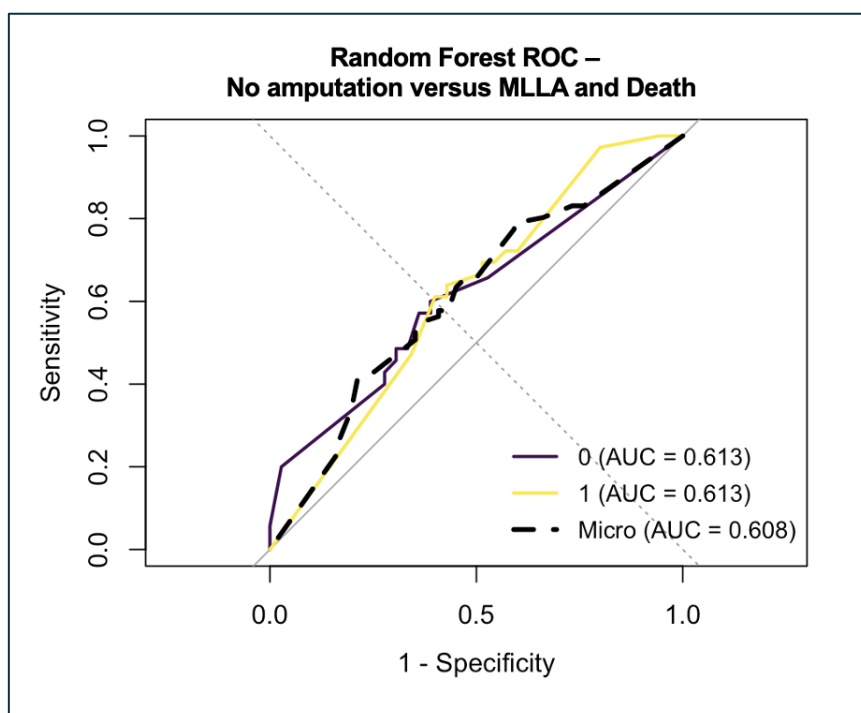


Figure 55: Random Forest ROC-AUC. No amputation versus MLLA and death. MLLA- major lower limb amputation.

3.9.5.2 Extreme Gradient Boosting (no amputation versus major lower limb amputation and death)

3.9.5.2.1 Model accuracy

XGB was applied to the no amputation versus MLLA and death group dataset. The model demonstrated an accuracy of 56.3% [95% CI 0.44-0.68; p-value 0.2]. This required 29 iterations to achieve the accuracy rate stated. The confusion matrix below from the test set demonstrates the model was good at predicting both classes but was not statistically significant (table 51). The model had a sensitivity of 45% and specificity of 66%. The positive predictive value was 0.57 and the negative predictive value was 0.56. The detection rate was low for class 0 i.e. the no-amputation group at 39%.

	Class 0 (no amputation)	Class 1 (MLLA and death)
Class 0 (no amputation)	16	12
Class 1 (MLLA and death)	19	24

Table 51: Extreme gradient boosting - confusion matrix. No amputation versus MLLA and death combined.

3.9.5.2.2 Model accuracy

This demonstrates the presence of Charcot as well as the absence of osteomyelitis as the predominant drivers in the overall outcome classes. The presence of ulcers, diabetic myopathy/sarcopenia, osteomyelitis and collection also had an influence but to a lesser extent. However, it does not offer good predictive power (table 52 and figure 55).

Feature	Gain	Cover	Frequency
Charcot (Yes)	0.27860354	0.24004565	0.21264368
OM (No)	0.23585343	0.21376074	0.20114943
Ulcer (Yes)	0.19288267	0.20937802	0.22413793
DMS (Yes)	0.11989044	0.14806931	0.18965517
OM (Yes)	0.09021403	0.06728864	0.06896552
Collection (Yes)	0.08255588	0.12145763	0.10344828

Table 52: Extreme gradient boosting - feature selection. No amputation versus MLLA and death combined. OM – osteomyelitis, DMS – diabetic myopathy/ sarcopenia.

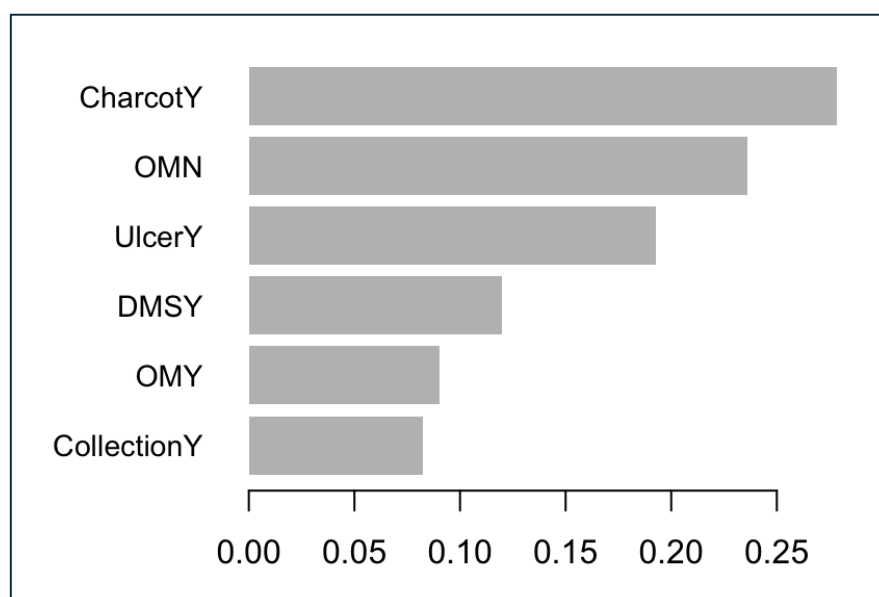


Figure 56: Extreme Gradient Boosting - Feature importance. No amputation versus MLLA and death combined.

3.9.5.2.3 ROC-AUC Extreme Gradient Boosting

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class (figure 56). For class 0 (no amputation group) and class 1 (all amputation group), the ROC-AUC was 0.614.

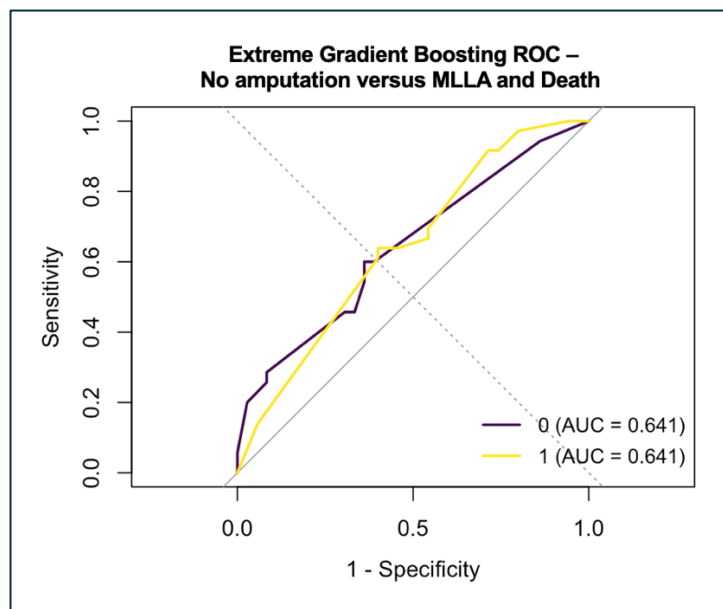


Figure 57: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1(MLLA and death combined).

3.9.5.3 Multilayer Perceptron (no amputation versus MLLA and death combined)

As before, the data training and test set was set at a ratio of 80:20. The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output.

3.9.5.3.1 Model accuracy

The model was 60.3% accurate using the training set but this dropped slightly to 60% on the test set. The MLP model performed best on when predicting both classes as shown in the confusion matrix in table 53. Two hidden layers were used. There were 7 units in the first hidden layer and 5 units in the second hidden layer. The cross-entropy was still high i.e. >1 which limits its use or any functionality.

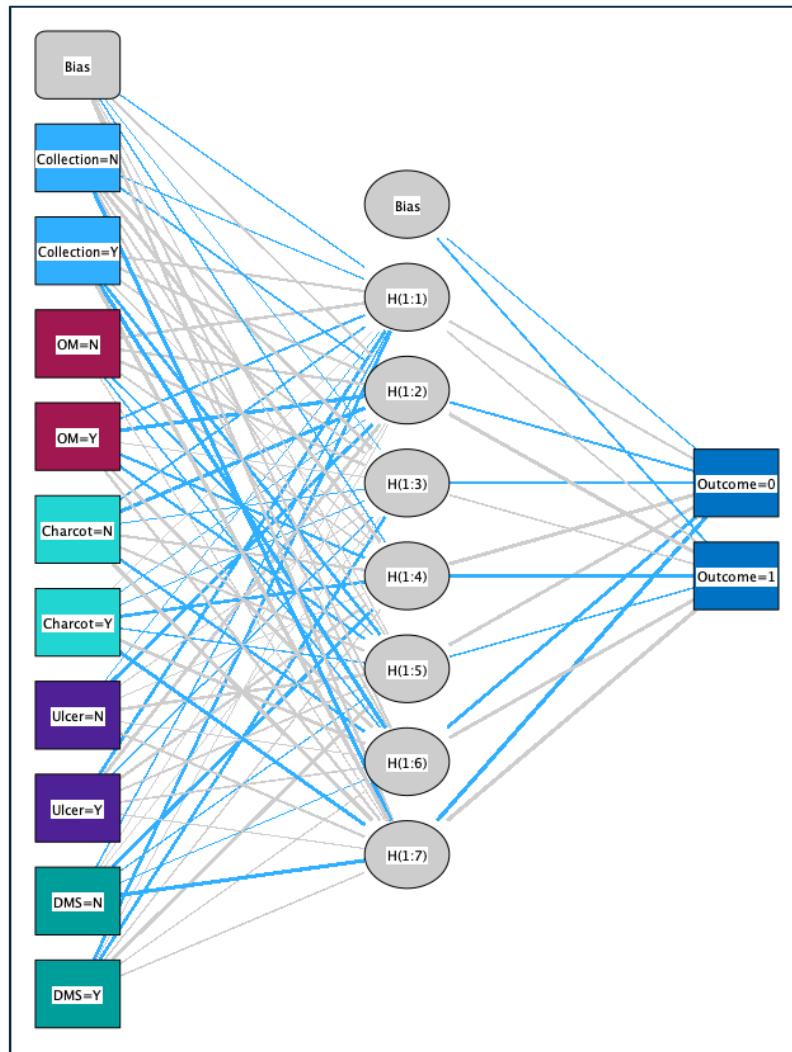


Figure 58: Multilayer perceptron output (no amputation versus MLLA and death combined). Class 0 (no amputation), Class 1 (MLLA and death).

	Class 0 (no amputation)	Class 1 (MLLA and death)
Class 0 (no amputation)	13	16
Class 1 (MLLA and death)	10	26

Table 53: Multilayer perceptron confusion matrix (no amputation versus MLLA and death combined).

3.9.5.3.2 ROC-AUC for Multilayer Perceptron

The ROC-AUC for both classes 0 and 1 was 0.64 (figure 58). Whilst this was slightly better, class 0 was still overpredicted.



Figure 59: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus MLLA and death combined (class1).

3.9.6 No amputation versus all amputation and death

A total of 418 patients were included. 176 patients were labelled as Class 0 (no amputation group). 242 patients were included in Class 1 (77 all amputations and 165 deaths).

3.9.6.1 Random Forest (no amputation versus all amputation and death)

3.9.6.1.1 Model accuracy

RaF was applied using the methodology discussed earlier. The model opted to try two variables at every split with 500 DTs. The out-of-bag error was 43.58% which was the percentage of predictions the model miscalculated. The confusion matrix for the training set (table 54) showed the model predicted Class 0 (no amputation group) 48 times with an error rate of 65%. Class 1 (all amputation and death) was classified correctly 141 times with an error rate of 27%.

	Class 0 (No amputation)	Class 1 (All amputations + Death)	Class error
Class 0 (No amputation)	48	93	0.6595745
Class 1 (All amputations + Death)	53	141	0.2731959

Table 54: Random Forest training set confusion matrix (class 0 = no amputation, class 1= all amputation and death).

The RaF test set confusion matrix demonstrates the model was good at predicting both classes (table 55). The model's accuracy was 61.45% [95% CI 0.5-0.7; p-value 0.29].

	Class 0 (no amputation)	Class 1 (All amputations + Death)
Class 0 (no amputation)	20	17
Class 1 (All amputations + Death)	15	31

Table 55: Random Forest test set confusion matrix (class 0 = no amputation, class 1= All amputations and death).

3.9.6.1.2 Performance table and feature importance

The performance table below shows the sensitivity, and specificity was low for class 0 (no amputation group). Class 0 was predicted correctly 54% of the time. A single outcome column was generated by R with Class 0 (no amputation group) as the "positive class" (table 56). The model correctly identifies the no amputation group almost 57% of the time and the combined all amputation and death group 64% of the time. The detection prevalence of 45% implies that the model predicts the no amputation group for 45% of the patients. The balanced accuracy which is an average of the sensitivity and specificity has a value of 0.6086.

	Class 0 (no amputation)
Sensitivity	0.5714
Specificity	0.6458
Positive Predictive Value (PPV)	0.5405
Negative predictive value (NPV)	0.6739
Prevalence	0.4217
Detection rate	0.2410
Detection Prevalence	0.4458
Balance Accuracy	0.6086

Table 56: RaF performance table (no amputation versus all amputations and death combined).

3.9.6.1.3 Feature importance

Feature importance did not change from the initial analysis with osteomyelitis having the most significant impact in the no amputation group (class 0) and the all-amputation and death group (class 1). The presence of ulcer also had a strong influence on class 0 but not on class 1. DMS and Charcot had a much weaker influence on both outcomes (figure 59 and table 57).

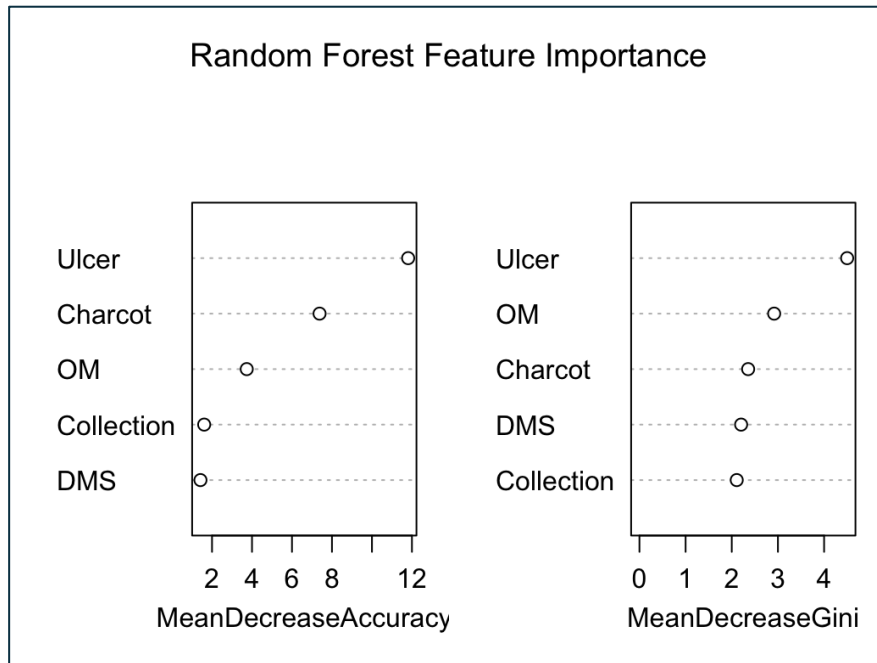


Figure 60: RaF feature importance (no amputation versus all amputation and death combined).

Variable	Class 0	Class 1	Mean decrease accuracy	Mean decrease Gini
	(no amp)	(All amputations and death)		
OM	13.231859	- 11.087525	3.736391	2.918023
Collection	1.571108	1.010480	1.608140	2.107459
Ulcer	16.810804	- 3.833878	11.810299	4.499226
Charcot	6.987478	4.652378	7.379916	2.354679
DMS	7.121164	- 5.774815	1.421203	2.202566

Table 57: RaF feature importance – Class 0 (no amputation), Class 1 (all amputation and death combined).

3.9.6.1.4 ROC-AUC for Random Forest

An ROC-AUC was generated to demonstrate the accuracy of the model for both classes (figure 60). For class 0 (no amputation group), the ROC-AUC was 0.631, For class 1 (all amputation group and death combined), the ROC-AUC was 0.631. The micro-AUC was 0.657.

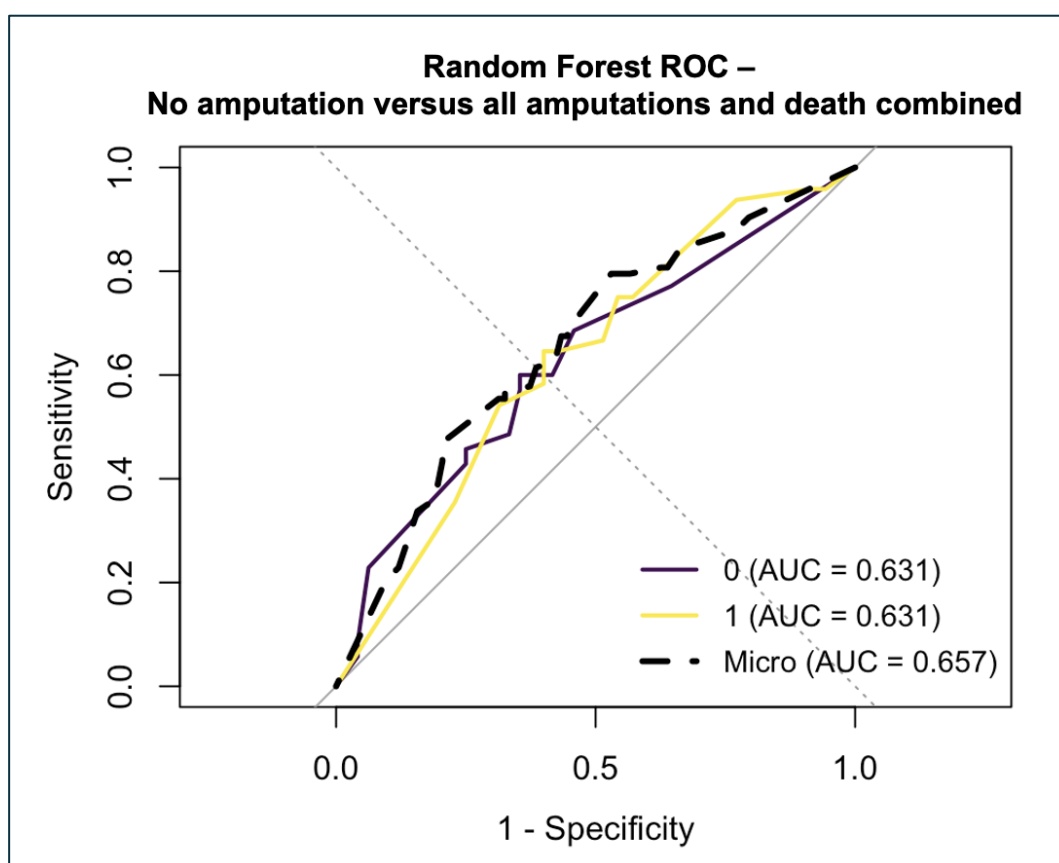


Figure 61: Random Forest ROC-AUC. No amputation versus all amputations and death.

3.9.6.2 Extreme Gradient Boosting (no amputation versus all amputation and death)

3.9.6.2.1 Model Accuracy

XGB was applied to the no amputation versus all amputation and death group dataset. The model demonstrated an accuracy of 59% [95% CI 0.48-0.69; p-value 0.46]. This required 28 iterations to achieve the accuracy rate stated. The confusion matrix below from the test set demonstrates the model was good at predicting both classes but was not statistically significant (table 58). The model had a sensitivity of 60% and specificity of 58%. The positive predictive value was 0.5 and the negative predictive value was 0.67. The detection rate was low for class 0 i.e. the no-amputation group at 49%.

	Class 0 (no amputation)	Class 1 (all amputation and death)
Class 0 (no amputation)	21	20

Class 1 (all amputation and death)	14	28
---	----	----

Table 58: Extreme gradient boosting - confusion matrix. No amputation versus all amputation and death combined.

3.9.6.2.2 Feature importance

This demonstrates the presence of ulcer was the main driver in the overall outcome classes. Other features had a lesser influence on the outcomes (table 59 and figure 61).

Feature	Gain	Cover	Frequency
Ulcer (Yes)	0.37386516	0.17317405	0.15706806
Charcot (Yes)	0.16347121	0.19805303	0.20418848
OM (No)	0.15193554	0.18202160	0.19895288
Collection (Yes)	0.14532538	0.16782905	0.15183246
DMS (Yes)	0.10354849	0.17913855	0.19895288
OM (Yes)	0.06185421	0.09978372	0.08900524

Table 59: Extreme gradient boosting - feature selection. No amputation versus all amputation and death combined. OM – osteomyelitis, DMS – diabetic myopathy/sarcopenia.

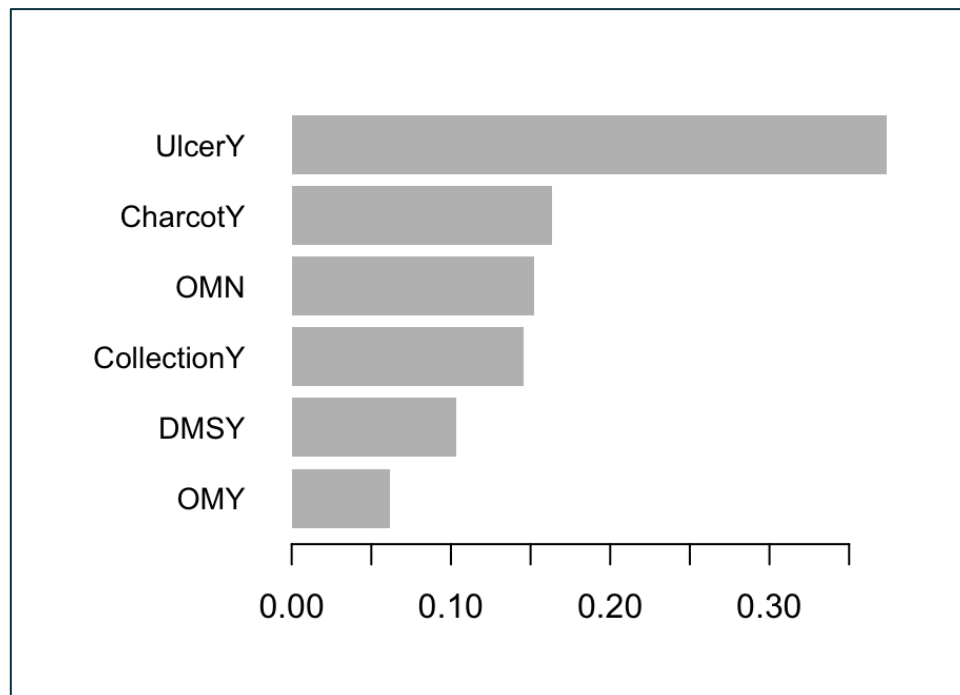


Figure 62: Extreme Gradient Boosting - Feature importance. No amputation versus all amputation and death combined.

3.9.6.2.3 ROC-AUC Extreme Gradient Boosting

A multiclass ROC-AUC was generated to demonstrate the accuracy of the model for each class (figure 62). For class 0 (no amputation group) and class 1 (all amputation group), the ROC-AUC was 0.633.

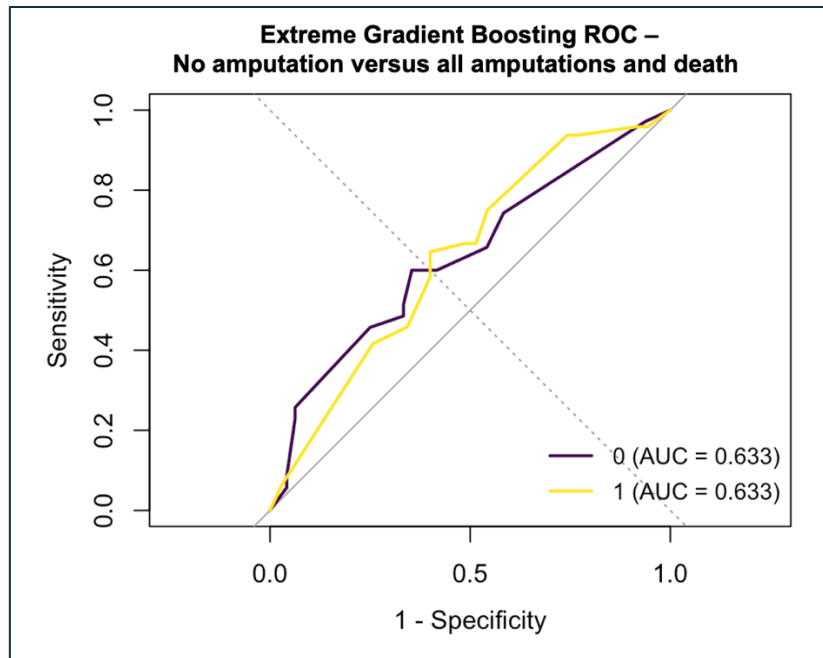


Figure 63: ROC-AUC Extreme Gradient Boosting. Class 0 (no amputation), Class 1(all amputations and death combined).

3.9.6.3 Multilayer Perceptron (no amputation versus all amputations and death combined)

As before, the data training and test set was set at a ratio of 80:20. The architecture setting was set at a minimum number of hidden layers of 1 and a maximum of 50. The lambda value was set at 0.0000001. The sigma value was kept at 0.0005 for the activation function between one neuron to the next. The 'SoftMax' function was used for the output.

3.9.6.3.1 Model accuracy

The model was 60.4% accurate using the training set but this increased to 67.5% on the test set. The MLP model performed best on when predicting both classes as shown in the confusion matrix in table 60. Two hidden layers were used. There were 7 units in the first hidden layer and 5 units in the second hidden layer. The cross-entropy was still high i.e. >1. The confusion matrix below shows that the model was better at classifying class 1 (all amputation and death group).

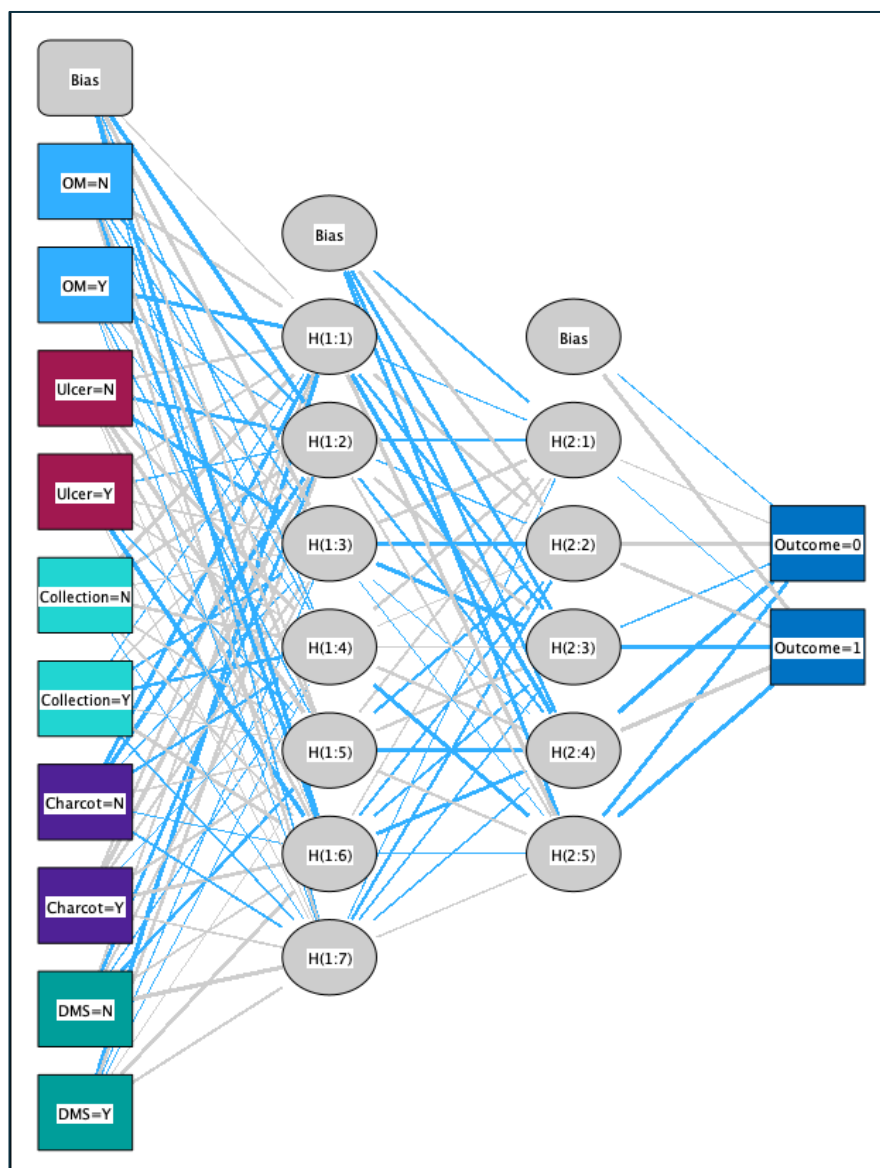


Figure 64: Multilayer perceptron output (no amputation versus all amputation and death combined). Class 0 (no amputation), Class 1 (all amputation and death).

	Class 0 (no amputation)	Class 1 (MLLA and death)
Class 0 (no amputation)	6	23
Class 1 (MLLA and death)	2	46

Table 60: Multilayer perceptron confusion matrix (no amputation versus all amputation and death combined).

3.9.6.3.2 ROC-AUC for Multilayer Perceptron

The ROC-AUC for both classes 0 and 1 was 0.617 (figure 64).

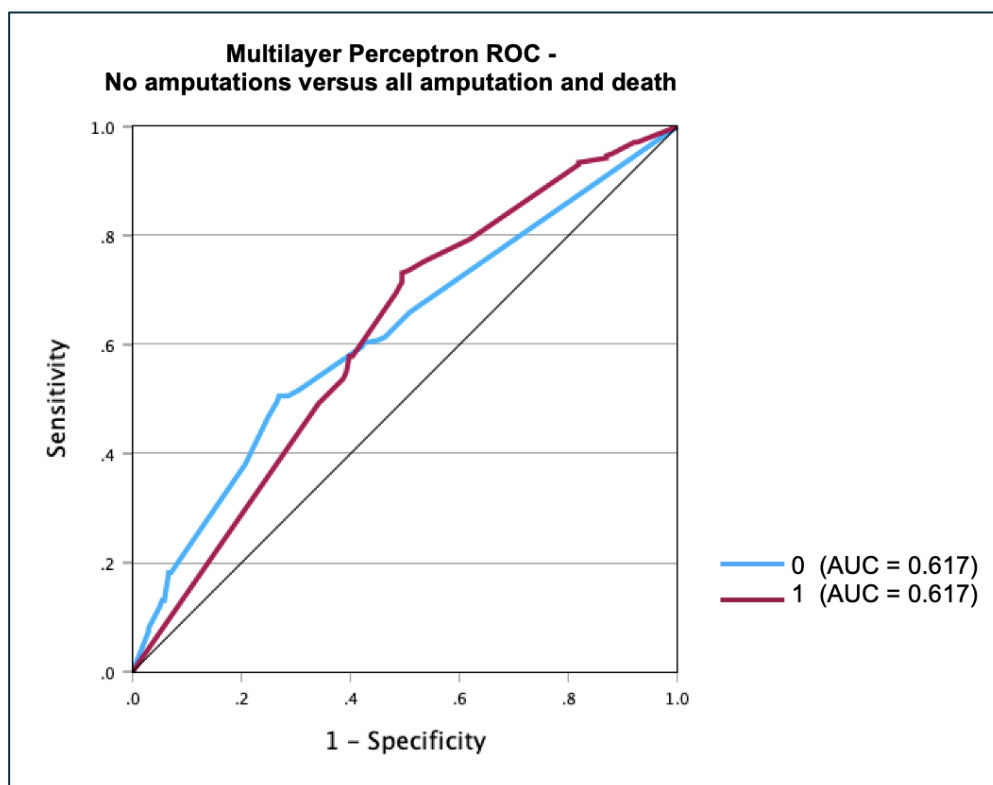


Figure 65: Multilayer Perceptron ROC-AUC for no amputation (class 0) versus all amputation and death combined (class1).

3.10 Discussion

The table below consolidates results of all three models assessing the initial intended outcomes as well as the additional composite analysis (table 61). RaF has a high accuracy rate but very poor AUC values and possibly overpredicts the majority group i.e. no amputation group. RaF has a high accuracy rate in predicting MLLA and death based on MRI reports but this is not statistically significant.

XGB has similar rates of accuracy across all groups and outcomes. It is statistically significant for predicting the 'no amputation versus minor and major amputation group' but not for any other group or sub-group. It can also be reproduced and may possibly perform better on a different dataset which is more balanced along with RaF. MLP had the lowest accuracy rate and has the highest cross entropy loss. This did improve in the major lower limb amputation and death group with lower cross entropy loss but still not statistically significant for clinical application. Furthermore, is not recursively reproducible given its black box nature.

The accuracy rate was also high for the MLLA versus death group but not statistically significant although this may be attributable to class imbalance. It would have been expected that as part of the analysis, the accuracy would have been high for the no amputation versus all other groups, however this was not the case. A recurring issue across the data was the class imbalance but, when adjusted for this, the ML performance counterintuitively deteriorated in RaF and XGB for all groups. This often meant the ML models overpredicted the predominant group. This imbalance is to be expected to some extent given the cohort who would have had the scan were DFU or DFD patients mostly. It is also worth considering the variability in reporting over the years.

Standard logistic regression and the AI models identified osteomyelitis and collections as a feature of importance in predicting the outcome even though DMS and ulcers were also equally well represented in the reports. Conversely, when comparing all amputations and death as an outcome, both RaF and XGB highlighted DMS along with OM as strong predicting features on MRI reports.

It would be interesting to see how the models would perform on a bigger and more varied test set with more varied representation of each feature. It is also worth considering why reports have not been used for prediction purposes. The salient feature being reported as part of the reason for the scan is more often than not osteomyelitis.

AMP	Groups	Accuracy	AUC	Feature Importance
RaF	No amp v minor amp v major amp	67%	No amp = 0.397 Minor amp = 0.481 Major amp = 0.42	OM Collection Ulcers
	No amputation v all amputations	65%	No amp = 0.664 All amp = 664	OM Collection
	MLLA v death	91%	MLLA = 0.552 Death = 0.552	OM Collection
	All amputations v death	61.7%	All amp = 0.557 Death = 0.557	DMS OM
	No amputation v MLLA and death	59.7%	No amp = 0.613 MLLA and death = 0.613	Charcot Ulcer OM
	No amputation v all amputation and death	61.45%	No amp = 0.631 All amp and death = 0.631	Ulcer Charcot
XGB	No amp v minor amp v major amp**	63.4%	No amp = 0.724 Minor amp = 0.779 Major amp = 0.439	Absence of OM Presence of Charcot
	No amputation v all amputations	68.29%	No amp = 0.714 All amp = 0.714	Absence of OM Presence of OM
	MLLA v death	68.57%	MLLA = 0.536 Death = 0.536	Absence of OM Presence of OM Presence of ulcer

	All amputations v death	57.45%	All amp = 0.605 Death = 0.605	Absence of OM Presence of DMS
	No amputation v MLLA and death	56.3%	No amp = 0.641 MLLA and death = 0.641	Presence of Charcot Absence of OM Presence of ulcer
	No amputation v all amputation and death	59%	No amp = 0.633 All amp and death = 0.633	Presence of ulcer
MLP	No amp v minor amp v major amp	60.6%	No amp = 0.531 Minor amp = 0.704 Major amp = 0.672	Not provided
	No amputation v all amputations	61.9%	No amp = 0.671 All amp = 0.671	
	MLLA v death	97.3%	MLLA = 0.676 Death = 0.676	
	All amputations v death	67.1%	All amp = 0.620 Death = 0.620	
	No amputation v MLLA and death	60%	No amp = 0.64 MLLA and death = 0.64	
	No amputation v all amputation and death	67.5%	No amp = 0.617 All amp and death = 0.617	

Table 61: Summary of applied models and results; Amp- Amputation, AUC – Area under the curve, DMS – Diabetic myopathy/Sarcopenia, MLLA – major lower limb amputation, MLP – Multilayer perceptron, OM – osteomyelitis, Raf – Random Forest, v – versus, XGB – Extreme Gradient Boosting, ** - statistically significant.

In recent years, the prevalence of DMS being reported has become increasingly common given its awareness as disease process. This has been discussed in Chapter one and has recently become a disease process with formal recognition in ICD-10. It is possible that the performance of the models reflect the heterogeneity and inconsistency in reporting. A

standardised reporting format for DFD related MRIs may be of benefit. However, there may be greater benefit in assessing the MRI images directly to evaluate their use in predicting limb loss and will be discussed in subsequent chapters.

Chapter 4: Applying AI models to MRI images of the foot

4.1 Introduction

The use of two- and three-dimensional imaging and AI in the realm of DFD has remained confined to wound/ulcer detection, recognition (ischaemic versus non-ischaemic lesions) and ulcer grading [90, 152-156]. It has also been used for classifying thermograms of patients with DFD [153]. Multiple models exist for object detection (you only look once – YOLO), object classification (Vision Transformer – ViT and Residual Network – ResNet) and semantic or spatial segmentation (no new U-Net – nnUNet). This chapter discusses the relevant machine learning models – ViT, ResNet and nnUNet and their application to the MRI images in this project.

4.1.1 Vision Transformer (ViT)

Convolutional Neural Networks (CNNs) were previously discussed in chapters one and three but will be discussed again briefly. Vision Transformers (ViT) is a fairly novel deep learning model introduced in 2020 as an alternative to CNNs. ViTs aim to classify, detect and segment images. ViT was the first transformer to have greater accuracy than pre-existing CNNs when applied to the industry standard ImageNet dataset (with over 15 million labelled high-resolution images in over 22,000 categories). This has started to shift the stronghold CNN's have maintained in this area until very recently.

In order to understand how ViTs work, a brief introduction to their origin is required. The history of ViT is grounded in the transformer architecture which was first proposed by Vaswani et al's team at Google in 2017, initially for natural language processing tasks [157].

Transformer is a neural network architecture in which text is split into sub-words or characters and converted to 'tokens'. Figure 53 demonstrates the structure of the transformer deep learning model as first described [157].

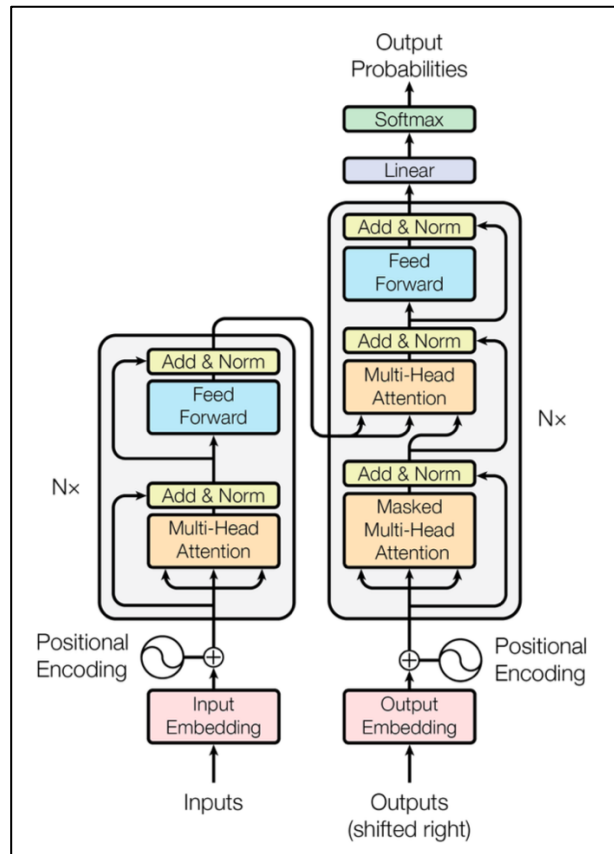


Figure 66: Transformer model architecture. As printed in the original paper by Vaswani et al [157].

Tokens are numerical representations of the text which are then converted into a vector using a learned word embedding table. A learned embedding table is a matrix with parameters which takes each token identity and maps it into a vector representation. The model “learns” and optimises these vectors during training to capture their structural meaning.

This is then processed through multiple layers via a multi-head self-attention mechanism which forms the cornerstone of the transformer architecture [157, 158]. The idea of self-attention allows the model to ascertain which part of the input to focus on when processing a token. Each token produces three vectors – Query (Q), Key (K) and Value (V) [157]. These vectors allow the attention score between tokens to be calculated using the following equation:

$$Attention(Q, K, V) = softmax \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

- **Q** = Query

- **K** = Key
- **V** = Value
- **QK^T** = Transpose of the Key Matrix i.e. the similarity between Q and K
- **Softmax** = adds weighting by converting similarities into probabilities

It runs these equations multiple times in parallel with different learned projections of Q, K and V, thus being called 'multi-head'. Each head learns a different relationship i.e. semantic, local and global. It then concatenates i.e. joins the outputs or vectors end-to-end along a specific dimension, not dissimilar to a stack, into a bigger matrix using the following equation [157]:

$$Multihead(Q, K, V) = Concatenate(head_1, \dots, head_h)W^o$$

- **Q** = Query
- **K** = Key
- **V** = Value
- **W^o** = learned weight matrix. This essentially projects the concatenated vector back to the original embedding dimension by learning how to collate the different heads into a unified representation.

Following attention, the concatenated and projected output goes through residual connections and layer normalisation (LayerNorm) which are essential to stabilise and train the model. These are explained below:

Residual Connections

This means that the original input is added back to the output so that the network has a direct path for the gradients to flow backward. This will be discussed again in the next section when discussing ResNet. In transformer, the output is added to the original input after multi-head attention and after the MLP block (figure 53).

Layer Normalisation (LayerNorm)

Allows for feature dimension normalisation to stabilise training, especially with deeper stacks. In transformer, it is applied before attention and MLP (figure 53).

Each token's representation is then refined by a multilayer perceptron, a feed forward network which has already been discussed in the previous chapter, to give an overall output.

In 2020, Dosovitskiy et al, also at Google, used the transformer architecture for image recognition at scale[159]. Using the transformer architecture for computer vision and image classification had not been done before [159]. This resulted in Vision Transformer (ViT). ViT followed similar principles and is demonstrated in figure 54 below.

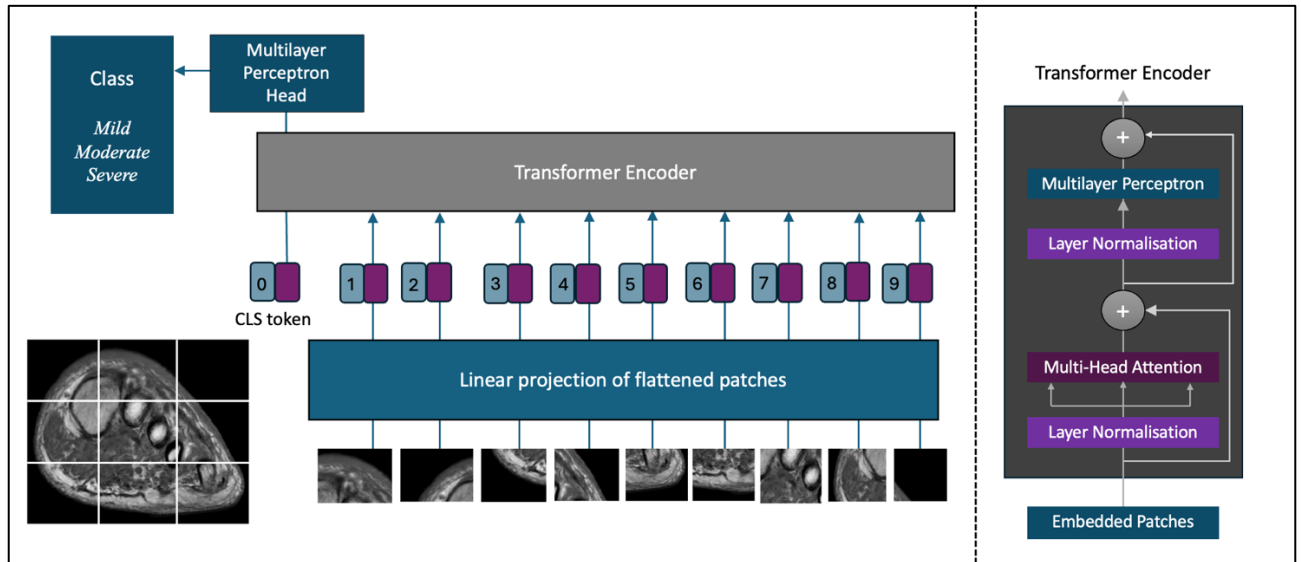


Figure 67: Architecture of Vision Transformer. Adapted from Dosovitskiy et al to demonstrate its use for this thesis [159]. CLS - classification token.

Images are stored as a matrix of pixel values on devices. An image is fed into the model with a set height, width and channels (red, green, blue) and is divided into patches. Each patch is flattened into a vector or 'token'. Assuming an image has 512x512 pixels (height x width) and all three channels i.e. red, green and blue. Using these image values, the tensor (a multidimensional array of numbers) is calculated as 512x512x3 which is 786,432 values. If each patch is 16x16 pixels, we need to calculate the number of patches along each dimension i.e. 512 divided by 16 which results in 32 patches across the height and 32 patches across the width. In order to establish the total number of tokens, these values are multiplied together i.e. 32x32=1024. This means the image is then split into 1024 tokens.

If an image is now in grayscale, measuring the intensity from black to white, such as a CT or MRI scan image but with the same pixels, the tensor would now be 512x512x1 which is 262,144 values. If each patch, again, is 16x16 pixels, the number of patches still remains unchanged but the number of values in each token changes such that an image with colour would have a token length of 786,432 whilst a grayscale image would have a token length of

262,144 values. Larger patches/tokens require significantly bigger embedding layers or stacks and more overall computation.

Linear projection of the flattened patches fixes/embeds the patches in a fixed dimension, Positional encoding or positional embedding helps the model to understand where each patch comes from (figure 54 – boxes 1-9) [158, 159].

This then undergoes the same process as the originally described transformer model including multi-head self-attention (understands the global relationship between the patches), residual connections, layer normalisation and multilayer perceptron.

Transformers process sequences of tokens; however, a single output vector is required for classification tasks. A classification token (CLS) prepends the other tokens. Using the same image example, a grayscale MRI image of 512x512 pixels and 16x16 patches and a token length of 262,144, with a prepended CLS would have 262,145 tokens as the input sequence. This was first proposed by Devlin et al, also at Google, in 2018 with BERT (bidirectional encoder representations from transformers) [160]. This introduced masked language modelling i.e. masking words and training the model to predict them, next sentence prediction as well as teaching the model to learn bidirectionally [160]. When applied to images, CLS gathers information about the image as a whole during self-attention. Once the information has passed through all the transformer encoder layers, a succinct summary of the overall input is neatly packaged in the final CLS embedding, which is then fed into the MLP and then the classification head. This then decides the label that an image should be allocated to i.e. this is an MRI of a foot with mild, moderate or severe sarcopenia.

4.1.2 Convolutional Neural Networks (CNN)

In order to understand ResNet and nnUnet, the concept of a convolutional neural network and how it works needs to be described. The concept was first described in Fukushima's 1980 article titled 'Neocognitron: a self-organising neural network model for a mechanism of pattern recognition unaffected by shift in position' [161]. Drawing on the hierarchy model of the visual nervous system in vertebrates, the article discussed how to "synthesise a

neural network model in order to endow it an ability of pattern recognition like a human being” [161].

The model, in its simplest form comprised of an input layer (photoreceptor array) followed by a cascade connection each of which was composed of two layers of cells connected in a cascade. The first layer was called simple or S-cells and were considered lower order hypercomplex cells whilst the second layer consisted of complex or C-cells with higher order hypercomplex cells. The “afferent synapses” to each S-cell were considered modifiable because of their plasticity. The model followed the rules of action potentials, with “neural cells” labelled as analogue types such that the inputs and outputs had positive numerical analogue values proportional to the “pulse density” of the firing action potential. Both S- and C-cells had shunting type inhibitory inputs as well [161]. This concept was applied to number recognition in this early application by Fukushima.

In 1998, LeCun et al attempted the first practical application of the early musings of CNNs to digit recognition using a dataset with 15,000 – 60,000 examples [162]. At the time, the most commonly accepted model for pattern recognition was a segmenter for extracting relevant features. The solution offered was to feed the system with raw pixel images and keep the pre-processing to a minimum [162].

The proposed methodology was gradient based learning and will become a key feature in the rationale for ResNet’s conception [162, 163]. An input is added to the neural network to produce an output. The prediction is compared to the correct answer to yield an error value. This is applied to all training examples and getting an average training loss based on the error value. This is known as forward pass. The aim is to find parameters and apply gradient descent or other optimisers to alter parameters in the direction which would minimise the average loss. The system is structured as a feed-forward network of modules through layers such that the output of each module is:

$$X_n = F_n(W_n, X_{n-1})$$

- X_n = output of the module
- W_n = trainable module parameters or weights
- X_{n-1} = output from the previous layer or convolutions

However, instead of computing gradients for every parameter from scratch, gradient descent does this in a recursive fashion. This is called backpropagation or backward pass which starts with the output loss and works back layer by layer to readjust the weights or parameters.

LeCun et al described the CNN for image recognition as we understand it today with LeNet-5 [162, 164]. CNNs are composed of three parts – convolutional layers, pooling layers and fully connected layers. A raw image is stored on a computing device as a matrix with numerical values and is fed into a CNN. A ‘kernel’ is a filter which resembles a smaller matrix with a smaller size than the original input image (figure 55).

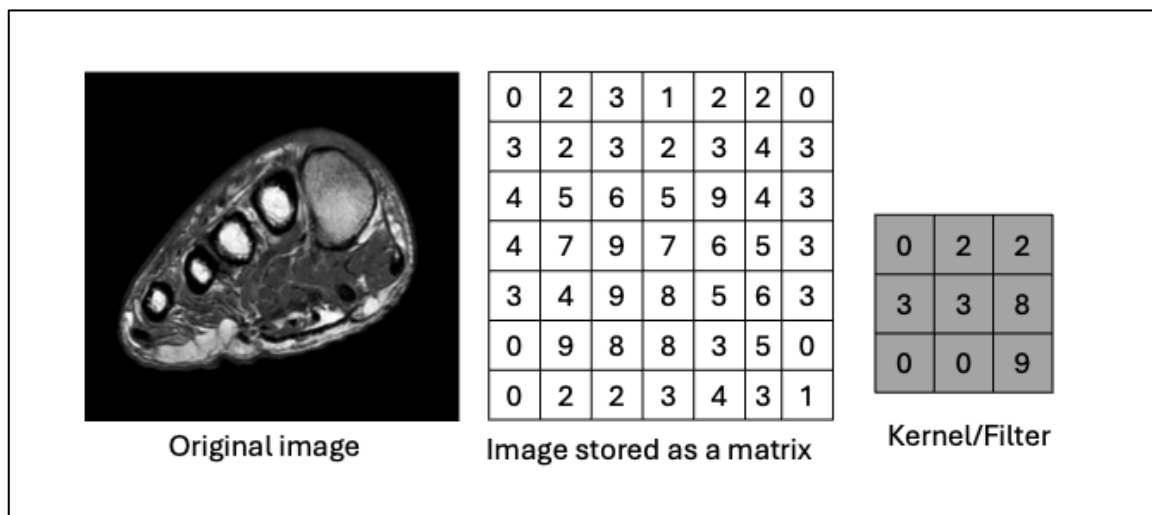


Figure 68: Imaging data stored as a numerical matrix on a computing device and a kernel/filter. Adapted from IBM to demonstrate its use in this thesis [164].

If an MRI image is 512x512, the kernel might be 3x3. The kernel scans across the image. This is called a convolution (figure 56). The number of pixels the kernel moves across horizontally and vertically each time is called a ‘stride’ and will become a relevant point as the discussion continues.

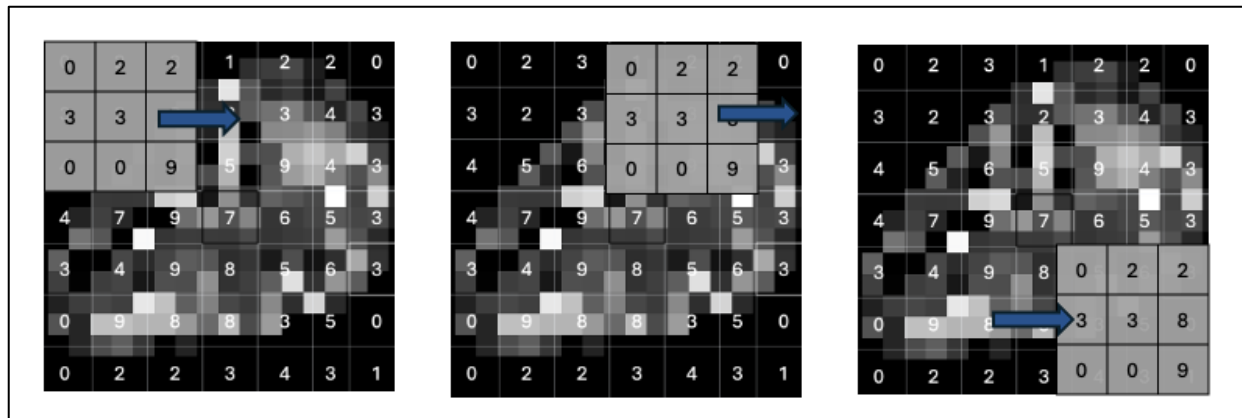


Figure 69: Schematic representation of a convolution. The kernel scans across the entire image to produce a convolution. Adapted from IBM to demonstrate its use in this thesis [164].

At each position, the kernel multiplies its weights with the corresponding pixel values when it detects them and produces a numerical sum value. Once the kernel has scanned the entire image with all the values, it produces a feature map [162-164]. Different kernels have different sizes and detect different features. Initial kernels are simple and detect patterns, edges and corners i.e. low-level features. Accumulated multiple feature maps gradually improve and detect more details and high-level features such as objects and patterns (figure 57).

An interesting feature of CNNs noted by Fukushima and LeCun was that if the input image was shifted, the feature map was also shifted by the same amount but left unchanged otherwise [161, 162]. Furthermore, the exact location of a particular feature became less importance once detected and only its approximate position relative to other features held any importance [162]. This is highly relevant as it forms the basis which allows for data augmentation and distortion and will also be used as part of the methodology in this thesis [140, 152, 162, 165].

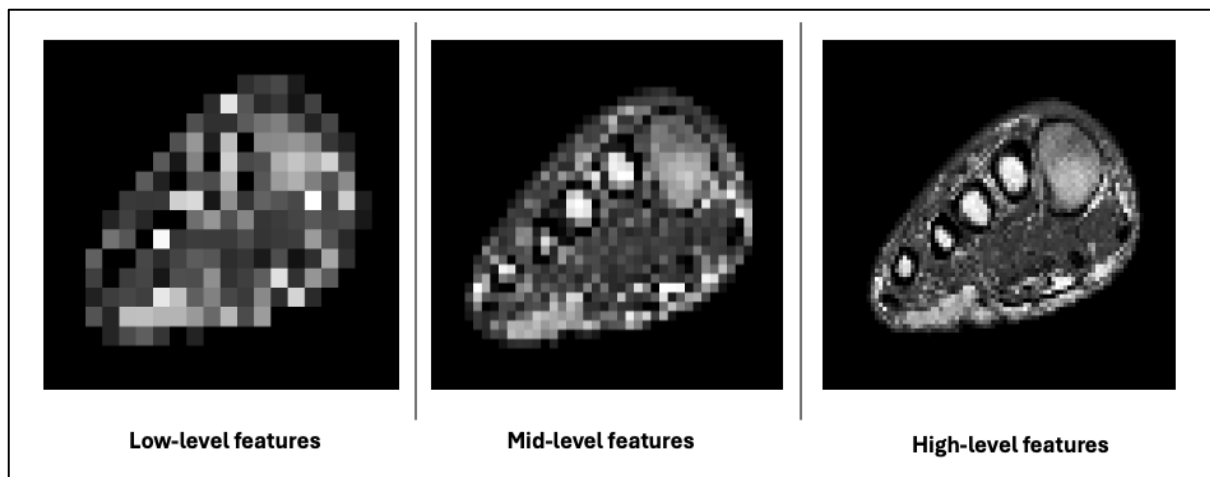


Figure 70: Feature maps through various convolutions. Adapted from LeCun et al to demonstrate its use for this thesis [162].

The pooling layer reduces the spatial size of the features maps but retains the salient features and discards the noise (figure 58). It also helps to reduce overfitting which has previously been discussed in chapter 3 (section 3.1.1). There are two types of pooling [164]:

- Max pooling: The kernel selects the pixel with the maximum value to the output array as it moves across the input. This is commonly used.
- Average pooling: The kernel calculates the average value within the receptive field which it then sends as the output.

The feature maps are leveraged and pooled to build more complex features maps. This is known as abstraction. Once the image has undergone a number of convolutions and pooling, the flattening layer converts the two- and three-dimensional feature maps into a one-directional vector which is then fed into the fully connected layer.

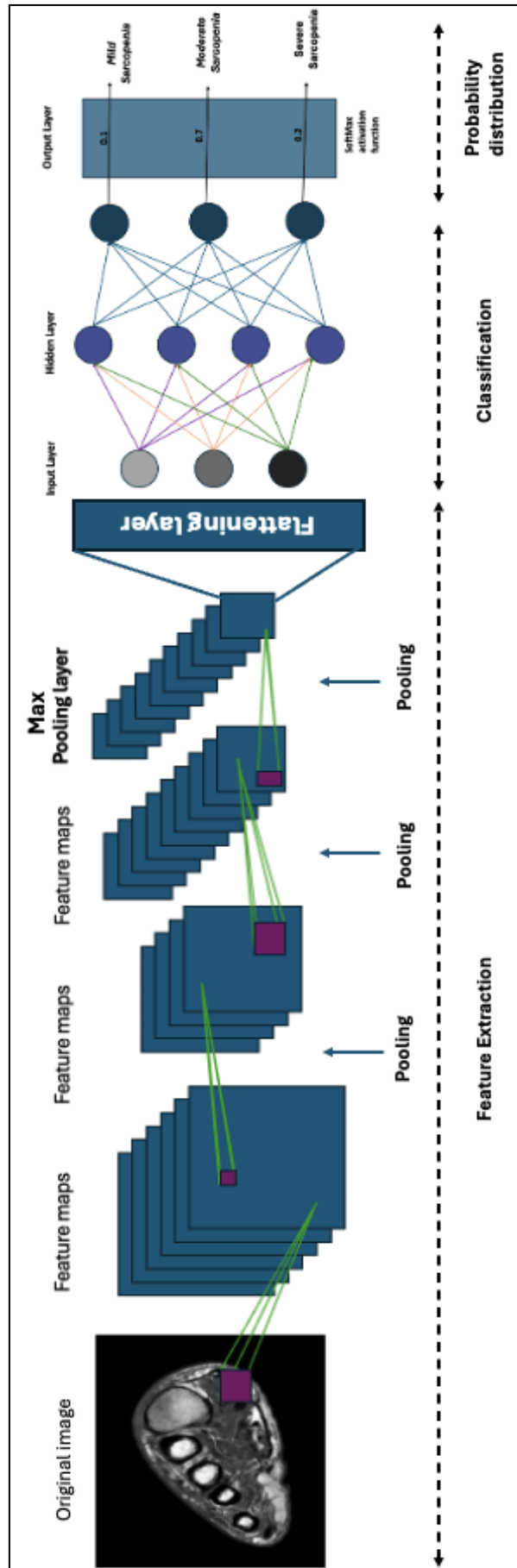


Figure 71: Schematic representation of a convolutional neural network. Adapted from LeCun et al and IBM to demonstrate its use in this thesis [162, 164].

This leads to the second part of the CNN called classification (figure 58). This is comprised of a fully connected feed forward network layer where each node in the output layer is connected to the node in the input layer. The one-dimensional pooled feature maps are put through the fully connected layers. These leverage a SoftMax activation function to provide a numerical probability of the input being classified into a particular category [164]. Another activation function which is also commonly used is rectified linear unit (ReLU) which decides if a neuron should activate/fire or not using the following formula [140]:

$$ReLU(x) = \max(0, x)$$

- x = function
- If $x > 0$, output = x
- If $x \leq 0$, output = 0

ReLU was proposed in a landmark paper by Alex Krizhevsky and his team in 2012 which initiated the global revolution for CNNs by significantly outperforming existing CNNs at the time[140]. The model, called AlexNet created a five convolutional layer with max pooling layers, three fully connected layers and a 1000 way SoftMax with 60 million parameters and 650,000 neurons [140]. This was trained on 1.2 million images [140]. The performance improved with ReLU, overlapping the pooling layers, data augmentation and dropout. Dropout involves resetting the output of each hidden neuron with a probability of 0.5 or below to zero. These neurons were essentially switched off or dropped out and therefore did not contribute to forward pass or backpropagation [140].

Krizhevsky et al did, however, note that the model's performance dropped significantly when a convolutional layer was removed [140]. The answer to this initially seemed simple – in order to increase accuracy, the number of convolutional layers needed to increase. In 2014, Simonyan et al found using very small convolutional kernels/filters and deeper CNNs (16-19 layers) significantly improved accuracy with their VGGNet [166]. Szegedy et al pushed this further in 2015 with GoogLeNet which was comprised of 27 layers including pooling layers [167]. However, as CNNs became deeper, a bigger problem emerged: vanishing gradient [167].

Vanishing gradient

Whilst a neural network is training, it adjusts weights using backpropagation which relies on gradients. In very deep networks, the ability to back propagate through all the layers is lost as gradients become smaller whilst being passed backwards and effectively becoming zero. This is known as vanishing gradient and is not unlike a meteor entering the atmosphere and disintegrating as it hurtles towards earth. By failing to backpropagate, the earlier layers fail to train as there is minimal feedback which takes a very long time to reach the earlier layers, if at all, or no feedback and therefore no adjustment to the weights in the earlier layers.

4.1.3 Residual Network (ResNet)

The issue of vanishing gradient was addressed with Residual Networks or ResNet proposed in the seminal paper by He et al at Microsoft in 2016 [168]. ResNet is a convolutional neural network originally intended for image classification and has gradually gained popularity in medical image classification tasks [152, 153, 155, 156]. It was the first CNN to address the degradation problem (i.e. stacking more layers leading to lower accuracy rates) using the concept of residual learning.

Residual Learning

In the normal CNNs, a function is applied to an input (x) and is mapped to an output (y) using a function which is the convolutional layer/weight layer as shown in figure 59a. In a residual network, a copy is made of the input with the assumption that all of the information required is already there, along with some of the residual learning from the previous layers (+), and then the difference between the input (x) and the output (y) is mapped as shown in figure 59b [168, 169].

Skip connections (dotted line in figure 59) mean that instead of learning a mapping through every single layer, it learns a residual function learned by several preceding stacked layers and then the input (x) is added back to the connections so that the information from the earlier layers is not lost.

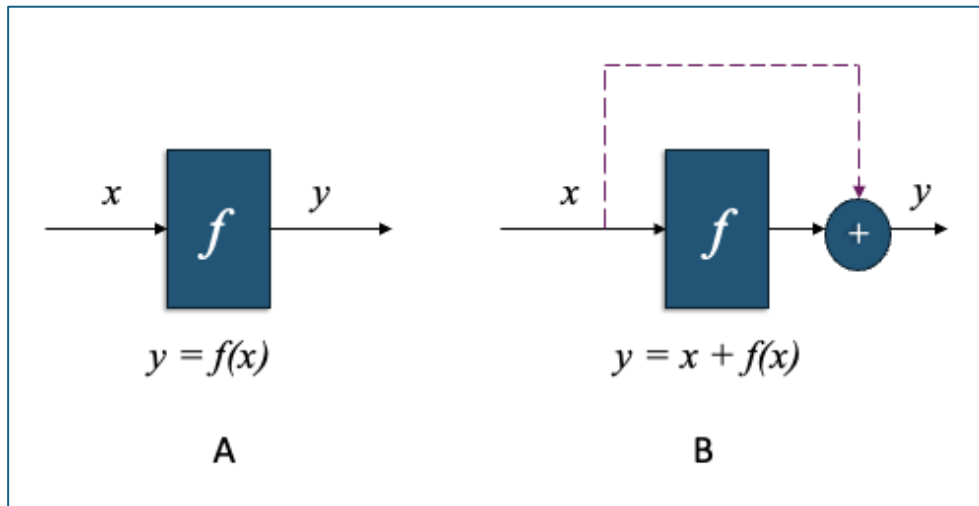


Figure 72: Schematic representation of normal CNN learning (A) versus residual learning/skip connections (B). Adapted from He, Brunton [168, 169].

When repeated, a basic ResNet block would resemble figure 60.

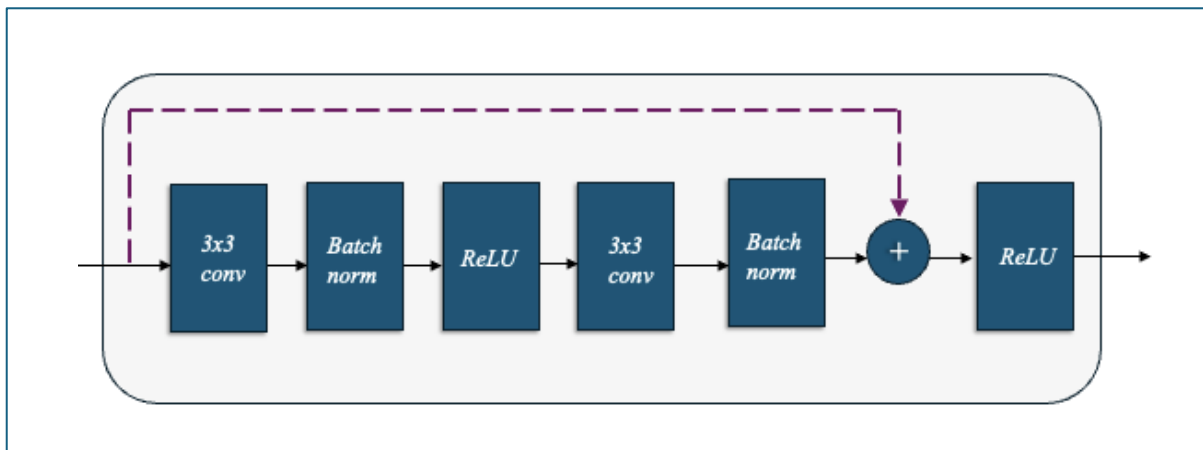


Figure 73: Schematic representation of a ResNet block. Adapted from He, Brunton [168, 169]. Conv – convolution, norm – batch normalisation, ReLU – rectified linear unit.

This addressed the issue of vanishing gradient and allowed He et al to produce a neural network with 152 layers which was a significantly deeper architecture compared to its predecessors, but otherwise follows the same pattern as a convolutional neural network with feature extraction and classification as shown in figure 58 earlier [168]. The figure below shows where ResNet fits in within the CNN.

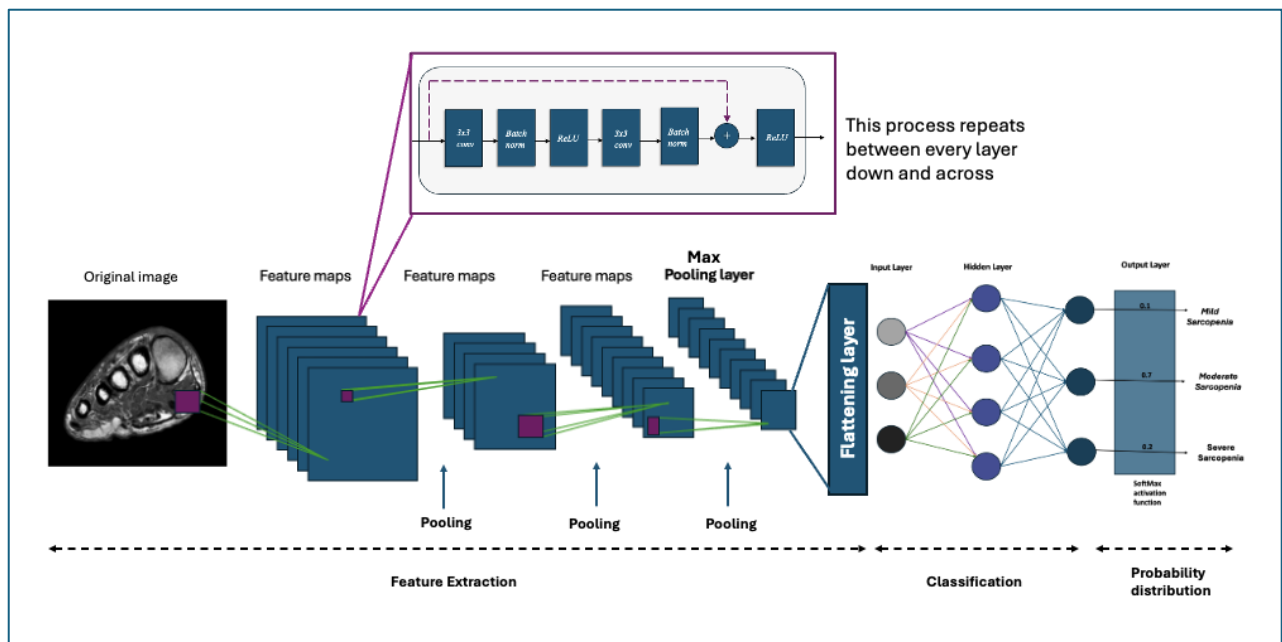


Figure 74: ResNet structural blocks within CNNs. Not adapted.

Bottleneck architecture

Another concept introduced here was a deeper bottleneck architecture for deeper networks. This means that as the layers approach the flattening layer, the size of the convolutions get smaller i.e. from 3x3 to 1x1 which allows for more efficient depth scaling [168].

Floating point operations (FLOPs)

A concept worth discussing with increasingly deeper ResNets is Floating Point Operations (FLOPs). FLOPs are a measure or metric of the computational complexity of a model. This includes all mathematical calculations including addition, subtraction, multiplication or division and are a hardware independent measure of computational complexity [168, 170].

ResNet variations

Several variations of ResNet are available. This is comprehensively summarised below in table 48:

Model	Layers	Residual block type	Parameters	FLOPs	Top-1 accuracy (ImageNet)
ResNet-18	18	Basic (2 convolutions per block)	11.7 million	1.8 billion	69-70%

ResNet-34	34	Basic (2 convolutions per block)	21.8 million	3.6 billion	73-74%
ResNet-50	50	Bottleneck (3 convolutions per block)	25.6 million	3.8 billion	76-77%
ResNet-101	101	Bottleneck (3 convolutions per block)	44.5 million	7.6 billion	77-78%

Table 62: Summary of ResNet versions. Partially adapted from He et al [168].

ResNet-18 and ResNet-34 are good for small datasets and offer low computation but fast inference. ResNet-50 is used quite commonly in medical research and offers a balance between accuracy and inference [152, 154]. ResNet-101 is slower to train but offers higher accuracy. These models are used in this thesis and will be discussed in the methodology and results later in this chapter.

4.1.4 U-Net and no-new-U-Net (nnU-Net)

U-Net is a convolutional neural network first proposed by Ronneberger et al in 2015 [165]. U-net is so called because of its architecture which resembles the 'U' of the alphabet (figure 62). It was designed for semantic segmentation in medical imaging. The paper segmented and trained on light microscopy images of Henrietta Lacks (HeLa) cells [165].

U-Net was necessitated by the need for segmenting on smaller datasets unlike CNNs which have been trained on datasets such as ImageNet with millions of labelled images. The other feature highlighted by Ronnegberger was the importance of localisation with the output. The aim of U-Net was to train with fewer images and provide more precise segmentations in scenarios such as cancer localisation [165].

U-Net follows an encoder-decoder architecture (figure 62). This can be divided into three steps and elaborates on the details of the figure:

- **Encoder [contracting path/downsampling]:** This is the left side of figure 62. The encoder aims to understand the image, its contents and which features in the image might be important. An image is shown to U-Net. The input size of the image is

572x572x1. This goes through the first 3x3 convolution and ReLU and the image output is 570x570 (blue arrow). It then goes through two further 3x3 convolutions, each with 64 feature channels i.e. 64 kernels/filters extracting 64 different patterns. The image is downsized to 568x568. It then max pools (red arrow), as we have already discussed in the CNN section and reduces the image size by half – 284x284 but doubles the number of feature channels to 128.

At the end of this, it max pools again and further reduces the size of the image (from 284 to 282, 280, 140, 138, 136, 68, 66, 64, 32, 30 and eventually to 28) and exponentially increases the number of feature channels (from 128 to 256, 512 and 1024). Increasing the number of channels helps to extract more abstract features. Throughout all of this, it follows the rules of skip connections as well.

- **Bottleneck:** This is the centre or trough of the U-shape. It is the deepest part of the structure with 1024 filters where convolutions are still being applied.
- **Decoder [expanding path/upsampling]:** This is the right section of figure 62. The decoder aims to make spatial sense of the objects in the image i.e. where the object is and how to recreate/reconstruct it. The kernels/filters now upconvolute (green arrow) and double the spatial resolution.

The grey arrows titled 'copy and crop' are essentially concatenating i.e. joining the outputs into a stack/bigger matrix. This has previously been discussed in the ViT section.

At each step on the left side, the image is up-sized and the number of channels is halved. At the last stage, a single 1x1 convolution (teal arrow) maps 64 channels to two classes i.e. HeLa cell versus background which is manifested as the segmentation map output. The final output image dimension is 388x388 due to the cropping.

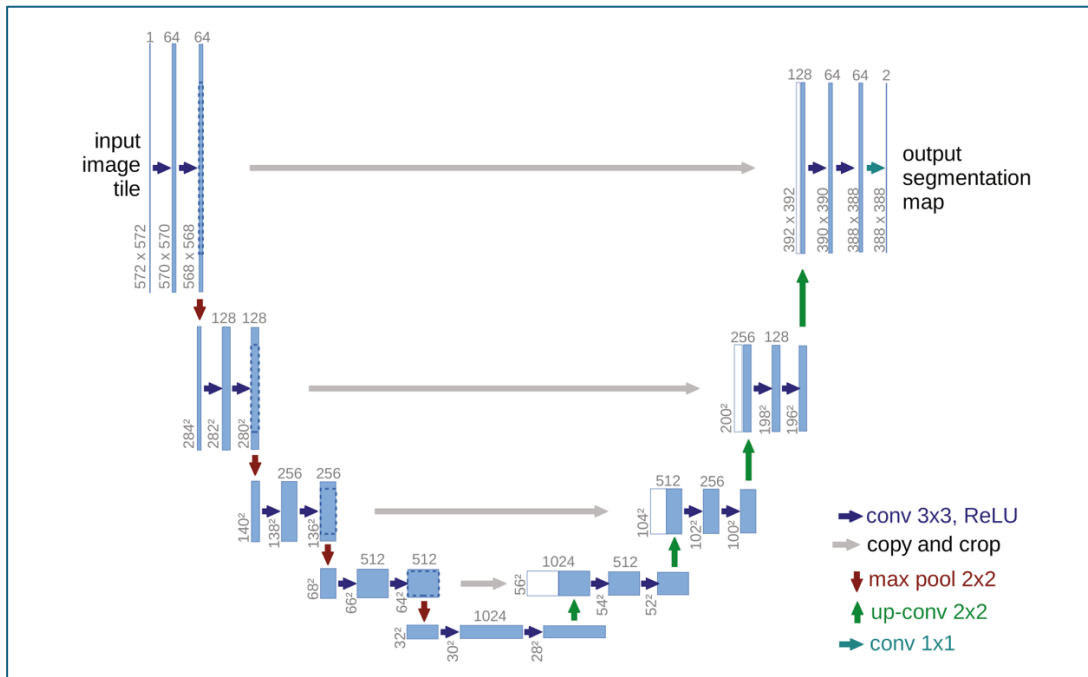


Figure 75: U-Net structure as proposed by Ronneberger et al. As printed in the original publication[165].

In 2020, Isensee et al created nnU-Net, providing an out-of-box empirical segmentation model with self-configuring including the preprocessing, network architecture, training and post-processing for a given task [110]. nnU-Net is capable of handling various types of biomedical imaging including CT, MRI, Histology and electron microscopy. This has already been discussed in chapter one. Kim et al used this for individual muscle segmentation (27 muscles) on 72 whole-thigh CT scans of patients with hip fractures for sarcopenia assessment and found to have DICE scores as high as 0.84 [113]. DICE scores have already been discussed extensively in chapter one, section eleven, sub-section four (1.11.4).

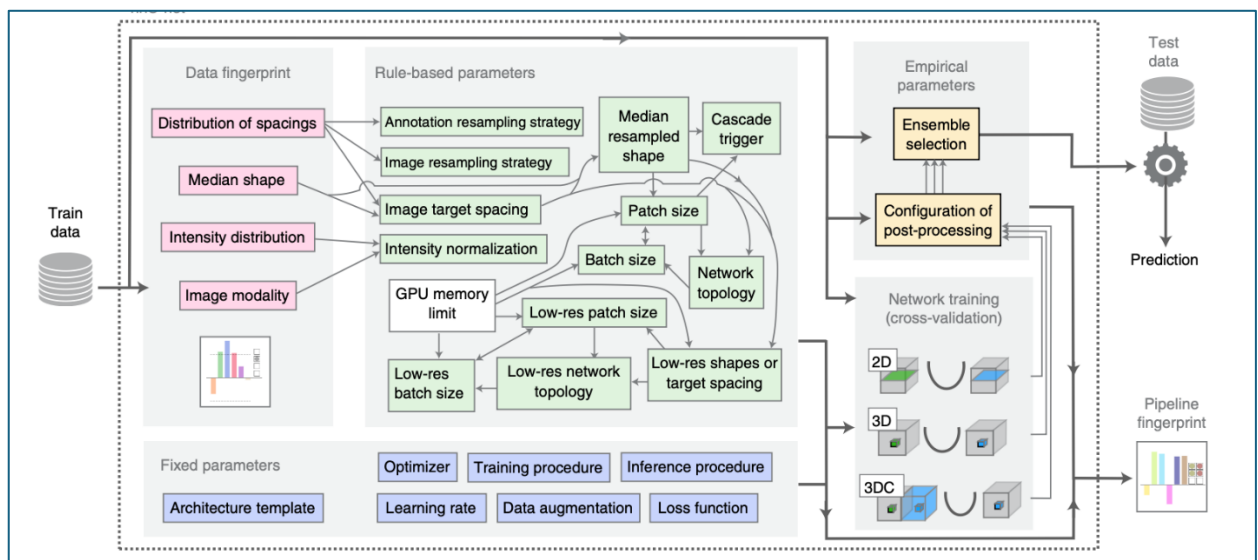


Figure 76: nnU-Net functions as described by Isensee et al. As printed in the original paper [110].

4.2 Aims

The aim of this chapter is to explore the potential applications of AI applied to MRI images for automated classification and severity grading of sarcopenia as a potential surrogate marker.

4.3 Methodology

It was essential to establish congruence in data labelling and segmentation prior to commencing any application of transformer or CNNs. Therefore, the inter-rater reliability and intraclass correlation co-efficient was measured.

4.3.1 Reliability and congruence in measurements

4.3.1.1 *Inter-rater reliability*

Inter-rater reliability is the degree to which two or more independent raters measure or grade the same phenomenon [171]. This is essential as agreement between raters in grading severity or in segmenting has an impact on the overall execution of the project, the reliability of the findings and the clinical applicability of the findings. This is, to some extent, dictated by the type of data.

50 two-dimensional images of the MRI scans at the anatomical region of interest (i.e. at the base of the 1st metatarsal) were selected at random from 853 images. Two raters – a Consultant Musculoskeletal Radiologist and a Vascular Registrar rated each image independently as showing mild, moderate or severe sarcopenia. Both assessors were blinded to the grading allocation (figure 53).

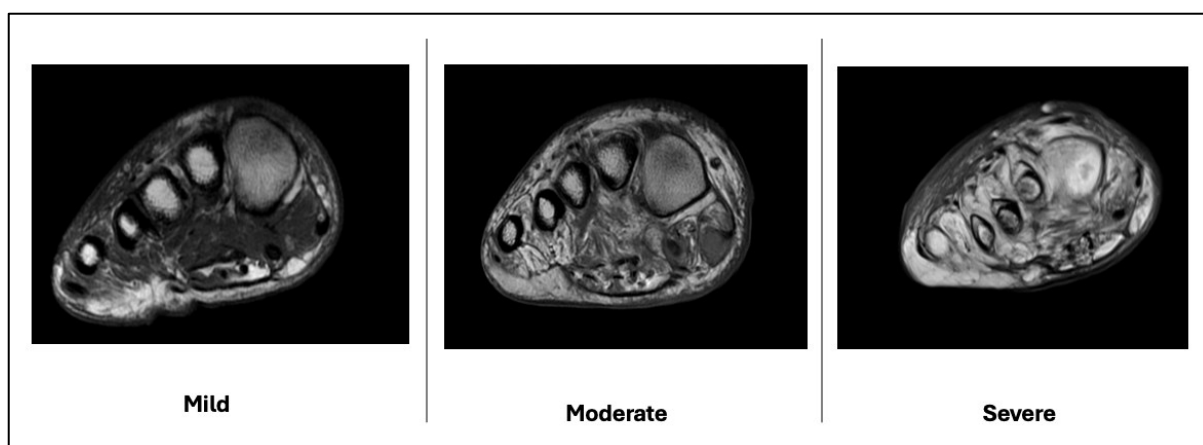


Figure 77: Grading of sarcopenia.

Given the categorical and ordinal nature of the data in recording the severity and the use of two raters, the most appropriate test for inter-rater reliability is weighted kappa which is slightly different to Cohen's kappa in that it accounts for the size or weight in the extent of disagreement. The kappa is a form of correlation co-efficient [171].

Within Cohen's weighted kappa, there are further discriminations – linear and quadratic. Linear weights penalise the extent of disparity between the categories allocated by two raters equally whilst quadratic weights penalise the disagreements in the extent of disparity exponentially between two raters.

Disagreement was noted in six out of the fifty scans. Each of these disagreements in allocating was by a single category i.e. mild to moderate or moderate to severe instead of mild to severe.

Using a Python extension within SPSS, Cohen's weighted kappa was calculated with quadratic weighting. The calculated value was 0.827 [95% CI 0.726-0.928; p-value <0.001]. Interestingly, the Cohen's weighted kappa was calculated to be 0.738 [95% CI 0.597-0.880; p-value <0.001].

The Landis and Koch kappa interpretation categories define kappa values as follows:

- <0-0.20 = No agreement
- 0.21-0.39 = Minimal
- 0.40-0.59 = Weak

- 0.6-0.79 = Substantial
- 0.8-1.0 = Almost perfect

Therefore, a value of 0.827 on a scale of minus one to one indicates an almost perfect agreement between the two raters whilst a value of 0.738 implies substantial agreement [171, 172].

4.3.1.2 Intraclass Correlation Co-efficient (ICC)

Intraclass correlation co-efficient is another measure of consistency between two or more measurements which are at baseline of the same quantity. It measures the similarity in grading from different raters and accounts for the difference between subjects and potential errors in measurement [173].

A two-way mixed effects models was utilised to calculate the ICC. Ordinal grading severity was converted to numerical values such that Mild=1, Moderate=2 and Severe=3.

The ICC was calculated using R version 4.5.0 using '**readxl**', '**writexl**' as well as the following other packages:

- **DescTools**: For descriptive statistics and exploratory data analysis.
- **psych**: For reliability analysis.
- **blandr**: To create Bland-Altman plots.

The complete, clean code is available in appendix 5. The ICC was 0.830 [95% CI 0.71-0.89; p-value <0.001]. Koo and Li suggest the following interpretations of reliability using ICC however, this is not applied by the scientific community to the same extent:

- <0.5 = Poor reliability
- 0.5-0.75 = Moderate reliability
- 0.75-0.9 = Good reliability
- >0.9 = Excellent reliability

Therefore, a value of 0.83 is indicative of good reliability which is of statistical significance between the measurements [173]. A clustered Bland Altman plot was also generated (figure 54). The purple band is almost at zero which implies significant lack of systematic bias between the two raters grading the severity of the MRI scans.

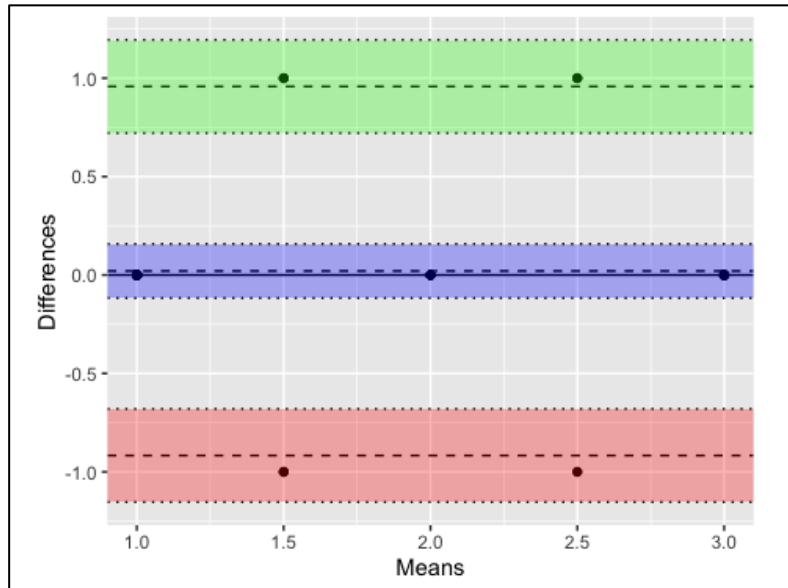


Figure 78: Bland-Altman plot. The green and pink lines are indicative of the upper and lower limits of agreement i.e. the standard deviation. The points are within these bands which demonstrate consistent differences within those bounds.

4.3.2 Methodology - ViT and ResNet

The basic preparation of data and pathway to run ViT and ResNet is the same and therefore the methods described below are apply to both models.

4.3.2.1 Creating the dataset

The use of XNAT to store the MRI images and process these has already been discussed in chapter two, (sections 13.2 and 14). Two-dimensional images of T1-weighted, coronal views of the foot at the base of the 1st metatarsal were captured in joint photographic experts group (JPEG) format for all 828 scans and cropped to initially include the surrounding black background. 4 of these scans could not be used/labelled therefore 824 were used in the end. Each scan image was labelled using the pseudonymised ID and a separate scan ID to ensure the images could only be traced back by a single person.

The new, anonymised images were processed using Roboflow inference graphical user interface on a secure integrated environment with Python capabilities and using an Apache 2.0 license (figure 66) [174].

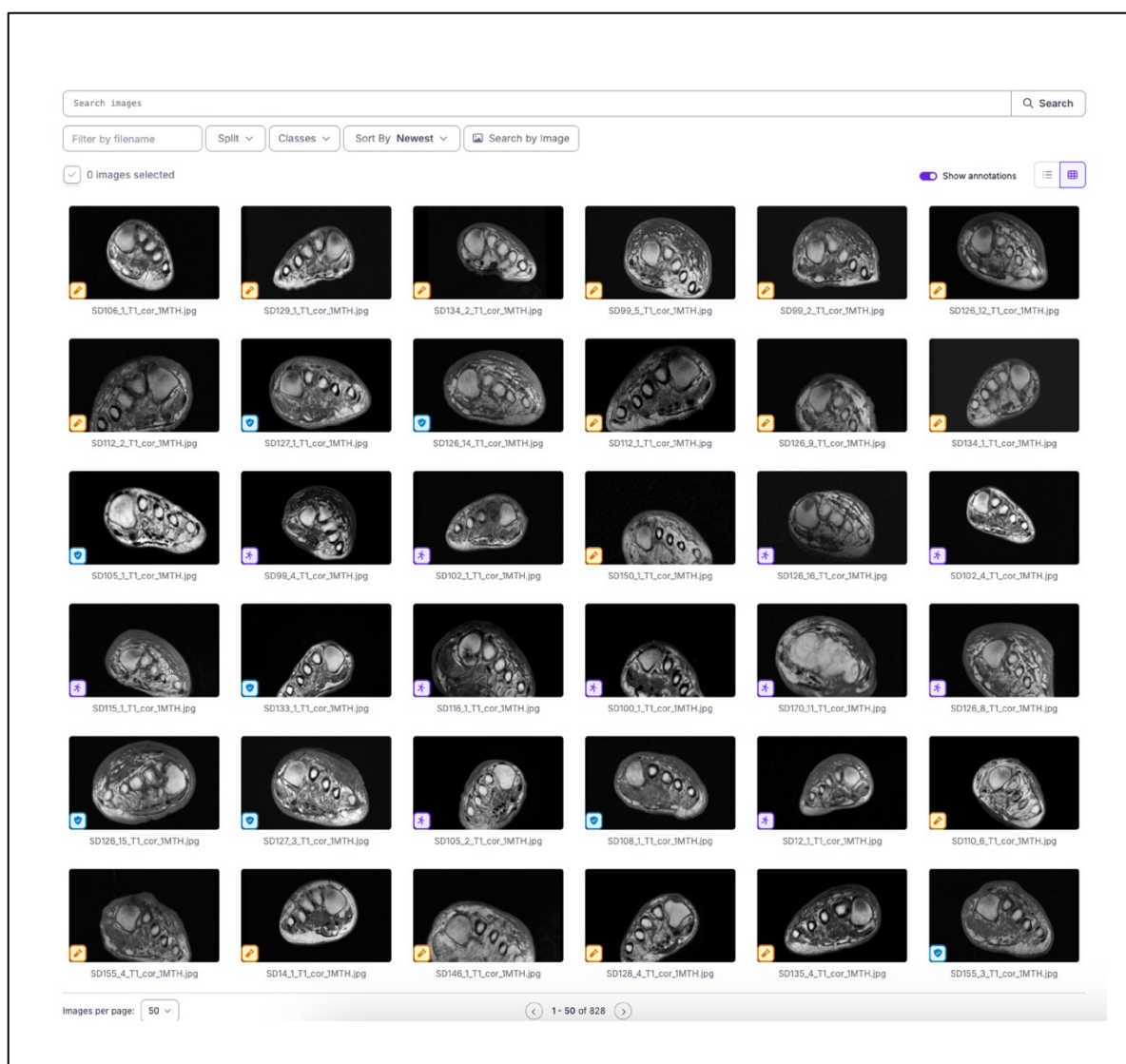


Figure 79: Collating the images to prepare the dataset. Full area of the foot was available for segmentation in each included case.

4.3.2.2 Data pre-processing and labelling

The images were pre-processed prior to labelling and annotating the images. Several steps were employed to pre-process the dataset prior to annotating the images which are summarised below in table 49.

Data Pre-processing	Purpose/Function
Auto-Orient	To standardise pixel ordering.

Discard exchangeable image file format (EXIF)	Removes identifiable content i.e. source of image, camera/scanner/computer used, location, etc.
Static Cropping	Crop image to the area of interest. 23-79% Horizontal region crop. 48-95% Vertical region crop.
Resizing	Image resized/stretched to 640x512 pixels. This allows to zoom into the area of interest.
Grayscale	Change all images to grayscale to ensure uniformity and a single channel to improve processing efficiency.

Table 63: Data pre-processing steps.

All 824 MRI images were reviewed individually and annotated/classified into three classes – mild, moderate and severe sarcopenia as shown below in figure 67.

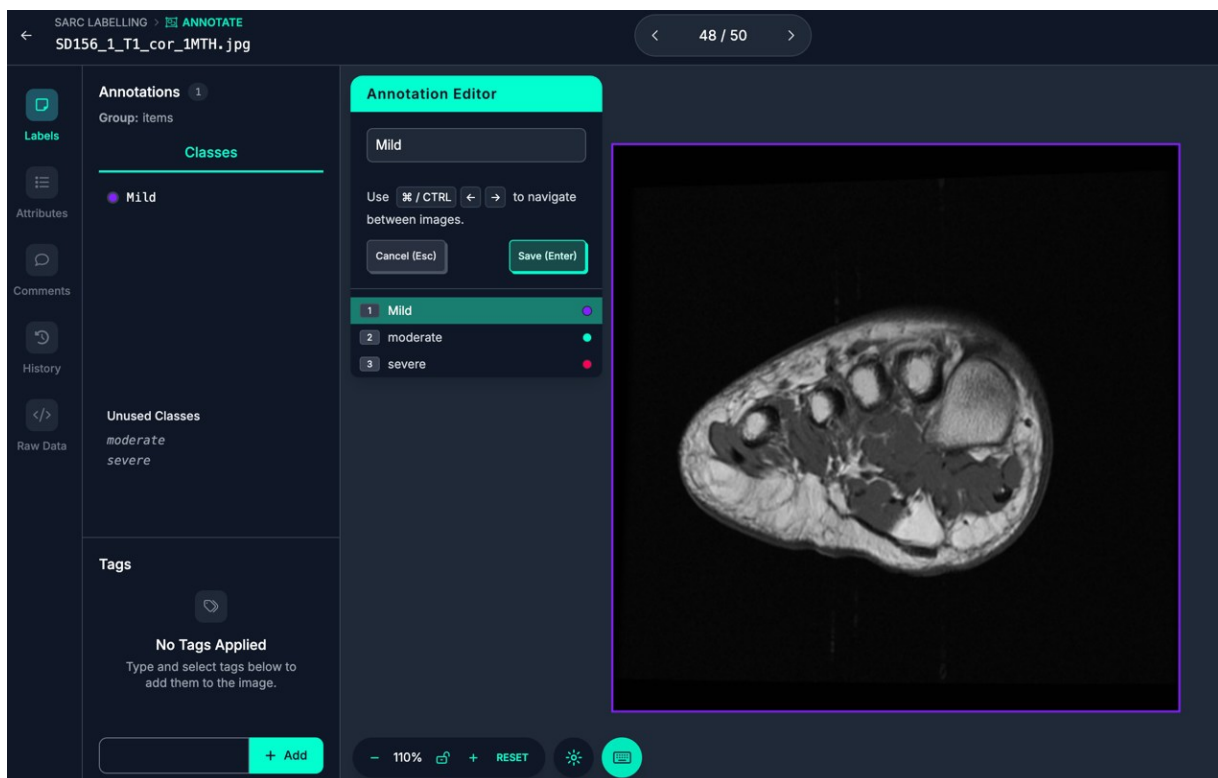


Figure 80: Image annotation of each MRI into three categories of sarcopenia severity - mild, moderate or severe.

4.3.2.3 Data Augmentation

Data augmentation has been discussed earlier in this chapter. In order to increase the overall number of images available for the dataset, a number of data augmentation strategies were utilised. The images were augmented in a way which would be similar to

real image changes at the time of the MRI. This included flipping the images vertically or horizontally, rotating each image by ninety degrees in a clockwise or anticlockwise fashion and zooming/cropping into an image [7% minimum zoom and 7% maximum zoom] (figure 68).

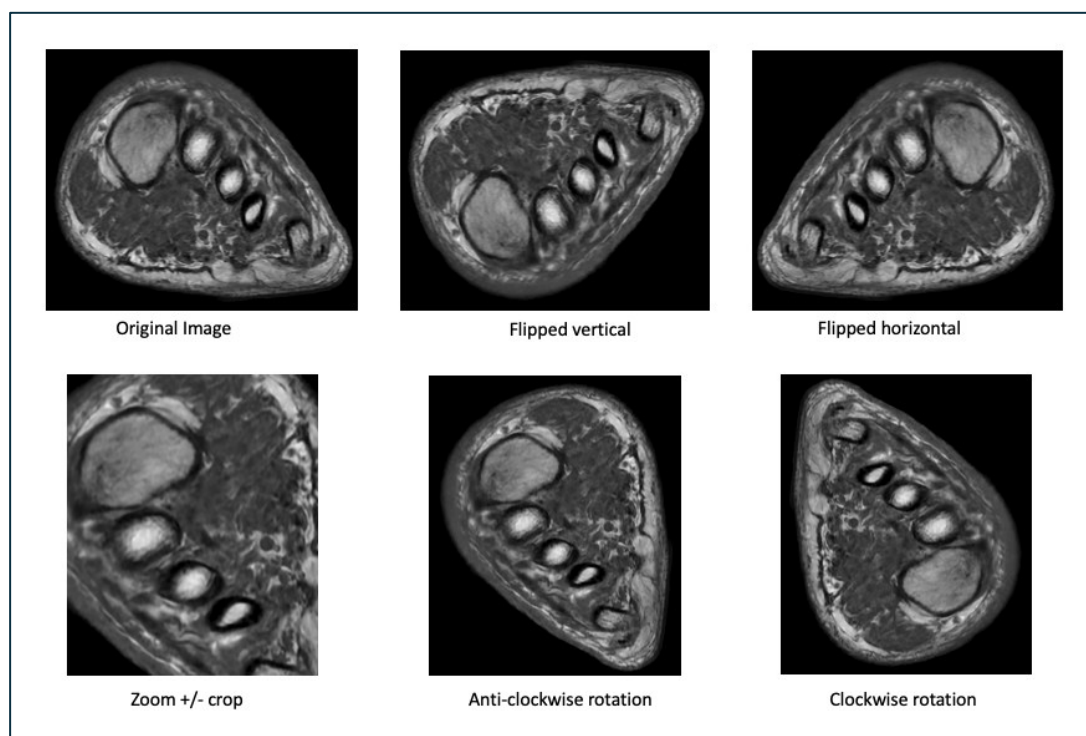


Figure 81: Data augmentation strategies prior to using ML models.

This served two functions – firstly, it helped to expand the dataset from 824 images to 1740 images for the models to work on. Secondly, it increased the number of images available for the validation and training set.

4.3.2.4 Data split and training the models

The data was split into a ratio of 70:20:10 (training:validation:testing) such that there were 1218 images in the training set, 348 images in the validation set and 174 images in the test set. The reason this split was used was due to the fact that a sufficient number of images were available after data augmentation to be available for each set which meant that a data split of 70:20:10 could be used instead of an 80:20 split to ensure there were enough cases for the test set to train on. ViT and ResNet were integrated within the Roboflow inference graphical user interface and deployed separately to classify the images as showing mild, moderate or severe sarcopenia.

Three versions of ViT were attempted and at least three versions were run for ResNet 18, ResNet 34, ResNet 50 and ResNet101 each. The first of each of these versions used raw, unprocessed data without any image modifications or augmentations with the original 824 images to establish a baseline. The remaining two versions of each of these models used processed images with data augmentation. The most accurate model output was selected for ViT and each ResNet version.

4.3.2.5 Results

In the original dataset which was annotated, 399 scans were labelled as mild, 269 as moderate and 165 as severe. The results discussed utilise a set of terms which are explained here prior to delving into the output for the models (ViT, ResNet-18, ResNet-34, ResNet-50 and ResNet-101).

Epoch

An epoch is a single complete pass through the entire dataset (of all MRI images) where every sample (an MRI image in this case) is passed through the model. One epoch involves a forward pass, loss calculation, backpropagation and update of the parameters and weights [175, 176]. An epoch is, in and of itself, a hyperparameter [176]. The number of epochs can vary and can be adjusted to help improve the model's performance at times but increasing the epochs too much risks overfitting whilst not having enough epochs risks underfitting, which has already been discussed in chapter 3. Increasing the number of epochs also significantly increases the training time [175]. This can be addressed with early stopping, a process whereby the model automatically stops when the model's performance begins to plateau or even decline on the validation set. This was used in XGB in the last chapter and is used here again. If the dataset is very large (not the case here), the number of samples can be broken down into smaller batches and the process is repeated on the smaller batches. This is called an iteration. One epoch has multiple iterations on the smaller batches [175].

Training Loss / Loss function

Loss function is an objective metric which measures the model's performance by calculating the difference or "loss" of the model's prediction from the ground truth [177]. If a majority

of the predictions are correct, the loss is small, however, if a majority of the model's predictions are wrong, the loss is large [177].

Accuracy / Validation accuracy

This is defined by the number of all the correct classifications, whether they were positive or negative and are defined using the equation below [178]:

$$Accuracy = \frac{\text{number of correct classifications}}{\text{number of total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **TP:** True positive
- **TN:** True negative
- **FP:** False positive
- **FN:** False negative

This has the same information as a confusion matrix. In an ideal model with a 100% accuracy, the model would have no false positives or false negatives. The general rules for interpreting accuracy are dictated by a score <60% indicating weak performance, 60-79% as reasonable performance, 80-89% as strong performance and 90-100% as excellent (this is rare, even in the AlexNet and ResNet papers as shown in table 48 in the section on ResNet earlier).

Recall / Probability of detection

Recall is the same as the true positive rate (TPR) i.e. the number of correct classifications which were actually correct and is calculated with the following calculation [178]:

$$Recall = \frac{\text{actual positives correctly classified}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

Recall improves when the number of false negatives is reduced.

Precision

Precision is the proportion of the model's positive classifications which were actually positive and is calculated using the following equation [178]:

$$Precision = \frac{\text{correctly classified positives}}{\text{all classified as positive}} = \frac{TP}{TP + FP}$$

Precision improves when the number of false positives decrease.

F1 score

The F1 score is defined as a “harmonic mean of the precision and recall” [178, 179]. The F1 score balances the precision and recall and can be used preferentially in datasets with significant class-imbalance [178]. The F1 score is between zero and one (ideal). It is calculated using the following equation:

$$F1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2(TP)}{2(TP + FP + FN)}$$

Confidence Threshold/Classification threshold

The confidence threshold is an adjustable cut-off value to decide if a classification predicted by the model is reliable enough to act on [180, 181]. An example of this is as follows. A model predicts that an MRI image has a 70% chance of showing severe sarcopenia. If the confidence threshold is set to 65%, the image is classified as severe sarcopenia. However, if the threshold is set to 90%, the image is not classified as severe sarcopenia as 70% is less than 90%. This usually involves some trade-off between the precision and recall, and will be discussed below. Higher thresholds might yield more false negatives because the model needs to be sure to a higher percentage that it is classifying correctly whilst a lower threshold makes more classifications but can also make more errors.

4.3.2.5.1 Results [ViT]

Table 50 below shows a summary of the ViT model and its performance. The first row shows the raw, unprocessed images which yields a low accuracy. After increasing the number of images in the dataset using data augmentation and pre-processing, the accuracy of the model increases to 78.1%.

An F1 score of almost 80% (79.9%) is considered strong performance in the classification task. The model demonstrates high precision (81.5%) and very good recall (78.7%) at a low

confidence threshold (27%). This also shows that the model is not overfitting and balancing the data input well.

Model Used	Type of dataset	Number of images	Accuracy	F1 Score	Precision	Recall	Confidence threshold
ViT	-Raw images -Unprocessed -No data augmentation	828	37.5%	33.6%	42.2%	47.2%	18%
ViT	-Processed images -After data augmentation	1740	78.7%	79.9%	81.5%	78.7%	27%

Table 64: Vision Transformer (ViT) summary of model metrics.

Figure 69 shows the automatically selected confidence threshold selected by the DL model at 27%. However, this was reset to 51% as shown in figure 70 without any loss in precision recall or accuracy. Beyond this confidence threshold, the precision continues to increase but at the expense of the recall.

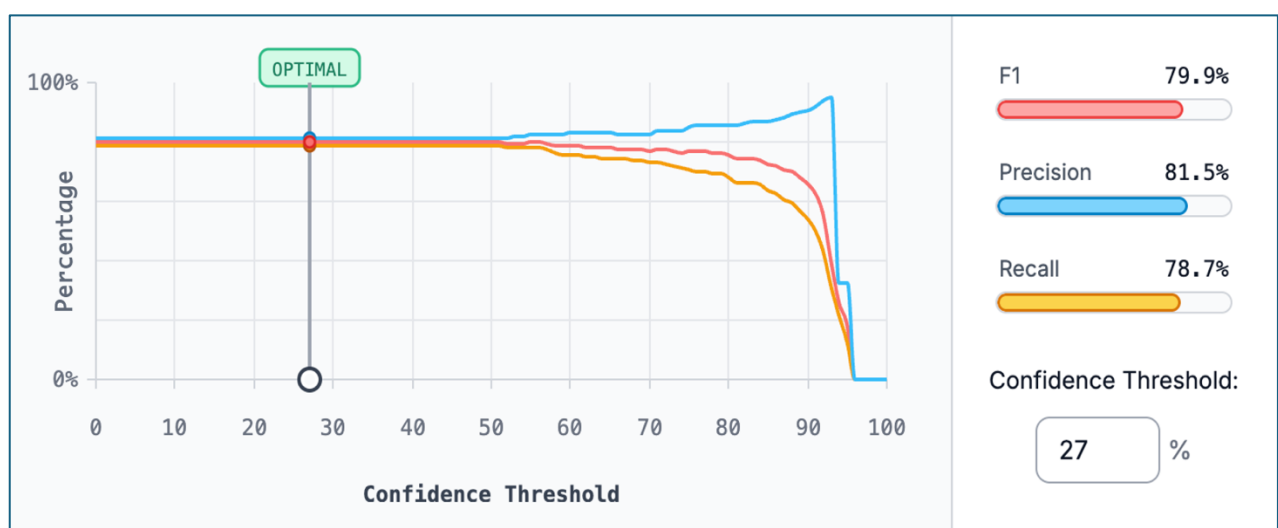


Figure 82: Vision Transformer (ViT) model performance metrics (automated).

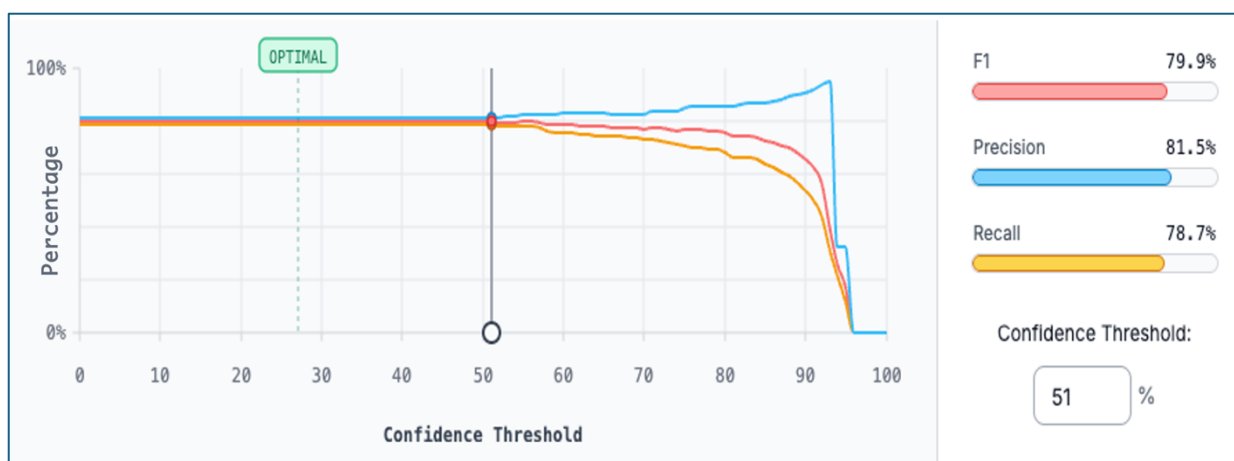


Figure 83: Vision Transformer (ViT) model performance metrics (adjusted).

The graph below (figure 71) is a visual representation of the training loss. The x-axis shows the number of epochs. The y-axis is the numerical representation of the error rate. There is evidence of high training loss when the model is applied. This drops significantly between the first and second epoch. After the fourth epoch, it remains fairly stable which means the ML model learned effectively in the earlier epochs. From epoch 5 onwards, the loss is almost zero and training the model with further epochs does not change this.

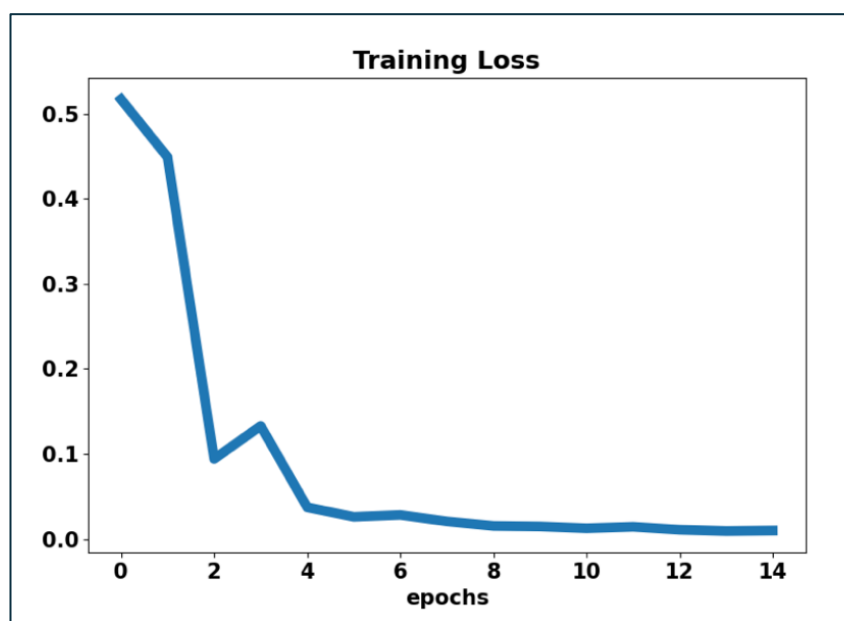


Figure 84: Vision Transformer (ViT) graph demonstrating training loss. Number of epochs (x-axis) and the training loss (y-axis).

Graph 72 below is a visual representation of the validation accuracy i.e. accuracy of the model on the validation set. The x-axis again illustrates the number of epochs taken for the

model to perform and the y-axis shows the proportion of correct predictions on the validation set (given as a value between zero and one).

The accuracy is high at the start, between the first and second epoch which shows efficient learning by the model. It then drops between the fourth and sixth epoch as the model adjusts to a learning rate step and starts to overfit. There is evidence of fluctuations between the sixth and twelfth epoch as the model adjusts before stabilising after the 12th epoch at 0.77-0.79 (77%-79%) and has generalised to the data.

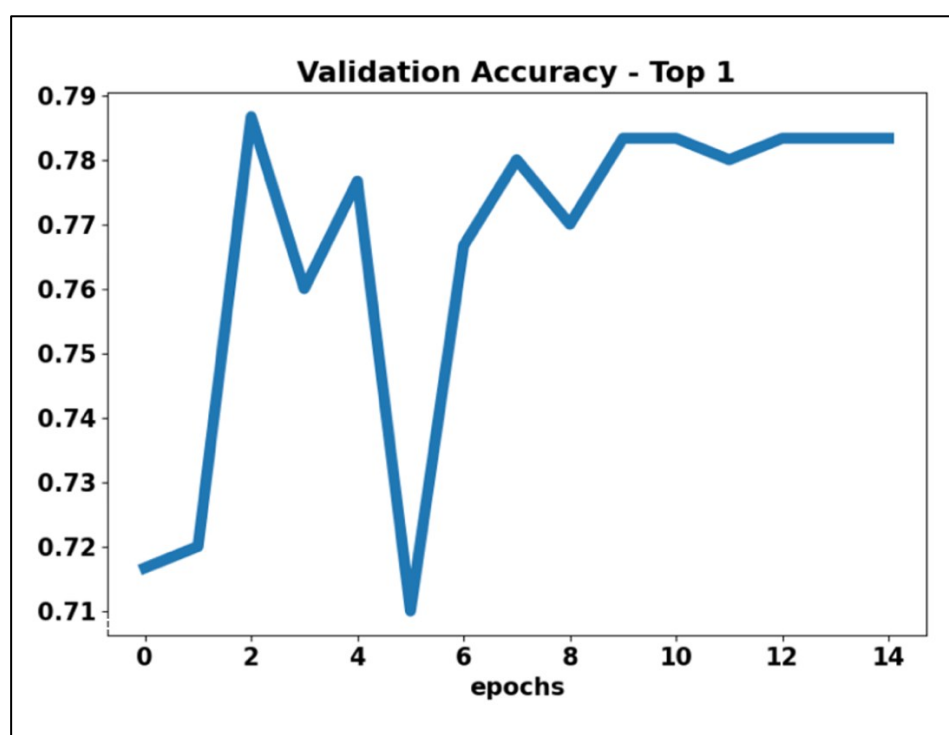


Figure 85: Vision Transformer (ViT) graph demonstrating validation accuracy. Number of epochs (x-axis) and validation accuracy value (y-axis).

Given the training loss is almost zero and the validation accuracy and F1 score, this demonstrates that the model fits the training data very well. Table 51 below shows the confusion matrix for the test set which shows good prediction for each classification category (mild, moderate and severe sarcopenia).

		Prediction			
Ground Truth		Moderate	Severe	Mild	False Negative
	Moderate	47	16	1	0
	Severe	6	70	21	0
	Mild	0	13	126	0
	False Positive	0	0	0	0

Table 65: Vision Transformer (ViT) confusion matrix of the test set.

4.3.2.5.1 Results [ResNet]

As discussed earlier in the chapter, ResNet-18 and ResNet-34 have a basic residual block with 2 convolutions per block, whilst ResNet-50 and ResNet-100 have a bottleneck architecture with 3 convolutions per block.

ResNet-101 was applied to the raw dataset with an accuracy of 74.4%, which is not significantly different to ResNet18 with data pre-processing and augmentation (75%).

The metrics for all five models are summarised in table 52. ResNet-34 had an accuracy rate similar to ViT earlier, whilst ResNet-50 and ResNet-101 had better performance with high accuracy rates of 79% and 80.7% respectively.

Model Used	Type of dataset	Number of images	Accuracy	F1 Score	Precision	Recall	Confidence threshold
ResNet 101	- Raw images - Unprocessed - No data augmentation	828	74.4%	64.3%	65.6%	63%	64%

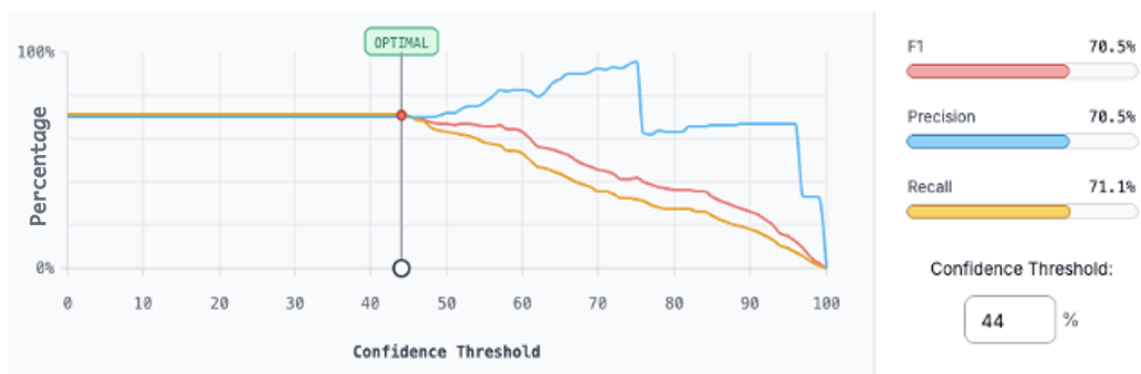
ResNet 18	- Processed Images - After data augmentation	1740	75%	70.5%	70.5%	71.1%	44%
ResNet 34	- Processed images - After data augmentation	1740	78.7%	78.9%	79.5%	78.5%	18%
ResNet 50	- Processed images - After data augmentation	1740	79%	74%	73.9%	75.3%	19%
ResNet 101	- Processed images - After data augmentation	1740	80.7%	69.3%	77.8%	66.8%	53%

Table 66: ResNet-18, ResNet-34, ResNet-50, ResNet-101 summary of model metrics.

Figures 73 and 74 illustrate the metrics for all four models. The confidence thresholds for ResNet-34 could be adjusted (from 18% to 41%) without any changes to the other values, as could the confidence threshold for ResNet-50 (from 19% to 36%). This was not possible for ResNet-18 and ResNet-101, as seen in the figures below.

The F1 value was highest for ResNet 34, as was the precision and recall. ResNet-18 demonstrates the most significant deterioration by increasing the confidence threshold, which all the other models maintain throughout. The lowest deviation between the precision and recall with increasing confidence threshold was seen in ResNet-101 followed by ResNet-34.

ResNet-18



ResNet-34

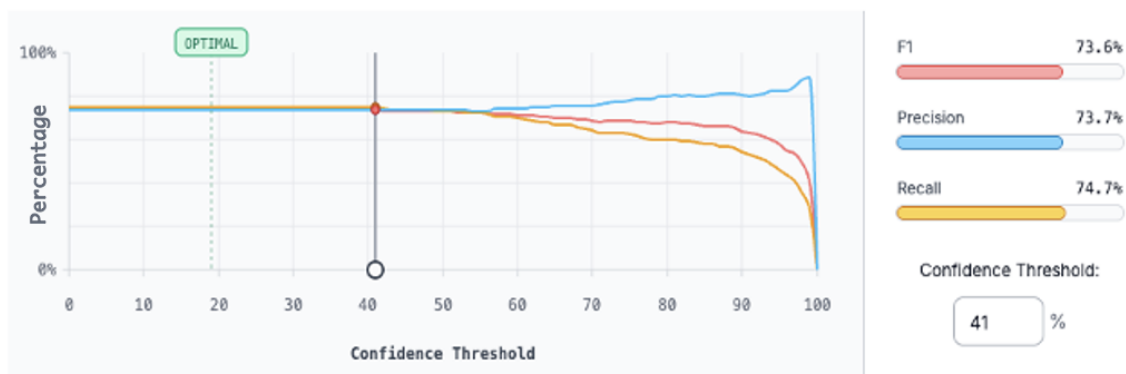
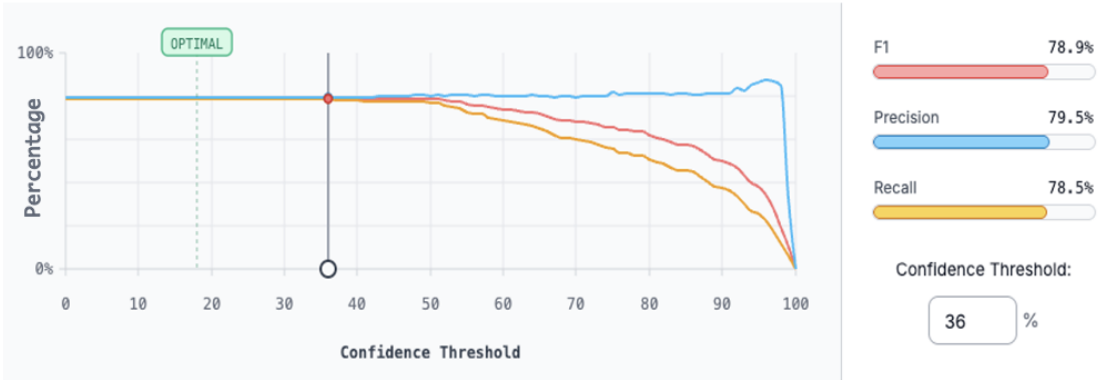


Figure 86: ResNet-18 and ResNet-34 model performance metrics (automated and adjusted).

ResNet-50



ResNet-101

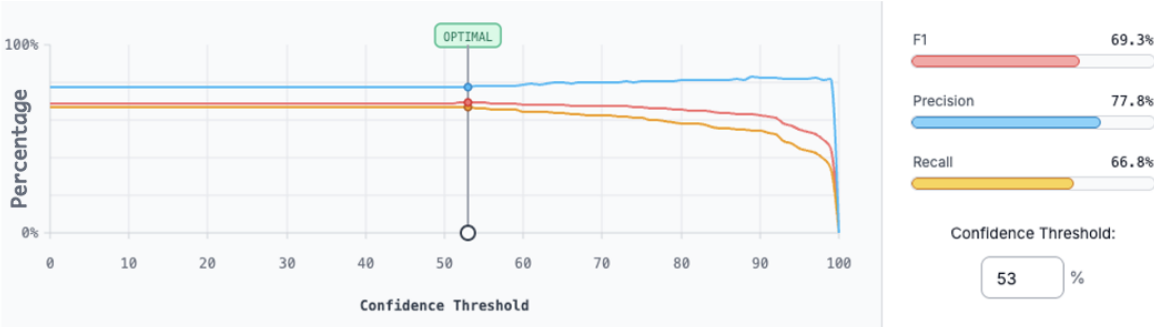


Figure 87: ResNet-50 and ResNet-101 model performance metrics (automated and adjusted).

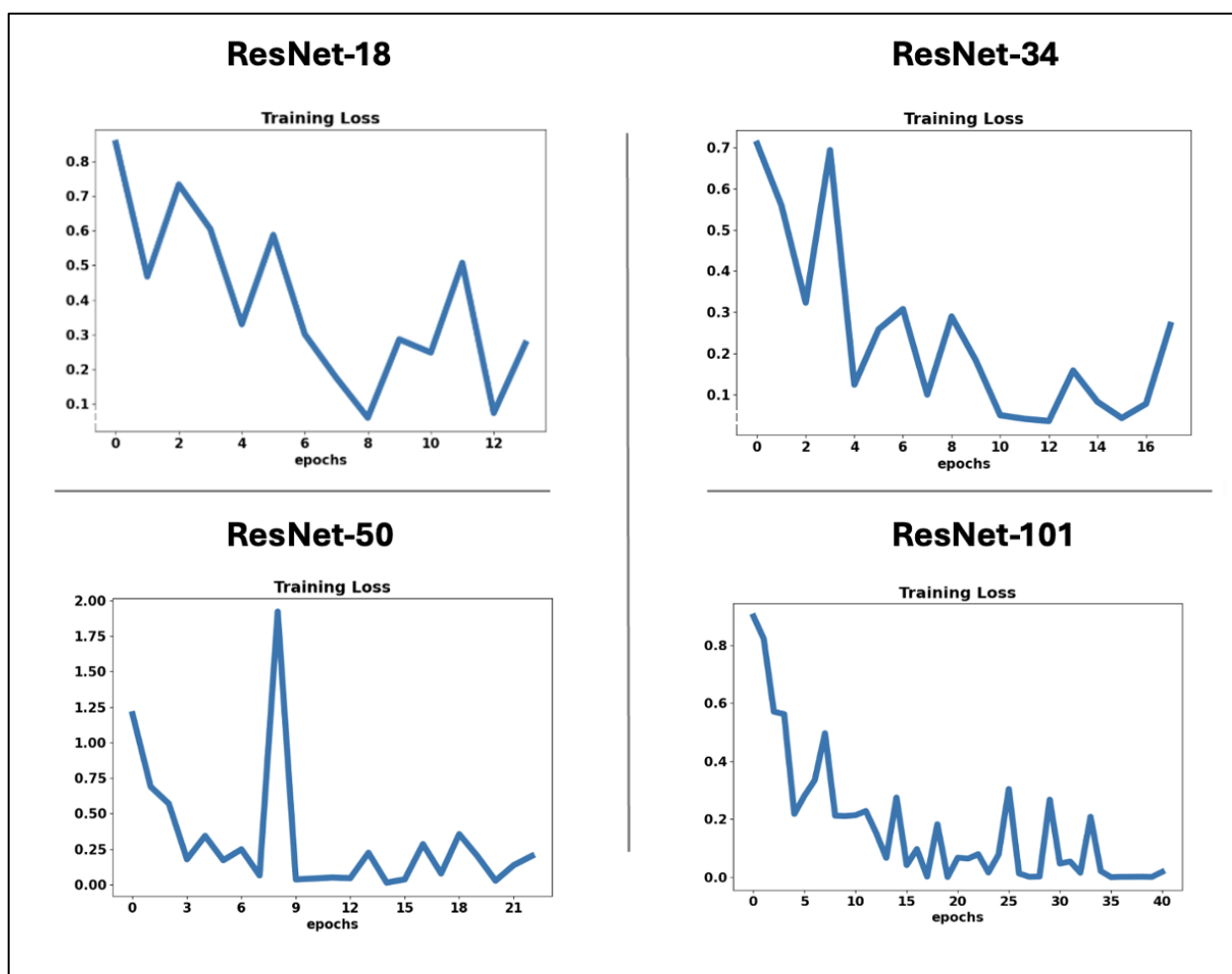


Figure 88: Graphs demonstrating training loss for ResNet-18, ResNet-34, ResNet-50 and ResNet-101. Number of epochs (x-axis) and the training loss (y-axis).

There was significant difference in the training loss for each model (figure 75). ResNet-18 shows a decrease in the training loss, reaching the lowest training loss around epoch eight and twelve, but then increases to around 0.3. The multiple spikes are indicative of instability in the model during the training phase. ResNet-34 has a slightly more gradual decline after epoch four, with the training loss reaching almost zero between epochs ten and twelve before getting to around between 0.2 and 0.3. The training loss is rapid in the early epochs, demonstrating efficient early training but then has a large spike between epoch six and nine, the reason for which is not known. Following this there are small oscillations, but the model does stabilise to a training loss value around 0.25. ResNet-101 has a rapid decline in the training loss in the early epochs, indicating the model is learning efficiently and optimising accordingly. It does have some fluctuations between epochs twenty-five and thirty-five, which shows overfitting but adjusts quickly and stabilises after epoch thirty-five, with the training loss near to zero which is excellent. The deeper models generally achieve

lower training losses as expected with more stability, besides the instability spike seen in ResNet-50.

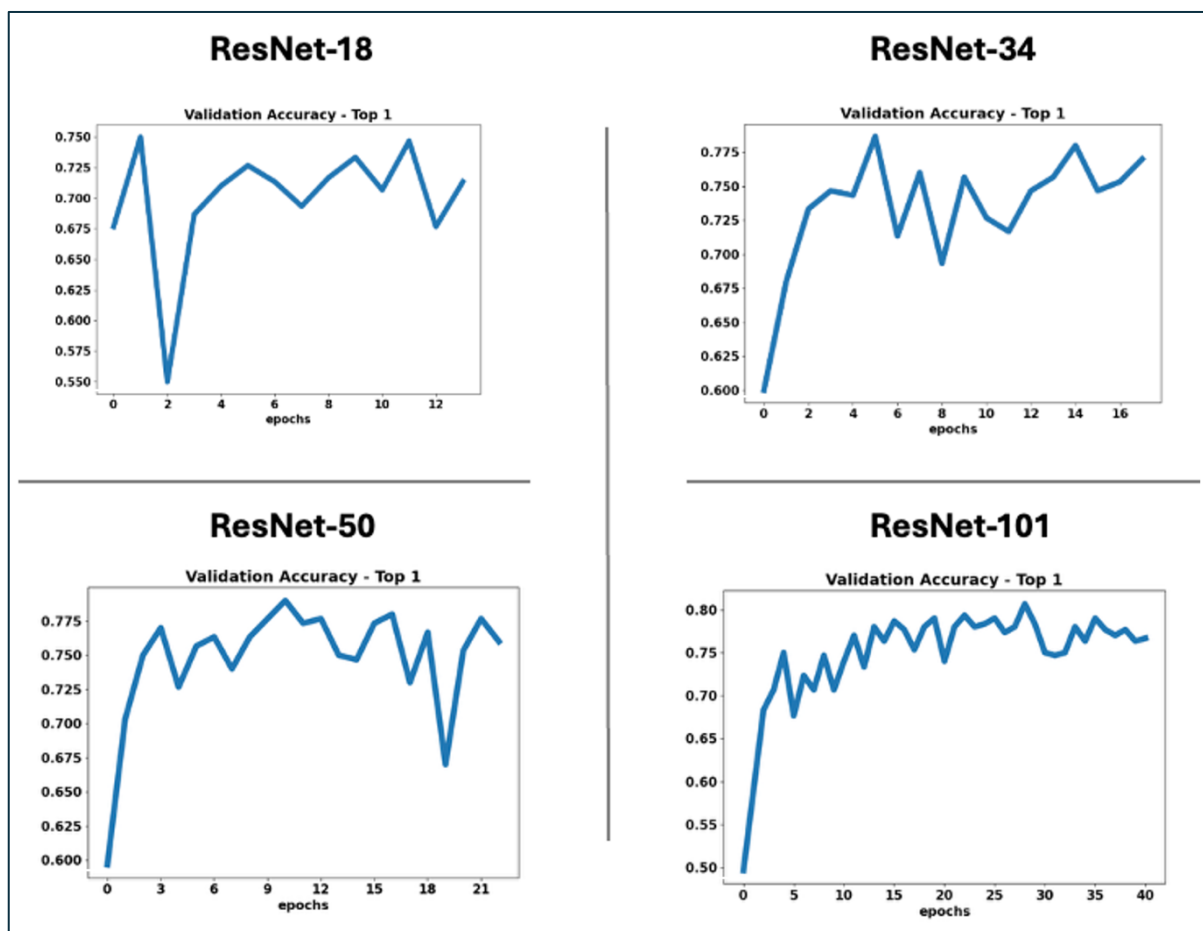


Figure 89: Graphs demonstrating the validation accuracy for ResNet-18, ResNet-34, ResNet-50 and ResNet-101. Number of epochs (x-axis) and validation accuracy (y-axis).

This is also reflected in the validation accuracy for each model (figure 76). ResNet-18 starts around 0.68 before showing a significant decline at epoch 2. The model then recovers but shows fluctuations throughout with variability, achieving a maximum average around 0.72. ResNet-34 demonstrates fluctuations in the model as well, before reaching a maximum average around 0.77. ResNet-50 also fluctuates in its accuracy with a downward spike between epochs eighteen and twenty-one. ResNet-101 begins with a lower validation accuracy, around 0.7, but consistently improves with small oscillations indicative of stability in learning and reaches an accuracy of around 0.8 in the end.

The training efficiency is also demonstrated in these graphs. Both ResNet-18 and ResNet-34 show rapid improvement in the earlier epochs but fail to improve any further. In

comparison, ResNet-101, a much deeper model, took significantly longer to train and learn (around forty epochs) but steadily improves, is more stable and accurate.

ResNet-18					
Ground Truth	Prediction				
		Moderate	Severe	Mild	False Negative
	Moderate	56	13	28	0
	Severe	23	112	1	3
	Mild	15	0	48	1
	False Positive	0	0	0	0

Table 67: Test set confusion matrix ResNet-18

ResNet-34					
Ground Truth	Prediction				
		Moderate	Severe	Mild	False Negative
	Moderate	49	0	15	0
	Severe	0	122	17	0
	Mild	10	18	69	0
	False Positive	0	0	0	0

Table 68: Test set confusion matrix ResNet-34.

ResNet-50					
Ground Truth	Prediction				
		Moderate	Severe	Mild	False Negative
	Moderate	125	14	0	0
	Severe	24	50	23	0
	Mild	0	10	54	0
	False Positive	0	0	0	0

Table 69: Test set confusion matrix ResNet-50.

ResNet-101					
Ground Truth	Prediction				
		Moderate	Severe	Mild	False Negative
	Moderate	56	1	37	3
	Severe	31	31	2	0
	Mild	7	0	131	1
	False Positive	0	0	0	0

Table 70: Test set confusion matrix ResNet-101.

The confusion matrices for each model also reflect this pattern (tables 53-56). All models generally performed very well in predicting mild cases from severe cases. Distinguishing between moderate and severe was evident in ResNet-18, but this was not the case with

ResNet-34 and ResNet-101. ResNet-50 performed well in classifying mild from moderate sarcopenia. ResNet-101 had the highest number of correct predictions for each category- mild, moderate and severe. Overall, ResNet-101 was the best performing model from all the ResNet versions but also took the longest in terms of computational resources and time.

4.3.3 Methodology - nnU-Net

nnU-Net has been discussed earlier in the chapter. All 824 MRI scans were uploaded to XNAT as discussed in chapter two. T1-weighted MRI images were selected as they provide good differentiation between fat and muscle. Using the XNAT viewer, the scan was loaded. The mask tool was selected (figure 77). The manual mask segmenting tool was selected (figure 78).

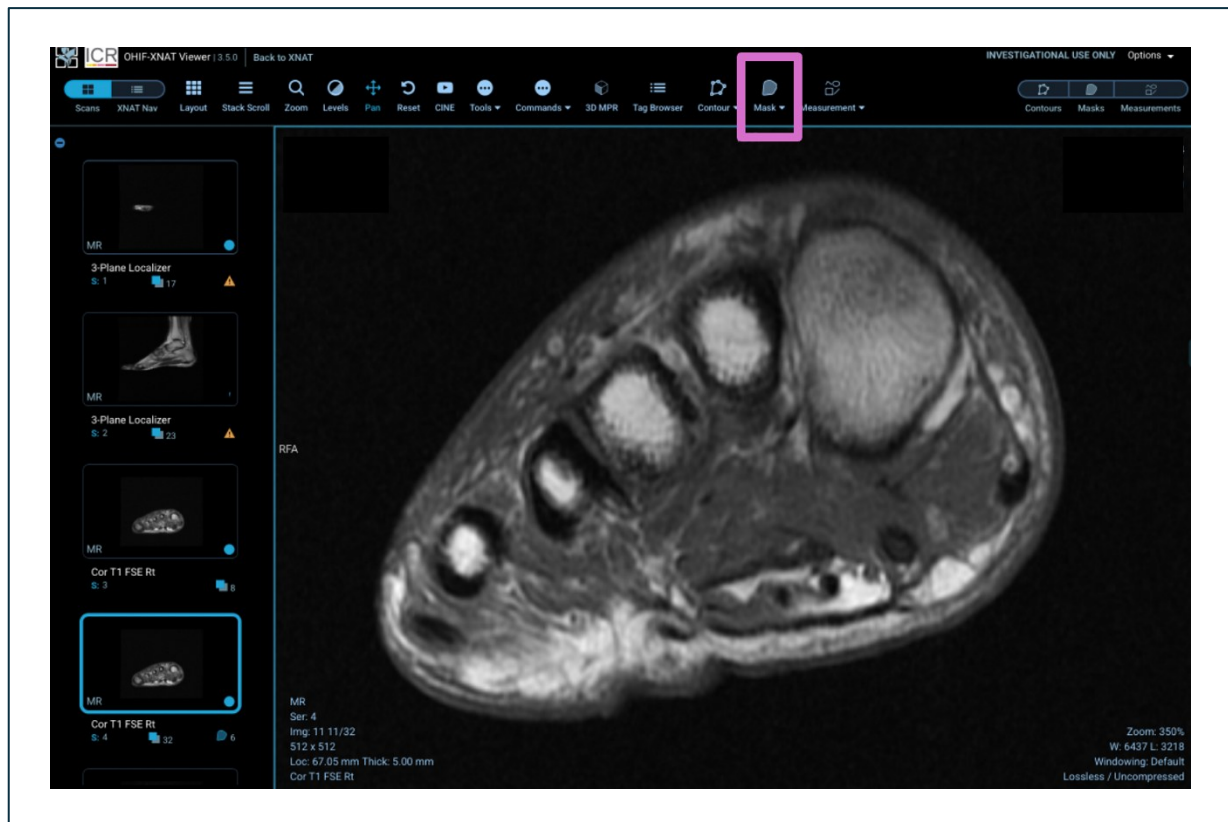


Figure 90: XNAT Viewer with the mask tool (purple box).

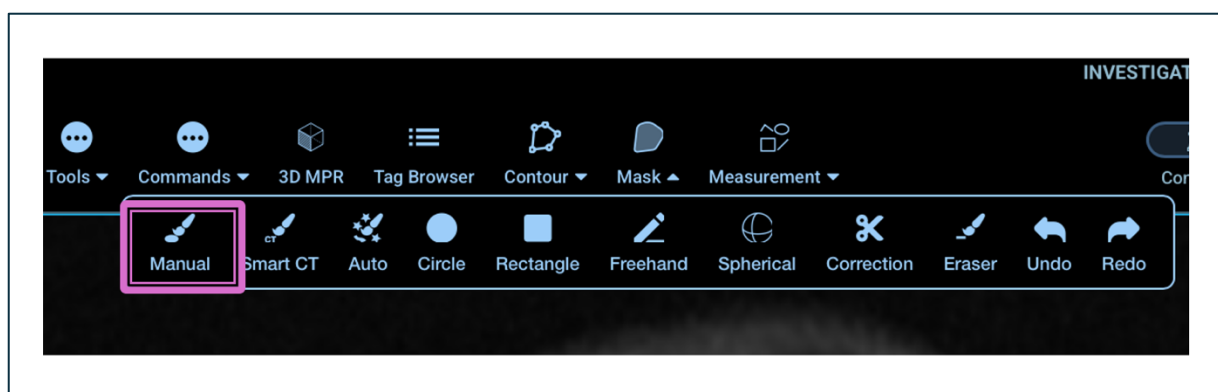


Figure 91: XNAT Viewer with the manual segmenting tool.

In the early segmentation trial sets, sagittal views from T2 scans were used, as well as T2 coronal views of the distal foot (figure 79). However, after reviewing the literature and discussing with radiology, certain rules were established to ensure uniformity whilst segmenting which nnU-Net could replicate.

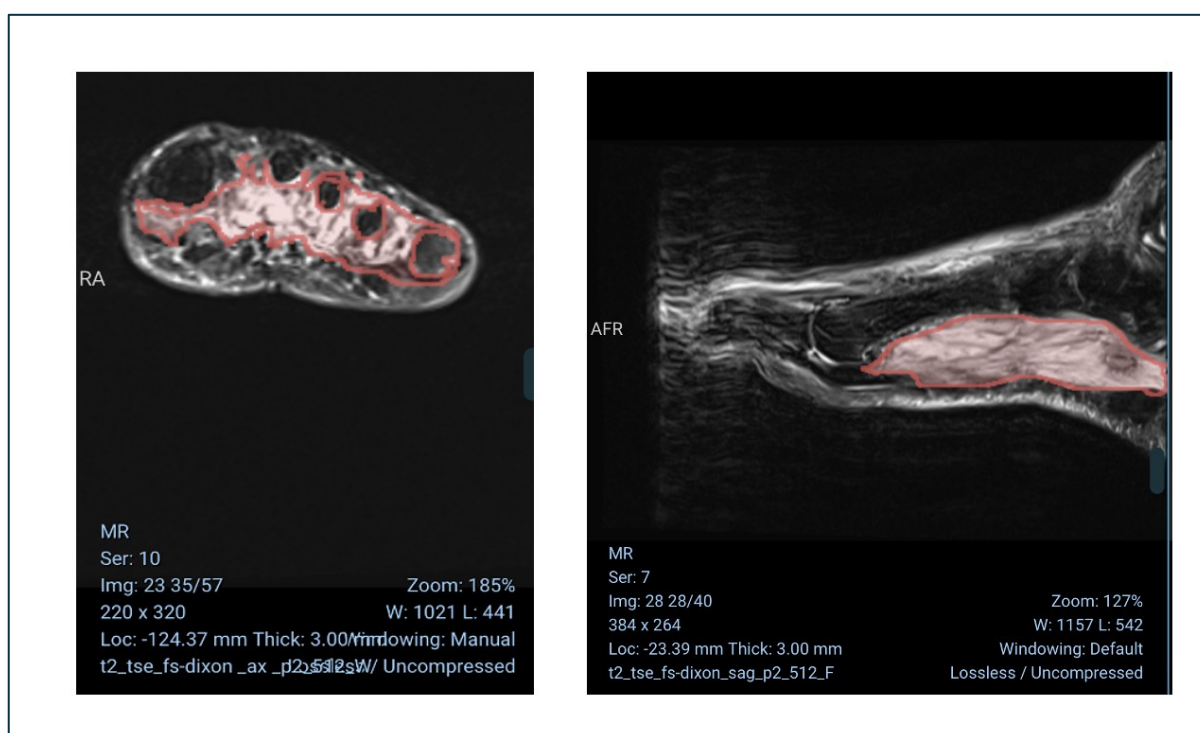


Figure 92: Early segmentation variations.

4.3.3.1 Segmentation rules

The anatomical landmark for segmenting the scans was set at the base of the 1st metatarsal. Blood vessels would be included in the segmentation region, but the flexor and extensor

tendon would be excluded. Post amputation regions were to be included, but areas with collections or bony destruction would be excluded.

4.3.3.2 Segmentation and checking

Using these rules, all 824 MRI scans were segmented manually as shown below (figure 80). A portion of the segmentation were double check by a musculoskeletal radiology consultant and adjusted, if required. Each segmentation also provides the volume, area, mean signal intensity and the standard deviation (figure 80 and 81). A portion of the plantar fat was also segmented to measure the mean signal intensity (figure 82). The rationale for this was to help calculate the pseudo fat fraction (which will be discussed in the next chapter). All the segmented scans were then downloaded along with their original scans to prepare to nnU-Net use.

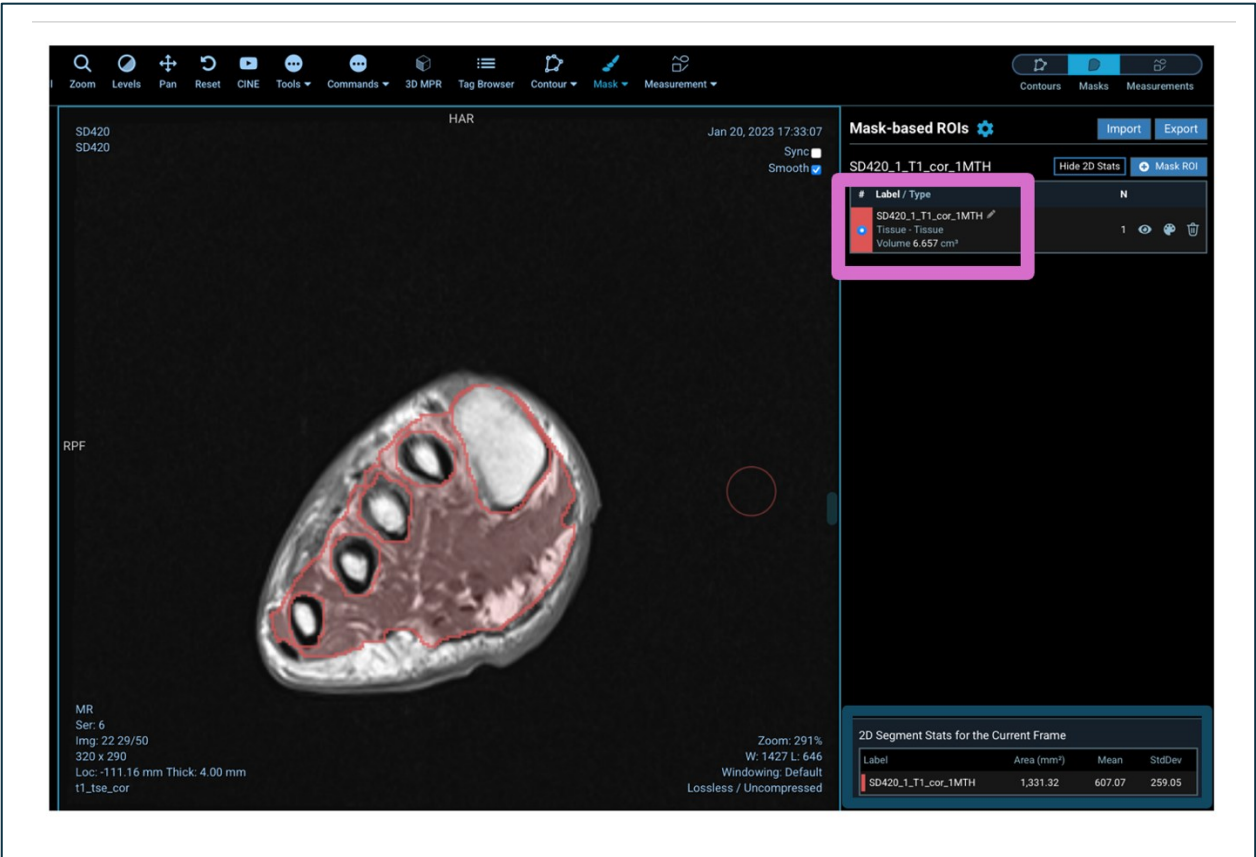


Figure 93: Region of interest segmented in red. The purple box on the top right provides the volume and the teal box in the lower right corner provides the area, mean signal intensity and the standard deviation.

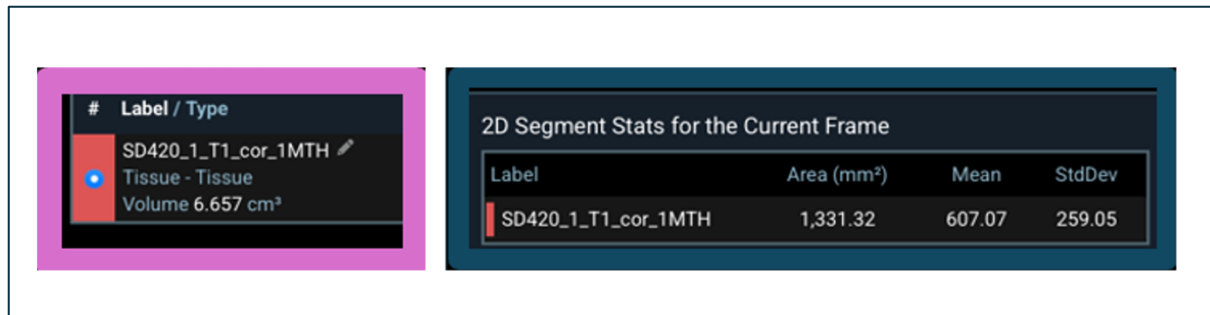


Figure 94: XNAT viewer with details of the segmented area. The purple box provides the volume, and the teal box provides the area, mean signal intensity and the standard deviation.

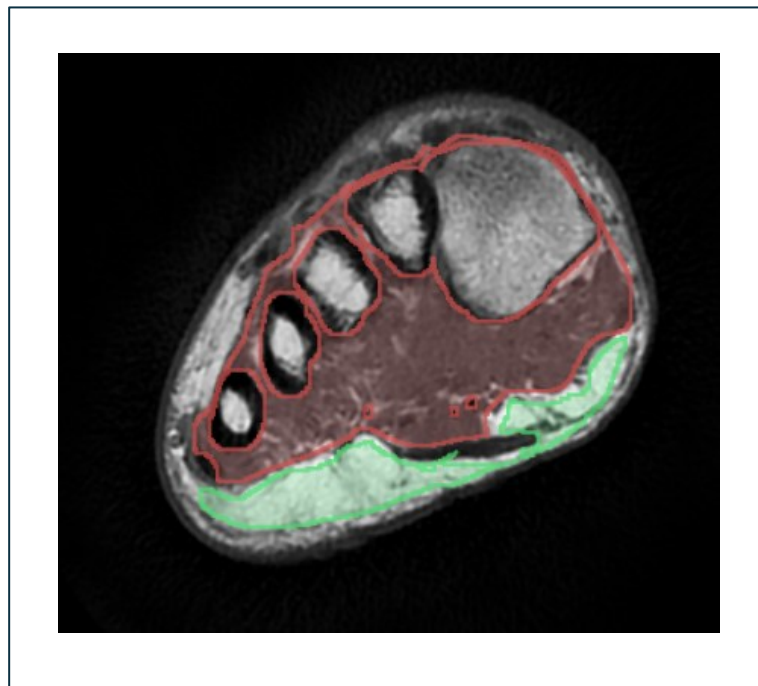


Figure 95: Final segmentation of the muscle (in red) and the plantar fat (in green).

4.3.3.4 Running nnU-Net

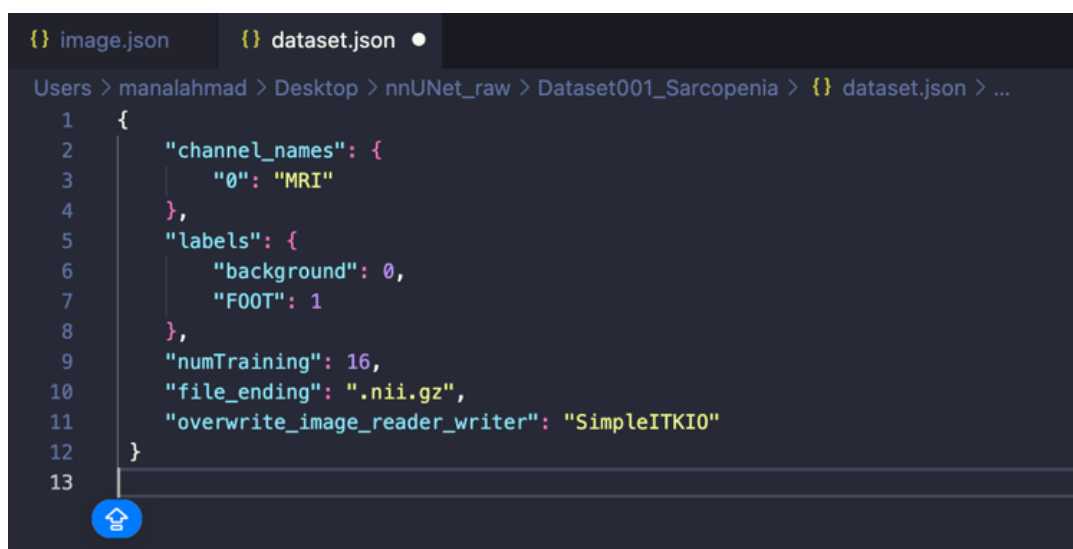
In keeping with the hardware requirements, an Intel graphics processing unit (GPU) with at least 64 gigabytes of unused random-access memory (RAM) was used with Microsoft Windows as the baseline software.

Python and PyTorch were installed in Anaconda. PyTorch is an open-source deep learning framework with tensor computation tools. Tensors have already been discussed earlier in this chapter. PyTorch leverages NVIDIA® GPUs via NVIDIA's compute unified device architecture (CUDA) technology for general purpose computing. In essence, it allows the computer to perform labour-intensive mathematical operations. nnU-Net (with Apache 2.0 licence) was installed via the GitHub repository [182].

Using the terminal window, three folders were created as required by the nnU-Net conventional architecture:

- nnUNet_raw: This contains all the scans prior to segmenting.
- nnUNet_preprocessed: This contains the segmented scans.
- nnUNet_results: This is where the predicted segmentations are added along with the rest of the output.

In another folder, three further folders were created. These were labelled 'imagesTr' (containing the training images), 'labelsTr' (with the segmented scans) and 'imagesTs' (test folder). The scans were re-labelled using the nnU-Net conventional naming. This follows the Dataset[IDENTIFIER]_[name] such that this dataset would be labelled as Dataset600_FOOT/predict_image_test/Dataset600_624.nii.gz, Further examples of the format for labelling the files is included in appendix 7. A JavaScript Object Notation (JSON) file titled dataset.json was created containing the metadata of the dataset (figure 83).



```
{
  "channel_names": {
    "0": "MRI"
  },
  "labels": {
    "background": 0,
    "FOOT": 1
  },
  "numTraining": 16,
  "file_ending": ".nii.gz",
  "overwrite_image_reader_writer": "SimpleITKI0"
}
```

Figure 96: JSON file for nnU-Net.

Each image could then be opened to review the scans as shown in figure 84 below. The first step was to allow the model to pre-process the images using the plan and preprocess coding 'nnUNetv3_plan_and_preprocess'. This is followed by the 'nnUNetv2_train' command. 100 epochs were used with 5-fold cross-validation. Cross validation is a method to evaluate the performance of a model. Instead of splitting the data once in a 70:20:10 or 80:30 split, the

data is split multiple times called ‘folds’ and these are then trained and tested multiple times. The final output then is an average of the performance from each round. The output included the DICE score, true positive, true negative, false negative and false positive values.

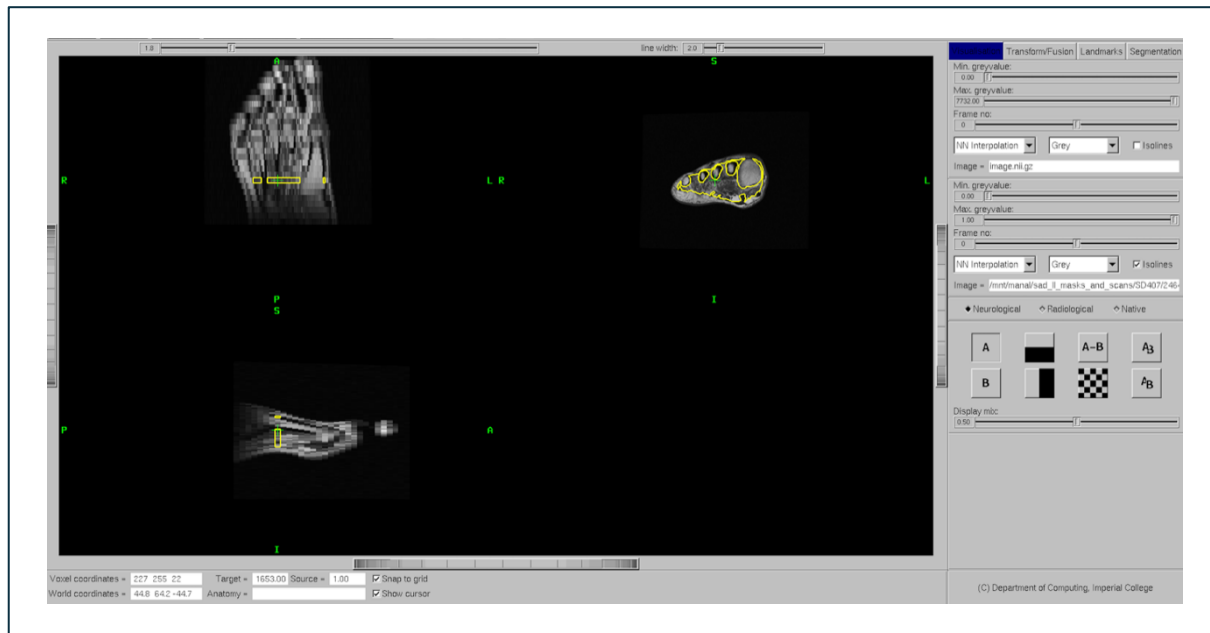


Figure 97: Image viewer prior to running nnU-Net.

Using this output, further analysis was undertaken using R packages ‘**tidyverse**’, ‘**readxl**’, ‘**knitr**’, ‘**gt**’ and ‘**scales**’. The complete, clean and reproducible code is available in appendix 6. This included descriptive statistics of the DICE scores, grouped summary statics by grouped performance. The metrics from the output were also used to calculate the accuracy, precision, recall, F1 score and the specificity.

4.3.3.5 Results [nnU-Net]

570 scans were used during training, 194 in the validation set and 62 in the test set. A summary table of the output is shown below in table 57. The complete output table with the corresponding dataset number and MRI file is available in appendix 7. The explanation for DICE scores has already been discussed extensively in chapter 1, section eleven, sub-section 4 (1.11.4).

The mean DICE score was 0.4 with a standard deviation of 0.29. The minimum DICE score was 0.0 and the maximum DICE score was 0.88. There were 5 complete fails (i.e. no true

positives) and 7 high performers. High DICE scores correspond to high true positive values and low false negative and false positive values whilst low DICE scores are associated with very low true positive and high false negative values. This implies that the lower DICE scores are driven by the missed positives.

DICE score	True Positive	True Negative	False Negative	False Positive
0.33993002405423100	6218	10717538	11225	12923
0.39399543653176400	8202	9141607	7095	18136
0.8042850264805010	13364	9417316	1773	4731
0.8124605080247690	6429	4709195	689	2279
0.6181800275753400	25108	10167492	3754	27262
0.17348245865596000	4537	9389416	18353	24878
0.0	0	10168301	24723	30592
0.5914731476237060	11487	7836965	2777	13091
0.512283570946066	14409	14113931	1766	25670
0.7772863966035730	10985	7322752	2903	3392
0.01778035248768970	399	9916990	20608	23475
0.5638870424770030	13149	10976560	8150	12189
0.24674943787271500	3786	9410283	8593	14522
0.25687019246114100	4024	8623445	11397	11886
0.08996142871481620	3137	9370580	37272	26195
0.27302916782541800	8345	12530128	10965	33474
0.18579559935434500	2187	6802581	5911	13257
0.09839437746414490	861	4701952	7455	8324
0.3084243627937770	3727	9416743	8854	7860
0.7729578631328290	14932	9151336	1206	7566
0.4781382003666220	14998	9389447	6816	25923
0.6304263715855450	8147	8370909	6169	3383
0.16963667820069200	1961	9153881	19152	46
0.22315945745326900	3307	8624421	10283	12741
0.13393826460341900	4133	9117458	17954	35495
0.04871109877528360	702	7049767	14681	12738
0.10679864548059400	2255	9135066	12570	25149
0.01654577347935330	349	11492499	14067	27421
0.0	0	2349560	0	5640
0.0	0	7325623	14409	0
0.02819831272391080	244	4544858	5450	11368

0.5314017366648950	29987	10402887	6016	46870
0.6945929887106360	14028	8362244	1049	11287
0.8401875418620230	12544	7584860	3229	1543
0.4148428731762070	8871	8878999	5315	19711
0.11856385558286900	3545	9643074	20084	32625
0.11625032639916400	3339	9120934	23572	27195
0.5859433962264150	18633	10440793	4847	21487
0.815016591550691	16456	8364682	1447	6023
0.28689065593864900	4564	10196363	9796	12893
0.6255498318161500	13297	7310816	2204	13715
0.8802335888211070	16882	8891420	683	3911
0.06290344003187670	1184	8876435	16809	18468
0.5232017217117790	15802	9392581	5915	22886
0.25827075491327300	6827	10964008	13275	25938
0.7330309665311230	14061	10723601	4126	6116
0.25669663984855700	4068	10720277	10133	13426
0.6648648648648650	17712	11760912	1638	16218
0.2180995052739660	4673	9661149	14892	18614
0.7416532451138290	17038	7835412	1792	10078
0.18181818181818200	911	3472490	3697	4502
0.8285025866345410	11691	9158509	717	4123
0.8295418740941460	13164	8632178	1364	4046
0.6196252667270760	10599	9151428	4890	8123
0.7800654542441690	18830	9145592	2061	8557
0.0	0	5112022	4164	3814
0.5690985619081020	6490	5881922	1928	7900
0.5492184838966050	7256	5879073	2601	9310
0.7838739011821760	3879	5302398	536	1603
0.0	0	8086512	20152	19800

Table 71: nnU-Net test set performance output.

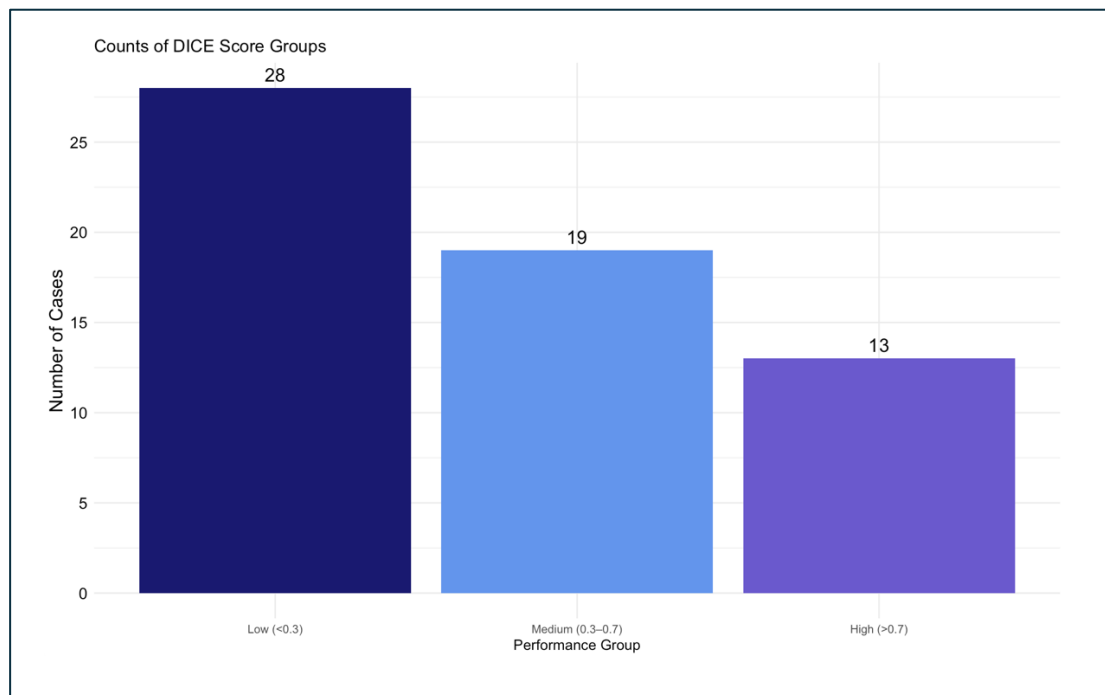


Figure 98: Bar chart of number of cases with low, medium and high DICE scores.

The scores were divided into three DICE score categories (low <0.3 , medium $0.3-0.7$ and high >0.7) to visualise the distribution (figure 85). 28 DICE scores were low (mean DICE score 0.13; standard deviation 0.1), 19 in the medium category (mean DICE score 0.54; standard deviation 0.11) and 13 DICE scores were >0.7 (mean DICE score 13; standard deviation 0.04).

Using the output from nnU-Net, the mean accuracy, precision, recall and specificity. The accuracy was 0.9974 but the values for precision, recall and F1 were 0.37, 0.49 and 0.43, respectively. The specificity was 0.99. The high accuracy and specificity means that the model performs well on segmenting the negative spaces i.e. the non-muscle areas and where it should not segment. However, the low precision, recall and F1 means that the model struggles to segment the defined areas of muscle.

There was a positive linear correlation between the precision and recall, however there is evidence of variability as the recall values increase with lower precision (figure 86). A positive correlation was also noted between DICE scores and precision which is, as expected, since better predicted segmentation results in a higher DICE score and, therefore, better precision (figure 87). The correlation between DICE scores and recall was also strongly positive (figure 88) and it is likely that recall has a significant influence on the

overall DICE score which is intuitive since the better a model is able to learn and adjust, the better it is likely to perform.

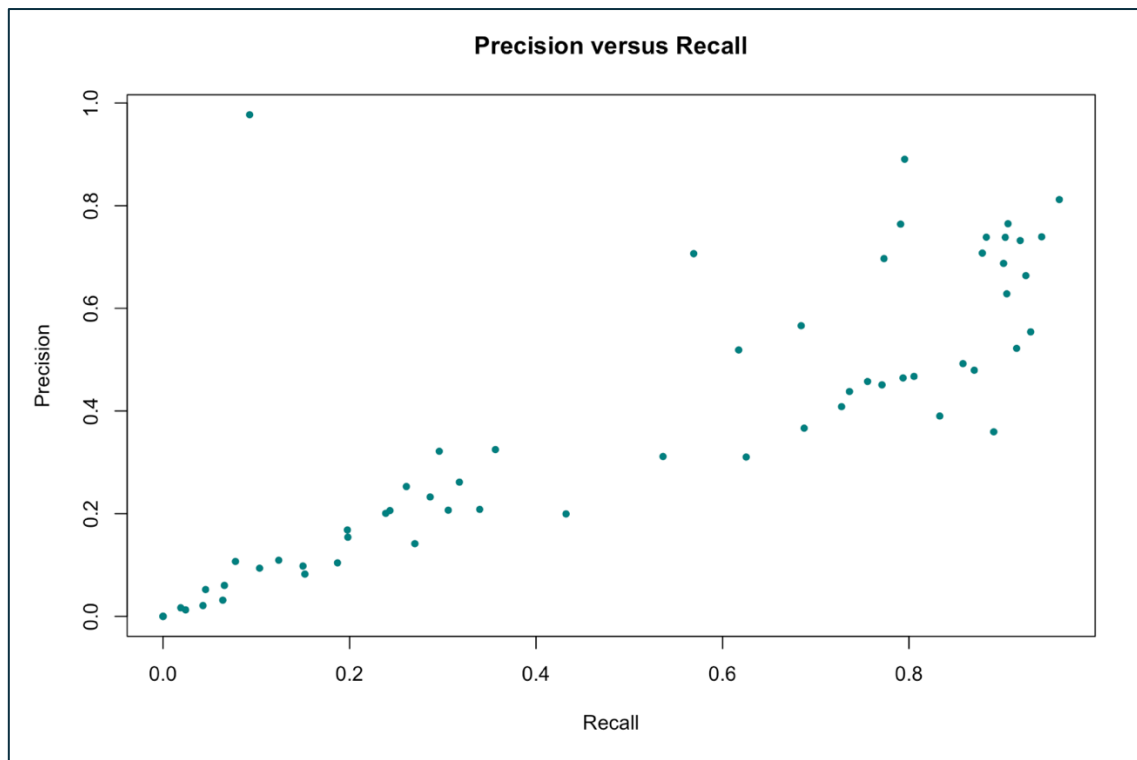


Figure 99: Scatterplot demonstrating the relationship between precision (x-axis) and recall (y-axis).

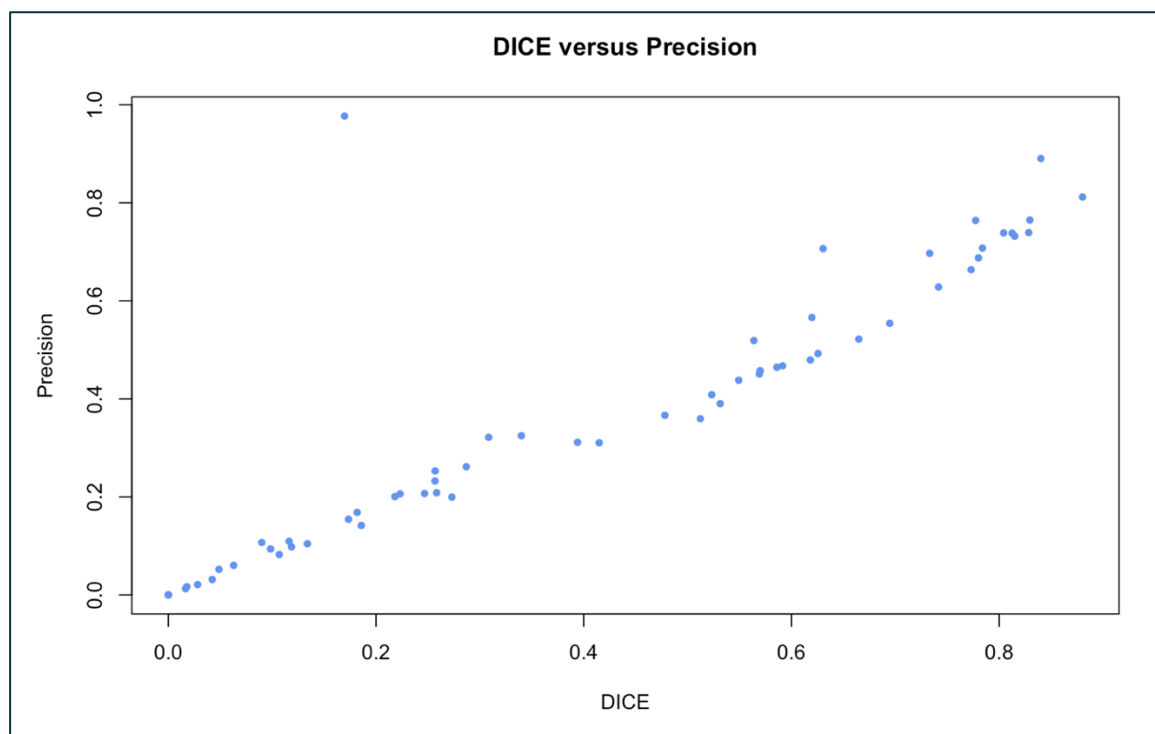


Figure 100: Scatterplot demonstrating the relationship between DICE scores (x-axis) and precision (y-axis).

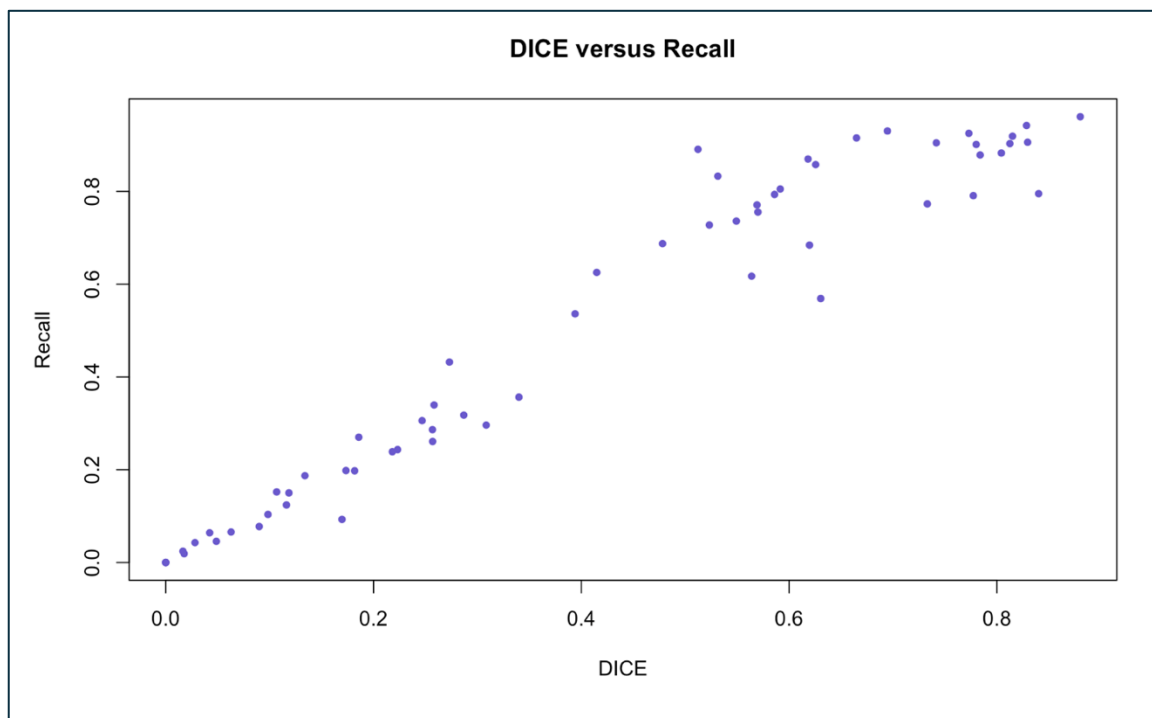


Figure 101: Scatterplot demonstrating the relationship between DICE scores (x-axis) and recall (y-axis).

The low/zero DICE score scans were also reviewed. The model was generally able to perform reasonably on very straightforward scans without significant variation and similar orientation, as shown in the picture on the left in figure 89. However, in scans where there were other findings or anatomical deviation as shown in the four tile pictures on the left in figure 89, the model completely faltered.

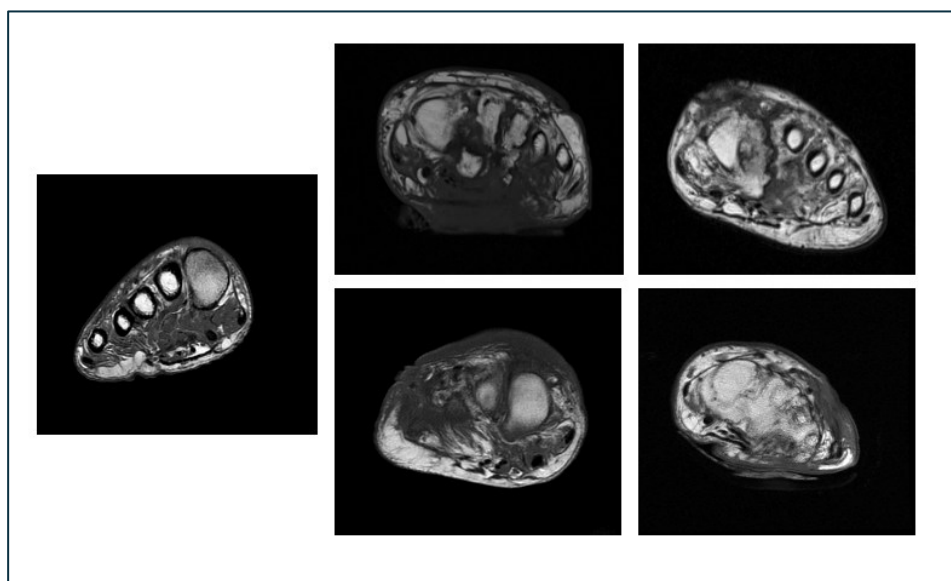


Figure 102: MRI scan variations.

4.4 Discussion

In this chapter, the application of deep learning models and their application in MRI classification and segmentation for sarcopenia was explored. The models have been applied to two- and three-dimensional MRI images of the foot.

ViT has an accuracy of 78.4% with high precision, recall and a more stable learning model. ResNet-101 has an accuracy of 80.7%, however a slightly lower precision and recall, a slightly less smooth learning curve but very performs well. Both are easy to deploy and use without significant pre-processing or training involved, and generate outputs which are easy to interpret.

nnU-Net shows the capability to potentially perform well but is not perfect and, at times, completely failed to segment. There may be a number of reasons for this. Clear anatomical delineation from the surrounding region with well demarcated anatomy can help to segment individual muscles. Segmenting more slices within each MRI can also potentially offer variability where the anatomy is more defined or demarcated.

Another aspect to consider is that other imaging modalities, such as CTs have fixed radiodensity measurements i.e. Hounsfield units. This is not the case with MRI and signal intensities can change. It is possible that a fixed two-dimensional jpeg picture of the MRI with fixed grayscale values replicates the same concept, which is why ViT and ResNet potentially performed well.

It is worth noting that the annotating of images for ViT and the versions of ResNet cannot be compared to the methodology for scan segmentation nor can the final result. One is for classification of severity, the other for segmenting a defined area. In order to proceed further with segmentations and their use, the model must first be able to define the area correctly. Whilst the significant variability in the DICE scores from the ground truth with variations in MRI scans impedes clinical applicability in patients with DFD, past or present, it may have some use in patients with relatively normal anatomy.

Furthermore, the dataset was exponentially increased for classification tasks which allowed for a bigger dataset to be evaluated, an advantage which could not be applied for segmentation tasks. ViT and ResNet-101 have the potential for clinical use but nnU-Net would need further exploration before it can be considered for any real clinical applicability.

Chapter 5: Summary of patient metadata

This chapter summarises the associated metadata described in chapter two. This is relevant prior to the discussion in chapter 6. It includes patient characteristics, diabetes related complications, presence of peripheral arterial disease and interventions. The values were calculated using R's **DescTools** package.

5.1 Patient characteristics

The patient characteristics are summarised in tables 72 and 73. The vast majority of patients included were Caucasian (54%) and male (72%) over the age of 60 which follows the national representative trend of patients with DFD [45]. 27% of the cohort were female. The mean age was 65 years. Other ethnicities including Afro-Caribbean/Black, Asian and Middle Eastern represented 13-14% of the overall cohort each respectively. 62% of all included patients were alive at 12 months from the time of their last scan. The mean patient weight was 85.1 kilograms. There was a fairly equal distribution of smokers and non-smokers, where the data was available with 28% of the patients who had never smoked, 21.8% who were current smokers and 23.6% were former smokers. Data regarding smoking status was not available for a fifth of the patients.

Patient characteristics	Category	Number (Percentage)
Sex	Male	280 (72.5%)
	Female	106 (27.5%)
Ethnicity	Afro-Caribbean/Black	57 (14.7%)
	Asian	54 (14%)
	Caucasian	193 (54%)
	Middle Eastern	13 (13%)
	Other	69 (17.9%)
Living status	Alive	241 (62.4%)
	Dead	145 (37.6%)
Smoking status	Never smoked	110 (28.5%)
	Current smoker	84 (21.8%)

	Former smoker	91 (23.6%)
	Not available	101 (21.8%)

Table 72: Patient characteristics (Table A).

Patient Characteristics	Mean (+/-SD) [95% CI]	Range
Age	65.2 years (13.37) [63.86-66.53]	25-94 years
Weight	85.10Kg (62) [78.37-91.83]	38.7-183Kg

Table 73: Patient characteristics (Table B). Kg – kilograms.

5.2 Diabetes related and other co-morbidities

These are summarised in table 74 below.

5.2.1 Diabetes-related

There was a predominant prevalence of patients with type 2 diabetes (85.4%) whilst only 10% of the included patients had type 1 diabetes and 4.6% were not classified as diabetic. Almost 86% had pre-existing neuropathy and 56% had retinopathy at the time of their first MRI scan.

The extent of nephropathy was also reviewed - 29.5% did not have any nephropathy/chronic kidney disease, 12.2% had CKD stage 1, 8.6% had CKD stage 2, 18.4% had CKD stage 3, 8.6% had CKD 4 and 0.5% had CKD stage 5. 16.7% of all included patients were dialysis dependent and 5.4% had a renal transplant.

The management of diabetes included dietary changes (4.5%), single oral anti-hyperglycaemic agent (18.3%), multiple oral anti-hyperglycaemic agents (23.2%), combination of oral anti-hyperglycaemic and insulin (24.46%) and insulin only (19.8%).

5.2.2 Cardiovascular disease

85.5% of all patients had pre-existing cardiovascular disease. The most common cardiovascular conditions included hypertension (75.6%), hypercholesterolaemia (44%) and coronary artery disease/ischaemic heart disease/myocardial infarction (54%). 20.5% had atrial fibrillation, 14.8% had heart failure and 10.1% had a history of coronary artery bypass grafting. 0.77% also had a history of endocarditis.

5.2.3 Cerebrovascular disease

Almost one-fifth of all patients had a history of cerebrovascular disease (18.3%). This included transient ischaemic attacks (5.6%), stroke without residual neurological deficit (6.3%) and stroke with residual neurological deficit (6.3%).

5.2.4 Respiratory disease

8% had a history of asthma, 9.6% had emphysema or chronic obstructive pulmonary disease, 2.6% had obstructive sleep apnoea, 2.6% had a history of tuberculosis. Interestingly, 57 patients (13.34%) also had a history of COVID-19.

5.2.5 Other medical and social conditions

Other social and medical aspects were also reviewed - 10.5% of all patients had a history of excessive alcohol consumption, 9% had a history of liver disease, 2.5% had a history of pancreatitis. 4.4% had a history of gout. 7% had a history of deep venous thrombosis, 6% had a history of chronic venous insufficiency which was treated. 5.6% had a known mental health condition.

Diabetes related	Category	Number (Percentage)
Neuropathy	No neuropathy	51 (13.4%)
	Pre-existing neuropathy	327 (85.6%)
	Developed neuropathy later	4 (1.0%)

Retinopathy	Retinopathy present	204 (55.9%)
	No Retinopathy	161 (44.1%)
Nephropathy	No CKD	109 (29.5%)
	CKD 1	45 (12.2%)
	CKD 2	32 (8.6%)
	CKD 3	68 (18.4%)
	CKD 4	32 (8.6%)
	CKD 5	2 (0.5%)
	Peritoneal Dialysis	12 (3.2%)
	Dialysis	50 (13.5%)
	Renal Transplant	20 (5.4%)
Cardiovascular disease	Absence of cardiovascular disease	56 (14.5%)
	Presence of cardiovascular disease	330 (85.5%)
	Hypertension	292 (75.6%)
	Ischaemic Heart Disease	68 (17.6%)
	Heart Failure	52 (13.5%)
	Atrial Fibrillation	79 (20.5%)
	Coronary artery disease/PCI	55 (14.2%)
	Left ventricular failure	5 (1.29%)
	Myocardial Infarction	86 (22.3%)
	Hypercholesterolaemia	170 (44%)
	Coronary artery bypass graft	39 (10.1%)
	Endocarditis	3 (0.77%)
Cerebrovascular disease	No cerebrovascular disease	309 (81.7%)
	Transient Ischaemic Attack	21 (5.6%)
	Stroke without neurological deficit	24 (6.3%)
	Stroke with neurological deficit	24 (6.3%)
Use of anti-hyperglycaemic	None	29 (7.67%)
	Diet controlled	17 (4.5%)
	Single oral anti-hyperglycaemic	69 (18.3%)
	Multiple oral anti-hyperglycaemics	88 (23.3%)
	Oral anti-hyperglycaemic and insulin	100 (26.46%)

	Insulin only	75 (19.8%)
--	--------------	------------

Table 74: Patient characteristics - diabetes related. CKD – chronic kidney disease, PCI – percutaneous coronary intervention.

5.3 Peripheral arterial disease

The findings pertaining to peripheral arterial disease is summarised below in table 75.

Thirty-nine percent of all patients had palpable pedal pulses on clinical examination and almost sixty percent had signals on the hand-held doppler. 36.77% had arterial disease confirmed on arterial duplex or diagnostic digital subtraction angiography. The most common distribution of disease was multilevel (25%) followed by crural (15%) and femoral (7.69%). Almost a quarter had endovascular intervention followed by hybrid intervention (5.8%) and open (2.35%). The vast majority of patients required multiple interventions.

Peripheral Arterial Disease	Category	Number (Percentage)
Clinical findings	Palpable pedal pulses	146 (39%)
	Signals on hand-held doppler	224 (59.9%)
	No pulses and no signals	4 (1.1%)
Level of arterial disease	No	270 (63.23%)
	Yes	157 (36.77%)
	- Iliac	2 (0.46%)
	- CFA/PFA/SFA	19 (7.69%)
	- Popliteal	2 (0.46%)
	- Crural	64 (15%)
	- Multilevel	70 (28%)
Intervention	No	292 (68.4%)
	Yes	135 (31.6%)
	- Endovascular	100 (23.4%)
	- Open	10 (2.3%)

	- Hybrid	25 (5.85%)
Number of interventions where applicable	Single Intervention	44 (32% of all interventions)
	Multiple Interventions	97 (72% of all interventions)

Table 75: Patient characteristics - peripheral arterial disease. CFA – common femoral artery, PFA – profunda femoris artery, SFA – superficial femoral artery.

5.4 Biochemical markers

The mean haemoglobin was 115 grams/litre but ranged from 75-176 grams/litre. This may be due to the chronicity of the disease process however, the mode was 94 grams/litre which is a lower value and in keeping with the discussion in chapter 2 – section eleven, sub-section 2 (2.11.2; page 82) pertaining to the complex interplay between haemoglobin levels and diabetes. The mean C-reactive protein was 52 which was raised. This is as expected as a majority of the scans would have been done to rule out covert and indolent infections. The mean albumin was 30 grams/litre, which was low, as expected in the patient cohort. The average glycated haemoglobin value was raised at 63.79mmol/mol. This is reflective of poor glycaemic control which is a contributing factor for DFD as discussed in the earlier chapters.

Biochemistry	Mean (+/-SD) [95% CI]	Range
Haemoglobin (Hb) [g/L]	115.65 (20.97) [112.9-118.36]	75-176
C-Reactive Protein (CRP)	52.316 (66.7) [44.8-59.8]	0.4-341
Albumin [g/L]	30 (7.8) [29.25-31.12]	4.6-46
HbA1c [mmol/mol]	63.79 (62) [78.37-91.83]	18-199

Table 76: Patient characteristics - biochemical markers.

5.5 MRI related and outcomes

79% of all patients had an infected ulcer at the time of their MRI which is intuitive as a vast majority of these scans would have been done when there was suspicion of osteomyelitis or collections. Using the signals intensities of the muscle and the plantar fat as discussed in chapter 4, section three, sub-section 3.2 (4.3.3.2), the pseudo fat fraction was calculated by dividing the plantar fat signal intensity by the muscle signal intensity and the fat signal intensity values. Fat fraction is defined as the signal arising from fat protons divided by the sum of the signals from fat and water protons using the following equation [183]:

$$\text{Fat percentage} = \frac{\text{Signal intensity of plantar fat}}{\text{signal intensities of the plantar fat} + \text{muscle (water)}}$$

The mean fat fraction value was 0.625 with a standard deviation of 0.068 and a range of 0.0013 to 0.833.

MRI Related	Category	Number (Percentage)
Presence of infection	Yes	299 (79%)
	No	79 (21%)
Pseudo fat fraction/Fat percentage	0.624913842 (0.068)	Range 0.0013-0.8327

Table 77: Patient characteristics - MRI/episode related.

40% of all patients had previous amputations. This included ipsilateral and contralateral amputations as shown in table 78. 23.4% previously had ipsilateral digit amputations and 4.7% had digit amputations not otherwise specified (from older records). 9.7% had previous contralateral amputations, 5.6% had ipsilateral transmetatarsal amputations, 4.45% had contralateral amputations, 4.9% had a contralateral below knee amputation and 3.6% had a previous contralateral above knee amputation.

The mean time from disease onset to the final outcome was approximately 16 months (range 0-156 months). The time between ipsilateral disease and contralateral disease was, on average 35.3 months.

Outcomes	Category	Number (Percentage)
Previous amputation	No	258 (60%)
	Yes	157 (40%)
	- Digit	20 (4.7%)
	- Digit Ipsilateral	58 (23.4%)
	- Digit Contralateral	24 (9.7%)
	- TMT Ipsilateral	14 (5.67%)
	- TMT Contralateral	11 (4.45%)
	- BKA Contralateral	21 (4.9%)
	- AKA Contralateral	9 (3.6%)
Time between onset and final outcome	16.4 months (23.6)	Range (0-156)
Time between ipsilateral and contralateral disease	35.3 months (37.7)	Range (1-156)

Table 78: Patient outcomes. TMT – transmetatarsal, BKA – Below knee amputation, AKA – Above knee amputation.

Chapter 6: Risk prediction

Risk prediction has previously been discussed in chapter one, section seven (1.7). A risk prediction model is often referred to as a prognostic model and aims to “accurately predict the risk of an outcome in the future by utilising the values of levels of multiple individual characteristics [53]. This chapter discusses the applied use of the sarcopenia measurements (classification and pseudo fat fractions) in combination with patient metadata from chapter five to explore the possibility of attempting to create a risk prediction model.

To begin with, 23 potential factors were considered however each of these had multiple sub-categories with missing values. Furthermore, the subclassifications of amputations with minimal numbers in each category added another layer of complexity to the statistical analysis. Therefore, a step-wise approach was taken. The amputation sub-groups were clustered into a binary outcome group i.e. amputation and no amputation.

6.1 Sarcopenia measurements and outcomes

6.1.1 Sarcopenia and all amputations as an outcome

Given the focus of this thesis was to explore sarcopenia as a surrogate marker/risk factor for lower limb amputation, a logistic regression model was used. Missing data was addressed using imputation. R packages ‘readxl’, ‘dplyr’ and ‘ggplot2’ were used. The complete and clean code is available in appendix 8.

Figure 102 below demonstrates the distribution of outcomes i.e. amputation versus no amputation according to the sarcopenia severity groups. There was a slightly higher cohort of patients with amputations in the severe sarcopenia group. The odds ratio for the outcome ‘amputation’ with the mild sarcopenia group was 0.75 [95% CI 0.54-1.03, p-value = 0.075), moderate sarcopenia was 1.01 [95% CI 0.61-1.68, p-value = 0.97] and severe sarcopenia was 1.21 [95% CI 0.66-2.21, p-value = 0.538]. Whilst there was a minimal increase in the odds ratio in the severe sarcopenia group, it was not statistically significant.

Overall, there was no statistical significance between the extent of sarcopenia and the outcome in terms of ‘all amputations’.

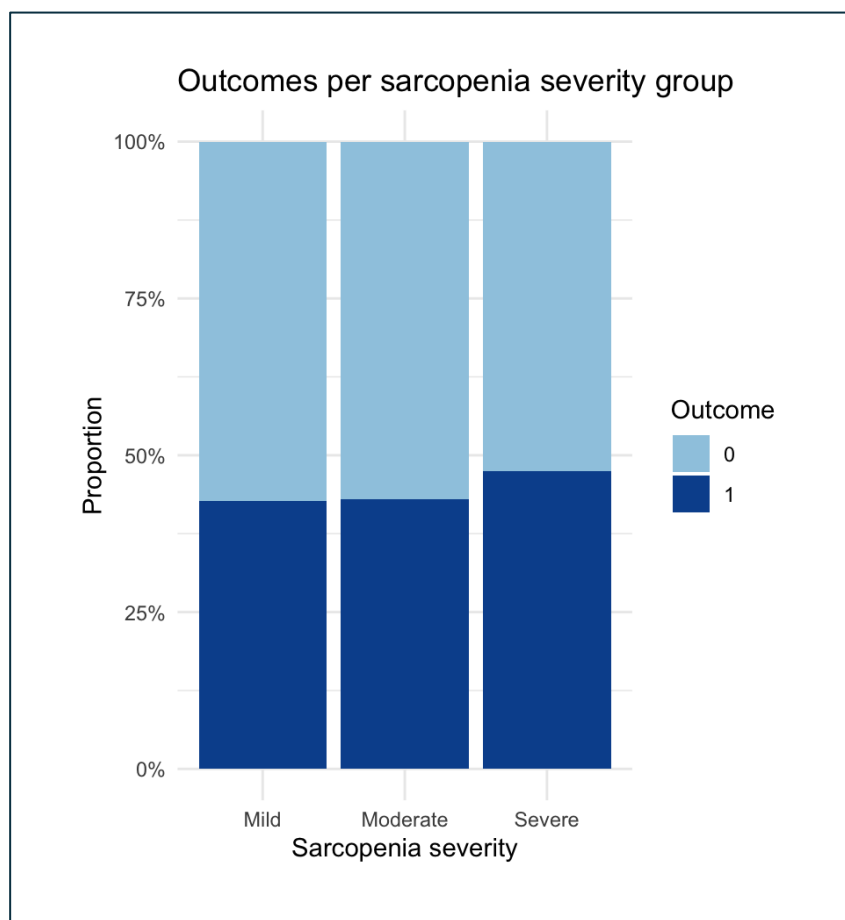


Figure 103: Outcomes per sarcopenia severity group. 0 = no amputation, 1= amputation.

6.1.2 Composite outcomes: sarcopenia severity and major lower limb amputations plus death

Further composite analysis was undertaken to explore if the sarcopenia severity is a risk factor for major lower limb amputation and death combined as an outcome. 176 were classified in the no amputation group and 183 were added to the (MLLA and death) group. The odds ratio for ‘MLLA and death combined’ with the mild sarcopenia group was 1.04 [95% CI 0.76-1.42, p-value = 0.818], moderate sarcopenia was 0.948 [95% CI 0.59-1.53, p-value = 0.827] and severe sarcopenia was 1.094 [95% CI 0.615-1.95, p-value = 0.76]. This is also illustrated in figure 103. The odds were slightly higher in the mild and severe group but neither were statistically significant.

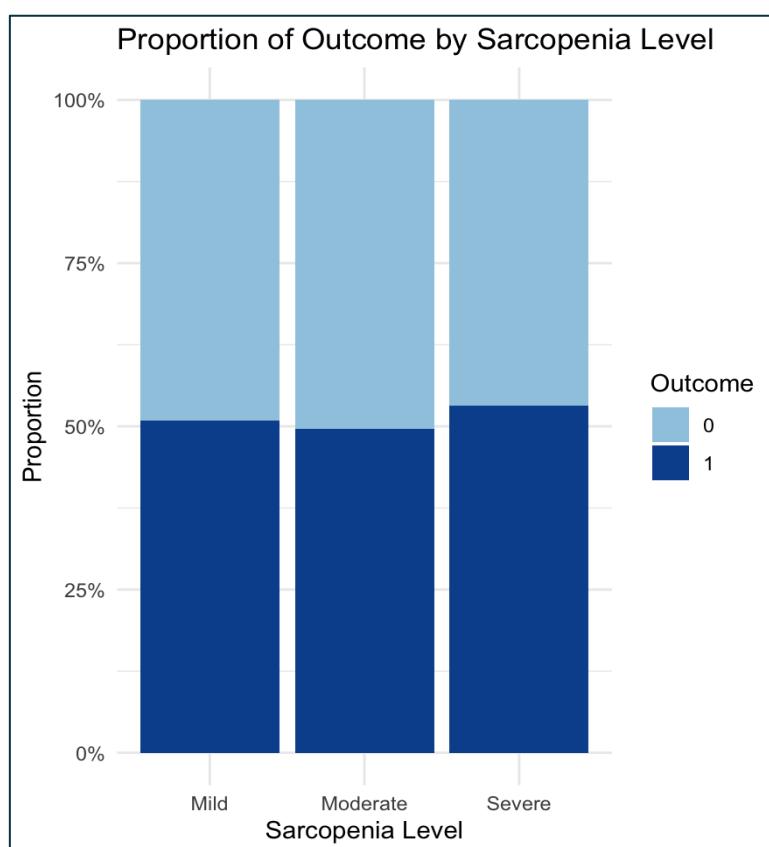


Figure 104: Outcomes per sarcopenia severity group. 0 = no amputation, 1= major lower limb amputation and death combined.

6.1.3 Composite outcomes: sarcopenia severity and all amputations and death

This was also evaluated by combining all amputations and death as a singular outcome.

There were 176 individuals in the no amputation group and 242 in the all-amputation and death group. The odds ratio for 'all amputations and death combined' with the mild sarcopenia group was 1.368 [95% CI 1.020-1.845, p-value = 0.0377], moderate sarcopenia was 1.22 [95% CI 0.76-1.96, p-value = 0.414] and severe sarcopenia was 0.974 [95% CI 0.559-1.7, p-value = 0.9] (figure 104).

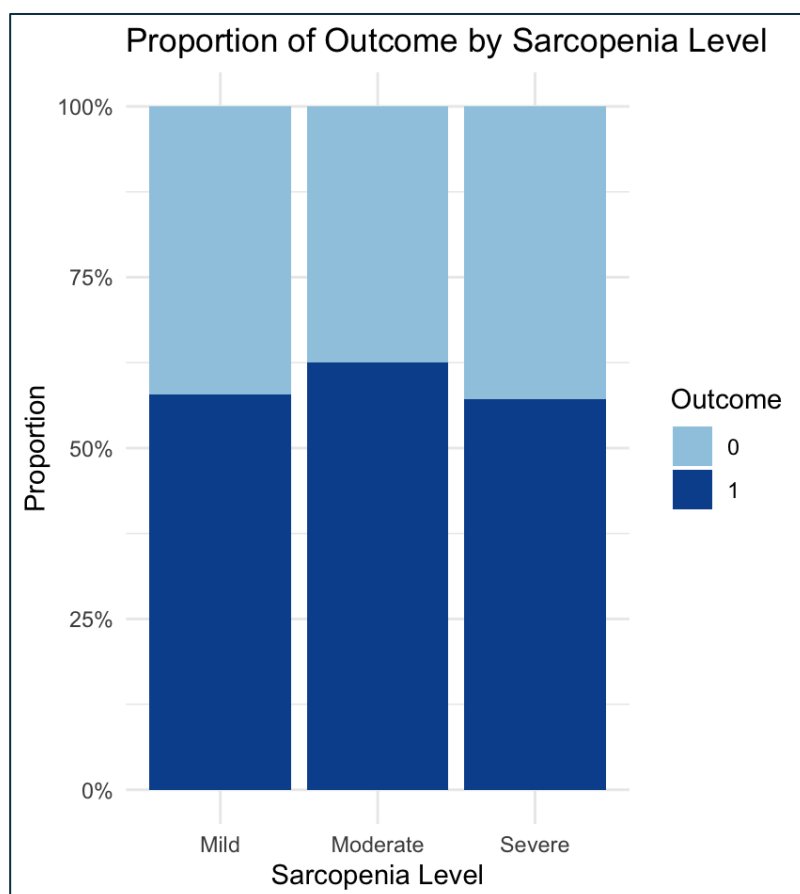


Figure 105: Outcomes per sarcopenia severity group. 0 = no amputation, 1= all amputation and death combined.

It is interesting to note that a greater proportion of patients with mild, moderate and severe sarcopenia were in the ‘major lower limb amputation and death’ and the ‘all amputation and death’ group when compared to all amputations versus no amputation group

6.1.4 Fat percentage values and all amputations as an outcome

The same R packages were used to explore the relationship between the fat percentage/pseudo fat fraction and the outcomes (appendix 9). The odds ratio for all amputation (major and minor) as an outcome was 1.09 [95% CI 0.04-42.2, p-value = 0.96]. Given the very wide confidence interval and high p-value, it is evident that the pseudo fat fraction is not a significant predictor for lower limb amputations as a whole.

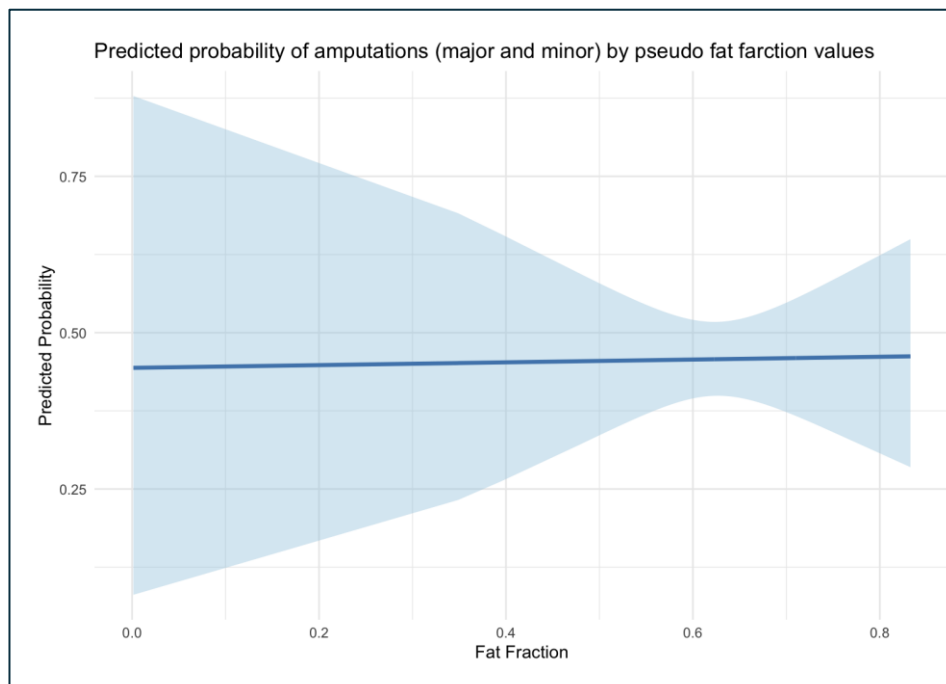


Figure 106: Logistic regression (Pseudo fat fraction and predicted probability of all amputations).

The pseudo fat fraction values were divided into tertiles to assess for any variance however, no change in the odds ratio was noted and the data was visualised to explore this by plotting the pseudo fat fraction values against the outcomes (figure 106). This demonstrates a concentration of values between 0.45 and 0.75 where almost all of the outcomes i.e. no amputation and all amputations have occurred, implying a skewed representation of data.

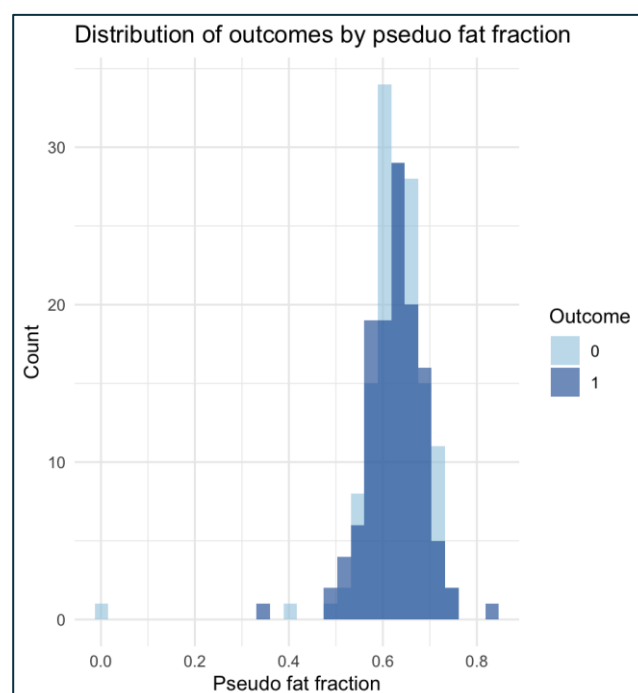


Figure 107: Visual chart demonstrating the pseudo fat fraction values and the outcomes. 0 = no amputation, 1 = all amputations.

6.1.5 Composite outcomes: pseudo fat fraction values and major lower limb amputations plus death

This was further explored by evaluating for any differences by combining outcomes with further composite analysis. The odds ratio for MLLA and death as an outcome was 40.16 [95% CI 0.9489-2734.6, p-value = 0.98].

Given this extremely wide confidence interval and lack of convergence due to instability in the model because of the distribution, Firth's logistic regression (penalised logistic regression which helps to reduce bias and stabilise the model) was used instead. The odds ratio was calculated from the likelihood ratio test values generated by Firth's logistic regression. The odds ratio was 32.2 [95% CI -0.16 – 7.71, p-value = 0.062]. Whilst not entirely statistically significant, there is a clear trend of the risk of MLLA or death increasing with increasing pseudo fat fraction values as visualised in figure 107.

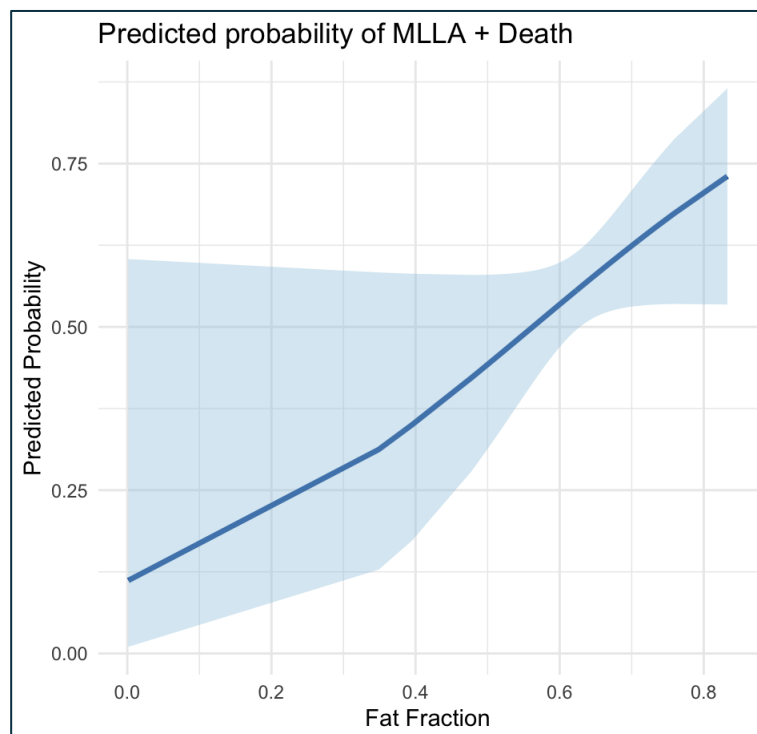


Figure 108: Logistic regression. Pseudo fat fraction and the probability of major lower limb amputation and death as a combined outcome.

6.1.5 Composite outcomes: pseudo fat fraction values and all amputations and death

The odds ratio for all amputation and death combined as an outcome was 0.35 [95% CI 0.01 – 11.32, p-value = 0.567]. The rationale for this output is not entirely clear. This was re-checked and visualised but could be due to the distribution of data or the number of individuals having minor amputations as being considered a “better outcome” as the expectation would be for the graph below (figure 108) to follow a similar pattern when comparing the combined MLLA and death cohort.

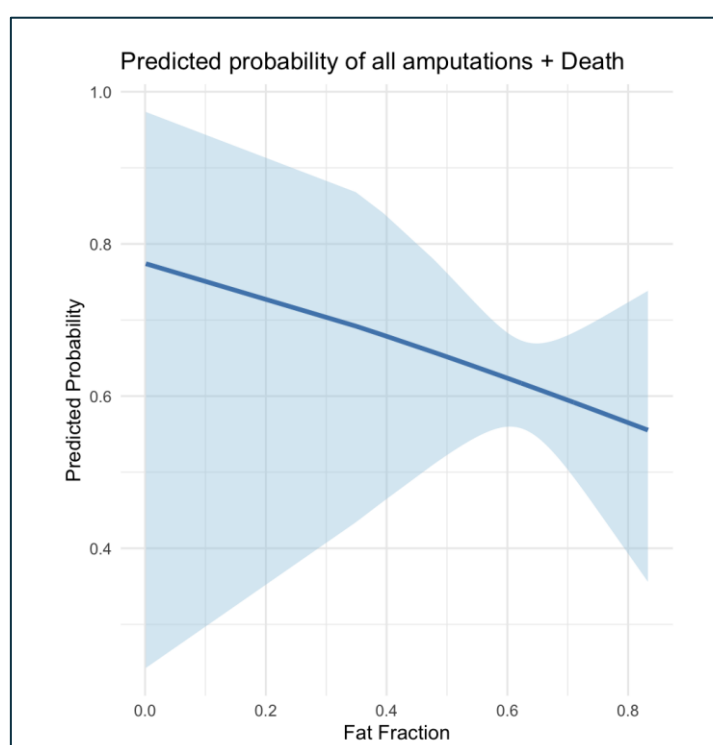


Figure 109: Logistic regression. Pseudo fat fraction and the probability of all amputations and death as a combined outcome.

6.2 Identifying risk factors

The next step was to identify factors/predictors from the gathered metadata which would have an influence on all amputations as an outcome. This included categorical and numerical variables. Using R, these were labelled as factors. The initial approach taken to manage missing data was to drop the rows with any missing values which significantly reduced the dataset. This was addressed with multivariate imputation by chained equations (MICE).

MICE is an iterative method for imputing missing values in a recursive fashion [184]. This is done by initially filling the missing values with estimates to create a “complete dataset”. It then uses chained equations for each variable with missing data and fits a regression model using all other variables in the dataset as the predictors as a form of predictive modelling [184]. This is repeated multiple time in a recursive fashion until the model stabilises. The advantage of using MICE is in its capability of handling different types of variables including numerical and categorical and its adaptability. It also uses multiple imputations such that a number of datasets are created as opposed to a single filled in dataset [184]. This is also available to use in R and was used with predictive mean matching.

Using standard logistic regression can lead to overfitting and overestimation of how well a model may be performing [185]. It also helps to address the issue of collinearity by reducing the co-efficient or multiple variables which might be correlated. This can be addressed with least absolute shrinkage and selection operator (LASSO) which is available in R with the ‘**glmnet**’ package. LASSO regression aims to identify variables which lead to a model with minimal prediction error. It is done by shrinking the co-efficient and setting certain co-efficients to zero thereby removing non-relevant predictors [185].

There are two methods for regularisation – L1 (LASSO) and L2 (Ridge). L1 adds the sum of the co-efficients as a penalty terms to the loss function whilst L2 add the squared sum of coefficients as the penalty term to the loss function [186]. L1 aims to simplify the model whilst L2 retains all the factors but these may not have a significant impact [186].

Three different methods were used to help identify features of importance to the outcome – LASSO, Ridge regression and Random Forest.

6.2.1 Lasso regression for selecting salient predictors for LLAs

6.2.1.1 Method

The following R packages were used - ‘**glmnet**’, ‘**readxl**’, ‘**dplyr**’, ‘**tidyr**’ and ‘**mice**’. The complete code is available in appendix 10. The set seed was set to 123 for reproducibility.

The alpha value for lasso regression was set to 1 (pure L1 regularisation). Cross-validation was set to 1. A cross-validation plot was also coded for.

6.2.1.2 Results

The LASSO co-efficient was -0.2939912 for the intercept i.e. the baseline. This means that none of the features were selected to have significant importance to the outcome. This should be interpreted with some caveat as LASSO is highly selective and predictors often need to be sufficiently strong to withstand the shrinkage.

Figure 109 demonstrates the cross validation of error with LASSO. The x-axis shows the lambda value/ penalty parameter on a log scale. The y-axis is the binomial deviance which is representative of the model fit or goodness-of-fit. The red line is the mean cross validation error and the vertical bars around it represent the error/standard deviation and reflects the error across each cross-validation fold.

On the left, the penalty is strong, and the binomial deviance is low. As the lambda penalty value relaxes across to the left of the figure, the binomial deviation increases with more error (vertical bars) indicating poor prediction. The minimum stable binomial deviance is approximately at the $-\log(\lambda)$ value of four.

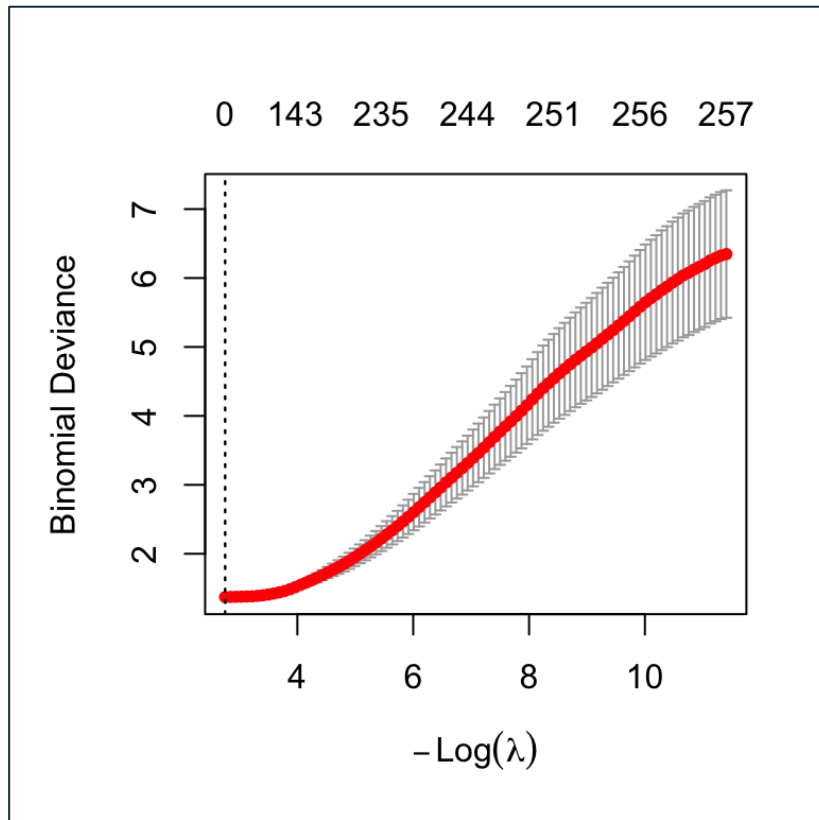


Figure 110: Cross validation error with LASSO.

6.2.2 Ridge regression for selecting salient predictors for LLAs

6.2.2.1 Method

The following R packages were used - **'glmnet'**, **'readxl'**, **'dplyr'**, **'tidyr'** and **'mice'**. The complete code is available in appendix 10. The set seed was set to 123 for reproducibility. The alpha value for lasso regression was set to 0 (pure L2 regularisation). Cross-validation was set to 10 again.

6.2.1.2 Results

The ridge co-efficients were in the following order: intercept i.e. baseline [-0.28883171], age [0.09886910], male sex [0.09867872], HbA1c [0.09856397] and weight [0.09837686]. Ridge regression aims to keep as many predictors as possible and distributes the weight across these. These co-efficients around 0.09 indicate moderate predictability.

6.2.3 Random Forest for selecting salient predictors for LLAs

6.2.3.1 Method

Random Forest has already been discussed throughout chapter 3 and will not be discussed again. The following R packages were used - ‘**randomForest**’, ‘**readxl**’, ‘**dplyr**’, ‘**tidyr**’ and ‘**mice**’. The complete code is available in appendix 10. The set seed was set to 123 for reproducibility and the number of decision trees was set to 500. The R code for this is available in appendix 11.

6.2.3.2 Results

Table 79 and figure 105 demonstrate the Random Forest model to highlight the following variables to have strong predictive power in the amputation outcome: age, renal status, CRP, haemoglobin, albumin and HbA1c as well as the pseudo fat fraction. Moderate predictors for amputation included ethnicity and smoking status.

The presence of pseudo fat fraction values was again not expected as it had previously demonstrated no correlation between amputation its score and therefore should be interpreted again with some reservation. The ROC-AUC value was 1 (figure 106).

Interestingly, whilst age was a predictor in ridge, it was not a strong predictor using RaF. Certain factors did converge between ridge regression and RaF including age and HbA1c.

Variable	Class 0 (no amputation)	Class 1 (amputation)	Mean decrease accuracy	Mean decrease Gini
Age	-1.9309643931	-3.8554255	-3.951772746	17.240051
Sex	1.2287461027	-0.1266487	0.637107841	2.504992
HbA1c	-1.1484638998	-2.7487629	-2.669655808	17.351085
Ethnicity	-0.5630633408	-0.8848437	-0.898233795	10.663008

Smoking	-0.5905041828	0.1732668	-0.261568867	8.588566
Neuropathy	-0.8663725767	0.8847048	-0.005401848	2.297235
PAD	1.6786917899	3.1031122	3.274140165	4.321387
Retinopathy	-0.0003178781	-1.1484026	-0.819227160	2.362012
CVA	-0.5547268334	-2.0521607	-1.804561353	4.475003
Renal	0.6949859552	0.5481978	0.956797104	17.295274
CVD	1.4472159306	-1.5461899	0.132615506	1.469279
CRP	-0.4046732877	-1.5563578	-1.420364182	17.264636
Infection	-0.1030273134	-1.6758378	-1.145739777	2.401969
Haemoglobin	-3.7803089894	-2.3402191	-4.250848537	15.855944
Albumin	1.9779114840	0.2577057	1.507271965	16.997006
SS	-1.0082823697	-2.7947824	-2.467555870	4.927611
PFF	-1.0891317681	-1.8410514	-1.944371154	17.287793

Table 79:Random Forest feature importance incorporating patient metadata. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction.

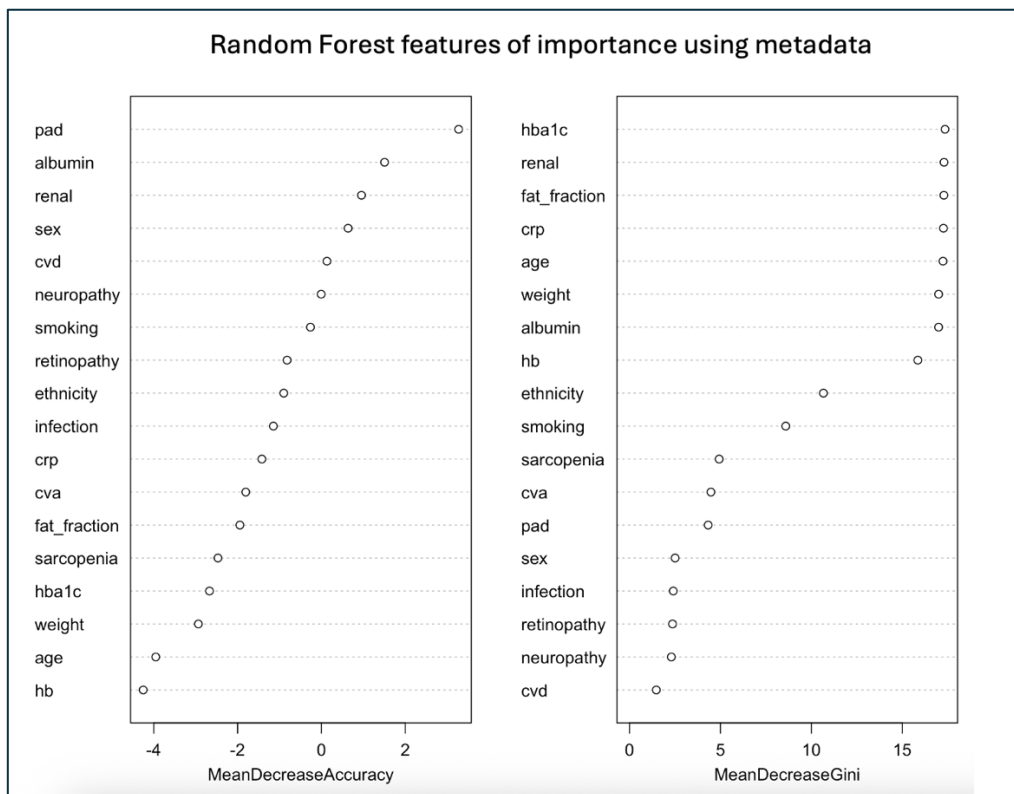


Figure 111: Visual chart of Random Forest feature importance incorporating patient metadata. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction.

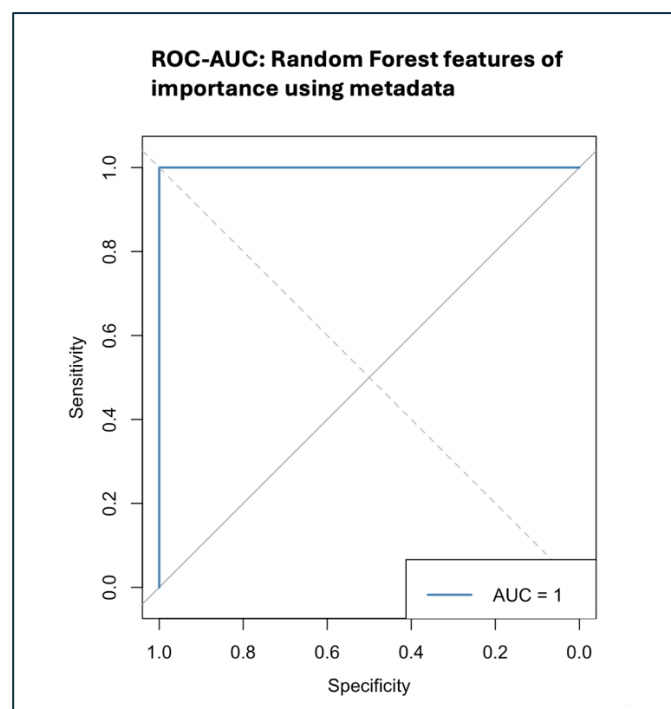


Figure 112: ROC-AUC Random Forest using metadata.

6.2.4 Extreme gradient boosting for selecting salient predictors for LLAs

6.2.4.1 Method

XGB has already been discussed throughout chapter four. The following R packages were used - 'caret', 'readxl', 'dplyr', 'tidyr', 'Matrix', 'xgboost', 'lattice' and 'ggplot'. The complete code is available in appendix 10. The set seed was set to 123 for reproducibility.

6.2.4.2 Results

The model accuracy was 44.6%. The model identified the following as features of importance with the following co-efficients – neuropathy [0.14], cerebrovascular disease [0.08], sex [0.078], ethnicity [0.062], haemoglobin [0.059], peripheral arterial disease [0.057], HbA1c [0.053], weight [0.052] and renal status [0.049]. The ROC-AUC value was 0.52. This is not entirely unexpected given that the dataset was smaller i.e. the number of complete cases was low and XGB does not usually perform well on small datasets.

6.2.5 LASSO and ridge regression for composite outcomes – (MLLA + death) and (all amputations + death)

The same methodology was utilised for composite outcomes. LASSO did not highlight any features for MLLA and death but did select a previous history of transient ischaemic attack to be associated with a higher likelihood of all amputations and death with a co-efficient value of 0.2406578 (from 0 to 1) p-value 0.013.

Ridge regression selected thirteen very weak predictors for MLLA and death as an outcome and these are summarised in table 80.

Ridge Regression Feature	MLLA + Death
Fat fraction	0.0192302477
Ethnicity (Middle Eastern)	-0.0048302822

Renal (CKD 2)	0.0036813526
CVA (stroke with neurological deficit)	-0.0034862668
Ethnicity (Black)	-0.0028087433
Renal (PD)	-0.0027574551
Ethnicity (Caucasian)	-0.0026733189
CVA (stroke no neurological deficit)	0.0024965440
Renal (CKD 4)	-0.0024435744
CVA (TIA)	0.0018187660
Infection (absence)	0.0015691338
Ethnicity (Asian)	-0.0014163090
Sarcopenia (Moderate)	-0.0012452566

Table 80: Ridge regression feature selection. Major lower limb amputation and death combined. CKD – chronic kidney disease, CVA – cerebrovascular accident, PD – peritoneal dialysis, TIA – transient ischaemic attack. The minus sign indicates it decreased the likelihood of the outcome.

Similarly, for all amputations and death as an outcome, fourteen very weak predictors were selected by ridge regression (table 81). Interestingly, CVA appeared in LASSO and Ridge regression models for the outcome of all amputations and death as a weak predictor.

Ridge Regression Feature	All amputations + Death
CVA (TIA)	0.0042935985
Renal (PD)	-0.0026820293
Renal (CKD 4)	-0.0018026829
CVD (Yes)	-0.0017387861
Renal (Dialysis)	0.0013945802
Renal (No CKD)	0.0013285363

No infection	0.0012622147
Renal (CKD 1)	-0.0010974056
Sarcopenia (Moderate)	0.0010859744
Fat fraction	-0.0007369780
Former smoker	0.0006415674
CVA (stroke with neurological deficit)	0.0006002937
Ethnicity (Middle Eastern)	0.0005848630
Signals on HHD	0.0005477987

Table 81: Ridge regression feature selection. All amputation and death combined. CKD – chronic kidney disease, CVA – cerebrovascular accident, CVD – cardiovascular disease, HHD- handheld doppler, PD – peritoneal dialysis, TIA – transient ischaemic attack. The minus sign indicates it decreased the likelihood of the outcome.

6.2.6 Random Forest for composite outcomes – MLLA + Death and All amputations + Death

Figure 112 below demonstrates the feature importance from the Random Forest model for the composite outcomes – MLLA and death on the left and all amputations and death on the left. The ROC-AUC for both models was 1. The fat fraction was highlighted as a feature of importance on the model for MLLA and death with a mean decrease Gini value of 14.15. This was followed by HbA1c and weight as other features of importance (table 82).

In terms of all amputations and death as an outcome, weight was highlighted as a feature of importance with a mean decrease Gini value of 14.15 followed by fat fraction with a mean decrease Gini value of 11.735341 and HbA1c (table 83).

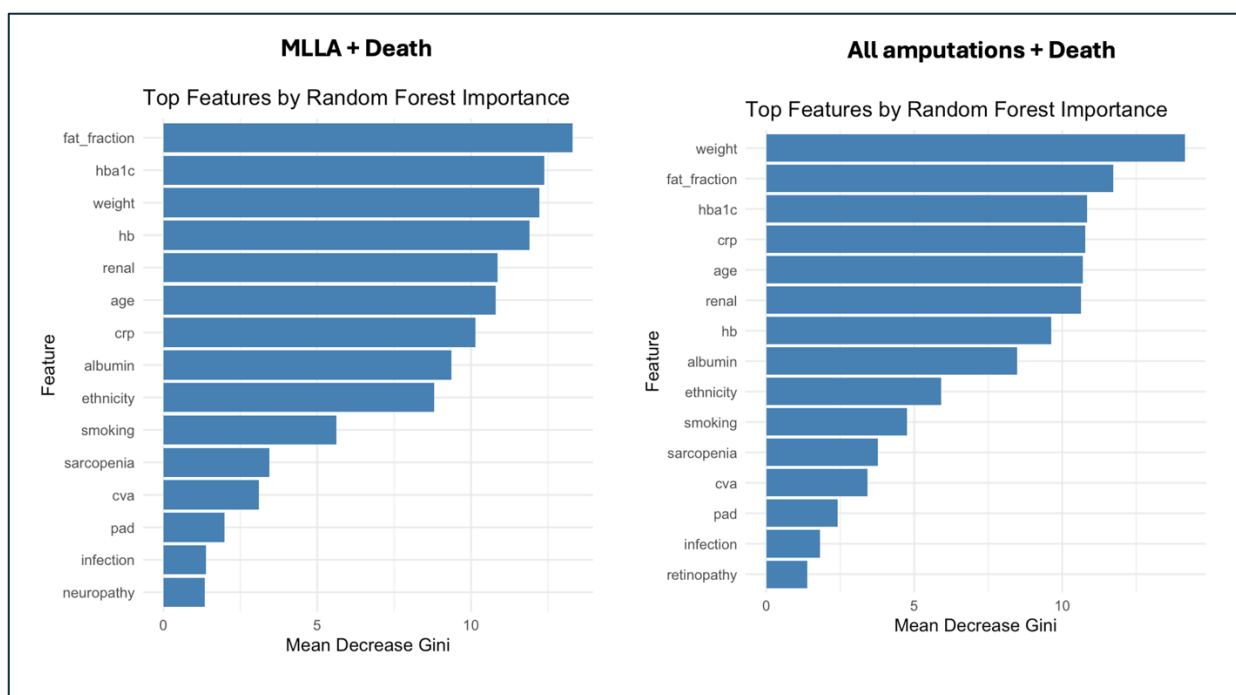


Figure 113: Random Forest feature importance. Visual chart of Random Forest feature importance incorporating patient metadata for composite outcomes. Left chart – Major lower limb amputation and death as the outcome. Right chart – all amputations and death as the outcome. HbA1c – glycated haemoglobin, PAD – peripheral arterial disease, CVA – cerebrovascular disease, CVD – cardiovascular disease, CRP – C-reactive protein, SS – sarcopenia severity, PFF – pseudo fat fraction.

Feature	Mean decrease Gini
Fat fraction	14.151453
HbA1c	11.735341
Weight	10.833869
Hb	10.781567
Renal	10.702256
Age	10.626331
CRP	9.624208
Albumin	8.481727
Ethnicity	5.900787
Smoking	4.752842
Sarcopenia	3.770557
CVA	3.424625
PAD	2.419024

Infection	1.812481
Neuropathy	1.378042

Table 82: RaF feature importance MLLA and Death.HbA1c – glycated haemoglobin, Hb – haemoglobin, CRP- C-reactive protein, CVA – cerebrovascular accident, PAD – peripheral arterial disease.

Feature	Mean decrease Gini
Weight	14.151453
Fat fraction	11.735341
HbA1c	10.833869
CRP	10.781567
Age	10.702256
Renal	10.626331
Hb	9.624208
Albumin	8.481727
Ethnicity	5.900787
Smoking	4.752842
Sarcopenia	3.770557
CVA	3.424625
PAD	2.419024
Infection	1.812481
Retinopathy	1.378042

6.3 Estimating survival

6.3.1 Univariate and Multivariate Cox proportional hazard

The time to the outcomes i.e. amputation versus no amputation were available for a subset of patients. A cox proportional hazard model was used to identify time-to-event predictors with adjustments for the other covariates.

6.3.1 Method

The data was recoded and the variables were converted to factors. The outcome classes were converted to binary outcomes such that 0 was no amputation and 1 was amputation. Using R's '**readxl**', '**survival**' and '**survminer**' packages, the cox model was applied. The clean R code is available in appendix 11.

6.3.2 Results

After cleaning the data, a sample size of 170 patients was available with clear time to event details in months.

Initial univariate cox model analysis was used to identify variables which demonstrated an association with the combined outcome (amputation/death). This identified several variables although none were statistically significant. Severe sarcopenia was a variable with a hazard ratio of 1.79 [95% CI 0.92-3.49, p-value 0.088], age HR 0.998 [95% CI 0.976-1.02, p-value 0.826], male sex HR 1.39 [95% CI 0.74-2.62, p-value = 0.3], HbA1c 0.995 [95% CI 0.985-1.005, p-value = 0.35], weight HR 0.998 [95% CI 0.994-1.003, p-value = 0.42], CRP HR 0.99 [95% CI 0.995-1.002, p-value 0.49], haemoglobin HR 1.01 [95% CI 0.997-1.022, p-value 0.14], albumin HR 0.99 [95% CI 0.96-1.036, p-value 0.94] and fat fraction HR 0.78 [95% CI 0.04-15.02, p-value = 0.866].

Using this information, a multivariate cox proportional hazard model was used. The output included the concordance index or the C-index which is similar to an ROC-AUC. A c-index value of 1 implies perfect prediction whilst a value of 0.5 or below is not dissimilar to random prediction [187]. The C-index value was 0.634 with a standard error of 0.04.

The logrank test compares the “observed number of events with the number expected to occur if two groups have the same underlying true survival function” [188]. The logrank test was 56.05 [p-value <0.0001] which indicates the significance of the model’s predictive value.

The output from the model is summarised in table 80 below. This includes the log hazard ration, the effect size, the standard error (SE) such that a small SE implies better precision and vice versa. CRP was a strong predictor, as expected, with a hazard ratio of 1.004 [p-value = 0.02]. This means that for every unit increase in the CRP, the risk of amputation increases by 0.38%.

Not having neuropathy (HR 0.04, p-value 0.007) or having pre-existing neuropathy (HR 0.03, p-value 0.003) significantly reduced the risk of amputation. It is possible that pre-existing neuropathy was considered as a baseline during the analysis and therefore considered similar to the non-neuropathic group as opposed to the group coded as developing neuropathy later.

Other factors associated with an increase in the hazard ratio included age, retinopathy, moderate and severe sarcopenia and the pseudo fat fraction values but these were not statistically significant.

Variable	Log HR	HR (effect size)	SE (CE)	p-value
Age	0.00471	1.004721	0.009874	0.63333
Sex (Male)	-0.42219	0.655609	0.318566	0.18508
HbA1c	-0.00595	0.994068	0.005643	0.29171
CRP	0.003883	1.003891	0.001671	0.02014
weight	-0.00738	0.992651	0.006069	0.22421
Smoking (Former Smoker)	-0.36181	0.696418	0.380035	0.34108
Smoking (Never Smoked)	-0.09571	0.908732	0.348851	0.78382

Neuropathy (No neuropathy)	-3.18282	0.041469	1.193326	0.00765
Neuropathy (Pre-existing)	-3.48038	0.030796	1.17422	0.00304
PAD (Pedal Pulses)	-0.99236	0.3707	1.085047	0.36041
PAD (Handheld Doppler Signals)	-0.88246	0.413766	1.076371	0.41231
Retinopathy (Yes)	0.111743	1.118226	0.268004	0.67672
Cardiovascular disease (Yes)	-0.37549	0.686952	0.396168	0.34323
Sarcopenia (Moderate)	0.368485	1.445543	0.297542	0.21556
Sarcopenia (Severe)	0.166494	1.181157	0.370109	0.65282
Pseudo fat fraction	0.526237	1.692551	0.33553	0.11679

Table 83: Cox proportional hazard model output and hazard ratios for risk of amputation (minor and major combined). Statistically significant results in bold. HR – Hazard ratio, HbA1c – glycated haemoglobin, CRP – C-reactive protein, PAD- peripheral arterial disease, SE(CE) – standard error of the coefficient/log hazard ratio.

The Kaplan-Meier curve below demonstrates the survival curve when comparing mild, moderate and severe sarcopenia groups. There is evidence of heavy overlap across the three groups with wide confidence intervals (shaded areas). What is evident though is that the survival is almost halved around the 24 month mark.

Furthermore, the log-rank test demonstrates a p-value of 0.56 which is not statistically significant. This is likely due to the smaller dataset which was available after the data had been cleaned which limits the interpretability.

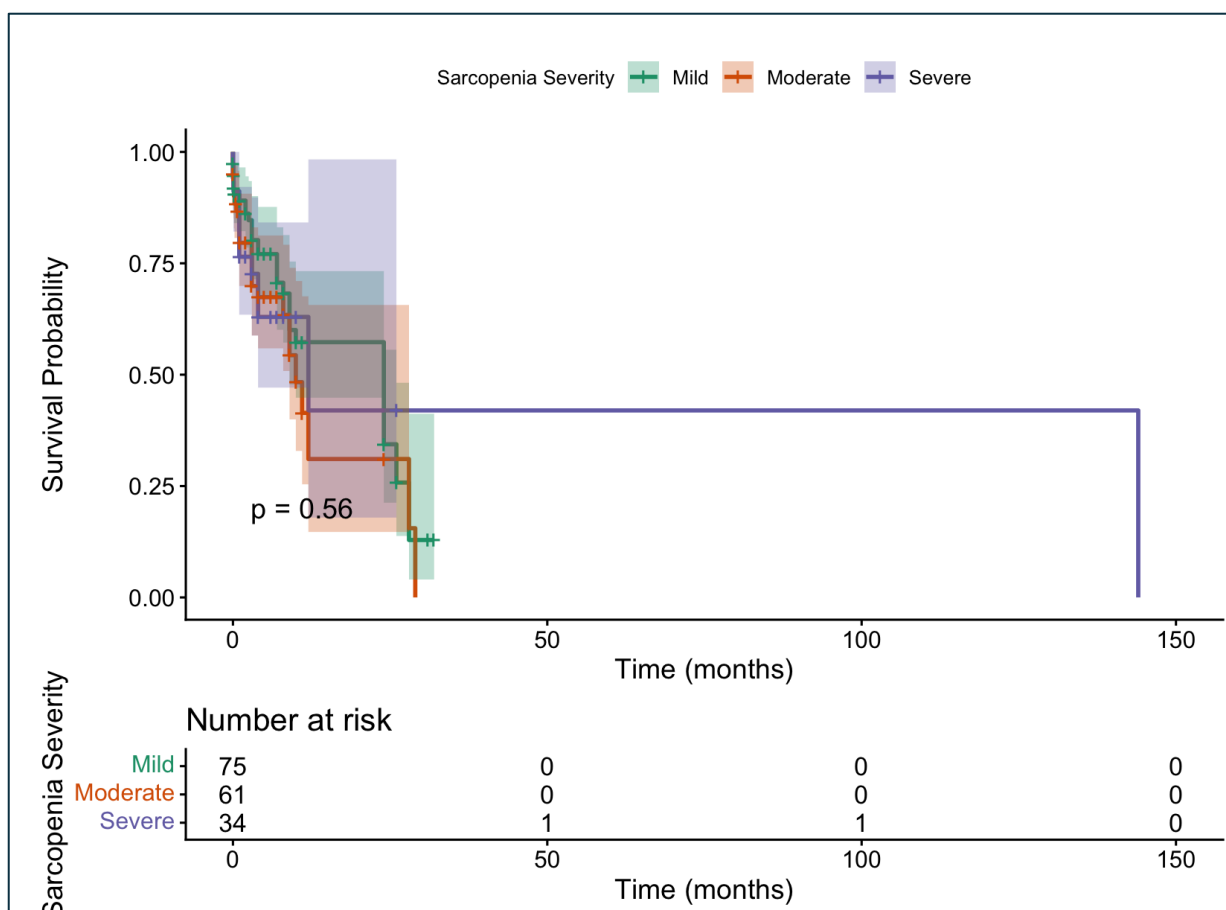


Figure 114: Kaplan – Meier curve. Sarcopenia severity (grouped into mild, moderate and severe) and survival.

Figure 114 below is a Kaplan-Meier curve demonstrating the time to event model with CRP which is grouped into tertiles. It demonstrates overlap across the three groups as well. The low CRP group follows a better survival trend whilst medium CRP has a worse survival trend as does the high CRP group.

The p-value of the log-rank test was 0.28 which implies that this is not statistically significant but clinically is intuitive as patients with high CRP would likely be unwell clinically which could be cause or the effect.

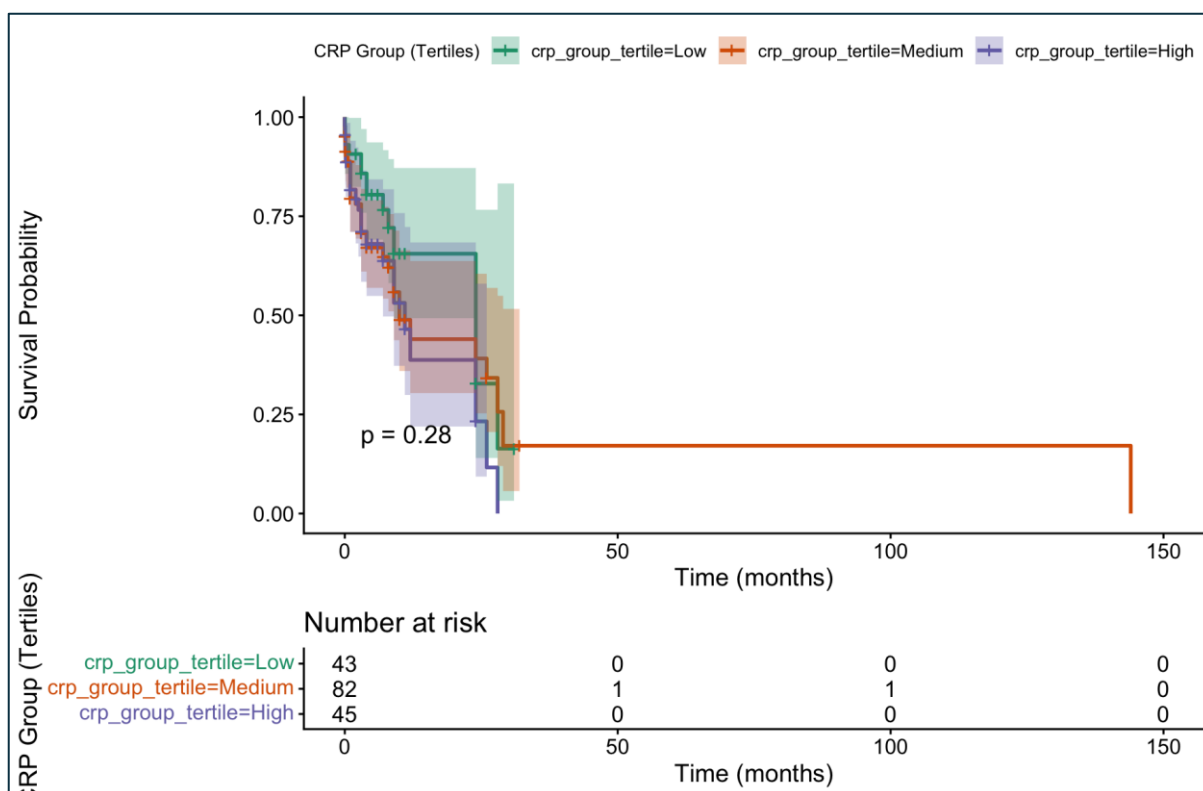


Figure 115: Kaplan- Meier curve - C-reactive protein (divided into tertiles) and survival.

6.4 Discussion

Sarcopenia severity and the pseudo fat fraction was combined with the patient metadata collected to further explore if this could be used as a surrogate marker or predictor for LLAs followed by MLLA and death (combined) and all amputation and death (combined).

Overall, there were more slightly more patients with severe sarcopenia in the all amputations group. There was also a higher distribution of individuals with the MLLA/death group and all amputation/death group with moderate and severe sarcopenia but this was not statistically significant. However, the odds ratio for 'all amputations and death combined' with the mild sarcopenia group was 1.368 [95% CI 1.020-1.845, p-value = 0.0377] with logistic regression. Severe sarcopenia was a variable with a hazard ratio of 1.79 [95% CI 0.92-3.49, p-value 0.088] on univariate cox proportional hazard model but did not appear on multivariate cox proportional models when combined with other variables which could possibly be attributed to smaller sample size, skewed distribution of the data or collinearity.

The pseudo fat fraction values did not demonstrate a correlation with all amputation or all amputation/death groups but did illustrate a trend of increasing odds of MLLA/death with

increasing pseudo fat fraction values (although not statistically significant). This may be due to the multimorbid nature of chronic diseases in the cohort. It could also be plausible that these patients may have had other amputations or be less mobile/bed bound over time which would also contribute to global sarcopenia as opposed to localised sarcopenic changes.

Multiple models were used to identify predictors which would be of value in creating a risk prediction model for all amputation followed by MLLA and death and all amputation and death as composite outcomes.

LASSO did not highlight any features for the all-amputation group or the combined MLLA and death group but did identify a previous history of transient ischaemic attacks to be associated with a higher likelihood of all amputations and death with a co-efficient value of 0.2406578 (from 0 to 1) p-value 0.013.

Ridge regression identified age, renal status, CRP, haemoglobin, albumin, HbA1c and pseudo fat fractions as predictors for all amputations however, ridge regression usually aims to retain as many predictors and distribute the weight as equally as possible. XGB had a low accuracy rate but identified neuropathy, cerebrovascular disease, sex, ethnicity, haemoglobin, peripheral arterial disease, HbA1c, weight and renal status. The low accuracy rate with XGB is not entirely unexpected as it tends to perform better on bigger datasets. Ridge regression identified multiple weak predictors for MLLA and death with the pseudo fat fraction having a co-efficient value of 0.0192302477 but not statistically significant. Interestingly, a previous history of TIA also appeared as a weak predictor for the all amputation and death group with a co-efficient value of 0.042935985. This may be possibly due to the multimorbid status of patients or the anatomical distribution of arterial disease which may include carotid disease.

RaF had high accuracy and identified age, renal status, CRP, haemoglobin, albumin and HbA1c as well as the pseudo fat fraction for all amputations. Fat fraction was highlighted as a feature of importance on the model for MLLA and death with a mean decrease Gini value of 14.15. This was followed by HbA1c and weight. In terms of all amputations and death as

an outcome, weight was highlighted as a feature of importance with a mean decrease Gini value of 14.15 followed by fat fraction with a mean decrease Gini value of 11.735341 and HbA1c.

The converging variables which appear in all three models for all amputations were Haemoglobin, HbA1c and renal status. The overlapping variables across ridge regression and RaF were CRP, age, albumin and the pseudo fat fraction.

Further analysis with a cox proportional hazard model found CRP and neuropathy to be strong predictors with statistical significance. Other features were also identified including age, retinopathy, moderate and severe sarcopenia and the pseudo fat fraction as predictors but these were not statistically significant. The Kaplan Meier curves found low CRP to be associated with better survival which is not unexpected. Similarly, mild sarcopenia had slightly better survival but both curves had wide confidence intervals.

Overall, it appears that CRP is a recurring and strong predictor. Other factors which appear to have a weaker influence on the outcome include age, renal status, haemoglobin, HbA1c, albumin and the pseudo fat fraction but these were not necessarily statistically significant. Not having neuropathy is a strong protective factor, which is not unexpected.

Chapter 7: Final Discussion and conclusion

7.1 Introduction (to the end)

This thesis set out to explore the potential applications of AI to help predict limb loss in patients with diabetic foot disease using MRIs. This was attempted by exploring the use of MRI reports, two-dimensional images of the foot through image classifications and three-dimensional images using segmentations. The results from these were incorporated with patient metadata to identify salient features which could be utilised as surrogate markers to predict the risk of limb loss. A secondary aim was to identify the extent of sarcopenia in patients with DFD.

Multiple deep learning models have been utilised throughout the thesis including random forest, extreme gradient boosting, multilayer perceptron models, vision transformers, ResNet and nnU-Net. Traditional statistical analyses have also been equally important. From descriptive statistics to Kruskal Wallis, inter-rater reliability, logistic regression, LASSO, ridge regression and cox proportional hazards to name a few. Each statistical analysis has provided some answers but added more questions along the way.

The NIHR good clinical practice module recommends using Gibb's reflective cycle to aid learning from experiences [189]. This comprises of six stages – description, feelings, evaluation, analysis, conclusion and action plan. Throughout the process of starting and hopefully, concluding this thesis, the reflective cycle has been an invaluable aid and which I will incorporate throughout this chapter. The project evolved along the way as different approaches were used to explore the use of multiple models to address the problem.

DFD is a complex disease process which includes multiple concurrent pathways affecting the body globally as discussed in chapter one. The presence of DFD is often a late presentation to a chronic pathological disease process and detecting early surrogate markers could help to delay or prevent this to some extent.

7.2 Summary of results and reflection

In chapter 3, MRI reports were explored as an avenue to help predict limb loss in patients. This included comparing minor and major amputation groups with individuals who did not have an amputation, comparing no amputation with all amputations, comparing major lower limb amputations with death, comparing all amputations with death, comparing no amputation with major lower limb amputations and death combined and comparing no amputation with all amputation and death combined. The aim was to identify features on MRI reports which would predict the risk.

Initial logistic regression identified collections and osteomyelitis as high-risk features. Random Forest, Extreme Gradient Boosting and Multilayer perceptron were used. Whilst not statistically significant, osteomyelitis was a recurring feature of importance although it

must be considered that the majority of the MRIs would have been done to rule out osteomyelitis therefore their presence and absence was more frequently reported.

The only model of statistical significance was extreme gradient boosting comparing the no amputation versus minor amputation versus major amputation group. The model had a 63.4% accuracy [95% CI 0.52-0.7; p-value <0.001]. The ROC-AUC values were 0.724 (no amputation group), 0.779 (minor amputation group) and 0.439 (major amputation group). Absence of osteomyelitis and presence of Charcot were identified as variables which influenced the outcome. This then warrants a discussion as to whether a certain uniform reporting format should be adopted to ensure consistency in reporting salient features when reviewing MRIs in patients with DFD.

In chapter 4, MRI images were utilised to explore how sarcopenia could be graded or measured. This included using two dimensional images and classification of severity using vision transformers (ViT) and ResNet models.

ViT had an accuracy of 78.7% with an F1 score of 79.9%. ResNet-50 had an accuracy of 79% with an F1 score of 74% and ResNet-100 had an accuracy of 80.7% with an F1 score of 69.3%.

The benefit of using two-dimensional images using ViT and ResNet confers certain benefits. All three models are relatively easy to train without a steep learning curve. Small datasets can be easily utilised and expanded using data augmentation to increase the number of images a model has to train on. The image types which may be used for classification tasks is not limited. The models can use radiology images, photographs, image-based reports, etc. The task of image classification requires minimal training. Furthermore, all models can be deployed easily and utilised for classification purposes. The applicability of this finding goes beyond the scope of this thesis and may be utilised for classification tasks in other areas of vascular surgery with images of varying complexity.

MRI scans of the foot were also segmented, and nnU-Net was applied. The mean DICE score was around 0.4. Whilst it did demonstrate some potential with DICE scores of up to 0.8 with

uncomplicated anatomy, it completely faltered in areas where this was distorted. This severely limits its clinical applicability at this moment in time given that the cohort of patients with DFD would likely not have unaltered anatomy.

Throughout the process of segmentations and running nnU-Net, certain learning points became evident to ensure future successful use of segmentations with nnU-Net which is, in and of itself, an incredibly valuable and powerful tool with potential. The area being segmented or measured needs to have a clear, defined region with an area for calculation on cross-sectional imaging. The area segmented should have some contrast to the surrounding region which is where other imaging modalities can be useful as they allow for uniformity in measurements of values and not just the area or diameter. Furthermore, multiple slices on fewer scans is likely to add more value than single slices on multiple scans. It does have potential for applicability in patients without distorted anatomy but segmenting normal MRI scans would not add any clinical benefit unless it is for a different pathology.

Manual segmentations also allowed for measurement of the pseudo fat fraction. This was combined with patient metadata along with grading of sarcopenia to evaluate its use as a predictor for LLAs. Severe sarcopenia was associated with a slightly higher number LLAs and the pseudo-fat fraction did appear as a weak predictor. However, neither sarcopenia severity nor pseudo fat fraction values were statistically significant predictors for LLAs alone. Haemoglobin, HbA1c and renal status were recurring predictors across all models. CRP, age, albumin and the pseudo fat fraction appeared in both RaF and ridge regression. CRP appeared to have the strongest association with LLAs. This is not entirely unexpected and follows the current clinical presentation and trajectory. Interestingly, a previous history of TIA's was a predictor on LASSO and ridge regression when evaluating composite outcomes.

Whilst sarcopenia severity was not a strong predictor for all amputations or the MLLA/death groups, the odds ratio for 'all amputations and death combined' with the mild sarcopenia group was 1.368 [95% CI 1.020-1.845, p-value = 0.0377] with logistic regression modelling. Severe sarcopenia was a variable with a hazard ratio of 1.79 [95% CI 0.92-3.49, p-value 0.088] on univariate cox proportional hazard modelling. This is somewhat reflective of

existing literature including a meta-analysis of fifty six articles from 2022 which found Overall, sarcopenia was associated with a significantly higher risk of mortality (Hazard Ratio 2.00 [95% CI: 1.71, 2.34]; Odds Ratio: 2.35 [95% CI: 1.64, 3.37]), and which was independent of population, sarcopenia definition, and follow-up period on subgroup analyses [190]. This is also in keeping with the discussion in chapter one, section nine point four (1.9.4) where sarcopenia and the risk of adverse outcomes including mortality was discussed.

The pseudo fat fraction values demonstrated a strong positive correlation in the MLLA/death group albeit not being statistically significant, but this was not the case with all amputations or all amputations/death when comparing outcomes. The positive correlation with MLLA and death may be due to the multimorbid nature of chronic diseases in the cohort. It could also be plausible that these patients may have wither had other amputations or be less mobile/bed bound over time which would also contribute to global sarcopenia as opposed to localised sarcopenic changes.

7.3 Limitations

This thesis is not without its limitations. The cohort of patients included adds bias to the overall project. A majority of patients who had MRIs were diabetic and had a clinical reason to warrant the imaging i.e. to rule out osteomyelitis, to check if osteomyelitis had resolved or to check for collections.

Whilst there were patients with mild sarcopenia, the more prevalent groups were patients with moderate or severe sarcopenia. This provides a slightly skewed patient cohort and creates class imbalance throughout the project which was accounted for whenever possible during the statistical analyses. It is possible that is there was a significant cohort of patients without sarcopenia being compared to patients with sarcopenia, the results may potentially be different.

The reporting format for MRIs does not follow a set pattern of reporting but a vast majority were reported by musculoskeletal radiologist with expertise in the area and therefore the presence and absence of features would be almost the same but variations in reporting cannot be accounted for.

7.4 Real-world implications and scope for future work

The findings from this particular thesis will not change clinical practice but they have created the space for exploring further applicability of these methods in DFD. This had not been done previously. XGB was accurate in the no amputation versus minor amputation versus major amputation group and ViT and ResNet-50 were accurate on the dataset used as part of the thesis however, they require external validation. Using MRI reports for predictive purposes is not onerous in terms of costs but using ViT and ResNet is computationally intensive.

Further consideration must also be given to the emerging themes from the PPI undertaken at the start of the project. It is apparent that patients want reliability when using AI but are open to its integration. Further PPI to present these findings and gain feedback could provide further direction depending on the sentiment of the patients. If these can be validated externally, the next step would be to evaluate the cost-effectiveness in integrating such models.

Alternatively, the findings from this thesis can be utilised to potentially shift the focus of the pathology being evaluated. The focus in this thesis revolves around exploring sarcopenia as a surrogate marker however, it is apparent that osteomyelitis and infection are also potentially more salient predictors. Therefore, ViT and ResNet models may be used for identifying the presence or absence of osteomyelitis on MRI scans.

The features highlighted as predictors in the preceding chapters for LLAs including CRP, osteomyelitis, previous history of TIAs, renal status, neuropathy could also be explored further on a bigger dataset which would provide this information including hospital episode statistics (HES) or biobank data.

7.5 Conclusion

To conclude, sarcopenia was not found to be a statistically significant surrogate marker in predicting lower limb amputations as a whole but did increase the odds of all amputation/death as a composite outcome with statistical significance. There was a strong

positive correlation with the pseudo fat fraction values and MLLA/death albeit not statistically significant.

XGB was found to be accurate in predicting no amputation versus minor versus major amputation on MRI reports.

Vision transformers and ResNet models have potential applicability in classifying mild, moderate and severe sarcopenia. nnU-Net yielded a low mean DICE score in segmenting the foot and has very limited applicability at this present moment in time.

References

1. Loriaux DL. Diabetes and the ebers papyrus: 1552 BC. *The Endocrinologist*. 2006;16(2):55-6.
2. MacFarlane I. Mathew Dobson of Liverpool (1735–1784) and the history of diabetes. *Practical Diabetes International*. 1990;7(6):246-8.
3. Poretsky L. *Principles of diabetes mellitus*. Poretsky L, editor. New York: Springer; 2010 Jan 24.
4. Aljohary H, Murad MA, Alfkey R. Diabetic Foot Ulcer: A Historical Overview. *Diabetic Foot Ulcers-Pathogenesis, Innovative Treatments and AI Applications*. 2024 Jan 5.
5. Laios K, Karamanou M, Saridaki Z, Androutsos G. Aretaeus of Cappadocia and the first description of diabetes. *Hormones*. 2012;11:109-13.
6. Allan FN. The writings of Thomas Willis, MD: diabetes three hundred years ago. *Diabetes*. 1953 Jan 1;2(1):74-8.
7. Cawley T. A singular case of diabetes, consisting entirely in the quality of the urine; with an inquiry into the different theories of that disease. *The London Medical Journal*. 1788;9(Pt 3):286.
8. Bliss M. *The Discovery of Insulin-the 25th anniversary*. University of Chicago Press; 2007.
9. Egan AM, Dinneen SF. What is diabetes?. *Medicine*. 2019 Jan 1;47(1):1-4.

10. International Diabetes Federation. IDF Diabetes Atlas. 11th ed. Brussels: IDF; 2025 [cited 2025 Apr 17]. Available from: <https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/>
11. National Institute for Health and Care Excellence (NICE). Type 2 diabetes: background information – prevalence. Clinical Knowledge Summaries. London: NICE; 2023 [cited 2025 Apr 17]. Available from: <https://cks.nice.org.uk/topics/diabetes-type-2/background-information/prevalence/>.
12. Chen H, Huang X, Dong M, Wen S, Zhou L, Yuan X. The Association Between Sarcopenia and Diabetes: From Pathophysiology Mechanism to Therapeutic Strategy. Diabetes, Metabolic Syndrome and Obesity. 2023;1541-54.
13. Chen HC, Sunardi, Liao BY, Lin CY, Akbari VB, Lung CW, Jan YK. Estimation of various walking intensities based on wearable plantar pressure sensors using artificial neural networks. Sensors. 2021 Sep 29;21(19):6513.
14. Zhou B, Rayner AW, Gregg EW, Sheffer KE, Carrillo-Larco RM, Bennett JE, Shaw JE, Paciorek CJ, Singleton RK, Pires AB, Stevens GA. Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. The Lancet. 2024 Nov 23;404(10467):2077-93.
15. Diabetes UK. Number of people living with diabetes in the UK tops 5 million for the first time. London: Diabetes UK; 2023 [cited 2025 Apr 17]. Available from: <https://www.diabetes.org.uk/about-us/news/number-people-living-diabetes-uk-tops-5-million-first-time>.
16. Office for Health Improvement and Disparities (OHID). National Diabetes Foot Care Report. National Cardiovascular Intelligence Network. London: OHID; 2021 [cited 2025 Apr 17]. Available from: <https://fingertips.phe.org.uk/static-reports/diabetes-footcare/national-diabetic-footcare-report.html> .
17. Zhao L, Hu H, Zhang L, Liu Z, Huang Y, Liu Q, et al. Inflammation in diabetes complications: molecular mechanisms and therapeutic interventions. MedComm. 2024;5(4):e516.
18. Ward MG, Li G, Hao M. Apoptotic β -cells induce macrophage reprogramming under diabetic conditions. Journal of Biological Chemistry. 2018 Oct 19;293(42):16160-73.
19. Costantini D. The role of organismal oxidative stress in the ecology and life-history evolution of animals. Springer; 2024 Sep 9.
20. Burton GJ, Jauniaux E. Oxidative stress. Best practice & research Clinical obstetrics & gynaecology. 2011 Jun 1;25(3):287-99.

21. Hartog JW, Voors AA, Bakker SJ, Smit AJ, van Veldhuisen DJ. Advanced glycation end-products (AGEs) and heart failure: pathophysiology and clinical implications. *European journal of heart failure*. 2007 Dec;9(12):1146-55.
22. Healthcare Quality Improvement Partnership. National Diabetes Inpatient Audit 2015. London: HQIP; 2016 [cited 2025 Aug 17]. Available from: <https://www.hqip.org.uk/wp-content/uploads/2018/02/national-diabetes-inpatient-audit-2015.pdf>.
23. Farmaki P, Damaskos C, Garmpis N, Garmpi A, Savvanis S, Diamantis E. Complications of the type 2 diabetes mellitus. *Current cardiology reviews*. 2020;16(4):249-51.
24. Fitridge R, Chuter V, Mills J, Hinchliffe R, Azuma N, Behrendt CA, Boyko EJ, Conte MS, Humphries M, Kirksey L, McGinigle KC. The intersocietal IWGDF, ESVS, SVS guidelines on peripheral artery disease in people with diabetes and a foot ulcer. *Diabetes/metabolism research and reviews*. 2024 Mar;40(3): e3686.
25. Van Netten JJ, Bus SA, Apelqvist J, Chen P, Chuter V, Fitridge R, Game F, Hinchliffe RJ, Lazzarini PA, Mills J, Monteiro-Soares M. Definitions and criteria for diabetes-related foot disease (IWGDF 2023 update). *Diabetes/Metabolism Research and Reviews*. 2024 Mar;40(3): e3654.
26. Diabetes UK. The cost of diabetes report. London: Diabetes UK; 2014 [cited 2025 Apr 17]. Available from: <https://www.diabetes.org.uk/resources-s3/2017-11/diabetes%20uk%20cost%20of%20diabetes%20report.pdf>.
27. Drake R, Vogl AW, Mitchell AW. Gray's anatomy for students E-book. Elsevier Health Sciences; 2009 Apr 4.
28. Jastifer JR. Intrinsic muscles of the foot: Anatomy, function, rehabilitation. *Physical therapy in sport*. 2023 May 1; 61:27-36.
29. Ahmad M, Mohamed A, Amiras D, Siracusa F, Shalhoub J, Davies AH. Sarcopenia in the foot on magnetic resonance imaging in patients with diabetes mellitus—a systematic review. *Clinical Diabetes and Endocrinology*. 2024 Oct 25;10(1):31.
30. Dalal S, Widgerow AD, Evans GR. The plantar fat pad and the diabetic foot—a review. *International Wound Journal*. 2015 Dec;12(6):636-40.
31. Bus, S.A., Maas, M., de Lange, A., Michels, R.P. and Levi, M., 2005. Elevated plantar pressures in neuropathic diabetic patients with claw/hammer toe deformity. *Journal of biomechanics*, 38(9), pp.1918-1925.
32. Bus SA, Maas M, Cavanagh PR, Michels RP, Levi M. Plantar fat-pad displacement in neuropathic diabetic patients with toe deformity: a magnetic resonance imaging study. *Diabetes care*. 2004 Oct 1;27(10):2376-81.

33. Bodman MA, Dreyer MA, Varacallo MA. Diabetic peripheral neuropathy. InStatPearls [Internet] 2024 Feb 25. StatPearls Publishing.
34. Vinik AI, Maser RE, Mitchell BD, Freeman R. Diabetic autonomic neuropathy. *Diabetes Care*. 2003;26(5):1553-79.
35. Ahmad M, Patel K, Davies AH, Shalhoub J. Noninvasive optical methods to assess tissue perfusion in patients with peripheral arterial disease and diabetes mellitus: A scoping review and discussion. *JVS-Vascular Insights*. 2024 Jan 1; 2:100034.
36. Wukich DK, Schaper NC, Gooday C, Bal A, Bem R, Chhabra A, Hastings M, Holmes C, Petrova NL, Santini Araujo MG, Senneville E. Guidelines on the diagnosis and treatment of active Charcot neuro-osteoarthropathy in persons with diabetes mellitus (IWGDF 2023). *Diabetes/Metabolism Research and Reviews*. 2024 Mar;40(3): e3646.
37. Baroni A, Russo T, Piccolo V. Relationship between local neuroimmune impairment and diabetic foot: the immunocompromised district theory. *International Journal of Dermatology*. 2014 Mar;53(3):263-6.
38. National Institute for Health and Care Excellence (NICE). Diabetic foot problems: prevention and management. NICE Guideline NG19. London: NICE; 2023 [cited 2025 Apr 17]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK553608/>.
39. Schaper NC, Van Netten JJ, Apelqvist J, Bus SA, Fitridge R, Game F, Monteiro-Soares M, Senneville E, IWGDF Editorial Board. Practical guidelines on the prevention and management of diabetes-related foot disease (IWGDF 2023 update). *Diabetes/Metabolism Research and Reviews*. 2024 Mar;40(3): e3657.
40. Monteiro-Soares M, Hamilton EJ, Russell DA, Srisawasdi G, Boyko EJ, Mills JL, Jeffcoate W, Game F. Guidelines on the classification of foot ulcers in people with diabetes (IWGDF 2023 update). *Diabetes/Metabolism Research and Reviews*. 2024 Mar;40(3): e3648.
41. Røikjer J, Monteiro-Soares M, Walton D, Iacopi E, Jirkovska J, Edmonds M, Trocha A, Jeffcoate W, Bus S, Diabetic Foot Study Group (DFSG) prevention of the first ulcer research group. External validation of the DIAFORA system to predict lower-extremity amputations in a prospective Danish cohort. *Diabetic Medicine*. 2025 Jan;42(1): e15443.
42. Chen P, Vilorio NC, Dhatariya K, Jeffcoate W, Lobmann R, McIntosh C, Piaggese A, Steinberg J, Vas P, Viswanathan V, Wu S. Guidelines on interventions to enhance healing of foot ulcers in people with diabetes (IWGDF 2023 update). *Diabetes/metabolism research and reviews*. 2024 Mar;40(3): e3644.
43. Diabetes UK. Why we're urging NHS England to invest in diabetes foot care. London: Diabetes UK; 2020 [cited 2025 Apr 17]. Available from: <https://www.diabetes.org.uk/about-us/news-and-views/diabetes-foot-care#:~:text=Someone%20living%20with%20diabetes%20is,to%20the%20feet%20and%20legs.>

44. Holman N, Young RJ, Jeffcoate WJ. Variation in the recorded incidence of amputation of the lower limb in England. *Diabetologia*. 2012 Jul;55(7):1919-25.
45. Office for Health Improvement and Disparities (OHID). National Diabetes Foot Care Report. National Cardiovascular Intelligence Network. London: OHID; 2021 [cited 2025 Apr 17]. Available from: <https://fingertips.phe.org.uk/static-reports/diabetes-footcare/national-diabetic-footcare-report.html>.
46. Rivera, J.A., et al., *Estimating recent US limb loss prevalence and updating future projections*. Archives of Rehabilitation Research and Clinical Translation, 2024. 6(4): p. 100376.
47. Horrocks M. Vascular surgery – GIRFT programme national specialty report. London: NHS England; 2018 [cited 2025 Apr 17]. Available from: https://gettingitrightfirsttime.co.uk/wp-content/uploads/2018/02/GIRFT_Vascular_Surgery_Report-March_2018.pdf.
48. Mottolini N. Diabetes and lower limb complications – a thematic review of clinical negligence claims. London: NHS Resolution; 2022 [cited 2025 May 17]. Available from: https://resolution.nhs.uk/wp-content/uploads/2022/06/Diabetes_and_Lower_Limb_Complications.pdf.
49. Hex N, Bartlett C, Wright D, Taylor M, Varley DJ. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabetic medicine*. 2012 Jul;29(7):855-62.
50. Hex, N., MacDonald, R., Pocock, J., Uzdzińska, B., Taylor, M., Atkin, M., Wild, S.H., Beba, H. and Jones, R., 2024. Estimation of the direct health and indirect societal costs of diabetes in the UK using a cost of illness model. *Diabetic Medicine*, 41(9), p.e15326.
51. NHS England. 2023-25 NHS Payment Scheme (amended). 2024/25 prices workbook. [cited 2025 July 10]. Available from: <https://www.england.nhs.uk/publication/2023-25-nhs-payment-scheme/>.
52. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Annals of internal medicine*. 2008 Nov 18;149(10):751-60.
53. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC medical research methodology*. 2014 Jan 8;14(1):3.
54. Kaka, A.S., Landsteiner, A., Ensrud, K.E., Logan, B., Sowerby, C., Ullman, K., Yoon, P., Wilt, T.J. and Sultan, S., 2023. Risk prediction models for diabetic foot ulcer development or amputation: a review of reviews. *Journal of Foot and Ankle Research*, 16(1), p.13.
55. Monteiro-Soares M, Boyko EJ, Jeffcoate W, Mills JL, Russell D, Morbach S, Game F. Diabetic foot ulcer classifications: a critical review. *Diabetes/metabolism research and reviews*. 2020 Mar;36: e3272.

56. Van Battum P, Schaper N, Prompers L, Apelqvist J, Jude E, Piaggese A, Bakker K, Edmonds M, Holstein P, Jirkovska A, Mauricio D. Differences in minor amputation rate in diabetic foot disease throughout Europe are in part explained by differences in disease severity at presentation. *Diabetic medicine*. 2011 Feb;28(2):199-205.
57. Pickwell K, Siersma V, Kars M, Apelqvist J, Bakker K, Edmonds M, Holstein P, Jirkovská A, Jude E, Mauricio D, Piaggese A. Predictors of lower-extremity amputation in patients with an infected diabetic foot ulcer. *Diabetes care*. 2015 May 1;38(5):852-7.
58. Mills Sr JL, Conte MS, Armstrong DG, Pomposelli FB, Schanzer A, Sidawy AN, Andros G, Society for Vascular Surgery Lower Extremity Guidelines Committee. The society for vascular surgery lower extremity threatened limb classification system: risk stratification based on wound, ischemia, and foot infection (WIFI). *Journal of vascular surgery*. 2014 Jan 1;59(1):220-34.
59. Senneville É, Albalawi Z, Van Asten SA, Abbas ZG, Allison G, Aragón-Sánchez J, Embil JM, Lavery LA, Alhasan M, Oz O, Uçkay I. IWGDF/IDSA guidelines on the diagnosis and treatment of diabetes-related foot infections (IWGDF/IDSA 2023). *Diabetes/metabolism research and reviews*. 2024 Mar;40(3): e3687.
60. Semelka RC. MRI: basic principles and applications. Wiley-Liss; 1995.
61. McRobbie DW, Moore EA, Graves MJ, Prince MR. MRI from Picture to Proton. Cambridge university press; 2017 Apr 13.
62. Major NM, Anderson MW. Musculoskeletal MRI E-Book. Elsevier Health Sciences; 2019 Oct 4.
63. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Cooper C, Landi F, Rolland Y, Sayer AA, Schneider SM. Sarcopenia: revised European consensus on definition and diagnosis. *Age and ageing*. 2019 Jan 1;48(1):16-31.
64. Izzo A, Massimino E, Riccardi G, Della Pepa G. A narrative review on sarcopenia in type 2 diabetes mellitus: prevalence and associated factors. *Nutrients*. 2021 Jan 9;13(1):183.
65. Vellas B, Fielding RA, Bens C, Bernabei R, Cawthon PM, Cederholm T, Cruz-Jentoft AJ, Del Signore S, Donahue S, Morley J, Pahor M. Implications of ICD-10 for sarcopenia clinical practice and clinical trials: report by the international conference on frailty and sarcopenia research task force. *The Journal of frailty & aging*. 2018 Jan;7(1):2-9.
66. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, Jang HC, Kang L, Kim M, Kim S, Kojima T. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. *Journal of the American Medical Directors Association*. 2020 Mar 1;21(3):300-7.

67. Voulgaridou G, Tyrovolas S, Detopoulou P, Tsoumana D, Drakaki M, Apostolou T, Chatziprodromidou IP, Papandreou D, Giaginis C, Papadopoulou SK. Diagnostic criteria and measurement techniques of sarcopenia: a critical evaluation of the up-to-date evidence. *Nutrients*. 2024 Feb 1;16(3):436.
68. Shaw SC, Dennison EM, Cooper C. Epidemiology of sarcopenia: determinants throughout the lifecourse. *Calcified tissue international*. 2017 Sep;101(3):229-47.
69. Mesinovic J, Zengin A, De Courten B, Ebeling PR, Scott D. Sarcopenia and type 2 diabetes mellitus: a bidirectional relationship. *Diabetes, metabolic syndrome and obesity: targets and therapy*. 2019 Jul 8:1057-72.
70. Chung SM, Moon JS, Chang MC. Prevalence of sarcopenia and its association with diabetes: a meta-analysis of community-dwelling Asian population. *Frontiers in medicine*. 2021 May 20; 8:681232.
71. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, Jang HC, Kang L, Kim M, Kim S, Kojima T. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. *Journal of the American Medical Directors Association*. 2020 Mar 1;21(3):300-7.
72. Feng L, Gao Q, Hu K, Wu M, Wang Z, Chen F, Mei F, Zhao L, Ma B. Prevalence and risk factors of sarcopenia in patients with diabetes: a meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*. 2022 May 1;107(5):1470-83.
73. Mesinovic J, McMillan LB, Shore-Lorenti C, De Courten B, Ebeling PR, Scott D. Metabolic syndrome and its associations with components of sarcopenia in overweight and obese older adults. *Journal of clinical medicine*. 2019 Jan 27;8(2):145.
74. Feng L, Gao Q, Hu K, Wu M, Wang Z, Chen F, Mei F, Zhao L, Ma B. Prevalence and risk factors of sarcopenia in patients with diabetes: a meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*. 2022 May 1;107(5):1470-83.
75. Fuggle N, Shaw S, Dennison E, Cooper C. Sarcopenia. *Best Practice & Research Clinical Rheumatology*. 2017 Apr 1;31(2):218-42.
76. Dent E, Morley JE, Cruz-Jentoft AJ, Arai H, Kritchevsky SB, Guralnik J, Bauer JM, Pahor M, Clark BC, Cesari M, Ruiz J. International clinical practice guidelines for sarcopenia (ICFSR): screening, diagnosis and management. *The Journal of nutrition, health and aging*. 2018 Dec 1;22(10):1148-61.
77. Tsouknidas, A., Michailidis, N. and Anagnostidis, K., 2012, February. Medical Imaging as a Bone Quality Determinant and Strength Indicator of the Femoral Neck. In *International*

Joint Conference on Biomedical Engineering Systems and Technologies (pp. 210-221). Berlin, Heidelberg: Springer Berlin Heidelberg.

78. Kim YK, Lee HS, Ryu JJ, In Lee H, Seo SG. Sarcopenia increases the risk for mortality in patients who undergo amputation for diabetic foot. *Journal of Foot and Ankle Research*. 2018 Jun 19;11(1):32.
79. Reinders I, Murphy RA, Brouwer IA, Visser M, Launer L, Siggeirsdottir K, Eiriksdottir G, Gudnason V, Jonsson PV, Lang TF, Harris TB. Muscle quality and myosteatosis: novel associations with mortality risk: the Age, Gene/Environment Susceptibility (AGES)-Reykjavik Study. *American journal of epidemiology*. 2016 Jan 1;183(1):53-60.
80. Cheng Q, Hu J, Yang P, Cao X, Deng X, Yang Q, Liu Z, Yang S, Goswami R, Wang Y, Luo T. Sarcopenia is independently associated with diabetic foot disease. *Scientific reports*. 2017 Aug 21;7(1):8372.
81. Hon KY, Bain M, Edwards S, Pena G, McMillan N, Fitridge R. The association of sarcopenia and frailty in diabetes-related foot disease: A 3-year prospective evaluation. *Journal of Foot and Ankle Research*. 2025 Mar;18(1): e70038.
82. Yang Q, Ni X, Zhang Y, Zhu B, Zeng Q, Yang C, Shi J, Zhang C, Cai J, Hu J, Li Q. Sarcopenia is an independent risk factor for all-cause mortality rate in patients with diabetic foot ulcers. *Frontiers in Nutrition*. 2023 Apr 11; 10:1097008.
83. Matsubara Y, Matsumoto T, Aoyagi Y, Tanaka S, Okadome J, Morisaki K, Shirabe K, Maehara Y. Sarcopenia is a prognostic factor for overall survival in patients with critical limb ischemia. *Journal of vascular surgery*. 2015 Apr 1;61(4):945-50.
84. Lin YH, Tsai PS, Hung CL, Beg MF, Yeh HI, Yun CH, Wu MT. The Role of Muscle Density in Predicting the Amputation Risk in Peripheral Arterial Disease: A Tissue Composition Study Using Lower Extremity CT Angiography. *Diagnostics*. 2025 Jun 5;15(11):1439.
85. Sivaharan A, Boylan L, Witham MD, Nandhra S, Centre TN. Sarcopenia in patients undergoing lower limb bypass surgery is associated with higher mortality and major amputation rates. *Annals of Vascular Surgery*. 2021 Aug 1; 75:227-36.
86. Luque-Linero P, Frutos-Reoyo EJ, Castilla-Guerra L, Rico-Corral MÁ, Salamanca-Bautista P, Garrachón-Vallo F. Sarcopenia as a Prognostic Factor for Critical Limb Ischemia: A Prospective Cohort Study. *Journal of Clinical Medicine*. 2025 Jul 31;14(15):5388.
87. Kykkänen E, Kaartinen I, Arponen O, Vuorlaakso M. Low imaging-detected muscle mass as a prognostic factor for overall and amputation-free survival in patients undergoing lower extremity amputation. *Journal of Diabetes and its Complications*. 2025 Jun 30:109119.
88. Andersen H, Gadeberg PC, Brock B, Jakobsen J. Muscular atrophy in diabetic neuropathy: a stereological magnetic resonance imaging study. *Diabetologia*. 1997 Aug;40(9):1062-9.

89. Andreassen CS, Jakobsen J, Ringgaard S, Ejksjaer N, Andersen H. Accelerated atrophy of lower leg and foot muscles—a follow-up study of long-term diabetic polyneuropathy using magnetic resonance imaging (MRI). *Diabetologia*. 2009 Jun;52(6):1182-91.
90. Bus SA, Yang QX, Wang JH, Smith MB, Wunderlich R, Cavanagh PR. Intrinsic muscle atrophy and toe deformity in the diabetic neuropathic foot: a magnetic resonance imaging study. *Diabetes care*. 2002 Aug 1;25(8):1444-50.
91. Bus SA, Maas M, Michels RP, Levi M. Role of intrinsic muscle atrophy in the etiology of claw toe deformity in diabetic neuropathy may not be as straightforward as widely believed. *Diabetes Care*. 2009 Jun 1;32(6):1063-7.
92. Cheuy VA, Commean PK, Hastings MK, Mueller MJ. Reliability and validity of a MR-based volumetric analysis of the intrinsic foot muscles. *Journal of Magnetic Resonance Imaging*. 2013 Nov;38(5):1083-93.
93. Greenman RL, Khaodhiar L, Lima C, Dinh T, Giurini JM, Veves A. Foot small muscle atrophy is present before the detection of clinical neuropathy. *Diabetes Care*. 2005 Jun 1;28(6):1425-30.
94. Lue B, Amaya J, Silva FD, Raspovic K, Xi Y, Chhabra A. Characterizing conventional ankle MRI findings of nerve and muscle changes in diabetic patients: a case–control study. *Skeletal radiology*. 2023 Feb;52(2):225-31.
95. Altunkaynak BZ, Onger ME, Altunkaynak ME, Ayranci E, Canan S. A brief introduction to stereology and sampling strategies: basic concepts of stereology. *NeuroQuantology*. 2012 Mar 1;10(1):31-43.
96. Andersen H, Gadeberg PC, Brock B, Jakobsen J. Muscular atrophy in diabetic neuropathy: a stereological magnetic resonance imaging study. *Diabetologia*. 1997 Aug;40(9):1062-9.
97. Bus SA, Maas M, Lindeboom R. Reproducibility of foot structure measurements in neuropathic diabetic patients using magnetic resonance imaging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*. 2006 Jul;24(1):25-32.

98. Lin YC, Wu J, Baltzis D, Veves A, Greenman R. MRI assessment of regional differences in phosphorus-31 metabolism and morphological abnormalities of the foot muscles in diabetes. *Journal of Magnetic Resonance Imaging*. 2016 Nov;44(5):1132-42.
99. Goutallier D, Postel JM, Bernageau J, Lavau L, Voisin MC. Fatty muscle degeneration in cuff ruptures: pre-and postoperative evaluation by CT scan. *Clinical Orthopaedics and Related Research*. 1994 Jul 1; 304:78-83.
100. Oxford University Press. Oxford English Dictionary [Internet]. Oxford: Oxford University Press; 1993 [cited 2023 Apr 8]. Available from: <http://www.oed.com/viewdictionaryentry/Entry/11125>
101. ReportLinker. Artificial intelligence in healthcare market with COVID-19 impact analysis by offering technology, end-use application, end-user and region – global forecast to 2026. Lyon: ReportLinker; 2021 [cited 2025 Aug 17]. Available from: <https://www.reportlinker.com/p04897122/Artificial-Intelligence-in-Healthcare-Market-by-Offering-Technology-Application-End-User-Industry-and-Geography-Global-Forecast-to.html>.
102. Ahmad M, Tan M, Bergman H, Shalhoub J, Davies A. The use of artificial intelligence in three-dimensional imaging modalities and diabetic foot disease: A systematic review. *JVS-Vascular Insights*. 2024 Jan 1; 2:100057.
103. Panesar A. Machine learning and AI for healthcare (pp. 1-73). Coventry, UK: A press. 2019.
104. Dua A. Artificial intelligence vs machine learning vs deep learning: what's the difference? San Francisco: Turing; 2022 [cited 2023 March 17]. Available from: <https://www.turing.com/blog/artificial-intelligence-vs-machine-learning-vs-deep-learning/>.
105. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016 Feb;278(2):563-77.
106. Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, Al-Omran M. Machine learning in vascular surgery: a systematic review and critical appraisal. *NPJ Digital Medicine*. 2022 Jan 19;5(1):7.
107. Goyal M, Reeves ND, Rajbhandari S, Ahmad N, Wang C, Yap MH. Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques. *Computers in biology and medicine*. 2020 Feb 1; 117:103616.

108. Anaya-Isaza A, Zequera-Diaz M. Fourier transform-based data augmentation in deep learning for diabetic foot thermograph classification. *Biocybernetics and Biomedical Engineering*. 2022 Apr 1;42(2):437-52.
109. Zhang D, Dong W, Guan H, Yakupu A, Wang H, Chen L, Lu S, Tang J. CT-angiography-based outcome prediction on diabetic foot ulcer patients: a statistical learning approach. *Diagnostics*. 2022 Apr 25;12(5):1076.
110. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*. 2021 Feb;18(2):203-11.
111. Krantz T, Jonker A. What Is Ground Truth in Machine Learning? IBM Think [Internet]. 2024 Dec 20 [cited 2025 Aug 21]. Available from: <https://www.ibm.com/think/topics/ground-truth>.
112. Bertels J, Eelbode T, Berman M, Vandermeulen D, Maes F, Bisschops R, Blaschko MB. Optimizing the DICE score and jaccard index for medical image segmentation: Theory and practice. In *International conference on medical image computing and computer-assisted intervention* 2019 Oct 10 (pp. 92-100). Cham: Springer International Publishing.
113. Kim HS, Kim H, Kim S, Cha Y, Kim JT, Kim JW, Ha YC, Yoo JI. Precise individual muscle segmentation in whole thigh CT scans for sarcopenia assessment using U-net transformer. *Scientific Reports*. 2024 Feb 8;14(1):3301.
114. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, Beam AL, Collins GS, Darzi A, Deeks JJ, ElZarrad MK. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health*. 2020 Oct 1;2(10): e549-60.
115. Bosanquet DC, Nandhra S, Wong KH, Long J, Chetter I, Hinchliffe RJ. Research priorities for lower limb amputation in patients with vascular disease. *medRxiv*. 2021 May 24:2021-05.
116. Hitzig SL, Mayo AL, Kayssi A, Viana R, MacKay C, Devlin M, Dilkas S, Domingo A, Hebert JS, Miller WC, Andrysek J. Identifying priorities and developing strategies for building capacity in amputation research in Canada. *Disability and Rehabilitation*. 2021 Sep 11;43(19):2779-89.
117. Atkin L, Hitchman L, Russell D. Research priorities in diabetic foot disease. *Diabetic Foot*. 2024;27(2):9.
118. Health Research Authority. Public involvement – Best Practice [Internet]. HRA; 2025 [cited 2025 Aug 22]. Available from: <https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement>.

119. UK Standards for Public Involvement Development Partnership. UK Standards for Public Involvement in Research [Internet]. NIHR; [cited 2025 Sep 11]. Available from: <https://sites.google.com/nihr.ac.uk/pi-standards/home>.
120. Cohen J. Statistical power analysis for the behavioral sciences. Routledge; 2013 May 13.
121. Raosoft, Raosoft sample size calculator [computer programme]; Raosoft, Inc., Seattle, WA. 2004.
122. Williams JR. The Declaration of Helsinki and public health. Bulletin of the World Health Organization. 2008;86:650-2.
123. Traveset A, Rubinat E, Ortega E, Alcubierre N, Vazquez B, Hernández M, Jurjo C, Espinet R, Ezpeleta JA, Mauricio D. Lower hemoglobin concentration is associated with retinal ischemia and the severity of diabetic retinopathy in type 2 diabetes. Journal of diabetes research. 2016;2016(1):3674946.
124. Kwon E, Ahn C. Low hemoglobin concentration is associated with several diabetic profiles. The Korean journal of internal medicine. 2012 Sep 1;27(3):273.
125. Katwal PC, Jirjees S, Htun ZM, Aldawudi I, Khan S. The effect of anemia and the goal of optimal HbA1c control in diabetes and non-diabetes. Cureus. 2020 Jun 3;12(6).
126. Guérin-Dubourg A, Catan A, Bourdon E, Rondeau P. Structural modifications of human albumin in diabetes. Diabetes & metabolism. 2012 Apr 1;38(2):171-8.
127. Cai YW, Zhang HF, Gao JW, Cai ZX, Cai JW, Gao QY, Chen ZT, Liao GH, Zeng CR, Chen N, Liu PM. Serum albumin and risk of incident diabetes and diabetic microvascular complications in the UK Biobank cohort. Diabetes & metabolism. 2023 Sep 1;49(5):101472.
128. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020 [cited 2025 Sep 11]. Available from: <https://www.R-project.org/>.
129. Shwartz-Ziv, R. and A. Armon, *Tabular data: Deep learning is not all you need*. arXiv. arXiv preprint arXiv:2106.03253, 2021.
130. IBM. What is a decision tree? [Internet]. IBM; 2024 [cited 2025 Sep 11]. Available from: <https://www.ibm.com/think/topics/decision-trees>.
131. Masters in Data Science. What is a decision tree? [Internet]. 2025 [cited 2025 Sep 11]. Available from: <https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/>.
132. Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32.

133. Wikipedia. Bootstrap aggregating [Internet]. 2025 [cited 2025 Sep 11]. Available from: https://en.wikipedia.org/wiki/Bootstrap_aggregating.
134. IBM. What is random forest? [Internet]. IBM; 2024 [cited 2025 Sep 11]. Available from: <https://www.ibm.com/think/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,denoted%20by%20the%20leaf%20node>.
135. GeeksforGeeks. Bagging vs Boosting in Machine Learning [Internet]. 2025 Jul 11 [cited 2025 Sep 11]. Available from: <https://www.geeksforgeeks.org/machine-learning/bagging-vs-boosting-in-machine-learning/>.
136. Kavlakoglu E, Rath E. What is XGBoost? [Internet]. IBM; 2024 [cited 2025 Sep 11]. Available from: <https://www.ibm.com/think/topics/xgboost>.
137. GeeksforGeeks. Bagging vs Boosting in Machine Learning [Internet]. 2025 Jul 11 [cited 2025 Sep 11]. Available from: <https://www.geeksforgeeks.org/machine-learning/bagging-vs-boosting-in-machine-learning/>.
138. Kotsiopoulos T, Sarigiannidis P, Ioannidis D, Tzovaras D. Machine learning and deep learning in smart manufacturing: The smart grid paradigm. *Computer Science Review*. 2021 May 1;40:100341.
139. Chan KY, Abu-Salih B, Qaddoura R, Al-Zoubi AM, Palade V, Pham DS, Del Ser J, Muhammad K. Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing*. 2023 Aug 7;545:126327.
140. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012;25.
141. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
142. Anaconda Documentation. *Anaconda Software Distribution* [Internet]. Anaconda Inc.; 2020 [cited 2025 Sep 11]. Available from: <https://docs.anaconda.com/>.
143. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-30.
144. Room C. Confusion matrix. *Mach Learn*. 2019;6:27.
145. J Junge MR, Dettori JR. ROC solid: Receiver operator characteristic (ROC) curves as a foundation for better diagnostic tests. *Global spine journal*. 2018 Jun;8(4):424-9.
146. Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. *PloS one*. 2018 Aug 6;13(8):e0201904.

147. MathWorks. Train sequence classification network using data with imbalanced classes. In: *Deep Learning Toolbox – Train Deep Neural Networks, Built-In Training* [Internet]. MathWorks; [cited 2025 Sep 11]. Available from: <https://uk.mathworks.com/help/deeplearning/ug/sequence-classification-using-inverse-frequency-class-weights.html>.
148. Hunter J. Random forest variable importance: Calculation, accuracy & examples [Internet]. Displayr; 2018 [cited 2025 Sep 11]. Available from: <https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/>.
149. Martínez-Taboada F, Redondo JI. Variable importance plot (mean decrease accuracy and mean decrease Gini). *PLoS One*. 2020;15(4):e0230799.
150. Ro Notast. Explaining predictions: Boosted trees post-hoc analysis (XGBoost) [Internet]. *R-Bloggers*; 2019 Oct 11 [cited 2025 Sep 11]. Available from: <https://www.r-bloggers.com/2019/10/explaining-predictions-boosted-trees-post-hoc-analysis-xgboost/>.
151. GeeksforGeeks. The role of softmax in neural networks: Detailed explanation and applications [Internet]. 2025 Jul 23 [cited 2025 Sep 11]. Available from: <https://www.geeksforgeeks.org/deep-learning/the-role-of-softmax-in-neural-networks-detailed-explanation-and-applications/>.
152. Kairys A, Pauliukienė R, Raudonis V, Ceponis J. Towards home-based diabetic foot ulcer monitoring: a systematic review. *Sensors*. 2023 Mar 30;23(7):3618.
153. Zhang J, Qiu Y, Peng L, Zhou Q, Wang Z, Qi M. A comprehensive review of methods based on deep learning for diabetes-related foot ulcers. *Frontiers in Endocrinology*. 2022 Aug 8;13:945020.
154. El-Kady AM, Abbassy MM, Ali HH, Ali MF. Advancing diabetic foot ulcer detection based on resnet and gan integration. *Journal of Theoretical and Applied Information Technology*. 2024 Mar 31;102(6):2258-68.
155. Wang C, Yu Z, Long Z, Zhao H, Wang Z. A few-shot diabetes foot ulcer image classification method based on deep ResNet and transfer learning. *Scientific Reports*. 2024 Dec 2;14(1):1-9.
156. Alkhalefah S, AlTuraiki I, Altwaijry N. Advancing Diabetic Foot Ulcer Care: AI and Generative AI Approaches for Classification, Prediction, Segmentation, and Detection. *InHealthcare* 2025 Mar 16 (Vol. 13, No. 6, p. 648). MDPI.
157. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:6000-10.
158. GeeksforGeeks. Vision Transformer (ViT) architecture [Internet]. 2025 Jul 23 [cited 2025 Sep 11]. Available from: <https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/>.

159. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.
160. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).
161. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics. 1980 Apr;36(4):193-202.
162. LeCun Y, Haffner P, Bottou L, Bengio Y. Object recognition with gradient-based learning. In Shape, contour and grouping in computer vision 1999 Oct 22 (pp. 319-345). Berlin, Heidelberg: Springer Berlin Heidelberg.
163. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 2002 Aug 6;86(11):2278-324.
164. IBM. What are convolutional neural networks? [Internet]. IBM Think; 2025 [cited 2025 Sep 11]. Available from: <https://www.ibm.com/think/topics/convolutional-neural-networks>.
165. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Cham: Springer international publishing.
166. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
167. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1-9).
168. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
169. Brunton S. Residual networks (ResNet) [Physics Informed Machine Learning] [Internet]. YouTube; 2024 [cited 2025 Sep 11]. Available from: <https://www.youtube.com/watch?v=w1UsKanMatM&t=275s>
170. Boesch G. Deep residual networks (ResNet, ResNet-50): A complete guide [Internet]. Viso.ai; 2023 Nov 14 [cited 2025 Sep 11]. Available from: <https://viso.ai/deep-learning/resnet-residual-neural->

[network/#:~:text=ResNet%2D101%20and%20ResNet%2D152&text=And%20even%20at%20increased%20network,15.3%2F19.6bn%20FLOPS\).](#)

171. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia medica*. 2012 Oct 15;22(3):276-82.

172. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *biometrics*. 1977 Mar 1:159-74.

173. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*. 2016 Jun 1;15(2):155-63.

174. Dwyer B, Nelson J, Hansen T, et al. Roboflow (Version 1.0) [software]. 2024.

175. Nebius Team. What is epoch in machine learning? Understanding its role and importance [Internet]. Nebius; 2024 Jul 23 [cited 2025 Sep 11]. Available from: <https://nebius.com/blog/posts/epoch-in-machine-learning>.

176. Edge Impulse. Epochs – Edge Impulse documentation [Internet]. [cited 2025 Sep 11]. Available from: <https://docs.edgeimpulse.com/knowledge/concepts/machine-learning/neural-networks/epochs>.

177. Bergmann D, Stryker C. What is a loss function? [Internet]. IBM Think; 2024 Jul 12 [cited 2025 Sep 11]. Available from: <https://www.ibm.com/think/topics/loss-function>.

178. Google. Classification: Accuracy, recall, precision, and related metrics | Machine Learning Crash Course [Internet]. Google Developers; 2025 Aug 25 [cited 2025 Sep 11]. Available from: <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>.

179. scikit-learn developers. *sklearn.metrics.f1_score* — scikit-learn documentation [Internet]. Version 1.7.2 [cited 2025 Sep 11]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

180. Google. Thresholds and the confusion matrix | Machine Learning Crash Course [Internet]. Google Developers; 2025 Aug 25 [cited 2025 Sep 11]. Available from: Google. *Thresholds and the confusion matrix | Machine Learning Crash Course [Internet]*. updated 2025 Aug 25; Available from: <https://developers.google.com/machine-learning/crash-course/classification/thresholding>.

181. scikit-learn developers. Tuning the decision threshold for class prediction — scikit-learn documentation [Internet]. Version 1.7.2 [cited 2025 Sep 11]. Available from: https://scikit-learn.org/stable/modules/classification_threshold.html.

182. MIC-DKFZ., *nnU-Net [Internet]*. GitHub repository. Available from: https://github.com/MIC-DKFZ/nnUNet/blob/master/documentation/how_to_use_nnunet.md#model-training.

183. Bray TJ, Chouhan MD, Punwani S, Bainbridge A, Hall-Craggs MA. Fat fraction mapping using magnetic resonance imaging: insight into pathophysiology. *The British journal of radiology*. 2017 Nov 1;91(1089):20170344.
184. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*. 2011 Mar;20(1):40-9.
185. Ranstam J, Cook JA. LASSO regression. *Journal of British Surgery*. 2018 Sep;105(10):1348-.
186. Nagpal A, Whitfield B. L1 and L2 regularization methods, explained | Built In [Internet]. Built In; 2024 Oct 3 [cited 2025 Sep 11]. Available from: <https://builtin.com/data-science/l2-regularization>.
187. Square Inc. C-index – PySurvival [Internet]. [cited 2025 Sep 11]. Available from: https://square.github.io/pysurvival/metrics/c_index.html.
188. Gallin JI, Ognibene FP, editors. *Principles and practice of clinical research*. Academic Press; 2012 Jul 10.
189. National Institute for Health and Care Research (NIHR). Rethinking your CPD: Reflections on reflection [Internet]. NIHR; 2018 Aug 22 [cited 2025 Sep 11]. Available from: <https://www.nihr.ac.uk/blog/rethinking-your-cpd-reflections-reflection>
190. Xu J, Wan CS, Ktoris K, Reijnierse EM, Maier AB. Sarcopenia is associated with mortality in adults: a systematic review and meta-analysis. *Gerontology*. 2022 Jul 27;68(4):361-76.

Appendix

Appendix 1 - R Code for Kruskal-Wallis Test

```

# Working directory
setwd("/Users/manalahmad/Desktop")

# Libraries
library(readxl)
library(dplyr)
library(ggplot2)

# Loading data from working directory
data <- read_excel("B3_CNN.xlsx")

# Normalising column names
names(data) <- tolower(names(data))

# Check for 'outcome' column
if (!"outcome" %in% names(data)) {
  stop("Missing 'Outcome' column (case-insensitive).")
}

# Define grouping variables i.e the 6 MRI features or predictors
group_vars <- setdiff(names(data), "outcome")

# Create folder to save plots
if (!dir.exists("kruskal_plots")) dir.create("kruskal_plots")

# Loop through grouping variables
for (group_var in group_vars) {
  # Skip invalid columns
  if (all(is.na(data[[group_var]])) || is.numeric(data[[group_var]])) next

  cat("\n\n-----", group_var, "-----\n")

  # Running the kruskal-wallis test
  formula_str <- as.formula(paste("outcome ~", group_var))
  result <- kruskal.test(formula_str, data = data)
  print(result)

  # Running post-hoc analysis if significant
  if (result$p.value < 0.05) {
    cat("\nPost-hoc pairwise Wilcoxon tests (Bonferroni corrected):\n")
    pairwise_result <- pairwise.wilcox.test(
      x = data$outcome,
      g = data[[group_var]],
      p.adjust.method = "bonferroni"
    )
    print(pairwise_result)
  }
}

```

Appendix 2 – Setting up Jupyter notebook workspace and downloading Python and Sci-kit learn + running logistic regression

```
[ ]: print ("Hello Manal")
```

```
[2]: !python --version
```

```
Python 3.11.7
```

```
[5]: import numpy as np
```

```
[3]: import torch
```

```
[4]: torch.__version__
```

```
[4]: '2.5.1'
```

```
data = [[1, 2], [3, 4]] ***
```

```
[11]: x_data
```

```
[11]: tensor([[1, 2],
           [3, 4]])
```

```
[1]: conda install graphviz python-graphviz
```

```
Channels:
- defaults
- pytorch
Platform: osx-arm64
Collecting package metadata (repodata.json): done
Solving environment: done
```

```
## Package Plan ##
```

```
environment location: /opt/anaconda3/envs/SAD-LL
```

```
added / updated specs:
```

```
- graphviz
- python-graphviz
```

```
The following packages will be downloaded:
```

```
The following packages will be downloaded:
```

package	build	size
cairo-1.16.0	h302bd0f_5	1.0 MB
expat-2.6.3	h313beb8_0	154 KB
font-ttf-dejavu-sans-mono-2.37	hd3eb1b0_0	335 KB
font-ttf-inconsolata-2.001	hcb22688_0	83 KB
font-ttf-source-code-pro-2.030	hd3eb1b0_0	654 KB
font-ttf-ubuntu-0.83	h8b1ccd4_0	1.5 MB
fontconfig-2.14.1	h6402c1e_3	252 KB
fonts-anaconda-1	h8fa9717_0	3 KB

```
[2]: pip install hiddenlayer
```

```
Collecting hiddenlayer
  Downloading hiddenlayer-0.3-py3-none-any.whl.metadata (703 bytes)
  Downloading hiddenlayer-0.3-py3-none-any.whl (19 kB)
Installing collected packages: hiddenlayer
Successfully installed hiddenlayer-0.3
Note: you may need to restart the kernel to use updated packages.
```

```
[2]: !pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in /opt/anaconda3/lib/python3.11/site-packages (1.2.2)
Requirement already satisfied: numpy>=1.17.3 in /opt/anaconda3/lib/python3.11/site-packages (from scikit-learn) (1.26.4)
Requirement already satisfied: scipy>=1.3.2 in /opt/anaconda3/lib/python3.11/site-packages (from scikit-learn) (1.11.4)
Requirement already satisfied: joblib>=1.1.1 in /opt/anaconda3/lib/python3.11/site-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/anaconda3/lib/python3.11/site-packages (from scikit-learn) (2.2.0)
```

```
[10]: pip install -U scikit-learn numpy pandas matplotlib jupyter seaborn
```

```
Requirement already satisfied: scikit-learn in /opt/anaconda3/lib/python3.11/site-packages (1.2.2)
Collecting scikit-learn
  Downloading scikit_learn-1.7.1-cp311-cp311-macosx_12_0_arm64.whl.metadata (11 kB)
Requirement already satisfied: numpy in /opt/anaconda3/lib/python3.11/site-packages (1.26.4)
Collecting numpy
  Downloading numpy-2.3.2-cp311-cp311-macosx_14_0_arm64.whl.metadata (62 kB)
    62.1/62.1 kB 2.7 MB/s eta 0:00:00
Requirement already satisfied: pandas in /opt/anaconda3/lib/python3.11/site-packages (2.1.4)
Collecting pandas
  Downloading pandas-2.3.2-cp311-cp311-macosx_11_0_arm64.whl.metadata (91 kB)
    91.2/91.2 kB 5.0 MB/s eta 0:00:00
Requirement already satisfied: matplotlib in /opt/anaconda3/lib/python3.11/site-packages (3.8.0)
Collecting matplotlib
  Downloading matplotlib-3.10.6-cp311-cp311-macosx_11_0_arm64.whl.metadata (11 kB)
Requirement already satisfied: jupyter in /opt/anaconda3/lib/python3.11/site-packages (1.0.0)
Collecting jupyter
  Downloading jupyter-1.1.1-py2.py3-none-any.whl.metadata (2.0 kB)
Requirement already satisfied: seaborn in /opt/anaconda3/lib/python3.11/site-packages (0.12.2)
```

```
[14]: pip install -U scipy joblib ipykernel
```

```
Requirement already satisfied: scipy in /opt/anaconda3/lib/python3.11/site-packages (1.11.4)
Collecting scipy
  Using cached scipy-1.16.1-cp311-cp311-macosx_14_0_arm64.whl.metadata (61 kB)
Requirement already satisfied: joblib in /opt/anaconda3/lib/python3.11/site-packages (1.2.0)
Collecting joblib
  Using cached joblib-1.5.2-py3-none-any.whl.metadata (5.6 kB)
Requirement already satisfied: ipykernel in /opt/anaconda3/lib/python3.11/site-packages (6.28.0)
Collecting ipykernel
  Using cached ipykernel-6.30.1-py3-none-any.whl.metadata (6.2 kB)
Requirement already satisfied: numpy<2.6,=>1.25.2 in /opt/anaconda3/lib/python3.11/site-packages (from scipy) (1.26.4)
Requirement already satisfied: appnope=>0.1.2 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.2)
Requirement already satisfied: comm=>0.1.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.2)
Requirement already satisfied: debugpy=>1.6.5 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (1.6.7)
Requirement already satisfied: ipython=>7.23.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (8.20.0)
Requirement already satisfied: jupyter-client=>8.0.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (8.6.0)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.5.0)
Requirement already satisfied: matplotlib-inline=>0.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.6)
Requirement already satisfied: nest-asyncio=>1.4 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (1.6.0)
Requirement already satisfied: packaging=>22 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (23.1)
Requirement already satisfied: psutil=>5.7 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.9.0)
Requirement already satisfied: pyzmq=>25 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (25.1.2)
Requirement already satisfied: tornado=>6.2 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (6.3.3)
Requirement already satisfied: traitlets=>5.4.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.7.1)
Requirement already satisfied: decorator in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (5.1.1)
Requirement already satisfied: jedi=>0.16 in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (0.18.1)
Requirement already satisfied: prompt-toolkit<3.1.0,>=3.0.41 in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (3.0.43)
Requirement already satisfied: pygments=>2.4.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (2.15.1)
Requirement already satisfied: stack-data in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (0.2.0)
Requirement already satisfied: pexpect>4.3 in /opt/anaconda3/lib/python3.11/site-packages (from ipython=>7.23.1->ipykernel) (4.8.0)
Requirement already satisfied: python-dateutil=>2.8.2 in /opt/anaconda3/lib/python3.11/site-packages (from jupyter-client=>8.0.0->ipykernel) (2.8.2)
Requirement already satisfied: platformdirs=>2.5 in /opt/anaconda3/lib/python3.11/site-packages (from jupyter-core!=5.0.*,>=4.12->ipykernel) (3.10.0)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /opt/anaconda3/lib/python3.11/site-packages (from jedi=>0.16->ipython=>7.23.1->ipykernel) (0.8.3)
Requirement already satisfied: ptyprocess=>0.5 in /opt/anaconda3/lib/python3.11/site-packages (from pexpect>4.3->ipython=>7.23.1->ipykernel) (0.7.0)
Requirement already satisfied: wcwidth in /opt/anaconda3/lib/python3.11/site-packages (from prompt-toolkit<3.1.0,>=3.0.41->ipython=>7.23.1->ipykernel) (0.2.5)
Requirement already satisfied: six=>1.5 in /opt/anaconda3/lib/python3.11/site-packages (from python-dateutil=>2.8.2->jupyter-client=>8.0.0->ipykernel) (1.16.0)
Requirement already satisfied: executing in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython=>7.23.1->ipykernel) (0.8.3)
Requirement already satisfied: asttokens in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython=>7.23.1->ipykernel) (2.0.5)
Requirement already satisfied: pure-eval in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython=>7.23.1->ipykernel) (0.2.2)
Using cached scipy-1.16.1-cp311-cp311-macosx_14_0_arm64.whl (20.9 MB)
Using cached joblib-1.5.2-py3-none-any.whl (308 kB)
Using cached ipykernel-6.30.1-py3-none-any.whl (117 kB)
Installing collected packages: scipy, joblib, ipykernel
  Attempting uninstall: scipy
    Found existing installation: scipy 1.11.4
    Uninstalling scipy-1.11.4:
      Successfully uninstalled scipy-1.11.4
  Attempting uninstall: joblib
    Found existing installation: joblib 1.2.0
    Uninstalling joblib-1.2.0:
      Successfully uninstalled joblib-1.2.0
```

```

Found existing installation: joblib 1.2.0
Uninstalling joblib-1.2.0:
  Successfully uninstalled joblib-1.2.0
Attempting uninstall: ipykernel
Found existing installation: ipykernel 6.28.0
Uninstalling ipykernel-6.28.0:
  Successfully uninstalled ipykernel-6.28.0
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
gensim 4.3.0 requires FuzzyTM>=0.4.0, which is not installed.
aext-panels-server 4.0.15 requires anaconda-cloud-auth, which is not installed.
Successfully installed ipykernel-6.30.1 joblib-1.5.2 scipy-1.16.1
Note: you may need to restart the kernel to use updated packages.

```

```
[1]: pip install ipykernel
```

```

Requirement already satisfied: ipykernel in /opt/anaconda3/lib/python3.11/site-packages (6.30.1)
Requirement already satisfied: appnope>=0.1.2 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.2)
Requirement already satisfied: comm>=0.1.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.2)
Requirement already satisfied: debugpy>=1.6.5 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (1.6.7)
Requirement already satisfied: ipython>=7.23.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (8.20.0)
Requirement already satisfied: jupyter-client>=8.0.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (8.6.0)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.5.0)
Requirement already satisfied: matplotlib-inline>=0.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (0.1.6)
Requirement already satisfied: nest-asyncio>=1.4 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (1.6.0)
Requirement already satisfied: packaging>=22 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (23.1)
Requirement already satisfied: psutil>=5.7 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.9.0)
Requirement already satisfied: pyzmq>=25 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (25.1.2)
Requirement already satisfied: tornado>=6.2 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (6.3.3)
Requirement already satisfied: traitlets>=5.4.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipykernel) (5.7.1)
Requirement already satisfied: decorator in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (5.1.1)
Requirement already satisfied: jedi>=0.16 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (0.18.1)
Requirement already satisfied: prompt-toolkit<3.1.0,>=3.0.41 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (3.0.43)
Requirement already satisfied: pygments>=2.4.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (2.15.1)
Requirement already satisfied: stack-data in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (0.2.0)
Requirement already satisfied: pexpect>4.3 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=7.23.1->ipykernel) (4.8.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/anaconda3/lib/python3.11/site-packages (from jupyter-client>=8.0.0->ipykernel) (2.8.2)
Requirement already satisfied: platformdirs>=2.5 in /opt/anaconda3/lib/python3.11/site-packages (from jupyter-client>=8.0.0->ipykernel) (3.10.0)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /opt/anaconda3/lib/python3.11/site-packages (from jedi>=0.16->ipython>=7.23.1->ipykernel) (0.8.3)
Requirement already satisfied: ptyprocess>=0.5 in /opt/anaconda3/lib/python3.11/site-packages (from pexpect>4.3->ipython>=7.23.1->ipykernel) (0.7.0)
Requirement already satisfied: wcwidth in /opt/anaconda3/lib/python3.11/site-packages (from prompt-toolkit<3.1.0,>=3.0.41->ipython>=7.23.1->ipykernel) (0.2.5)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.11/site-packages (from python-dateutil>=2.8.2->jupyter-client>=8.0.0->ipykernel) (1.16.0)
Requirement already satisfied: executing in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=7.23.1->ipykernel) (0.8.3)
Requirement already satisfied: asttokens in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=7.23.1->ipykernel) (2.0.5)
Requirement already satisfied: pure-eval in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=7.23.1->ipykernel) (0.2.2)
Note: you may need to restart the kernel to use updated packages.

```

```
[7]: !python -m ipykernel install --user --name=sklearn-env --display-name "Python (sklearn-env)"
```

```
Installed kernelspec sklearn-env in /Users/manalahmad/Library/Jupyter/kernels/sklearn-env
```

```
[1]: import sys
print(sys.executable)
```

```
/opt/anaconda3/bin/python
```

```
[5]: import pandas as pd
```

```

# Load the uploaded Excel file
file_path = "/Users/manalahmad/Desktop/B3_CNN.xlsx"
excel_file = pd.ExcelFile(file_path)

# Display sheet names to understand the structure of the file
sheet_names = excel_file.sheet_names
sheet_names

```

```
[5]: ['Sheet1']
```

```

[7]: # Load the data from the first sheet
df = excel_file.parse(sheet_names[0])

# Display the first few rows to understand the structure
df.head()

```

```
[7]:
```

	OM	Collection	Ulcer	Charcot	DMS	Outcome
0	Y	N	Y	N	Y	1
1	Y	N	Y	N	Y	1
2	Y	N	Y	N	Y	0
3	N	N	N	N	Y	0
4	Y	N	N	N	N	1

```

[21]: # Imports
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split

```

```
[23]: !pip show scikit-learn
      !pip show pandas
```

```
Name: scikit-learn
Version: 1.7.1
Summary: A set of python modules for machine learning and data mining
Home-page:
Author:
Author-email:
License:
Location: /opt/anaconda3/lib/python3.11/site-packages
Requires: joblib, numpy, scipy, threadpoolctl
Required-by: batchgenerators, imbalanced-learn, nnunetv2
Name: pandas
Version: 2.3.2
Summary: Powerful data structures for data analysis, time series, and statistics
Home-page:
Author:
Author-email: The Pandas Development Team <pandas-dev@python.org>
License: BSD 3-Clause License
```

```
[31]: # Convert 'Y'/'N' to binary zeros and ones
df_binary = df.copy()
label_encoder = LabelEncoder()
for col in df_binary.columns[:-1]: # excluding Outcome
    df_binary[col] = label_encoder.fit_transform(df_binary[col])

# Split into features and target
X = df_binary.drop("Outcome", axis=1)
y = df_binary["Outcome"]

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Fit logistic regression
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
report = classification_report(y_test, y_pred, output_dict=True, zero_division=0)

# Coefficients for feature importance
feature_importance = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': model.coef_[0],
    'Absolute Coefficient': abs(model.coef_[0])
}).sort_values(by='Absolute Coefficient', ascending=False)

report_df = pd.DataFrame(report).transpose()
feature_importance
feature_importance.to_csv("feature_importance.csv", index=False)
from IPython.display import FileLink
FileLink("feature_importance.csv")
```

```
[31]: feature_importance.csv
```

```
[ ]:
```

Appendix 3 - R Code for Random Forest

```
# Working directory
setwd("/Users/manalahmad/Desktop")

# Install packages
install.packages("randomForest")

# Setup
set.seed(42)

needed <- c("readxl", "dplyr", "caret", "randomForest")
to_install <- setdiff(needed, rownames(installed.packages()))
if (length(to_install)) install.packages(to_install, repos = "https://cloud.r-
project.org")

library(readxl)
library(dplyr)
library(caret)
library(randomForest)

# Loading data from working directory
data_path <- "B3_CNN.xlsx"

# Rename outcome column
outcome_col <- "Outcome"

df <- read_excel(data_path)
stopifnot(outcome_col %in% names(df))

predictor_cols <- setdiff(names(df), outcome_col)

#Cleaning and encoding
df <- df |>
  filter(!is.na(.data[[outcome_col]])) |>
  mutate(
    !!outcome_col := as.factor(.data[[outcome_col]])
  ) |>
  mutate(
    across(all_of(predictor_cols), \(x) if (is.numeric(x)) x else as.factor(x))
  )

label_map <- levels(df[[outcome_col]])

# Training and testing data splitting (stratified)
set.seed(42)
train_idx <- caret::createDataPartition(df[[outcome_col]], p = 0.80, list = FALSE)

train_data <- df[train_idx, ]
test_data <- df[-train_idx, ]

# Fit Random Forest
rf_fit <- randomForest(
  formula = reformulate(predictor_cols, outcome_col),
  data = train_data,
  ntree = 500, # number of trees
  mtry = floor(sqrt(length(predictor_cols))), # predictors tried at each split
  importance = TRUE
)

print(rf_fit)

# Prediction and evaluation
pred_lbl <- predict(rf_fit, newdata = test_data)

cm <- caret::confusionMatrix(data = pred_lbl, reference = test_data[[outcome_col]])
print(cm)

# Feature selection by importance
imp <- importance(rf_fit)
```

```

print(imp)

varImpPlot(rf_fit, n.var = min(15, length(predictor_cols)),
  main = "Random Forest Feature Importance")

# Multiclass Area under the curve - Receiver Operating Characteristics AUC-ROC for
Random Forest
needed_roc <- c("pROC", "viridis")
to_install <- setdiff(needed_roc, rownames(installed.packages()))
if (length(to_install)) install.packages(to_install, repos = "https://cloud.r-
project.org")
library(pROC)
library(viridis)

# Ensure we have probabilities for each class on the TEST set
# (randomForest returns a matrix with one column per class)
p_mat <- predict(rf_fit, newdata = test_data, type = "prob")

# Make sure columns align with outcome levels
label_map <- levels(train_data[[outcome_col]])
stopifnot(all(label_map %in% colnames(p_mat)))
p_mat <- p_mat[, label_map, drop = FALSE]

truth <- factor(test_data[[outcome_col]], levels = label_map)
num_class <- length(label_map)

# Compute multiclass ROC-AUC per class defined as 0[no amputation], 1[minor
amputation], 2[major amputation]
roc_list <- vector("list", num_class)
auc_vec <- numeric(num_class)
cols <- viridis(num_class, option = "D")

for (i in seq_len(num_class)) {
  y_i <- as.integer(truth == label_map[i]) # 1 for class i, 0 otherwise
  roc_i <- pROC::roc(response = y_i, predictor = p_mat[, i],
    quiet = TRUE, direction = "<")
  roc_list[[i]] <- roc_i
  auc_vec[i] <- pROC::auc(roc_i)
}

# Plot per-category ROC-AUC curve for each category
plot(
  roc_list[[1]],
  col = cols[1],
  lwd = 2,
  legacy.axes = TRUE,
  main = "Random Forest: Multiclass ROC",
  print.auc = FALSE
)
abline(a = 0, b = 1, lty = 3, col = "gray60")

if (num_class > 1L) {
  for (i in 2:num_class) {
    lines(roc_list[[i]], col = cols[i], lwd = 2)
  }
}

legend_labels <- sprintf("%s (AUC = %.3f)", label_map, auc_vec)

# Micro-average ROC-AUC
# Stack all multiclass curves into one ROC
y_stack <- unlist(lapply(seq_len(num_class), function(i) as.integer(truth ==
label_map[i])))
score_stack <- as.numeric(c(p_mat)) # column-wise stack to match y_stack

roc_micro <- pROC::roc(response = y_stack, predictor = score_stack,
  quiet = TRUE, direction = "<")
auc_micro <- pROC::auc(roc_micro)

```

```

# Adding the micro-average curve
lines(roc_micro, col = "black", lwd = 3, lty = 2)

# Legend with per-class/category + micro AUC
legend_labels <- c(legend_labels, sprintf("Micro (AUC = %.3f)", as.numeric(auc_micro)))
legend_cols   <- c(cols, "black")
legend("bottomright", legend = legend_labels, col = legend_cols,
      lwd = c(rep(2, num_class), 3), lty = c(rep(1, num_class), 2),
      cex = 0.9, bty = "n")

cat("Macro-average AUC:", round(mean(auc_vec, na.rm = TRUE), 3), "\n")
cat("Micro-average AUC:", round(as.numeric(auc_micro), 3), "\n")

```

Appendix 4 - R Code for Extreme Gradient Boosting

```
# Working directory
setwd("/Users/manalahmad/Desktop")

# Install packages
install.packages("caret")
install.packages("ggplot2")
install.packages("lattice")

# Libraries
library(readxl)
library(dplyr)
library(Matrix)
library(xgboost)
library(caret)
library(lattice)
library(ggplot2)

set.seed(42)

# Loading data from working directory
df <- read_excel("B3_CNN.xlsx")

# Rename outcome column
outcome_col <- "Outcome"
stopifnot(outcome_col %in% names(df))

# Predictors = all columns except outcome
predictor_cols <- setdiff(names(df), outcome_col)
stopifnot(length(predictor_cols) > 0)

# Cleaning and encoding
df <- df |>
  filter(!is.na(.data[[outcome_col]])) |>
  mutate(
    !!outcome_col := as.factor(.data[[outcome_col]])
  ) |>
  mutate(
    across(all_of(predictor_cols), \(x) if (is.numeric(x)) x else as.factor(x))
  )

label_map <- levels(df[[outcome_col]])
num_class <- length(label_map)
y_all <- as.integer(df[[outcome_col]]) - 1L

# One-hot encode predictors to ensure train/test use the same columns
form <- reformulate(term.labels = predictor_cols, intercept = FALSE)
x_all <- Matrix::sparse.model.matrix(form, data = df)
x_all <- Matrix::Matrix(x_all, sparse = TRUE)

# Stratified data split using the 'caret' library
set.seed(42)
train_idx <- caret::createDataPartition(df[[outcome_col]], p = 0.80, list = FALSE)
train_idx <- as.integer(train_idx)
test_idx <- setdiff(seq_len(nrow(df)), train_idx)

# Data Matrix objects (stores features and labels for training/prediction purposes)
X_train <- x_all[train_idx, , drop = FALSE]
X_test <- x_all[test_idx, , drop = FALSE]

dtrain <- xgb.DMatrix(data = X_train, label = y_all[train_idx])
dtest <- xgb.DMatrix(data = X_test, label = y_all[test_idx])

# Class weights for imbalance
class_counts <- table(y_all[train_idx])
class_wts <- as.numeric(max(class_counts) / class_counts)
train_wts <- class_wts[y_all[train_idx] + 1L]
xgboost::setinfo(dtrain, "weight", train_wts)
```

```

# Parameters (binary vs multiclass)
if (num_class == 2L) {
  params <- list(
    objective = "binary:logistic",
    eval_metric = c("logloss", "auc"),
    eta = 0.05,
    max_depth = 4,
    min_child_weight = 15,
    subsample = 0.8,
    colsample_bytree = 0.5,
    lambda = 15.0,
    alpha = 20.0
  )
} else {
  params <- list(
    objective = "multi:softprob",
    num_class = num_class,
    eval_metric = "mlogloss",
    eta = 0.05,
    max_depth = 4,
    min_child_weight = 15,
    subsample = 0.8,
    colsample_bytree = 0.5,
    lambda = 15.0,
    alpha = 20.0
  )
}

watchlist <- list(train = dtrain, valid = dtest)

# Train and early stopping
bst <- xgb.train(
  params = params,
  data = dtrain,
  nrounds = 1000,
  watchlist = watchlist,
  early_stopping_rounds = 30,
  verbose = 1
)

cat("Best iteration:", bst$best_iteration, "\n")

# Prediction and evaluation
if (num_class == 2L) {
  p_test <- predict(bst, dtest)
  pred_lbl <- factor(ifelse(p_test > 0.5, label_map[2], label_map[1]),
    levels = label_map)
} else {
  p_test <- predict(bst, dtest)
  p_mat <- matrix(p_test, ncol = num_class, byrow = TRUE)
  pred_idx <- max.col(p_mat) - 1L
  pred_lbl <- factor(label_map[pred_idx + 1L], levels = label_map)
}

truth <- factor(label_map[xgboost::getinfo(dtest, "label") + 1L], levels = label_map)

cm <- caret::confusionMatrix(data = pred_lbl, reference = truth)
print(cm)

# Feature importance selection by neural network
imp <- xgb.importance(model = bst, feature_names = colnames(x_all))
print(head(imp, 15))
xgb.plot.importance(imp)

# Multiclass Area under the curve - Receiver Operating Characteristics AUC-ROC)
# Requires probabilities --> multi:softprob
if (num_class < 2L) stop("ROC needs at least 2 classes.")

# Ensure probabilities available for each class for the test set

```

```

if (num_class == 2L && params$objective == "binary:logistic") {
  # Convert binary problem to a 2-column matrix for consistency
  p_test <- predict(bst, dtest)
  p_mat <- cbind(1 - p_test, p_test)
  colnames(p_mat) <- label_map
} else {
  # Multiclass
  stopifnot(params$objective == "multi:softprob")
  p_test <- predict(bst, dtest)
  p_mat <- matrix(p_test, ncol = num_class, byrow = TRUE)
  colnames(p_mat) <- label_map
}

truth <- factor(label_map[xgboost::getinfo(dtest, "label") + 1L], levels = label_map)

# Installing relevant ROC package to calculate and generate ROC curves
library(pROC)

# Building multiclass ROC. Classes defined as 0[no amputation], 1[minor amputation],
# 2[major amputation]
roc_list <- vector("list", length = num_class)
auc_vec <- numeric(num_class)

# Colour palette for ROC curve - can use RColorBrewer/hcl.colors/viridis
cols <- viridis(num_class, option = "D")

for (i in seq_len(num_class)) {
  y_i <- as.integer(truth == label_map[i])
  roc_i <- pROC::roc(response = y_i, predictor = p_mat[, i], quiet = TRUE, direction =
"<")
  roc_list[[i]] <- roc_i
  auc_vec[i] <- pROC::auc(roc_i)
}

# Plotting the first ROC category/class i.e. 0[no amputation]
plot(
  roc_list[[1]],
  col = cols[1],
  lwd = 2,
  legacy.axes = TRUE,
  main = "Multiclass ROC",
  print.auc = FALSE
)
abline(a = 0, b = 1, lty = 3, col = "gray60")

# Add remaining categories by looping
if (num_class > 1L) {
  for (i in 2:num_class) {
    lines(roc_list[[i]], col = cols[i], lwd = 2)
  }
}

# Adding the legends and labels
legend_labels <- sprintf("%s (AUC = %.3f)", label_map, auc_vec)
legend("bottomright", legend = legend_labels, col = cols, lwd = 2, cex = 0.9, bty =
"n")

# Calculating the average ROC-AUC
macro_auc <- mean(auc_vec, na.rm = TRUE)
cat("Macro-average AUC:", round(macro_auc, 3), "\n")

# Micro-average ROC-AUC
y_stack <- unlist(lapply(seq_len(num_class), function(i) as.integer(truth ==
label_map[i])))
score_stack <- as.numeric(c(p_mat)) # column-wise stack to match y_stack

roc_micro <- pROC::roc(response = y_stack, predictor = score_stack, quiet = TRUE,
direction = "<")

```

```

auc_micro <- pROC::auc(roc_micro)

# Add micro-average curve to the plot
lines(roc_micro, col = "black", lwd = 3, lty = 2)

# Update legend to include micro-average
legend_labels <- c(legend_labels, sprintf("Micro (AUC = %.3f)", auc_micro))
legend_cols <- c(cols, "black")
legend("bottomright", legend = legend_labels, col = legend_cols, lwd = c(rep(2,
num_class), 3),
      lty = c(rep(1, num_class), 2), cex = 0.9, bty = "n")

cat("Micro-average AUC:", round(as.numeric(auc_micro), 3), "\n")

```

Appendix 5 - R Code for intraclass correlation and Bland-Altman plot

```

## Working directory
setwd("/Users/manalahmad/Desktop")

##Packages
install.packages("blandr")
install.packages("lme4")
install.packages("DescTools")

##Libraries
library(readxl)
library(psych)
library(blandr)
library(writexl)
library(DescTools)

## Data
TRT_RV <- read_excel("/Users/manalahmad/Desktop/TRT.xlsx")
View(TRT_RV)
attach(TRT_RV)

##correlation
cor.test(DA, MA)

##Ensure no missing values
ratings <- na.omit(TRT_RV[, c("DA", "MA")])

##Compute all available ICC types
icc_result <- ICC(ratings, type = "all")

print(icc_result)

##Bland Altman Plot
blandr.draw(TRT_RV$DA, TRT_RV$MA)

```

Appendix 6 – R Code for analysing the nnU-Net output

```
## Working directory
setwd("/Users/manalahmad/Desktop")

##Packages
install.packages("tidyverse")
install.packages("readxl")
install.packages("knitr")
install.packages("gt")
install.packages("scales")

##Libraries
library(readxl)
library(tidyverse)
library(writexl)
library(knitr)
library(gt)
library(scales)

## Data
df <- read_excel("/Users/manalahmad/Desktop/nnunetoutput.xlsx")
View(df)
attach(df)

## DICE scores made into a tibble
df <- tibble(
  dice = c(
    0.33993002405423100,0.39399543653176400,0.8042850264805010,0.8124605080247690,
    0.6181800275753400,0.17348245865596000,0.0,0.5914731476237060,0.512283570946066,
    0.7772863966035730,0.01778035248768970,0.5638870424770030,0.24674943787271500,
    0.25687019246114100,0.08996142871481620,0.27302916782541800,0.18579559935434500,
    0.09839437746414490,0.3084243627937770,0.7729578631328290,0.4781382003666220,
    0.6304263715855450,0.16963667820069200,0.22315945745326900,0.13393826460341900,
    0.04871109877528360,0.10679864548059400,0.01654577347935330,0.0,0.0,
    0.02819831272391080,0.5314017366648950,0.6945929887106360,0.8401875418620230,
    0.4148428731762070,0.11856385558286900,0.11625032639916400,0.5859433962264150,
    0.815016591550691,0.28689065593864900,0.6255498318161500,0.8802335888211070,
    0.06290344003187670,0.5232017217117790,0.25827075491327300,0.7330309665311230,
    0.25669663984855700,0.6648648648648650,0.2180995052739660,0.7416532451138290,
    0.18181818181818200,0.8285025866345410,0.8295418740941460,0.6196252667270760,
    0.7800654542441690,0.0,0.5690985619081020,0.5492184838966050,0.7838739011821760,
    0.0
  )
)

## Summary statistics table
summary_tbl <- df %>%
  summarise(
    Mean = mean(dice),
    Median = median(dice),
    SD = sd(dice),
    Min = min(dice),
    Q25 = quantile(dice, 0.25),
    Q75 = quantile(dice, 0.75),
    Max = max(dice),
    Failures = sum(dice == 0),
    HighPerformers = sum(dice > 0.8)
  )

## Summary table using knitr::kable
kable(summary_tbl, digits = 3, caption = "Statistical Summary of DICE Scores")

## Summary table (viewable table) with gt
summary_tbl %>%
  gt() %>%
  tab_header(
    title = "Statistical Summary of DICE Scores"
  ) %>%
  fmt_number(
```

```

    columns = c(Mean, Median, SD, Min, Q25, Q75, Max),
    decimals = 3
  )

## Grouped summary statistics
df <- df %>%
  mutate(
    Performance = case_when(
      dice < 0.3 ~ "Low (<0.3)",
      dice >= 0.3 & dice <= 0.7 ~ "Medium (0.3-0.7)",
      dice > 0.7 ~ "High (>0.7)"
    )
  )

## Grouped summaries
summary_grouped <- df %>%
  group_by(Performance) %>%
  summarise(
    N = n(),
    Mean = mean(dice),
    Median = median(dice),
    SD = sd(dice),
    Min = min(dice),
    Q25 = quantile(dice, 0.25),
    Q75 = quantile(dice, 0.75),
    Max = max(dice),
    .groups = "drop"
  )

## knitr::kable package summary
kable(summary_grouped, digits = 3, caption = "Summary of DICE Scores by performance groups")

## gt package summary
summary_grouped %>%
  gt() %>%
  tab_header(
    title = "Grouped Summary of DICE Scores"
  ) %>%
  fmt_number(
    columns = c(Mean, Median, SD, Min, Q25, Q75, Max),
    decimals = 2
  )

## gt table
summary_grouped %>%
  gt() %>%
  tab_header(
    title = "Summary of DICE Scores by performance groups"
  ) %>%
  fmt_number(
    columns = c(Mean, Median, SD, Min, Q25, Q75, Max),
    decimals = 2
  )

## Bar chart by groups
df <- df %>%
  mutate(
    Performance = factor(
      Performance,
      levels = c("Low (<0.3)", "Medium (0.3-0.7)", "High (>0.7)")
    )
  )

ggplot(df, aes(x = Performance, fill = Performance)) +
  geom_bar() +
  scale_fill_manual(values = c(
    "Low (<0.3)" = "#6A0DAD", # Purple
    "Medium (0.3-0.7)" = "#BA55D3", # Magenta
  ))

```

```

    "High (>0.7)" = "#008080" # Teal
  )) +
  labs(
    title = "Counts of DICE Score Groups",
    x = "Performance Group",
    y = "Number of Cases"
  ) +
  theme_minimal()

## Plot histogram of DICE scores
df <- read_excel("/Users/manalahmad/Desktop/nnunetoutput.xlsx")
colnames(df)

## DICE Histogram
hist(
  df$dice,
  breaks = 20,
  col = "#008080",
  border = "white",
  main = "Histogram of DICE Scores",
  xlab = "DICE Score",
  ylab = "Frequency"
)

## Adding the mean and median lines
abline(v = mean(df$dice, na.rm = TRUE), col = "#6A0DAD", lwd = 2, lty = 2)
abline(v = median(df$dice, na.rm = TRUE), col = "#BA55D3", lwd = 2, lty = 2)

legend("topright", legend = c("Mean", "Median"),
      col = c("#6A0DAD", "#BA55D3"), lty = 2, lwd = 2)

## Calculating metrics to calculate the precision, recall, specificity and F1 score
df <- read_excel("/Users/manalahmad/Desktop/nnunetoutput.xlsx")
colnames(df)

df$accuracy <- (df$TP + df$TN) / (df$TP + df$TN + df$FP + df$FN)
df$precision <- ifelse((df$TP + df$FP) > 0, df$TP / (df$TP + df$FP), NA)
df$recall <- ifelse((df$TP + df$FN) > 0, df$TP / (df$TP + df$FN), NA)
df$specificity <- ifelse((df$TN + df$FP) > 0, df$TN / (df$TN + df$FP), NA)
df$F1 <- ifelse((df$precision + df$recall) > 0,
  2 * (df$precision * df$recall) / (df$precision + df$recall),
  NA)

head(df)

## Summary statistics for all performance metrics
summary(df[, c("dice", "accuracy", "precision", "recall", "specificity", "F1")])

plot(df$recall, df$precision, col = "#008080", pch = 20,
  main = "Precision versus Recall",
  xlab = "Recall", ylab = "Precision")

## DICE versus Precision graph
plot(df$dice, df$precision, col = "#6495ED", pch = 19,
  main = "DICE versus Precision",
  xlab = "DICE", ylab = "Precision")

## DICE versus Recall graph
plot(df$dice, df$recall, col = "#6A5ACD", pch = 19,
  main = "DICE versus Recall",
  xlab = "DICE", ylab = "Recall")

## DICE versus F1 graph
plot(df$dice, df$F1, col = "#4682B4", pch = 19,
  main = "DICE versus F1 Score",
  xlab = "DICE", ylab = "F1 Score")

```

Appendix 7 – nnU-NET output

DICE	tru ePo s	true Neg	fals eNe g	fals ePo s	num Pred	nu mR ef	groundTruthFile	predictedFile
------	-----------------	-------------	------------------	------------------	-------------	----------------	-----------------	---------------

0	0.33993002 405423100	621 8	1071 7538	112 25	129 23	191 41	174 43	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_570.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_570.nii.gz
1	0.39399543 653176400	820 2	9141 607	709 5	181 36	263 38	152 97	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_571.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_571.nii.gz
2	0.80428502 64805010	133 64	9417 316	177 3	473 1	180 95	151 37	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_572.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_572.nii.gz
3	0.81246050 80247690	642 9	4709 195	689	227 9	870 8	711 8	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_573.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_573.nii.gz
4	0.61818002 75753400	251 08	1016 7492	375 4	272 62	523 70	288 62	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_574.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_574.nii.gz
5	0.17348245 865596000	453 7	9389 416	183 53	248 78	294 15	228 90	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_575.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_575.nii.gz
6	0.0	0	1016 8301	247 23	305 92	305 92	247 23	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_576.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_576.nii.gz
7	0.59147314 76237060	114 87	7836 965	277 7	130 91	245 78	142 64	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_577.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_577.nii.gz
8	0.51228357 0946066	144 09	1411 3931	176 6	256 70	400 79	161 75	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_578.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_578.nii.gz
9	0.77728639 66035730	109 85	7322 752	290 3	339 2	143 77	138 88	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_579.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_579.nii.gz
10	0.01778035 248768970	399	9916 990	206 08	234 75	238 74	210 07	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_580.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_580.nii.gz
11	0.56388704 24770030	131 49	1097 6560	815 0	121 89	253 38	212 99	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_581.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_581.nii.gz
12	0.24674943 787271500	378 6	9410 283	859 3	145 22	183 08	123 79	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_582.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_582.nii.gz
13	0.25687019 246114100	402 4	8623 445	113 97	118 86	159 10	154 21	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_583.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_583.nii.gz
14	0.08996142 871481620	313 7	9370 580	372 72	261 95	293 32	404 09	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_584.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_584.nii.gz
15	0.27302916 782541800	834 5	1253 0128	109 65	334 74	418 19	193 10	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_585.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_585.nii.gz
16	0.18579559 935434500	218 7	6802 581	591 1	132 57	154 44	809 8	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_586.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_586.nii.gz
17	0.09839437 746414490	861	4701 952	745 5	832 4	918 5	831 6	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_587.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_587.nii.gz
18	0.30842436 27937770	372 7	9416 743	885 4	786 0	115 87	125 81	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_588.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_588.nii.gz
19	0.77295786 31328290	149 32	9151 336	120 6	756 6	224 98	161 38	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_589.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_589.nii.gz
20	0.47813820 03666220	149 98	9389 447	681 6	259 23	409 21	218 14	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_590.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_590.nii.gz
21	0.63042637 15855450	814 7	8370 909	616 9	338 3	115 30	143 16	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_591.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_591.nii.gz
22	0.16963667 820069200	196 1	9153 881	191 52	46	200 7	211 13	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_592.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_592.nii.gz
23	0.22315945 745326900	330 7	8624 421	102 83	127 41	160 48	135 90	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_593.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_593.nii.gz
24	0.13393826 460341900	413 3	9117 458	179 54	354 95	396 28	220 87	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_594.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_594.nii.gz
25	0.04871109 877528360	702	7049 767	146 81	127 38	134 40	153 83	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_595.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_595.nii.gz
26	0.10679864 548059400	225 5	9135 066	125 70	251 49	274 04	148 25	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_596.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_596.nii.gz
27	0.01654577 347935330	349	1149 2499	140 67	274 21	277 70	144 16	/mnt/manal/nnUNet_raw/Dataset600_F OOT/mask_test/Dataset600_597.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/ predict_image_test/Dataset600_597.nii.gz

28	0.0	0	2349 560	0	564 0	564 0	0	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_598.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_598.nii.gz
29	0.0	0	7325 623	144 09	0	0	144 09	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_599.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_599.nii.gz
30	0.02819831 272391080	244	4544 858	545 0	113 68	116 12	569 4	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_600.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_600.nii.gz
31	0.53140173 66648950	299 87	1040 2887	601 6	468 70	768 57	360 03	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_601.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_601.nii.gz
32	0.69459298 87106360	140 28	8362 244	104 9	112 87	253 15	150 77	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_602.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_602.nii.gz
33	0.84018754 18620230	125 44	7584 860	322 9	154 3	140 87	157 73	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_603.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_603.nii.gz
34	0.41484287 31762070	887 1	8878 999	531 5	197 11	285 82	141 86	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_604.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_604.nii.gz
35	0.11856385 558286900	354 5	9643 074	200 84	326 25	361 70	236 29	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_605.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_605.nii.gz
36	0.11625032 639916400	333 9	9120 934	235 72	271 95	305 34	269 11	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_606.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_606.nii.gz
37	0.58594339 62264150	186 33	1044 0793	484 7	214 87	401 20	234 80	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_607.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_607.nii.gz
38	0.81501659 1550691	164 56	8364 682	144 7	602 3	224 79	179 03	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_608.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_608.nii.gz
39	0.28689065 593864900	456 4	1019 6363	979 6	128 93	174 57	143 60	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_609.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_609.nii.gz
40	0.62554983 18161500	132 97	7310 816	220 4	137 15	270 12	155 01	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_610.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_610.nii.gz
41	0.88023358 88211070	168 82	8891 420	683	391 1	207 93	175 65	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_611.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_611.nii.gz
42	0.06290344 003187670	118 4	8876 435	168 09	184 68	196 52	179 93	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_612.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_612.nii.gz
43	0.52320172 17117790	158 02	9392 581	591 5	228 86	386 88	217 17	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_613.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_613.nii.gz
44	0.25827075 491327300	682 7	1096 4008	132 75	259 38	327 65	201 02	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_614.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_614.nii.gz
45	0.73303096 65311230	140 61	1072 3601	412 6	611 6	201 77	181 87	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_615.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_615.nii.gz
46	0.25669663 984855700	406 8	1072 0277	101 33	134 26	174 94	142 01	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_616.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_616.nii.gz
47	0.66486486 48648650	177 12	1176 0912	163 8	162 18	339 30	193 50	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_617.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_617.nii.gz
48	0.21809950 52739660	467 3	9661 149	148 92	186 14	232 87	195 65	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_618.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_618.nii.gz
49	0.74165324 51138290	170 38	7835 412	179 2	100 78	271 16	188 30	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_619.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_619.nii.gz
50	0.18181818 181818200	911	3472 490	369 7	450 2	541 3	460 8	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_620.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_620.nii.gz
51	0.82850258 66345410	116 91	9158 509	717	412 3	158 14	124 08	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_621.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_621.nii.gz
52	0.82954187 40941460	131 64	8632 178	136 4	404 6	172 10	145 28	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_622.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_622.nii.gz
53	0.61962526 67270760	105 99	9151 428	489 0	812 3	187 22	154 89	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_623.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_623.nii.gz
54	0.78006545 42441690	188 30	9145 592	206 1	855 7	273 87	208 91	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_624.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_624.nii.gz
55	0.0	0	5112 022	416 4	381 4	381 4	416 4	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_625.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_625.nii.gz

56	0.56909856 19081020	649 0	5881 922	192 8	790 0	143 90	841 8	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_626.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_626.nii.gz
57	0.54921848 38966050	725 6	5879 073	260 1	931 0	165 66	985 7	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_627.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_627.nii.gz
58	0.78387390 11821760	387 9	5302 398	536	160 3	548 2	441 5	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_628.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_628.nii.gz
59	0.0	0	8086 512	201 52	198 00	198 00	201 52	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_629.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_629.nii.gz
60	0.56994675 54587160	185 72	9128 441	601 0	220 17	405 89	245 82	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_630.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_630.nii.gz
61	0.04227122 905977100	124 4	1173 8866	181 59	382 11	394 55	194 03	/mnt/manal/nnUNet_raw/Dataset600_FOOT/mask_test/Dataset600_631.nii.gz	/mnt/manal/nnUNet_raw/Dataset600_FOOT/predict_image_test/Dataset600_631.nii.gz

Appendix 8 – R code for logistic regression and sarcopenia severity

```
# Working directory
setwd("/Users/manalahmad/Desktop")

#Install packages
install.packages("readxl")
install.packages("dplyr")

#Load packages
library(readxl)
library(dplyr)

#Load dataset
cc_data <- read_excel("cc.xlsx")

#clean data and handle missing values
cc_clean <- cc_data |>
  mutate(
    outcome = factor(outcome),
    sarcopenia = factor(sarcopenia)
  ) |>
  filter(
    !is.na(outcome),
    !is.na(sarcopenia)
  )

# Logistic regression
logit_model <- glm(
  outcome ~ sarcopenia,
  data = cc_clean,
  family = binomial
)

# Output with odds ratio and confidence interval
summary(logit_model)

exp(
  cbind(
    OR = coef(logit_model),
    confint(logit_model)
  )
)
```

Appendix 9 – R code for logistic regression and pseudo fat fraction

```

# Working directory
setwd("/Users/manalahmad/Desktop")

# Setup
library(readxl)
library(dplyr)

# Loading data from the working directory
cc_data <- read_excel("cc.xlsx")

# clean data and handle missing values
cc_clean <- cc_data |>
  mutate(
    outcome = factor(outcome)
  ) |>
  filter(
    !is.na(outcome),
    !is.na(fat_fraction)
  )

# Logistic regression
logit_model <- glm(
  outcome ~ fat_fraction,
  data = cc_clean,
  family = binomial
)

# Summary and odds ratios
summary(logit_model)

exp(
  cbind(
    OR = coef(logit_model),
    confint(logit_model)
  )
)

# visualise data
library(ggplot2)

# prediction data across pseudo fat fractions
prediction_data <- cc_clean |>
  select(fat_fraction) |>
  distinct() |>
  arrange(fat_fraction)

# Add predicted probabilities and confidence intervals
prediction <- predict(
  logit_model,
  newdata = prediction_data,
  type = "link",
  se.fit = TRUE
)

prediction_data <- prediction_data |>
  mutate(
    fit_link = prediction$fit,
    se_link = prediction$se.fit,
    predicted_prob = plogis(fit_link),
    lower = plogis(fit_link - 1.96 * se_link),
    upper = plogis(fit_link + 1.96 * se_link)
  )

# Plots
ggplot(prediction_data, aes(x = fat_fraction, y = predicted_prob)) +
  geom_line(color = "#08519c", size = 1.2) +
  geom_ribbon(

```

```

aes(ymin = lower, ymax = upper),
fill = "#9ecae1",
alpha = 0.4
) +
labs(
  title = "Predicted probability of amputation by pseudo fat fraction values",
  x = "Fat Fraction",
  y = "Predicted Probability"
) +
theme_minimal()
levels(cc_clean$outcome)

```

Appendix 10 – R code for feature importance with patient metadata using LASSO, Ridge regression and Random Forest

```

# Working directory
setwd("/Users/manalahmad/Desktop")

# Install packages
install.packages("readxl")
install.packages("dplyr")
install.packages("tidyr")
install.packages("mice")
install.packages("randomForest")
install.packages("glmnet")

# Setup
library(readxl)
library(dplyr)
library(tidyr)
library(mice)
library(glmnet)

cc_data <- read_excel("cc.xlsx")

# Convert categorical variables to factors
cc_data <- cc_data %>%
  mutate(across(c(sex, ethnicity, smoking, neuropathy, pad,
                  retinopathy, cva, renal, cvd, infection, sarcopenia),
              as.factor))

# Imputation settings using mice and predictive mean matching
set.seed(123)
imputed <- mice(cc_data, m = 5, method = "pmm")
data_imputed <- complete(imputed, 1)

data_imputed <- data_imputed %>% select(-Record_id)
data_imputed <- na.omit(data_imputed)

data_imputed$outcome <- as.factor(data_imputed$outcome)
stopifnot(!any(is.na(data_imputed$outcome)))

# Model matrix
x <- model.matrix(outcome ~ ., data = data_imputed)[, -1]
y <- data_imputed$outcome

# LASSO Regression with 10 fold cross validation. Change alpha value to zero for ridge
regression
set.seed(123)
cv_lasso <- cv.glmnet(
  x, y,
  family = "binomial",
  alpha = 1,
  nfolds = 10,
  type.measure = "deviance"
)

# Visual of cross validation error with lasso
plot(cv_lasso)

# Final model with best lambda (automates the ideal lambda value)
best_lambda <- cv_lasso$lambda.min
lasso_model <- glmnet(x, y, family = "binomial", alpha = 1, lambda = best_lambda)

# Print findings
selected_vars <- coef(lasso_model)
print(selected_vars[selected_vars != 0])
#####

# Setup
library(readxl)
library(dplyr)
library(mice)

```

```

library(glmnet)
library(randomForest)

setwd("/Users/manalahmad/Desktop")
cc_data <- read_excel("cc.xlsx")

# Convert categorical variables to factors
cc_data <- cc_data %>%
  mutate(across(c(sex, ethnicity, smoking, neuropathy, pad,
                  retinopathy, cva, renal, cvd, infection, sarcopenia),
              as.factor))

# imputation
set.seed(123)
imputed <- mice(cc_data, m = 5, method = "pmm", seed = 123)
data_imputed <- complete(imputed, 1)

data_imputed <- data_imputed %>% select(-Record_id)
data_imputed <- na.omit(data_imputed)
data_imputed$outcome <- as.factor(data_imputed$outcome)

# glmnet model matrix
x <- model.matrix(outcome ~ ., data = data_imputed)[, -1]
y <- data_imputed$outcome

# LASSO Regression ( $\alpha = 1$ )
set.seed(123)
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
lasso_model <- glmnet(x, y, family = "binomial", alpha = 1, lambda =
cv_lasso$lambda.min)
cat("\nLASSO selected variables:\n")
print(coef(lasso_model)[coef(lasso_model) != 0])

# Ridge Regression ( $\alpha = 0$ )
set.seed(123)
cv_ridge <- cv.glmnet(x, y, family = "binomial", alpha = 0)
ridge_model <- glmnet(x, y, family = "binomial", alpha = 0, lambda =
cv_ridge$lambda.min)
cat("\nTop Ridge coefficients:\n")
print(coef(ridge_model)[order(abs(coef(ridge_model)), decreasing = TRUE)][1:10])

# Random Forest
set.seed(123)
rf_model <- randomForest(outcome ~ ., data = data_imputed, importance = TRUE, ntree =
500)
cat("\nRandom Forest variable importance:\n")
print(importance(rf_model))
varImpPlot(rf_model)

ridge_coefs <- coef(ridge_model)
top_vars <- as.matrix(ridge_coefs)[order(abs(ridge_coefs[,1]), decreasing = TRUE)
[1:10], , drop = FALSE]
print(top_vars)

rownames(coef(lasso_model))

```

Appendix 11 – R code for Cox proportional hazard

```
## Working directory
setwd("/Users/manalahmad/Desktop")

##Packages
install.packages("readxl")
install.packages("survival")
install.packages("survminer")

# Load required packages
library(survival)
library(survminer)

# Read the dataset
data <- read.csv("cox_data_cleaned.csv")

# Convert relevant variables to factors
data$sex <- as.factor(data$sex)
data$smoking_status <- as.factor(data$smoking_status)
data$neuropathy <- as.factor(data$neuropathy)
data$pad <- as.factor(data$pad)
data$retinopathy <- as.factor(data$retinopathy)
data$cardiovascular_disease <- as.factor(data$cardiovascular_disease)
data$sarcopenia_severity <- as.factor(data$sarcopenia_severity)

# Fit the Cox model
cox_model <- coxph(Surv(time_to_event, event) ~ age + sex + hba1c + crp + weight +
  smoking_status + neuropathy + pad + retinopathy +
  cardiovascular_disease + sarcopenia_severity + fat_fraction,
  data = data)

# View summary
summary(cox_model)
```

Appendix 12 – PhD related funding

- Imperial College Healthcare NHS Trust - Imaging Research Committee – “Pump Priming” Research Grant £10,000 - November 2023
- Lord Cairns Fund - £500 - October 2024
- Venous Symposium physician in training scholarship - [Meeting fees and accommodation - £1240] - February 2025
- Biobank early researcher ‘getting started’ platform credits - £1000 - May 2025

Appendix 13 – Publications during the PhD

Directly related to the PhD

- Ahmad M, Patel K, Davies AH, Shalhoub J. Noninvasive optical methods to assess tissue perfusion in patients with peripheral arterial disease and diabetes mellitus: A scoping review and discussion. *JVS-Vascular Insights*. 2024 Jan 1;2:100034.
- Ahmad M, Tan M, Bergman H, Shalhoub J, Davies A. The use of artificial intelligence in three-dimensional imaging modalities and diabetic foot disease: A systematic review. *JVS-Vascular Insights*. 2024 Jan 1;2:100057.
- Ahmad M, Mohamed A, Amiras D, Siracusa F, Shalhoub J, Davies AH. Sarcopenia in the foot on magnetic resonance imaging in patients with diabetes mellitus—a systematic review. *Clinical Diabetes and Endocrinology*. 2024 Oct 25;10(1):31.

Other publications during the PhD

- Ahmad M, Tan M, Abuarqoub M, Patel K, Siracusa F, Shalhoub J, Davies AH. Carotid webs: A review of diagnosis and management strategies in current literature. *J. Vasc. Soc. Great Br. Irel*. 2025;4:99-110.
- Soh CL, Pinho C, Ahmad M, Davies AH. Duplex Ultrasound Scan Can Prevent Leg Amputation in Severe Postthrombotic Syndrome: A Case Report. *Case Reports in Vascular Medicine*. 2025;2025(1):5555757.
- Ahmad M, El-Boghdady M. International Women’s Day: insights from surgeons in training. *RCSEd Surgeon’s News*. 2025 Mar;:48–49.

- Tran MT, Ahmad M, Patel K, Argyriou O, Davies A, Shalhoub J. Comparing virtual reality and simulation to teach the assessment and management of acute surgical scenarios: A pilot study. Health Science Reports. 2024 Jul;7(7):e2245.
- Crisp J, Ahmad M, Crockett S, Mohamed A, Hamady M, Bernstein O, Shalhoub J. Spontaneous bilateral superficial femoral artery pseudoaneurysms and a unilateral posterior tibial artery aneurysm in an immunocompromised patient. Clinical Case Reports. 2024 Mar;12(3):e8686.
- Patel KD, Ahmad M, Tan M, Onida S, Davies A. A systematic review evaluating imaging techniques to determine chronicity of deep vein thrombosis. Phlebology. 2025 Jul 16:02683555251358915.

Appendix 14 – Oral presentations from PhD related work

Meeting	Title
• European Congress of Radiology 2026 [Vienna, Austria]	Identifying predictors for limb loss and death in patients with diabetic foot disease by combining patient metadata and MRI based imaging findings.
• European Congress of Radiology 2026 [Vienna, Austria] • Surgical Research Society 2026 [London, UK]	Using vision transformer to classify sarcopenia severity on two-dimensional magnetic resonance images of the foot in patients with diabetic foot disease.
• European Congress of Radiology 2026 [Vienna, Austria]	Using deep learning models (AI) on diabetic foot disease related MRI reports of the foot to identify salient predictors for limb loss.

• Surgical Research Society 2026 [London, UK]	
• Surgical Research Society 2026 [London, UK]	Using nnU-Net to segment MRI images of the foot in patients with diabetic foot disease.
• Surgical Research Society 2026 [London, UK]	Using Residual Network (ResNet) convolutional neural networks to classify sarcopenia severity on two-dimensional magnetic resonance images of the foot in patients with diabetic foot disease.
• Charing Cross Symposium 2024 [London, UK]	The use of artificial intelligence in three-dimensional imaging modalities and diabetic foot disease - a systematic review.
• Surgical Research Society 2024 [Cambridge, UK]	Evaluating the role and efficacy of virtual reality and simulation in teaching acute and trauma scenarios – a systematic review
• International Symposium on diabetic foot 2023 [The Hague, Netherlands]	Non-invasive optical methods to assess tissue perfusion in patients with peripheral arterial disease and diabetes - a review.
• Charing Cross Symposium 2023 [London, UK]	Non-invasive optical methods to assess tissue perfusion in patients with peripheral arterial disease and diabetes - a systematic review.

• Charing Cross Symposium 2023 • [London, UK]	Lower Limb sarcopenia on magnetic resonance imaging in patients with diabetes - a systematic review.
--	--

Appendix 15 – Poster presentations from PhD related work

Meeting	Title
• European Congress of Radiology 2026 • [Vienna, Austria]	Using Residual Network (ResNet) convolutional neural networks to classify sarcopenia severity on two-dimensional magnetic resonance images of the foot in patients with diabetic foot disease.
• European Congress of Radiology 2026 • [Vienna, Austria]	Using nnU-Net to segment MRI images of the foot in patients with diabetic foot disease.
• Vascular Society Meeting 2022 • [Belfast, UK] • American Venous Forum 2024 • [Florida, USA]	Non-invasive optical methods to assess tissue perfusion in patients with peripheral arterial disease and diabetes - a systematic review.
• Vascular Society Meeting 2022 • [Belfast, UK]	Lower Limb sarcopenia on magnetic resonance imaging in patients with diabetes - a systematic review.

• International Symposium on diabetic foot 2023 [The Hague, Netherlands] • Surgical Research Society 2024 [Cambridge, UK]	
• Surgical Research Society 2024 [Cambridge, UK]	The use of artificial intelligence in three-dimensional imaging modalities and diabetic foot disease - a systematic review.