

AI Evaluations of Sarcoma From Magnetic Resonance Imaging

Arnar Evgení Gunnarsson

Thesis of 30 ECTS credits submitted to the Department of Engineering at Reykjavík University in partial fulfillment of the requirements for the degree of Philosophy Doctor

October 17, 2025

Thesis Committee:

Paolo Gargiulo, Supervisor
Professor, Reykjavik University, Iceland

Halldór Jónsson Jr., Supervisor
Professor Emeritus, University of Iceland, Iceland

Marco Cristani, Examiner
Professor, Università di Verona, Italy

Arsenio Cutolo, Examiner
Associate Professor, Università Degli Studi Napoli Federico II, Italy

Copyright © 2025 Arnar Evgení Gunnarsson 

This document is licensed under a Creative Commons Attribution 4.0 (CC BY-NC-SA) (<http://creativecommons.org/licenses/by-nc-nd/4.0/>) license. However, this license does not cover the Reykjavik University logo (word- and figurative mark). The Reykjavik University logo is a registered trademark and remains the intellectual property of Reykjavik University. Any reuse, redistribution, or modification of the Reykjavik University logo is strictly prohibited without explicit written permission from Reykjavik University.

You may copy and redistribute the material in any medium or format, provide appropriate credit, link to the license and indicate what changes you made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You may not use the material for commercial purposes. If you remix, transform or build upon the material, you may not distribute the modified material.

The images or other third party material in this thesis are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Contents

Contents	iii
List of Figures	vii
List of Tables	xiii
1 Introduction	3
1.1 Sarcomas In General	3
1.2 Sarcomas In the Study	4
1.3 Symptoms	8
1.4 Diagnostics, Grading and Staging	10
1.5 Novel Diagnostic Techniques	11
1.6 Treatment	12
1.7 Directions of Treatment Improvements	13
2 Toward New Assessment in Sarcoma Identification and Grading Using Artificial Intelligence Techniques	17
2.1 Introduction	18
2.2 Methods	19
2.2.1 Data	20
2.2.2 Segmentation	21
2.2.3 Data Pre-Processing	22
2.2.4 Feature Extraction	22
2.2.5 Feature Selection	24
2.2.6 Binary Classification	24
2.2.7 Grade Classification	26
2.3 Results	27
2.3.1 Binary Classification: Healthy–Pathological	27
2.3.1.1 Original Features	28
2.3.1.2 Wavelet Features	29
2.3.1.3 Combined Features	30

2.3.2	Grade Classification: Sarcoma Grade	31
2.3.2.1	Original Features	31
2.3.2.2	Wavelet Features	33
2.3.2.3	Combined Features	34
2.4	Discussion	36
2.4.1	Dataset Complexity	37
2.4.2	Image Transform Algorithms	39
2.4.3	Dataset Imbalance	39
2.4.4	Clinical Sarcoma Grading	39
2.5	Conclusions	40
3	An Explainable Radiomics-Based Classification Model for Sarcoma Diagnosis	41
3.1	Introduction	42
3.2	Materials and Methods	43
3.2.1	Dataset	43
3.2.1.1	Sarcoma Cases	43
3.2.1.2	Healthy Cases	44
3.2.2	Methods	45
3.2.2.1	Segmentation and Features Extraction	46
3.2.2.2	Hyperparameter Tuning and Random Forest classification	49
3.2.2.3	Model Explainability	50
3.3	Results	51
3.3.1	Results of the Hyperparameter Tuning	51
3.3.2	Final Random Forest classification	51
3.3.3	Model Explainability	53
3.3.3.1	Feature Importance Results	53
3.3.3.2	Local Interpretability with LIME	54
3.3.4	Performance comparison	55
3.4	Discussion	57
3.5	Conclusions	61
4	A Machine Learning Approach to Sarcoma Diagnostics and Staging	63
4.1	Introduction	63
4.1.1	Medical Imaging	63
4.1.2	Radiomics	64
4.1.2.1	General Workflow	64
4.1.2.2	Feature Classes	65
4.1.2.3	Image Transforms	65
4.1.2.4	Quantitative Capacity and Data Integration	65
4.1.2.5	Applications in Oncology and Beyond	66

4.1.2.6	Challenges and Considerations	66
4.1.2.7	Emergence of Machine Learning	67
4.1.2.8	Empirical Applications of Radiomics	67
4.1.3	Tumor Habitat Analysis	67
4.1.3.1	General Workflow	68
4.1.3.2	Applications	69
4.1.4	Machine Learning	70
4.1.4.1	ML Pipeline	70
4.1.4.2	ML in Sarcoma Research	71
4.1.5	Research Aims	73
4.2	Materials and Methods	74
4.3	DICOM Image Dataset	74
4.3.1	Segmentation	76
4.3.2	Image and Mask Post-Processing	76
4.3.2.1	Normalization	77
4.3.2.2	SLIC	77
4.3.3	Habitat Generation	78
4.3.4	Feature Extraction	79
4.3.4.1	Radiomic Feature Extraction	79
4.3.4.2	Habitat Feature Extraction	82
4.3.5	Learning Models	83
4.3.6	Feature Selection	84
4.3.6.1	SHAP	84
4.3.6.2	Recursive Feature Elimination	85
4.3.7	Model Training and Testing	87
4.3.7.1	Tissue Classification	91
4.3.7.2	Grade Classification	92
4.3.7.3	Sub-Group Classification	93
4.4	Results	95
4.4.1	DICOM Image Dataset and Segmentation	95
4.4.2	Image and Mask Post-Processing	97
4.4.3	Feature Extraction and RFE	99
4.4.3.1	Radiomic Feature Space	99
4.4.3.2	Habitat Feature Space	107
4.4.4	Model Training and Testing	108
4.4.4.1	Tissue Classification	108
4.4.4.2	Grade Classification	113
4.4.4.3	Sub-Group Classification	115
4.4.4.4	Habitat Classification	118
4.5	Discussion	121
4.5.1	DICOM Image Database and Segmentation	121
4.5.2	Image And Mask Post-Processing	123

4.5.3	Habitat Generation	124
4.5.4	Feature Extraction	125
4.5.5	Feature Selection	125
4.5.6	Model Training and Testing	126
4.5.6.1	Tissue Classification	126
4.5.6.2	Grade Classification	128
4.5.6.3	Sup-Group Classification	129
4.5.6.4	Habitat Classification	130
4.5.6.5	Final Take	131
4.5.7	Sarcoma Manager	132
4.5.8	Roadmap to Deployment	139
4.5.9	Improvements	140
4.5.10	Notable Attempts	142
5	Conclusion and Future Directions	145
	Bibliography	149
A	Image Transforms	167
B	Simple Linear Iterative Clustering (SLIC)	173
C	Results	175
C.1	Figures of Results	175
C.1.1	Tissue Classification	175
C.1.2	Grade Classification	191
C.1.3	Sub-Group Classification	199
C.2	Tables of Results	215
C.2.1	Tissue Classification	215
C.2.2	Grade Classification	219
C.2.3	Sub-Group Classification	221

C.2.4	Habitat Classification	223
-------	----------------------------------	-----

List of Figures

1.1	MRI slices of different sarcoma sub-types from the IceSG dataset with tumor regions shown by the red bounding box: (a) CS (b) OS (c) DFSP (d) STS (e) ES (f) GCTB.	6
1.2	Examples of biopsy images taken from different sarcoma sub-types.	8
2.1	Visual representation of the steps in the workflow.	20
2.2	Confusion matrix setup. (a) For binary classification, true negative (TN) indicates tissue correctly predicted as healthy, true positive (TP) indicates tissue correctly predicted as pathological, false negative (FN) indicates pathological tissue predicted as healthy and false positive (FP) indicates pathological tissue predicted as healthy. The micro average displays the number of samples classified in integers, and macro average displays classification in ratios. (b) For grade classification, true grade (TG_i) is the number of samples belonging to class i and predicted as such, and false grade ($FG_{i,j}$) corresponds to samples predicted as grade i but belonging to grade j . The micro average shows the total number of classified data points, and the macro average shows number of classified patients.	26
2.3	MRI slices: (a) A case of OS where the machine uses specific settings and contrast media to brighten the tumor region and suppress fat response. (b) Another patient (with liposarcoma in this case) where the tumor region is dark in contrast and fat tissue is not suppressed.	37
2.4	An example of how healthy and pathological tissue can overlap in terms of appearance. The left two images show healthy tissue, and the rightmost image shows a tumor region.	38
2.5	Variance in image quality. (a) A grainy slice with relatively high noise. (b) A corrupted image. (c) An example of uneven lighting distribution.	38
3.1	Example of T1-weighted MRI scan variants.	45
3.2	Example of T2-weighted MRI scan variants.	46

3.3	Example of tumor and healthy tissue segmentation in axial T1-weighted MRI of the thighs. (A) shows the original MRI with a unilateral soft tissue sarcoma. (B) displays the manual segmentation of the tumor (blue), while (C) shows the corresponding healthy tissue regions (light blue) from the contralateral limb, used as negative samples.	47
3.4	Workflow: segmentation of MRI scans, radiomic feature extraction, Random Forest classification after NCV, and model explainability using Feature Importance and LIME.	47
3.5	Confusion matrix for classification on the hold-out test set.	53
3.6	LIME Explanation: True = 0, Predicted = 0 (Correct Classification).	55
3.7	LIME Explanation: True = 1, Predicted = 1 (Correct Classification).	56
3.8	LIME Explanation: True = 0, Predicted = 1 (Misclassification).	56
3.9	LIME Explanation: True = 1, Predicted = 0 (Misclassification).	57
4.1	An example of image transforms applied to a slice of an MRI scan of a sarcoma tumor.	66
4.2	The general radiomics feature extraction method. (a) displays the raw image (b) the highlighted areas represent: healthy region in blue, sarcoma in green and the edema in red (c) the numerical results acquired from extracting 4 features from the different highlighted areas.	68
4.3	The overall workflow methodology in the research, depicting a high-level description.	74
4.4	The general workflow for generating habitats.	78
4.5	The overall radiomic feature extraction protocol.	79
4.6	Workflow for extracting features from the generated habitats.	82
4.7	The training workflow for all tasks. The only differences between tasks are what the prediction aim is and the initial full feature data table: extracted features from the full sarcoma mask, extracted features from the SLIC segments and statistical features extracted from the MSI matrices of generated habitats. SHAP ranks features, feature elimination removes the bottom ranked features, remaining features are trained and tested on for all models, this is done recursively until few features remain. Result metrics are calculated at each recursion iteration.	86
4.8	A more detailed workflow of the model training and testing component (see figure 4.7 for an overview). Patients are split into folds, based on criteria set by the training task (sarcoma vs. healthy, grade and subgroup classifications)	87
4.9	A depiction of the tests made for classifying healthy vs. pathological tissue.	92
4.10	A depiction of the tests made for classifying sarcoma grade.	93
4.11	A depiction of the tests made for classifying sub-groups.	93

4.12	(a) An example of a corrupt image in the IceSG dataset (b) the mask of the sarcoma and healthy regions on a corrupt image. (c) Another example of a corrupt image (d) The resulting mask of the second corrupt image.	96
4.13	(a) An example of a bone sarcoma, where the border of the tumor is very unclear (b) The resulting mask (c) Another example of a sarcoma appearing very similar to surrounding anatomy (d) The resulting mask.	97
4.14	Examples of other difficulties in the segmentation process. (a) An MRI slice that seems to be a mix of two slices (b) Visual artefacts in an image slice (c) An example of a relatively noisy and grainy image (d) Uneven lighting or response throughout the image.	98
4.15	An example of a raw image slice and resulting SLIC processed super pixels on masks. (a) displays the original raw image slice (b) displays the masks; green is healthy, red is sarcoma and blue is edema (c) the resulting SPs.	98
4.16	An example of a generated habitat (a) the original raw image slice of only the sarcoma (b) the SPs generated from the SLIC process, with an approximate volume of $30mm^3$, displayed in color due to the large number of SPs generated in this step (c) the resulting habitat generated.	99
4.17	AUC-ROC score against the number of feature applied for each model using the SvAH dataset.	109
4.18	ROC curve of the highest scoring run: logistic regression at 100 features for tissue classification using the SvAH dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	109
4.19	AUC-ROC score against the number of feature applied for each model using the SvAH-SP dataset.	110
4.20	ROC curve of the highest scoring run: CatBoost at 90 features for tissue classification using the SvAH-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	110
4.21	AUC-ROC score against the number of feature applied for each model using the SvCH dataset.	111
4.22	ROC curve of the highest scoring run: logistic regression at 90 features for tissue classification using the SvCH dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	111
4.23	AUC-ROC score against the number of feature applied for each model using the SvCH-SP dataset.	112
4.24	ROC curve of the highest scoring run: logistic regression at 100 features for tissue classification using the SvCH-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	112
4.25	AUC-ROC score against the number of feature applied for each model for grade classification.	113

4.26	ROC curve of the highest scoring run: SGD Classifier at 60 features for grade classification using the grade dataset. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title. .	114
4.27	AUC-ROC score against the number of feature applied for each model for grade classification using SPs.	114
4.28	ROC curve of the highest scoring run: ridge classifier at 10 features for grade classification using the grade-SP dataset. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title. .	115
4.29	AUC-ROC score against the number of feature applied for each model for sub-group classification.	116
4.30	ROC curve of the highest scoring run: ridge classifier at 149 features for sub-group classification using the SG dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	116
4.31	AUC-ROC score against the number of feature applied for each model for sub-group classification using SPs.	117
4.32	ROC curve of the highest scoring run: SGD classifier at 100 features for sub-group classification using the SG-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.	117
4.33	AUC-ROC score against the number of feature applied for each model for grade classification using habitats.	118
4.34	ROC curve of the highest scoring run: SGD classifier at 14 features for grade classification using habitats. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title.	119
4.35	AUC-ROC score against the number of feature applied for each model for sub-group classification using habitats.	119
4.36	ROC curve of the highest scoring run: SGD classifier at 18 features for sub-group classification using habitats. The 95% confidence interval of the AUC-ROC score is shown in the legend.	120
4.37	MRI slices: (a) shows a case of OS where the machine uses specific settings and contrast media to brighten the tumor region and suppress fat response. (b) displays another patient, liposarcoma in this case, where the tumor region is dark in contrast and fat tissue is not suppressed. .	122
4.38	An example of how healthy and pathological tissue can overlap in terms of appearance. Left two images show healthy tissue, right most image shows a tumor region.	123
4.39	Variance in image quality: (a) Shows a grainy slice, with relatively high noise (b) Displays a corrupted image (c) An example of uneven lighting distribution.	123
4.40	Sarcoma Manager.	133
4.41	Example 1 of a patient viewed in the Sarcoma Manager program. . . .	135
4.42	Example 2 of a patient viewed in the Sarcoma Manager program. . . .	135
4.43	Example 3 of a patient viewed in the Sarcoma Manager program. . . .	136

4.44	Example 4 of a patient viewed in the Sarcoma Manager program. . . .	136
4.45	Example 5 of a patient viewed in the Sarcoma Manager program. . . .	137
4.46	Example 6 of a patient viewed in the Sarcoma Manager program. . . .	137
4.47	The roadmap towards deployment.	139
A.1	An example of all transforms applied to a single medical image slice of sarcoma.	171
B.1	Example of the product of the SLIC algorithm applied to a medical image	174
C.1	Tissue classification accuracy: SvAH dataset	175
C.2	Tissue classification macro F1-score: SvAH dataset	176
C.3	Tissue classification macro precision: SvAH dataset	176
C.4	Tissue classification macro recall: SvAH dataset	177
C.5	Tissue classification micro F1-score: SvAH dataset	177
C.6	Tissue classification micro precision: SvAH dataset	178
C.7	Tissue classification micro recall: SvAH dataset	178
C.8	Tissue classification accuracy: SvAH-SP dataset	179
C.9	Tissue classification balanced accuracy: SvAH-SP dataset	179
C.10	Tissue classification macro F1-score: SvAH-SP dataset	180
C.11	Tissue classification macro precision: SvAH-SP dataset	180
C.12	Tissue classification macro recall: SvAH-SP dataset	181
C.13	Tissue classification micro F1-score: SvAH-SP dataset	181
C.14	Tissue classification micro precision: SvAH-SP dataset	182
C.15	Tissue classification micro recall: SvAH-SP dataset	182
C.16	Tissue classification accuracy: SvCH dataset	183
C.17	Tissue classification balanced accuracy: SvCH dataset	183
C.18	Tissue classification macro F1-score: SvCH dataset	184
C.19	Tissue classification macro precision: SvCH dataset	184
C.20	Tissue classification macro recall: SvCH dataset	185
C.21	Tissue classification micro F1-score: SvCH dataset	185
C.22	Tissue classification micro precision: SvCH dataset	186
C.23	Tissue classification micro recall: SvCH dataset	186
C.24	Tissue classification accuracy: SvCH-SP dataset	187
C.25	Tissue classification balanced accuracy: SvCH-SP dataset	187
C.26	Tissue classification macro F1-score: SvCH-SP dataset	188
C.27	Tissue classification macro precision: SvCH-SP dataset	188
C.28	Tissue classification macro recall: SvCH-SP dataset	189
C.29	Tissue classification micro F1-score: SvCH-SP dataset	189
C.30	Tissue classification micro precision: SvCH-SP dataset	190
C.31	Tissue classification micro recall: SvCH-SP dataset	190
C.32	Grade classification accuracy: G dataset	191

C.33	Grade classification balanced accuracy: G dataset	191
C.34	Grade classification macro F1-score: G dataset	192
C.35	Grade classification macro precision: G dataset	192
C.36	Grade classification macro recall: G dataset	193
C.37	Grade classification micro F1-score: G dataset	193
C.38	Grade classification micro precision: G dataset	194
C.39	Grade classification micro recall: G dataset	194
C.40	Grade classification accuracy: G-SP dataset	195
C.41	Grade classification balanced accuracy: G-SP dataset	195
C.42	Grade classification macro F1-score: G-SP dataset	196
C.43	Grade classification macro precision: G-SP dataset	196
C.44	Grade classification macro recall: G-SP dataset	197
C.45	Grade classification micro F1-score: G-SP dataset	197
C.46	Grade classification micro precision: G-SP dataset	198
C.47	Grade classification micro recall: G-SP dataset	198
C.48	Sub-group classification accuracy: SG dataset	199
C.49	Sub-group classification balanced accuracy: SG dataset	199
C.50	Sub-group classification macro F1-score: SG dataset	200
C.51	Sub-group classification macro precision: SG dataset	200
C.52	Sub-group classification macro recall: SG dataset	201
C.53	Sub-group classification micro F1-score: SG dataset	201
C.54	Sub-group classification micro precision: SG dataset	202
C.55	Sub-group classification micro recall: SG dataset	202
C.56	Sub-group classification accuracy: SG-SP dataset	203
C.57	Sub-group classification accuracy: SG-SP dataset	203
C.58	Sub-group classification macro F1-score: SG-SP dataset	204
C.59	Sub-group classification macro precision: SG-SP dataset	204
C.60	Sub-group classification macro recall: SG-SP dataset	205
C.61	Sub-group classification micro F1-score: SG-SP dataset	205
C.62	Sub-group classification micro precision: SG-SP dataset	206
C.63	Sub-group classification micro recall: SG-SP dataset	206
C.64	Grade classification accuracy using habitats	207
C.65	Grade classification balanced accuracy using habitats	207
C.66	Grade classification macro F1-score using habitats	208
C.67	Grade classification macro precision using habitats	208
C.68	Grade classification macro recall using habitats	209
C.69	Grade classification micro F1-score using habitats	209
C.70	Grade classification micro precision using habitats	210
C.71	Grade classification micro recall using habitats	210
C.72	Sub-group classification accuracy using habitats	211
C.73	Sub-group classification balanced accuracy using habitats	211
C.74	Sub-group classification macro F1-score using habitats	212

C.75 Sub-group classification macro precision using habitats	212
C.76 Sub-group classification macro recall using habitats	213
C.77 Sub-group classification micro F1-score using habitats	213
C.78 Sub-group classification micro precision using habitats	214
C.79 Sub-group classification micro recall using habitats	214

List of Tables

2.1 Dataset demographics. Included sarcoma subtypes are chondrosarcoma (CS), dermatofibrosarcoma protuberans (DFSP), Ewing sarcoma (ES), giant cell tumor of bone (GCTB), osteosarcoma (OS) and soft tissue sarcoma (STS).	21
2.2 Types of wavelet sub-band passes and captured details [57].	23
2.3 The combination of image transforms for each run.	23
2.4 Binary classification results. Both macro-averaged and micro-averaged metrics are displayed.	28
2.5 Number of features chosen for each feature group for the binary classifier for the original transform group.	28
2.6 Feature importance considering purely the original transform group for binary classification.	29
2.7 Number of features chosen for each feature group for the binary classifier for the wavelet transform group.	30
2.8 Feature importance considering only the wavelet transform group for binary classification.	30
2.9 Feature importance considering both original and wavelet transform groups for binary classification.	31
2.10 Number of features chosen from each feature group for the binary classifier of all transform groups.	31
2.11 Macro- and micro-averaged results for grade classification using original features.	32
2.12 Number of features selected when considering the original transform group for grade classification.	32
2.13 Feature importance considering purely the original feature groups for grade classification.	33

2.14	Macro- and micro-averaged results for grade classification using wavelet features.	33
2.15	Number of features selected when considering the wavelet transform group for grade classification.	34
2.16	Feature importance considering only wavelet transform feature groups for grade classification.	34
2.17	Macro- and micro-averaged results for grade classification using all transform groups.	35
2.18	Number of features selected when considering both original and wavelet transform groups for grade classification.	35
2.19	Feature importance considering both the original and wavelet transform feature groups for grade classification.	36
3.1	Number of MRI exams by sequence type, combining contrast-enhanced variants.	44
3.2	3D Wavelet Transform Components.	48
3.3	Hyperparameter combinations and performance metrics across outer folds.	52
3.4	Performance metrics from classification on the hold-out test set. . . .	52
3.5	Top 10 most important features ranked by Random Forest.	53
3.6	Comparative performance of classification algorithms on the balanced binary dataset.	58
4.1	Dataset demographics. Included sarcoma subtypes are chondrosarcoma (CS), dermatofibrosarcoma protuberans (DFSP), ewing sarcoma (ES), giant cell tumor of bone (GCTB), osteosarcoma (OS) and soft tissue sarcoma (STS).	75
4.2	Types of image transforms applied in the research.	80
4.3	Feature classes and their description	81
4.4	Summary of features extracted from normalized MSI matrices $\mathcal{L}$	83
4.5	Metrics used for the assessment of model performance and their respective definitions.	91
4.6	A list of all ML tests applied in the research. Note: each of the 10 tests go through the feature ranking, feature selection, feature elimination and cross-validation independently.	94
4.7	Selected features for highest scoring set of features for binary classification for all healthy against sarcoma masks. Best model was LogisticRegression with 100 features and a AUC-ROC score of 0.999.	100
4.8	Selected features for highest scoring set of features for binary classification for all healthy against sarcoma masks processed with SPs. Best model was CatBoost with 90 features and a AUC-ROC score of 0.885. . . .	100

4.9	Selected features for highest scoring set of features for binary classification for comparable healthy against sarcoma masks. Best model was RidgeClassifier with 247 features and a AUC-ROC score of 0.974. . . .	101
4.10	Selected features for highest scoring set of features for binary classification for comparable healthy against sarcoma masks processed with SPs. Best model was LogisticRegression with 100 features and a AUC-ROC score of 0.865.	101
4.11	Selected features for highest scoring set of features for grade classification for whole masks. Best model was SGDClassifier with 60 features and a AUC-ROC score of 0.690.	102
4.12	Selected features for highest scoring set of features for grade classification for masks processed with SPs. Best model was RidgeClassifier with 10 features and a AUC-ROC score of 0.670.	102
4.13	Selected features for highest scoring set of features for sub-group classification for whole masks. Best model was RidgeClassifier with 149 features and a AUC-ROC score of 0.863.	103
4.14	Selected features for highest scoring set of features for sub-group classification for masks processed with SPs. Best model was SGDClassifier with 100 features and a AUC-ROC score of 0.802.	103
4.15	The Jaccard index of the top 100 for all models of the SvAH dataset. .	104
4.16	The Jaccard index of the top 100 for all models of the SvAH-SP dataset.	104
4.17	The Jaccard index of the top 100 for all models of the SvCH dataset. .	105
4.18	The Jaccard index of the top 100 for all models of the SvCH-SP dataset.	105
4.19	The Jaccard index of the top 100 for all models of the G dataset. . . .	105
4.20	The Jaccard index of the top 100 for all models of the G-SP dataset. . .	106
4.21	The Jaccard index of the top 100 for all models of the SG dataset. . .	106
4.22	The Jaccard index of the top 100 for all models of the SG-SP dataset. .	106
4.23	The Jaccard index from the top 100 features pooled from each model of a dataset compared amongst datasets.	107
4.24	Jaccard index when comparing top 10 ranking habitat features for all models when classifying sarcoma grade.	107
4.25	Jaccard index when comparing top 10 ranking habitat features for all models when classifying sarcoma sub-group.	108
C.1	AUC-ROC score for all models and tested number of features for tissue classification with all healthy against sarcoma masks. The highest score for each model is in bold.	215
C.2	AUC-ROC score for all models and tested number of features for tissue classification with all healthy against sarcoma masks processed with SPs. The highest score for each model is in bold.	216

C.3	AUC-ROC score for all models and tested number of features for tissue classification with comparable healthy against sarcoma masks. The highest score for each model is in bold.	217
C.4	AUC-ROC score for all models and tested number of features for tissue classification with comparable healthy against sarcoma masks processed with SPs. The highest score for each model is in bold.	218
C.5	AUC-ROC score for all models and tested number of features for grade classification with whole masks. The highest score for each model is in bold.	219
C.6	AUC-ROC score for all models and tested number of features for grade classification with masks processed with SPs. The highest score for each model is in bold.	220
C.7	AUC-ROC score for all models and tested number of features for subgroup classification with whole masks. The highest score for each model is in bold.	221
C.8	AUC-ROC score for all models and tested number of features for subgroup classification with masks processed with SPs. The highest score for each model is in bold.	222
C.9	AUC-ROC scores obtained from habitat analysis when predicting grade.	223
C.10	AUC-ROC scores obtained from habitat analysis when predicting subgroup.	224

AI Evaluations of Sarcoma From Magnetic Resonance Imaging

Arnar Evgení Gunnarsson

October 17, 2025

Abstract

Background: Sarcomas are a rare and heterogeneous group of malignant tumors, making early detection and diagnosis a high priority. Diagnosis traditionally relies on expert interpretation of radiological imaging and histopathological biopsies, processes that are time-consuming, subjective, prone to inter-observer variability and are invasive.

Methods: This research investigates the potential of artificial intelligence (AI), specifically radiomics and machine learning (ML), to support sarcoma diagnosis, characterization and grading based on MRI scans. Quantitative features were extracted from multiple image transforms, including original, wavelet, Laplacian of Gaussian, square, square root, logarithm, exponential, gradient, and local binary pattern transforms. From these images, first-order statistics and texture descriptors (GLCM, GLSZM, GLRLM, NGTDM) were computed. A diverse set of ML models were evaluated, including Random Forest, Logistic Regression, SGDCClassifier, Ridge Classifier, LightGBM, XGBoost, and CatBoost. Models were trained on three primary tasks: (1) binary classification of healthy vs. sarcoma tissue, (2) sarcoma grading based on the FNCLCC system, and (3) differentiation between bone and soft tissue sarcomas. In addition, a habitat generation framework was introduced, clustering radiomic segments into biologically meaningful subregions to enhance grade and sub-group classification. Model performance was assessed using AUC-ROC, accuracy, and macro- and micro-averaged F1-score, precision, and recall.

Results: For binary classification of healthy vs. sarcoma tissue, models performed strongly, achieving AUC-ROC scores ranging from 0.829 to 0.999 depending on data preprocessing, with consistently high metric values overall. This demonstrates that ML methods are well suited to distinguishing healthy from pathological tissue. Grade classification achieved AUC-ROC scores ranging from 0.618 to 0.690 using radiomics and between 0.520-0.591 using habitats, with metrics generally above baseline for three-class classification. While performance remains modest, results indicate meaningful correlation between radiomic inputs and FNCLCC grade. Sub-group classification of bone vs. soft tissue sarcomas achieved AUC-ROC scores of 0.750-0.863 using radiomics and 0.585-0.707 using habitats, indicating that ML methods can effectively aid in sarcoma sub-type differentiation.

Conclusions: The findings demonstrate that combining multiple image transforms with radiomics for diagnostic, characterization and grading classification is

feasible and effective. Advanced ML models substantially improve performance across diagnostic and prognostic tasks. Furthermore, the habitat generation approach offers additional biological interpretability and can enhance grade and subgroup prediction. These results highlight the promise of AI-driven radiomics pipelines to support clinical decision making in sarcoma diagnosis and hopefully earlier management and extended survival.

Keywords: Artificial Intelligence, Machine Learning, Sarcoma, Diagnostics, Sarcoma Grading, Radiomics, Habitats.

Gervigreindarmat á Sarkmein Út Frá Segulómrannsóknum

Arnar Evgení Gunnarsson

October 17, 2025

Útdráttur

Bakgrunnur: Sarkmein eru sjaldgæf og fjölbreytt tegund illkynja æxla og því er mikið forgangsatriði að uppgötva það og staðfesta sem fyrst. Hefðbundin greining byggir á mannlegri sérfræðitúlkun myndgreininga og ífarandi vefjasýna, ferlar sem eru tímafrekir, og háðir huglægu mati mismunandi sérfræðinga.

Aðferð: Þessi rannsókn skoðar möguleika gervigreindar og geislatölufræði (e. radiomics) til að styðja lýsingu, greiningu, og gráðun á sarkmeinum. Magnbundnir eiginleikar (e. features) voru unnir úr fjölbreyttum tegundum af umbreyttum segulómmyndum, sérstaklega þeim upprunalegu, en einnig bylgjugreiningu (wavelet), Laplace af Gaussian síun, fernings-, ferningsrót-, lógaritma-, veldis-, halla- og staðbundnum tvíundarmynstrum (e. local binary pattern). Úr þessum myndum voru reiknuð fyrstu stigs tölfræðileg gildi og áferðarlýsingsgildi (e. texture descriptors) eins og GLCM, GLSZM, GLRLM og NGTDM. Margar tegundir af gervigreindarlíkönum voru metin, þar á meðal Random Forest, Logistic Regression, SGDClassifier, Ridge Classifier, LightGBM, XGBoost og CatBoost. Líkönin voru þjálfuð með þrjú meginverkefni í huga: (1) tvíflokkagreiningu (e. binary classification) á heilbrigðum og sarkmeina vef, (2) stigun sarkmeina samkvæmt FNCLCC-kerfinu, og (3) greining milli beina- og mjúkvessarkmeina. Að auki var kynnt ferli fyrir myndun svokallaðra „habitat“-svæða, þar sem geislatölufræðilegir bútar voru flokkaðir í líffræðilega þýðingarmikil undirsvæði til að bæta stigun og undirflokkun. Afköst líkananna voru metin með AUC-ROC, nákvæmni (e. accuracy), og meðaltölum F1-stuðuls, næmni (e. recall) og nákvæmni (e. precision) fyrir bæði macro- og micro-mælikvarða.

Niðurstöður: Við tvíflokkagreiningu á heilbrigðum vef og sarkmeina vef sýndu líkönin góða frammistöðu, með AUC-ROC gildum á bilinu 0,829 til 0,999 eftir því hvaða forvinnsla gagna var notuð, og náði almennt háum mæligildum yfir heildina. Þetta sýnir að vélvæðingar aðferðir (ML) henta vel til flokkunar á milli heilbrigðs og sjúklegs vefs. Gráðun sarkmeina (e. grading) náði AUC-ROC gildum á bilinu 0,618 til 0,690 þegar notast var við geislatölufræðilega greiningu og á bilinu 0,520 til 0,591 þegar notuð voru svokölluð „habitat“-svæði, með mæligildum sem almennt voru yfir grunnviðmiðum fyrir þríflokkagreiningu. Þótt frammistaðan sé enn hófleg, benda niðurstöðurnar til marktæks sambands milli geislatölufræðilegra eiginleika og FNCLCC-gráðu. Undirflokkagreining milli beina- og mjúkvessarkmeina náði AUC-ROC gildum á bilinu 0,750 til 0,863 með geislatölufræðilegri greiningu og 0,585 til 0,707 með „habitat“-aðferð, sem gefur til kynna að vélvæðingar aðferðir geti verið árangursríkar við aðgreiningu undirflokka sarkmeina.

Ályktanir: Niðurstöðurnar sýna að samsetning margra myndbreytinga með geislatölufraðilegri greiningu er bæði raunhæf og árangursrík fyrir greiningu, einkenningu og gráðun sarkmeina. Háþróaðar gervigreindarlíkanir bæta verulega frammistöðu í bæði greiningar- og forspárverkefnum. Að auki veitir „habitat“-aðferðin aukna líffræðilega túlkunarmöguleika og getur eflt nákvæmni í gráðun og undirflokkagreiningu. Þessar niðurstöður undirstrika möguleika gervigreindarstuðlaðra geislatölufraðilegra kerfa til að styðja við klínískar ákvarðanir og þannig mögulega flýta fyrir greiningu og meðhöndlun sarkmeina og þannig mögulega bæta einnig lífslíkur.

Efnisorð: gervigreind, sarkmein, greining sarkmeina, gráðun sarkmeina, geislatölufraði eiginleikagreining, „habitat“-svæði, segulómrannsókn sarkmeina

Acknowledgments

I would like to express my gratitude to the Institute of Biomedical and Neural Engineering at Reykjavik University for giving me the opportunities to work on impactful research and giving me the platform to deepen my understanding of the natural world. I thank Professor Paolo Gargiulo for his unfaltering patience and support throughout the past years of my research, and for all of the knowledge, tools and skills that I have developed since joining his highly skilled team. I would also like to personally thank Halldór Jónsson Jr., MD for providing me with the opportunity of conducting research on such an extensive database of patients that he has meticulously prepared for over 20 years, and his excitement and positivity regarding the work produced throughout the research. I attribute the results of my work to my two supervisors and doubt this work could have been finished without their outstanding support.

This research was funded by Reykjavik University and the Icelandic Ministry of Higher Education.

Prologue

Sarcomas and cancers in general are a cause of fear for many in modern society. Research in the sector has been super charged with advancements in technology, but still presents many difficulties in diagnosis, staging and treatments.

This thesis presents all of the work carried out on research on the Icelandic Sarcoma Group (IceSG) database, a collection of over 200 patients that have suffered from sarcoma in the Icelandic healthcare system since 2004. This work contributes to the ongoing battle against cancers through the development of tools and pipelines to support sarcoma diagnosis, characterization and staging. These works act as a foundation for future research and improvements using this novel database of patients.

This thesis dissertation is organized into the following chapters:

Chapter 1: presents an introduction to sarcoma from a clinical perspective. Sarcomas and their diverse clinical presentations are discussed. Novel techniques and methods for diagnosis, grading, staging and treatment are reviewed.

Chapter 2: is the first article written throughout the research of the IceSG database. This article presents a preliminary analysis conducted, using fundamental strategies that are built on in later chapters. These include the application of image transforms and calculations of radiomics data, which are processed using by a random forest classifier to predict whether a data sample is a sarcoma or healthy tissue as well as predicting sarcoma FNCLCC grade.

Chapter 3: is the second article written on the database. It too presents preliminary methods used in later chapters including the creation of a dataset comparing sarcoma infected regions with regions of comparable healthy anatomy and feature analysis methods.

Chapter 4: builds on the previous research and goes into much more detail. Here, more image transforms are included in the radiomics feature extraction pipeline. Analyses are made at multiple size scales, from analyzing small segments of tissue to entire slices. In addition to predicting whether tissue samples are sarcoma or not

and grading, sarcoma sub-group (soft tissue sarcoma or bone sarcoma) is predicted using the same pipelines. A novel method for predicting sarcoma grade and sub-group through the fabrication of habitats is tested. This method create multiple regions within a sarcoma tumor to measure interactions and heterogeneity.

Chapter 5: states the conclusions drawn throughout the research regarding the potential of machine learning in supporting clinical decision making regarding sarcoma.

Chapter 1

Introduction

1.1 Sarcomas In General

Sarcomas are a rare and heterogeneous group of malignant tumors that originate in mesenchymal stem cells (MSCs), which are progenitor cells capable of differentiating into connective tissue components such as bone, muscle, fat, and cartilage. These tumors account for less than 1% of adult cancers but comprise a much higher percentage in pediatric populations[1]. The prevailing hypothesis suggests that sarcomas arise when MSCs acquire initiating mutations that disrupt normal differentiation and self-renewal, rendering them susceptible to further oncogenic events[2]. Due to MSCs ability to differentiate into multiple cell types, the clinical presentation of sarcomas is incredibly diverse.

Sarcomas are broadly categorized into soft tissue sarcomas (STS) and bone sarcomas (BS), each encompassing numerous histological sub-types with distinct molecular and clinical profiles. Their vast heterogeneity, spanning over 100 sub-types, presents significant challenges in diagnosis, classification, and treatment planning[1]. Accurate diagnosis often requires a combination of imaging, histopathological analysis, and increasingly, molecular profiling[3]. Treatment strategies are similarly complex and may include surgery, radiotherapy and chemotherapy, and often times multimodal combinations of these[1].

Sarcoma was first distinguished from carcinomas through the development of cellular pathology in the mid 19th century on the basis of their tissue origin[4] and has since been further categorized and described into sub-types. The earliest confirmed case of sarcoma in a hominid comes from a specimen labelled SK 7923 dating back 1.6-1.8 million years ago from an archeologic site Swartkrans, South Africa[5]. The study of SK 7923 shows that hominids have been struggling with these malignant diseases throughout our entire history. The oldest discovered case of sarcoma comes from a dinosaur specimen, namely Centrosaurus Apertus,

dating back 75-77 million years [6], which highlights the profound evolutionary history of this malignancy, suggesting that the fundamental biological mechanism behind sarcoma development emerged early in vertebrate history.

1.2 Sarcomas In the Study

The sarcomas considered in this thesis are soft tissue sarcoma (STS), osteosarcoma (OS), giant cell tumor of bone (GCTB), chondrosarcoma (CS), Ewing sarcoma (ES) and dermatofibrosarcoma protuberans (DFSP). The classification of these is often nuanced and each contains its own sub-types.

STS is a general umbrella term for a wide range of malignant tumors originating in soft tissue with over 50 histologic sub-types. They arise from mesenchymal tissues such as fat (liposarcoma), muscle (leiomyosarcoma or rhabdomyosarcoma), peripheral nerves (malignant peripheral nerve sheath tumors), and fibrous connective tissue (fibrosarcoma). These tumors most frequently occur in the extremities (59%), followed by the torso (19%), retroperitoneum (15%), and head and neck regions (9%). Approximately, 6% of STS occur in patients below 19 years of age, 13% from ages 20-39, 30% from ages 40-59, 37% from ages 60-79 and 14% aged 80 and above. Clinically, STS often present as deep-seated, painless, and progressively enlarging masses, which can delay diagnosis until they reach substantial size or affect adjacent structures. Epigenetic changes and mutations have been correlated with the development of STS with common genes including TP53, CDKN2A, retinoblastoma 1 (RB1), neurofibromin 1 (NF1) and ATP-dependent helicase (ATRAX). The development of metastatic disease occurs in approximately 40-50% of patients, of which the median survival is only around 12 months. Recurrence of STS varies greatly depending on depth, size and grade. Basic resection can result in up to 90% recurrence, but fortunately have dropped in selected types to 7-15% due to recent advances in medical imaging, neoadjuvant therapy, adjuvant radiation, and limb salvage procedures[7]–[12].

OS is a malignant tumor of bone, characterized by the production of osteoid, the unmineralized and organic component of the bone, by malignant mesenchymal cells. It most frequently arises in the metaphyseal regions of long bones, such as the distal femur, proximal tibia, and proximal humerus, areas of rapid growth during adolescence. OS displays a bimodal age distribution, with approximately 75% of cases diagnosed in patients under 25 years of age, and a secondary incidence peak in individuals over 65 years, often associated with prior radiation exposure or underlying conditions like Paget's disease. While histologically diverse, the most common sub-type is high-grade intramedullary osteosarcoma, which is biologically aggressive. The 5-year survival rate is at 60%, with local recurrence rates relatively low at approximately 5-10%, and the estimated long-term survival for these patients is less than 20%. The main driving force of epigenetic dysreg-

ulation in OS is through DNA methylation and RNA modifications, methylation has been shown to decrease robustness of tumor suppressor genes and microRNAs including miR-486-5p, miR-149, miR-195 contributing to OS cell proliferation, migration, and metastasis. Long non-coding RNAs (lncRNAs) such as THAP9-AS1 and HOTAIR further modulate gene expression by recruiting DNA methyltransferases to tumor suppressor loci, amplifying malignant traits. Clinically, OS typically presents with localized bone pain that may be initially attributed to trauma, followed by the appearance of swelling or a palpable mass[9], [13]–[16].

GCTB is a benign but locally aggressive primary bone tumor and therefore classified with sarcomas. Histologically it is characterized by a proliferation of neoplastic mononuclear stromal cells interspersed with reactive multinucleated giant cells with morphology akin to osteoclast. It mostly arises in the epiphyseal regions of long bones, particularly the distal femur, proximal tibia, and distal radius, often presenting as a metaphyseal lesion extending into the epiphysis. GCTB accounts for approximately 20% of benign skeletal tumors and 5% of all primary bone tumors, with a peak incidence between 20 and 50 years of age and a slight female predominance. Clinically, patients often present with localized pain, swelling, joint effusion, and restricted motion of the affected joint, sometimes accompanied by weight-bearing difficulties. Diagnosis relies on radiographic imaging, including plain X-rays, CT, and MRI, and lesions are commonly graded using the Campanacci system. Although GCTB is generally benign, it has a strong tendency for local recurrence, observed in up to 25% of cases overall and up to 50% following curettage. GCTB is primarily driven by epigenetic alterations, most notably H3F3A G34W/L mutations in the neoplastic stromal cells, which disrupt normal histone methylation (H3K36me3 and H3K27me3) and lead to widespread changes in gene expression, splicing, and cellular behavior. These changes do not affect the DNA sequence itself but alter chromatin structure and transcriptional regulation, contributing to increased proliferation, migration, and recurrence, despite the tumor's benign classification. In rare cases, GCTB can undergo malignant transformation or metastasize, particularly when arising in atypical locations such as the spine, skull, or larynx. Classification of malignant GCTB is split into 2 sub-types primary with a 5-year survival rate of 87% and secondary with a 5-year survival rate of 32%[17]–[19].

CS is a malignant cartilaginous neoplasm characterized by the production of atypical hyaline cartilage and a wide range of morphological features, including irregular lobules of cartilage, nuclear atypia, permeation of the cortical bone and marrow, and frequent calcifications. It most commonly arises *de novo* within normal bone (conventional or primary CS), though secondary forms can emerge from pre-existing benign lesions such as enchondromas or osteochondromas. The tumor typically affects adults over 50 years of age, with a slight male predominance, and frequently involves the axial skeleton (especially the pelvis, scapula, sternum, and ribs) followed by the proximal femur and humerus. Patients often

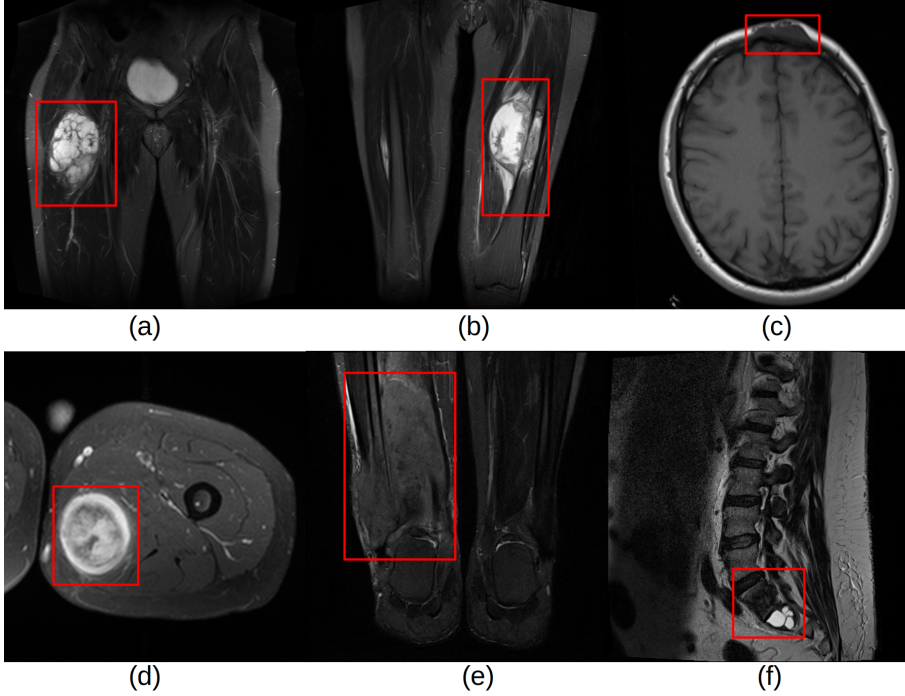


Figure 1.1: MRI slices of different sarcoma sub-types from the IceSG dataset with tumor regions shown by the red bounding box: (a) CS (b) OS (c) DFSP (d) STS (e) ES (f) GCTB.

present with chronic pain and swelling, occasionally accompanied by neurological symptoms when lesions occur in the skull base. It most commonly arises in the axial skeleton, particularly the pelvis, scapula, sternum, and ribs, followed by long bones such as the proximal femur and proximal humerus and accounts for approximately 20% of all primary malignant bone tumors. Local recurrence and metastasis are closely tied to histological grade, with dedifferentiated CS exhibiting the highest recurrence and metastatic potential. CS exhibits epigenetic dysregulation primarily through DNA hypermethylation and histone modifications, which contribute to tumor progression by silencing tumor suppressor genes such as P16INK4a and RUNX3. This hypermethylation may be driven by mutant IDH1/2-induced accumulation of the oncometabolite D2HG, though alternative mechanisms are also implicated. Survival outcomes depend on severity, with 5-year survival rates of 83% for low-grade tumors and 53% for intermediate to high-grade CS. No study exists on the overall recurrence of CS, although studies on smaller patient data-sets estimate an approximate 19-25% recurrence of CS [20]–

[24].

ES is an aggressive malignant tumor of both bone and soft tissue, most commonly affecting adolescents and young adults, and accounting for 10-15% of all primary bone sarcomas. It originates from mesenchymal progenitor cells and encompasses a family of tumors including classic Ewing sarcoma of bone, extraskeletal ES, Askin tumor, and primitive neuroectodermal tumors. ES typically arises in the pelvis, axial skeleton, and long bones such as the femur, and patients frequently present with localized pain, swelling, and sometimes systemic symptoms such as fever or weight loss. The disease predominantly affects individuals between 10-25 years of age, with a slight male predominance. Recurrence is common, affecting approximately 25% of patients with initially localized disease, and metastases commonly spreading to the lungs, bone, and bone marrow. At the genomic level, ES is driven by characteristic chromosomal translocations, most notably $t(11;22)(q24;q12)$, that lead to the formation of EWS-FLI1 fusion genes, which act as aberrant transcription factors altering chromatin accessibility and global gene expression. Despite advancements in treatments improving the 5-year survival rate to over 70% in localized disease, survival drops to less than 30% in patients with lung metastases and under 20% with skeletal or marrow involvement[25].

Dermatofibrosarcoma Protuberans (DFSP) is a rare, low- to intermediate-grade cutaneous soft tissue sarcoma characterized by slow but locally aggressive growth and a high rate of local recurrence. Histologically, DFSP consists of uniform spindle cells in a storiform pattern, and is strongly CD34 positive. The tumor arises from a specific chromosomal translocation, $t(17;22)(q22;q13)$, which results in a COL1A1-PDGFB fusion gene that drives overexpression of PDGFB and autocrine activation of PDGFR β , promoting tumor proliferation. Clinically, DFSP typically presents as a firm, plaque-like lesion on the trunk (47%), lower extremities (20%), or upper extremities (18%), with a male-to-female ratio of approximately 3:2 and a median age of diagnosis around 39 years. The tumor may remain indolent for years, with some patients reporting symptoms persisting for decades prior to diagnosis. Approximately 10-15% of DFSPs contain a fibrosarcomatous component (DFSP-FS), which is associated with more aggressive behavior, deeper tissue infiltration, increased mitotic activity, and a higher risk of metastasis. Recurrence occurs in up to 50% of cases if margins are positive after excision, but drops below 10% with wide local excision. Genomic changes observed in DFSP-FS include microsatellite instability and p53 mutations, which may represent late events in tumor progression. Adjuvant radiotherapy significantly reduces recurrence rates, particularly in cases with close or positive margins, achieving local control in 80-85% of such patients. Metastasis is rare ($\leq 5\%$) but more likely in DFSP-FS, particularly in cases with inadequate resection. While chemotherapy has limited efficacy, targeted therapy with tyrosine kinase inhibitors such as imatinib has demonstrated clinical benefit in patients with unresectable or metastatic disease. Five-

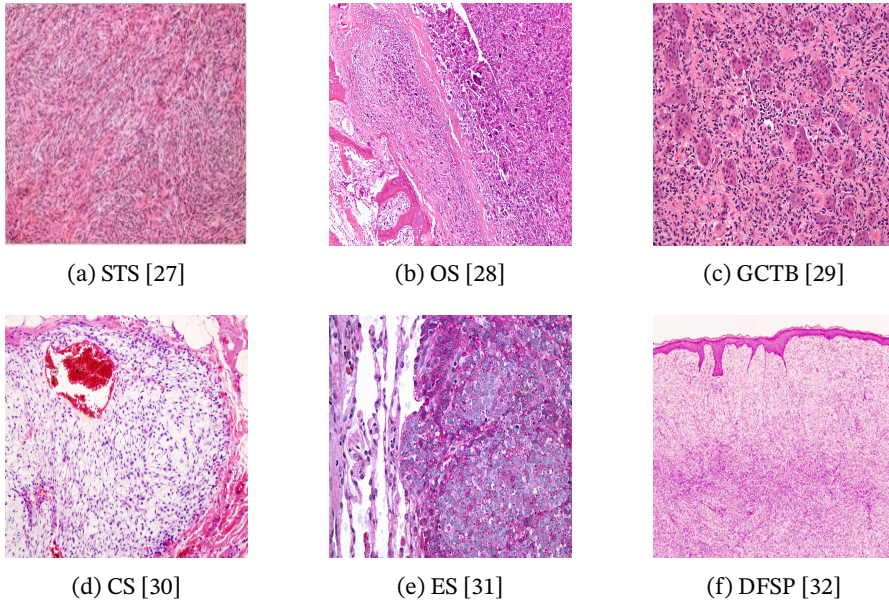


Figure 1.2: Examples of biopsy images taken from different sarcoma sub-types.

year local control rates exceed 80% for DFSP but drop below 30% for DFSP-FS, highlighting the importance of histologic sub-type and surgical margins in determining outcome[26].

1.3 Symptoms

Symptoms of sarcoma can range, and often times go unnoticed during the early stages. As the tumor grows it may hit a nerve causing patients pain. In certain places patients might notice a swelling or a lump growing. They can also limit range of motion around joints near its epicenter. Fevers of unknown origin have been reported. Sarcoma growth also inhibits the structural integrity of bones and may cause them to break for no apparent reason.

STS most commonly present as painless, asymptomatic masses that develop insidiously and often remain unnoticed until they reach a considerable size [7]. Tumors located in the distal extremities are usually detected earlier due to visibility and space constraints, whereas those in proximal or retroperitoneal regions may grow significantly before causing symptoms. As STS tumors expand centrifugally, they can compress adjacent structures, leading to pain, edema, neurological symptoms (especially with retroperitoneal involvement), and gastrointestinal dis-

turbances such as bowel obstruction [33].

Osteosarcoma typically presents as localized pain and swelling, often near the metaphysis of long bones such as the distal femur, proximal tibia, or proximal humerus. The pain may initially be intermittent and worsen at night or with activity, often mistaken for growth-related discomfort or sports injuries in young individuals. As the tumor progresses, swelling becomes more apparent, and function of the affected limb may decline. In more advanced cases, pathologic fractures can occur. Approximately 20% of patients present with metastatic disease at diagnosis, primarily to the lungs, which can sometimes be asymptomatic or manifest as respiratory issues in more severe cases [13].

GCTB commonly arises in the meta-epiphyseal regions of long bones, particularly around the knee, such as the femur, tibia, and radius. Clinically, patients often present with pain, swelling, joint effusion, restricted range of motion, and in more severe cases, impaired weight-bearing ability due to joint involvement [17].

CS are usually localized and include chronic pain and swelling, often present for months or years prior to diagnosis. The growth is generally slow, but may eventually lead to neurological symptoms if located near the spine or skull base. On imaging, chondrosarcomas appear as large, lobulated lesions with calcifications and cortical destruction, aiding in diagnosis alongside clinical presentation [20].

ES typically presents in adolescents and young adults with localized pain, swelling, or stiffness, often worsening at night. More than 50% of patients report intermittent pain over the affected site, which commonly includes the pelvis, axial skeleton, and long bones such as the femur. Systemic symptoms such as fever, weight loss, or fatigue, are more likely in cases with metastatic spread. When the tumor affects weight-bearing bones, patients may present with pathological fractures, while pelvic lesions may manifest as back pain. In approximately 20% of cases, metastatic disease is present at diagnosis, with the lungs or pleura being common sites of spread. Clinical findings may also include signs like petechiae in cases of bone marrow involvement, or neurologic symptoms if the central nervous system is affected [25].

DFSP typically presents as a slowly enlarging, asymptomatic cutaneous lesion that initially appears as a pink or violet-red plaque, often with telangiectatic overlying skin. It is usually fixed to the dermis but is mobile over deeper tissues and only develops nodular features later in its course. The disease tends to follow an indolent growth pattern, with symptom duration ranging from months to decades (mean of 6.4 years in one series). Lesions are typically superficial and less than 5 cm in diameter, but may rarely invade deeper structures in advanced cases[26].

1.4 Diagnostics, Grading and Staging

The protocol or path to diagnosis may differ between doctors and treatment centers, but begins in general with image examination following a doctor's visit. Most commonly, ultrasound and MRI are used in STS and conventional x-ray and CT in BS. The radiologist may confirm a tumor on the images, but on the suspicion of a sarcoma they may ask for a second opinion, most commonly by referral to a sarcoma center. Once the tumor has been diagnosed or suspicion raised as sarcoma, a core needle biopsy is made, and the sample is examined histologically to determine the sarcoma type and grade, using the French classification system FNCLCC, set by the Federation Nationale des Centres de Lutte Contre le Cancer (FNCLCC) [34]. For staging the severity of the disease, a PET scan is then taken of the patient's entire body to figure out whether or not the tumor has metastasized. A common rule of thumb used in this environment is that around 10% of patients have developed metastasis at the time of diagnosis. Once the diagnosis and staging has been finalized a meeting between a radiologist, oncologist and surgeon is held to discuss the best method for treatment. Due to the diverse nature of sarcoma, treatment may vary between centers and especially between countries. For the most common sarcomas, special treatment protocols are followed, but for rare types, treatment is individualized. The surgeon operates on the patient if no metastasis is discovered, otherwise the oncology department handles the patient through radiation therapy, chemotherapy or a combination. During operation the surgeon removes the tumor along with an extra centimeter of tissue to ensure that all malignant tumor cells are removed as well. In cases where the tumor lies on sensitive areas such as bone or vital organs, the patient is sent in for radiotherapy focused around the intersection of the tumor and sensitive region. This is the general procedure for handling sarcoma patients in Iceland, and is standardized based on the best practice guidelines on tested and effective therapies.

According to the FNCLCC guidelines the biopsy sample is examined by a professional to measure firstly tumor differentiation or how well the sample resembles non-pathological tissue behavior, a score of 1-3, higher for undifferentiated. The second step measures the amount of necrosis or the amount of dead and decaying tissue within the tumor, 0 points for no necrosis, 1 point if necrosis is less than 50% of the tumor volume and 2 points if it is 50% or above. The final measure is the mitotic count which is a measure of the speed at which tumor cells divide and grow; the sample is viewed in a high power field microscope and 10 regions are viewed, if 0-9 mitoses are counted in the 10 regions the patient receives a score of 1, a score of 2 is given if the count is from 10-19 and 3 points if it is greater than 19. The scores from these individual tests is tallied and a sarcoma grade is given grade 1 with a score of 2-3 points, grade 2 with a score of 4-5 and grade 3 with a score of 6-8.

Grading of soft tissue sarcomas, such as via the FNCLCC system, remains a

fundamentally subjective process, inherently limited by inter-observer variability and sampling constraints. The core biopsy provides only a small portion of the tumor, which may not fully capture its biological heterogeneity, particularly in terms of necrosis, mitotic activity, and cellular differentiation [35]. Mitotic count, for example, is dependent on accurate identification of mitotic figures under high-power microscopy, a process known to vary between pathologists, especially in borderline cases [36]. Similarly, estimation of tumor necrosis and degree of differentiation can be affected by sampling location and interpretive bias [37]. These limitations underscore the demand for complementary diagnostic tools, such as quantitative imaging and computational pathology, to enhance reproducibility and prognostic accuracy in sarcoma grading.

1.5 Novel Diagnostic Techniques

Recent advances in sarcoma diagnostics have increasingly focused on molecular and epigenetic profiling as tools for early detection, prognosis, and treatment stratification. One of the most actively explored areas is epigenetic modification, particularly DNA methylation, histone alterations, and non-coding RNA regulation. In OS for example, aberrant DNA methylation has been linked to malignant transformation, tumor progression, and chemotherapy resistance. Specific microRNAs, such as *miR-486-5p*, *miR-149*, and *miR-195*, have been shown to be downregulated due to promoter hypermethylation, and their restoration through demethylating agents like 5-Aza-2-deoxycytidine leads to reduced cell proliferation and metastatic potential [16].

Histone modifications also serve as promising diagnostic indicators. Trimethylation of histone H3 lysine 27 (H3K27me3) and other marks, such as H4K20me3, are involved in regulating oncogene and tumor suppressor gene expression. Demethylation of H3K27 via KDM6A/B enzymes has been correlated with increased tumor aggressiveness and chemotherapy resistance in OS. Furthermore, elevated levels of the m6A RNA methylation writer METTL3 have been associated with the stabilization of oncogenic transcripts such as *DRG1* and *ATAD2*, making METTL3 a target of interest for molecular diagnostics and therapy [16].

Another novel approach involves the use of three-dimensional (3D) bioprinted tumor models. These systems replicate the tumor microenvironment more accurately than traditional two-dimensional cultures, allowing for better evaluation of DNA methylation states and associated gene expression changes. Recent studies using 3D OS models have identified hundreds of differentially methylated and expressed genes compared to standard in vitro models, highlighting the diagnostic value of this technology for understanding epigenetic landscapes in sarcoma [16].

LncRNAs such as *HOTAIR* and *THAP9-AS1* have emerged as critical regulators of gene silencing through methyltransferase recruitment. Their differential

expression correlates with clinical outcomes and may serve as biomarkers for prognosis and therapeutic response. In addition, epigenetic changes across multiple sarcoma sub-types including soft tissue sarcomas, are being explored as diagnostic and predictive tools, with the potential to guide personalized treatment and overcome resistance mechanisms [11].

While traditional imaging remains standard for grading certain sub-types like Giant Cell Tumor of Bone (GCTB), the trend in sarcoma diagnostics is clearly shifting toward molecular techniques that provide greater specificity and insight into tumor biology. As these tools become more integrated into clinical practice, they offer the potential to enhance diagnostic precision and improve patient outcomes across the heterogeneous landscape of sarcoma [17].

1.6 Treatment

The standard treatment for STS is multimodal and depends on tumor-specific factors such as histologic sub-type, grade, size, and location, as well as patient age and comorbidities. The cornerstone of management is wide surgical excision aimed at achieving negative margins, often combined with radiation therapy, either as neoadjuvant or adjuvant treatment, to improve local control. Chemotherapy remains a controversial component of STS treatment; while it may be considered for certain high-grade or metastatic cases, no universally accepted chemotherapy protocol exists. In extremity STS, limb-sparing procedures are preferred over amputation when feasible, and treatment is best managed within specialized multidisciplinary teams to optimize outcomes and reduce recurrence risk [7], [10].

The standard treatment for OS involves a multimodal approach that combines neoadjuvant chemotherapy, surgical resection of the primary tumor, and adjuvant chemotherapy according to a specific protocol. Chemotherapeutic regimens typically include high-dose methotrexate, doxorubicin, cisplatin, and ifosfamide. Limb-salvage procedures are also preferred in OS over amputation when feasible and oncologically safe. The completeness of surgical resection, achieving negative margins, is one of the most important prognostic factors. For patients with pulmonary metastases, surgical removal of lung nodules may improve survival outcomes.

Despite aggressive treatment, the prognosis in both STS and OS remains poor for patients with metastatic or recurrent disease, underscoring the critical importance of early diagnosis and complete surgical remission [15]. However, surgical intervention remains the cornerstone of GCTB treatment. The type of surgery, typically curettage or wide resection, is guided by the tumor's grade and anatomical location. While curettage is less invasive and preserves joint function, it is associated with higher recurrence rates. Additional strategies such as chemical adjuvants (e.g., phenol or liquid nitrogen) and bone grafting or cementing are of-

ten used to reduce recurrence risk. In cases where surgery is not feasible, such as in central (spinal or pelvic) tumors, or when the tumor is deemed inoperable, radiotherapy has been employed, although it carries a risk of malignant transformation and collateral damage[17], [19].

The most common method of treating CS is surgical resection with adequate margins. Wide surgical excision is recommended, especially for intermediate- to high-grade tumors (grades 2 and 3), as incomplete resections are associated with higher rates of local recurrence. Chemotherapy and radiotherapy are largely ineffective due to the tumor's low mitotic rate, poor vascularity, and dense cartilage matrix, although high-dose radiation may be considered in select unresectable cases. Outcomes are significantly influenced by tumor grade and location, with pelvic lesions demonstrating worse prognosis and greater recurrence risk [21], [22], [24].

The standard treatment for Ewing sarcoma involves a multimodal approach consisting of chemotherapy and local treatment (surgery and/or radiotherapy), depending on the result of staging and the selected protocol. A common regimen includes alternating cycles of VDC (vincristine, doxorubicin, cyclophosphamide) and IE (ifosfamide, etoposide). Chemotherapy is typically initiated before local therapy and resumed postoperatively or concurrently with radiation. Surgical resection is preferred when feasible (negative margins), particularly for tumors in expendable bones, while radiotherapy is used for unresectable tumors or unsecure/positive margins. Local control strategies are adapted based on tumor location, size, histologic response (bad/good responder), and patient-specific factors [25].

The mainstay of DFSP treatment is complete surgical excision with wide margins. This approach yields high local control rates, generally exceeding 90% if margins are negative. However, close or positive surgical margins significantly increase the risk of local recurrence. In such cases, adjuvant radiotherapy either pre- or post-operatively has been shown to improve local control, with 7-10-year control rates ranging from 75-95% in various studies. DFSP rarely metastasizes, but advanced cases or those with fibrosarcomatous transformation (DFSP-FS) may spread to the lungs, warranting chest imaging in some patients[26].

1.7 Directions of Treatment Improvements

Novel therapeutic strategies in STS are increasingly leveraging advances in molecular biology and immunotherapy. Epigenetic modulation represents a promising frontier, with efforts focusing on the inhibition of histone deacetylases (HDACs), polycomb repressive complex 2 (PRC2), and DNA methyltransferases. These interventions aim to re-sensitize tumors to chemotherapy and overcome drug resistance mechanisms. For example, CDKN2A loss and associated promoter methy-

lation have been explored as targets for cyclin-dependent kinase 4/6 inhibitors like palbociclib. Similarly, synovial sarcoma-specific translocations (e.g., SS18-SSX) offer potential targets for epigenetic agents. In terms of immunotherapy, PD-1 inhibitors and checkpoint blockade have shown some activity in select subtypes such as undifferentiated pleomorphic sarcoma and alveolar soft part sarcoma. Moreover, regional therapies like isolated limb perfusion (ILP) and infusion (ILI), using agents such as TNF- α and melphalan, have demonstrated limb salvage potential in otherwise unresectable recurrent STS, though these approaches remain limited to specialized centers or clinical trials [10], [11].

Recent research into OS treatment has focused heavily on understanding and targeting the epigenetic landscape of the disease. Epigenetic mechanisms such as DNA methylation, histone modifications, and non-coding RNAs (e.g., lncRNAs and miRNAs) are being investigated for their roles in tumor progression, metastasis, and chemoresistance. For instance, hypermethylation of tumor suppressor genes and deregulation of miRNAs like miR-149 and miR-195 are associated with aggressive tumor phenotypes. Novel therapeutic strategies aim to reverse these changes using demethylating agents such as 5-Aza or by targeting writer, reader, and eraser proteins in the methylation and histone modification pathways. Promising targets include DNMT1, METTL3, and histone demethylases like KDM6B, all of which have shown preclinical efficacy in reducing tumor growth and restoring chemosensitivity [16].

Denosumab, a monoclonal antibody targeting RANKL, represents a significant advancement in the management of GCTB, especially in surgically challenging cases or when joint preservation is desired. Denosumab works by inhibiting the RANK/RANKL pathway that drives osteoclastogenesis in GCTB, leading to reduced giant cell formation and tumor regression. However, concerns have been raised regarding its long-term use, including the potential for malignant transformation and challenges in defining surgical margins post-treatment due to histological changes. Ongoing research is also exploring the role of H3F3A gene mutations in GCTB, which may aid in more accurate diagnostics and personalized treatment strategies [18], [19].

Given the limited efficacy of conventional chemotherapy and radiotherapy in chondrosarcoma, recent studies have explored the tumor's molecular and epigenetic underpinnings to identify new therapeutic avenues. Mutations in the isocitrate dehydrogenase (IDH1/IDH2) genes are commonly found in central chondrosarcomas, leading to altered methylation patterns and a global DNA hypermethylation phenotype. These findings have stimulated interest in epigenetic therapies, such as DNA methyltransferase and histone deacetylase inhibitors, which are being evaluated for their potential to reverse tumor-promoting gene expression profiles. Although still in early phases, these approaches offer hope for patients with advanced or unresectable tumors resistant to standard treatment [20], [21].

In relapsed or metastatic Ewing sarcoma, where standard therapies yield poor

outcomes, novel approaches are under investigation. These include high-dose ifosfamide (found superior to several other regimens in the rEECur trial), targeted therapies such as cabozantinib (a tyrosine kinase inhibitor with reported anti-tumor activity), and agents targeting specific molecular features like the EWS-FLI1 fusion protein (e.g., TK216, currently in clinical trials). Immunotherapy agents like pembrolizumab and nivolumab have not shown promising results, likely due to low PDL-1 expression and minimal tumor-infiltrating lymphocytes. Additionally, proton therapy is gaining traction for its ability to spare normal tissue and reduce late toxicity, especially in pediatric patients and those with tumors near critical structures[25].

For patients with unresectable, recurrent, or metastatic DFSP, especially those harboring the COL1A1-PDGFB fusion, targeted molecular therapy with tyrosine kinase inhibitors has emerged as a promising approach. Notably, imatinib (STI571), which targets the PDGFR- β pathway, has been shown to induce tumor regression and apoptosis in DFSP cells in both clinical and in vitro studies. These findings support its use in select cases and highlight ongoing exploration of PDGF antagonists as potential future therapies[26].

Despite the many recent developments in the field, accurately predicting the development and prognosis remains difficult and obstacles remain for surgeons and clinicians [38].

Although, most sarcoma have no known cause, there are a few factors that could increase the risk of developing sarcoma. Exposure to radiotherapy may increase the chance of developing a sarcoma. Some genetic disorders have been shown to increase the risk of sarcoma such as Von Recklinghausen's disease, Gardner syndrome, Werner syndrome, tuberous sclerosis, nevoid basal cell carcinoma syndrome, Li-Fraumeni syndrome and retinoblastoma. Exposure to certain chemicals such as vinyl chloride monomer, dioxin and arsenic might lead to sarcoma. Long-term swelling in arms and legs may also heighten the risk of sarcoma developing [39].

Chapter 2

Toward New Assessment in Sarcoma Identification and Grading Using Artificial Intelligence Techniques

Abstract: Background/Objectives: Sarcomas are a rare and heterogeneous group of malignant tumors, which makes early detection and grading particularly challenging. Diagnosis traditionally relies on expert visual interpretation of histopathological biopsies and radiological imaging, processes that can be time-consuming, subjective and susceptible to inter-observer variability. **Methods:** In this study, we aim to explore the potential of artificial intelligence (AI), specifically radiomics and machine learning (ML), to support sarcoma diagnosis and grading based on MRI scans. We extracted quantitative features from both raw and wavelet-transformed images, including first-order statistics and texture descriptors such as the gray-level co-occurrence matrix (GLCM), gray-level size-zone matrix (GLSZM), gray-level run-length matrix (GLRLM), and neighboring gray tone difference matrix (NGTDM). These features were used to train ML models for two tasks: binary classification of healthy vs. pathological tissue and prognostic grading of sarcomas based on the French FNCLCC system. **Results:** The binary classification achieved an accuracy of 76.02% using a combination of features from both raw and transformed images. FNCLCC grade classification reached an accuracy of 57.6% under the same conditions. Specifically, wavelet transforms of raw images boosted classification accuracy, hinting at the large potential that image transforms can add to these tasks. **Conclusions:** Our findings highlight the value of combining multiple radiomic features and demonstrate that wavelet transforms significantly enhance classification performance. By outlining the potential of AI-based approaches in

sarcoma diagnostics, this work seeks to promote the development of decision support systems that could assist clinicians.

Keywords: sarcoma; radiomics; image transform; machine learning; classification

2.1 Introduction

Sarcomas are rare malignant tumors that develop in connective tissues such as bones, muscles, fat and cartilage. Due to their aggressive nature and histological diversity, sarcomas present significant clinical challenges in both diagnosis and treatment. Despite advances in oncology, accurately predicting the development and progression of sarcomas remains a major obstacle for surgeons and clinicians [38].

One primary challenge is diagnosing sarcomas, due to their variable clinical presentations. These tumors can appear in diverse and often nonspecific forms, leading to frequent delays in diagnosis and subsequent therapeutic intervention [40], [41]. Due to the rarity of sarcoma tumors, accurate diagnosis is often delayed, resulting in patients suffering a higher mortality rate. Therefore, developing methods for early detection and diagnosis is crucial to improve the quality of life and survivability of patients.

When diagnosing a sarcoma tumor, the practitioner determines a grade representing the severity of the cancer. This scale has three grades: 1, 2 and 3, with each grade increasing in severity. This grading is a critical factor for prognosis and treatment decisions and still largely relies on traditional histopathological evaluation methods [42]. Imaging exams play a crucial role in the diagnostic process for sarcomas. Conventional X-rays and CT scans are essential for evaluating bone sarcomas, while ultrasound and MRI are key for assessing soft tissue sarcomas. These imaging techniques form the foundation of research, providing essential data for accurate diagnosis and classification of these complex tumors.

Although histopathological evaluation is the standard practice, it has significant limitations. The subjectivity inherent in interpretation of results by pathologists, combined with the complex characteristics of sarcomas, can lead to variations in diagnosis. Nearly 70% of sarcomas are larger than 5 cm at the time of diagnosis, underscoring the impact of diagnostic delays [40]. Furthermore, the similar morphological features of certain sarcoma subtypes increase the risk of human error.

To address these challenges, genetic diagnosis has emerged as a promising solution to enhance the accuracy and speed of sarcoma detection [43]. Genetic analysis can reveal specific mutations and molecular characteristics often missed by traditional methods, offering a more personalized approach to treatment. However, despite its potential, genetic diagnosis alone can struggle with the complex

heterogeneity of sarcomas and may not always provide enough information for accurate malignancy grading or real-time decision-making [44]. This is where artificial intelligence (AI) comes in as a powerful complementary tool. While genetic data identifies key molecular drivers, AI can analyze imaging data to predict malignancy grades and detect subtle patterns missed by the human eye, bridging the gap where genetic analysis may fall short. Together, AI and genetic diagnosis have the potential to revolutionize sarcoma management, improving both the accuracy and speed of diagnosis and ultimately enhancing patient outcomes [45]. Given these challenges faced by both radiologists and pathologists, the integration of artificial intelligence (AI) into the diagnostic and classification process for sarcomas could offer innovative solutions. AI has the potential to standardize and significantly improve the accuracy of sarcoma diagnoses by reducing evaluation variability and providing supportive tools for these specialties. In fact, AI is already being implemented in various areas of healthcare, particularly in diagnostics, where it assists in interpreting medical images such as CT scans, MRIs and X-rays with remarkable accuracy [46], [47]. Additionally, AI is being used in the detection of diabetic retinopathy [48] and neurological conditions like Alzheimer's disease [49], enabling earlier intervention and improving patient outcomes.

The overall aim of this study was to analyze different radiomic features and the role they play in classifying healthy and pathological tissue, as well as sarcoma severity. This was done by training a random forest (RF) model to first classify healthy and pathological tissue, followed by training it to classify severity. This research aimed to demonstrate the feasibility and effectiveness of AI in supporting the diagnosis of sarcomas, highlighting the potential for integrating these technologies into routine clinical practice. We hope to contribute to improved patient management by providing an innovative tool that is complementary to existing diagnostic methods.

2.2 Methods

The workflow for this paper is outlined in Figure 2.1 and highlights the main steps. First, the raw magnetic resonance imaging (MRI) data are segmented into 3 masks—sarcoma, edema and healthy tissue—to produce masks. Then, simple linear iterative clustering (SLIC) segmentation is applied to the masks to generate super pixels (SPs). This is followed by radiomics feature extraction of each SP to generate a table of feature values, which are applied to raw image SPs and wavelet-transformed SPs. A feature selection step is applied separately for the different feature transform groups, and remaining features are finally used for the classification of healthy and pathological tissue (binary classification) and the severity or the grade of the sarcoma (grade classification) by applying a random forest (RF) model. Relevant metrics such as accuracy, precision, recall and confusion matrices

are calculated, along with feature importance, to assess and analyze the results.

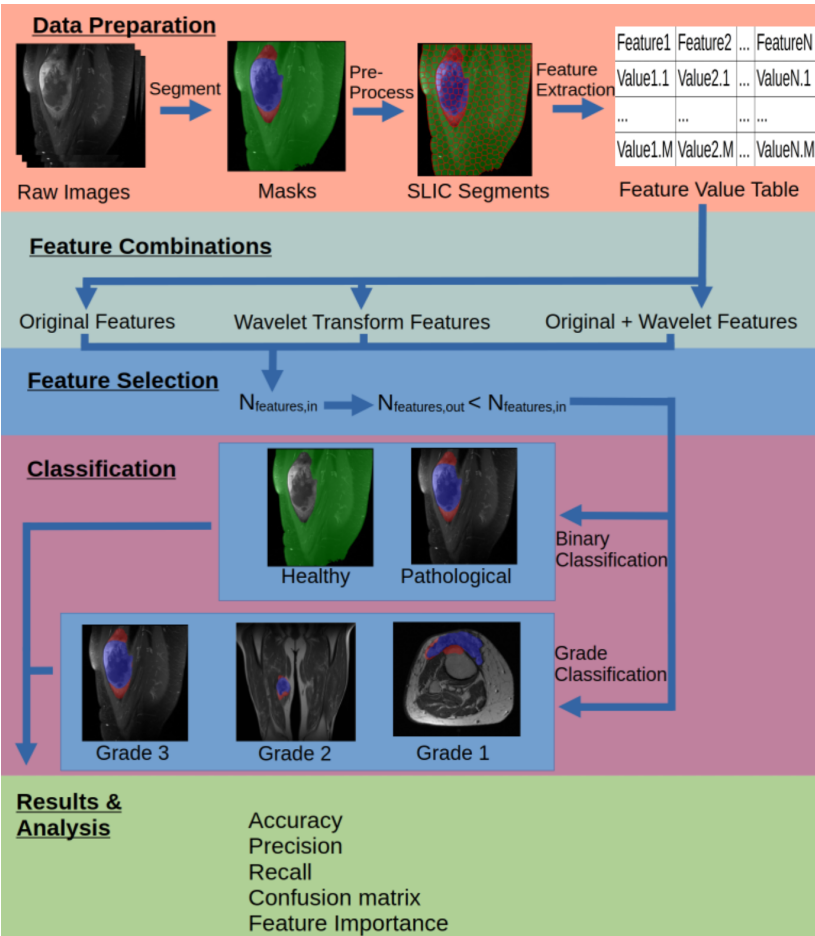


Figure 2.1: Visual representation of the steps in the workflow.

2.2.1 Data

This study used the IceSG (Icelandic Sarcoma Group) dataset, a collection of sarcoma cases assembled by Landspítali University Hospital and that is not publicly accessible. The use of the patient data was reviewed by the Landspítali Ethics Committee (5/2024 Sýndarvefjafræði í útlimasarkmeinum) and the Icelandic Scientific Ethics Committee (9 January 2024 tilv. 16 Sýndarvefjafræði í útlimasarkmeinum) and approved.

For the current work, 101 patients' MRI scans are considered for their ability to capture soft tissue details. If a patient has scans from multiple perspectives (axial, coronal, and/or sagittal), they are all included. Both T1 and T2 weighted scans are considered, with a mix of contrast media and with fat suppression both on and off. All sarcoma subtypes are considered. Out of the patient cohort, 19 are diagnosed with grade 1, 16 with grade 2 and 66 with grade 3. For each patient, the severity of prognosis is available at the time of the scan according to the French classification (FNCLCC) [50], which assigns a grade from 1 to 3 to the sarcoma, with 3 being the most severe.

Table 2.1 displays general demographics among age groups of the dataset, including sarcoma subtypes, FNCLCC grade and sex distributions.

Table 2.1: Dataset demographics. Included sarcoma subtypes are chondrosarcoma (CS), dermatofibrosarcoma protuberans (DFSP), Ewing sarcoma (ES), giant cell tumor of bone (GCTB), osteosarcoma (OS) and soft tissue sarcoma (STS).

Age Group		0–19		20–39		40–59		60–79		80+	
Subtype											
STS	4	3.96%	10	9.90%	19	18.81%	22	21.78%	10	9.90%	
OS	10	9.90%	0	0.00%	3	2.97%	1	0.99%	0	0.00%	
CS	0	0.00%	1	0.99%	6	5.94%	0	0.00%	1	0.99%	
ES	5	4.95%	4	3.96%	1	0.99%	1	0.99%	0	0.00%	
DFSP	0	0.00%	1	0.99%	0	0.00%	0	0.00%	0	0.00%	
GCTB	1	0.99%	0	0.00%	1	0.99%	0	0.00%	0	0.00%	
Grade											
1	3	2.97%	0	0.00%	10	9.90%	3	2.97%	3	2.97%	
2	1	0.99%	2	1.98%	3	2.97%	8	7.92%	2	1.98%	
3	16	15.84%	14	13.86%	17	16.83%	13	12.87%	6	5.94%	
Sex											
male	12	11.88%	8	7.92%	17	16.83%	14	13.86%	8	7.92%	
female	8	7.92%	8	7.92%	13	12.87%	10	9.90%	3	2.97%	

2.2.2 Segmentation

Segmentation is the process of marking regions of an image into a group to create a mask. In this study, the sarcoma tumor was segmented, as was the surrounding edema (where available), both of which were labeled as pathological tissue. The remaining tissue was segmented and labeled as non-pathological tissue; therefore, the healthy dataset was retrieved from the sarcoma patients. The orientation of scans was not standardized during data gathering; therefore, all orientations are considered, including axial, coronal and sagittal scans. All orientations available

for each patient were segmented, so there are up to 3 image sets segmented per patient.

Mimics (Materialize: Leuven, Belgium [51]) was used as the segmentation software. All segmentations were performed by hand and reviewed by a specialist. After segmentation, the raw images and resulting masks were exported for further processing.

2.2.3 Data Pre-Processing

After segmentation, 2946 images contained at least some trace of sarcoma and/or edema. There were 566 available images of grade 1 sarcoma, 395 of grade 2 and 1916 of grade 3. Due to the relatively small size of the dataset, a sampling technique was introduced to expand it for a purpose to that described by Karim Ham-moudi et al. [52].

Simple linear iterative clustering (SLIC) is an iterative and algorithmic method that splits an image into a user-defined number of segments based on similarity measures in pixel intensity [53]. Using this method, it is possible to further split the regions of interest into more data points for further processing, similar to Z. Khatun et al. [54], who segmented Achilles tendons using SLIC. In this manner, smaller details are captured, and the issue of a small dataset is somewhat addressed. The newly generated segments are referred to as super pixels (SPs). The target size of SPs is set to 500 pixels. This size should offer a good balance between capturing smaller and larger details. It should be noted that due to the nature of the SLIC algorithm, the actual number of pixels in each segment may vary. This research considers image slices individually; therefore, all generated super pixels are 2-dimensional.

The sampling method was applied separately for the pathological region and the non-pathological region. The technique generated approximately 40,000 pathological and 820,000 non-pathological data points, with roughly 5900 grade 1 data points, 3600 for grade 2 and 30,000 for grade 3. The approximate ratio of data points per patient is 312, 228 and 457 for grades 1, 2 and 3, respectively.

2.2.4 Feature Extraction

To acquire any meaningful information from the images and segmented regions, radiomics were employed to extract relevant features. Namely, the Python (version 3.9) PyRadiomics package [55] was applied for its vast set of extractable features that contain multiple feature categories. Feature groups include first-order statistics describing the distribution of pixel intensity (e.g., mean, variance and skewness). Gray-Level Co-occurrence Matrix (GLCM) features calculate features based on gray-level intensities and intensity relationships through the comparison of pixel pairs. Gray-Level Size-Zone Matrix (GLSZM) features are similar to GLCM

features but compare regional intensities instead of pair-wise relationships. Gray-Level Run-Length Matrix (GLRLM) features measure how pixels of similar intensity group together. Neighboring Gray Tone Difference Matrix (NGTDM) features focus on differences between pixels and their neighbors, quantifying details such as contrast. Shape features are omitted from this study, as only 2-dimensional SPs are generated; therefore, only 2-dimensional features can be extracted, whereas shape requires 3-dimensional image data.

Wavelet functions are often used to extract higher order features from images. They decompose an image into different components representing different frequency bands and have been shown to increase classification performance [56]. It is possible to apply a low-pass and a high-pass wavelet filter in each dimension (X and Y in this case) along each image, which creates new feature classes. The same features are extracted from the original raw image and the transformed images. Therefore, with the 4 combinations of wavelet transforms, the number of extracted features increases by five times. Table 2.2 shows the types of wavelet transforms and the details they capture. Features extracted from the raw image are referred to as original features, and those extracted from the wavelet transforms are referred to as wavelet features.

Table 2.2: Types of wavelet sub-band passes and captured details [57].

Sub-Band Symbol	X-Axis	Y-Axis	Details
LL	Low-pass	Low-pass	Approximation of original image
LH	Low-pass	High-pass	Images' horizontal details
HL	High-pass	Low-pass	Images' vertical details
HH	High-pass	High-pass	Images' diagonal details

All in all, a total of 465 features were extracted and analyzed for their classification ability (see Table 2.3).

Table 2.3: The combination of image transforms for each run.

Image Transform Group	Description	Number of features
1. Original	Raw image features	93 features
2. Wavelet	Wavelet-transformed features	279 features
3. Wavelet + Original	Raw and wavelet-transformed features	465 features

2.2.5 Feature Selection

The feature selection step is important to increase model robustness and computational efficiency. Given a large number of features, some may not correlate with the feature being predicted and can introduce noise, making classification more difficult. The chosen model for this study is the random forest (RF) classifier, as it is generally robust, accurate and scalable and has an internal feature-ranking mechanism that computes the importance of every feature [58]. RF has also seen many applications in sarcoma research and is supported by the literature, it is therefore a good candidate for this novel approach [59], [60].

As RFs are not very sensitive to highly correlated and low-variance features, the maximum allowable correlation between features is set to 0.99, and only a single correlated feature of a set is used. Once highly correlated features have been filtered out, the model is trained on the rest of the features and on the entire dataset to acquire a feature ranking.

To find the best set of features, a 10-fold cross validation is applied, splitting the dataset into 10 groups patient-wise, using 9 for training and 1 for testing. Features are eliminated, from lowest importance to highest, until an optimal model has been trained based on the average area under the receiver operating curve (AUC-ROC) score across all 10 folds. The AUC-ROC score is a measure of how well the model discriminates between the predicted classes or how well the classes are separated. The set of features that results give the best AUC-ROC score is used for subsequent training.

2.2.6 Binary Classification

After data pre-processing, the dataset is split into 2 groups: healthy and pathological, with the aim of analyzing which radiomic features best distinguish between healthy and pathological tissue. The healthy cohort of data points is randomly selected from the the SPs generated from the healthy mask component of the diagnosed patients. Each patient contributes an equal number of pathological and healthy data points. In this manner, a balanced dataset for training and testing is ensured.

A leave-1-out cross-validation using an RF classifier is applied such that the training dataset is maximized and as much variance as possible is captured by it and explained by the model.

Three types of analyses are conducted, using different feature combinations, to explore the effects that they have on the overall results. The first run includes only raw image (original) features. The second considers only wavelet-transformed features, omitting the LL wavelet transform, as it is an approximation of the raw image. The third run combines original features with the wavelet features. Refer to Table 2.3 for the combination of image transform spaces for each run. Feature

selection is applied independently on each combination of image transforms, selecting the only features used for the three binary classifiers; therefore, the number of optimal features can differ.

The accuracy, precision, recall and the confusion matrix (see Figure 2.2a for confusion matrix description) are the metrics calculated to assess each feature set, and feature importance is calculated to analyze their effects in classification. Averaged macro and micro metrics are also calculated. Macro metrics give equal weight to each patient during calculations, and micro metrics give equal weight to each data point during calculations. The macro and micro confusion matrices are expressed as follows:

$$\mathbb{C}_{Ma} = \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{\mathbb{C}_i}{\sum_{j,k} c_{i,j,k}} \quad (2.1)$$

$$\mathbb{C}_{Mi} = \sum_i^p \mathbb{C}_i \quad (2.2)$$

where N_p is the total number of patients, $\mathbb{C}_i$ is the confusion matrix of the i^{th} patient and $c_{i,j,k}$ denotes the value at line j and row k of the i^{th} patient's confusion matrix.

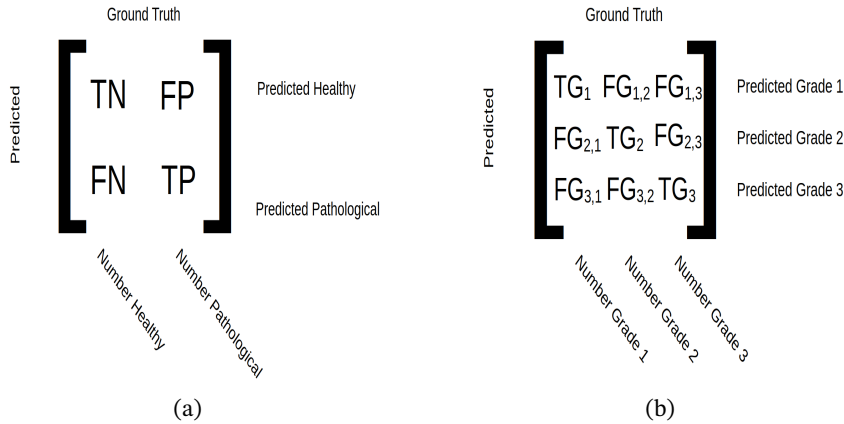


Figure 2.2: Confusion matrix setup. **(a)** For binary classification, true negative (TN) indicates tissue correctly predicted as healthy, true positive (TP) indicates tissue correctly predicted as pathological, false negative (FN) indicates pathological tissue predicted as healthy and false positive (FP) indicates pathological tissue predicted as healthy. The micro average displays the number of samples classified in integers, and macro average displays classification in ratios. **(b)** For grade classification, true grade (TG_i) is the number of samples belonging to class i and predicted as such, and false grade ($FG_{i,j}$) corresponds to samples predicted as grade i but belonging to grade j . The micro average shows the total number of classified data points, and the macro average shows number of classified patients.

2.2.7 Grade Classification

Following the FNCLCC guidelines, there are 3 potential severities of sarcoma prognosis: grades 1, 2 and 3. These are based off of tumor differentiation, mitotic count and tumor necrosis. As the aim is to extract radiomic data from the patient MRI scans to analyze which features might explain sarcoma severity, only the grade is applied for prediction purposes, along with extracted radiomics data.

The dataset comprises an imbalanced cohort of patients: 19 patients receiving a grade 1 diagnosis, 16 with grade 2 and 66 patients with grade 3 sarcoma. This issue is tackled using a balanced random forest model [61] such that the models do not overly adhere to one specific class.

As for binary classification, three analyses are compiled (see Table 2.3). The first only considers original features; the second only considers wavelet transforms, excluding the LL wavelet transform; and the third is a combination of the feature sets. Again, feature selection is applied independently on each combination of image transforms, so the number of features can differ for each transform combination trial. Macro and micro accuracy, precision, recall and the confusion

matrix (see Figure 2.2b for confusion matrix description) are calculated and used for assessment, as is the feature importance. In this case, macro-averaged calculations depend on how the patient was classified (based on majority rules), and micro-averaged calculations depend on how each individual segment was classified.

The micro confusion matrix is calculated using Equation (2.2), and the macro confusion matrix is calculated as follows:

$$\mathbb{C}_{Ma} = \sum_i^{N_p} \mathbb{C}'_i \quad (2.3)$$

where

$$\mathbb{C}'_i : c'_{i,j,k} = \begin{cases} 1, & c_{i,j,k} = \max(\mathbb{C}_i) \\ 0, & \text{otherwise} \end{cases} \quad (2.4)$$

where subscript i, j, k denotes the values of the j^{th} row and k^{th} column in the i^{th} patient's confusion matrix.

2.3 Results

The presented results are based on the leave-one-out cross-validation applied to the images processed from the IceSG dataset. After SLIC was applied to each image to generate super pixels (SPs), features were extracted from the raw image, and the four wavelet transform passes. A random forest classifier was applied first to the raw (original) images, then to the wavelet-transformed images and, finally, to the combination of the two to first distinguish between healthy and pathological tissue (binary classification), then to classify the severity or grade of the prognosis. Macro- and micro-averaged accuracy, precision, recall and confusion matrices, along with feature importance, were calculated and used to quantify each run.

The resulting accuracy, precision and recall for the different combinations of features show a general improvement when using more features. First-order statistical and GLCM features were generally the most important, and GLSZM and NGTDM were generally the lowest ranked feature groups. No single transform feature group performed significantly better throughout all runs, with importance varying depending on the situation.

2.3.1 Binary Classification: Healthy–Pathological

Table 2.4 displays the overall results of the binary classification results. The macro- and micro-averaged accuracy, precision and recall are displayed for original, wavelet and combined feature groups.

Table 2.4: Binary classification results. Both macro-averaged and micro-averaged metrics are displayed.

	Macro-Averaged			Micro-Averaged		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Original	72.66%	76.80%	68.80%	68.70%	72.14%	67.79%
Wavelet	70.16%	72.75%	66.68%	64.79%	66.86%	62.27%
Combined	76.02%	80.88%	71.82%	71.88%	75.82%	68.05%

2.3.1.1 Original Features

When considering only original features, macro-averaged accuracy, precision and recall of 72.66%, 76.80% and 68.80% were obtained, respectively. The micro-averaged accuracy, precision and recall were 68.70%, 72.14% and 67.79%, respectively. The following confusion matrices were obtained:

$$\begin{aligned} \text{Macro Confusion Matrix:} & \begin{bmatrix} 0.3826 & 0.1174 \\ 0.1560 & 0.3440 \end{bmatrix} \\ \text{Micro Confusion Matrix:} & \begin{bmatrix} 27686 & 12088 \\ 12810 & 26964 \end{bmatrix} \end{aligned}$$

Table 2.5 displays the number of features chosen from each feature group, and Table 2.6 displays the cumulative importance for each feature group. A total of 47 features were selected, with all feature groups being represented. First order ranked as the most important, followed by GLCM, GLDM, GLRLM, GLSZM and NGTDM.

Table 2.5: Number of features chosen for each feature group for the binary classifier for the original transform group.

Feature Group	Selected Features
First Order	14
GLCM	9
GLDM	8
GLRLM	6
GLSZM	5
NGTDM	5
Total	47

Table 2.6: Feature importance considering purely the original transform group for binary classification.

Feature Group	Original
First Order	35.60%
GLCM	16.75%
GLDM	15.95%
GLRLM	12.23%
GLSZM	9.58%
NGTDM	9.90%
Total	100.00%

2.3.1.2 Wavelet Features

Considering pure wavelet features, omitting the LL wavelet pass of the transform resulted in a macro-averaged accuracy, precision and recall of 70.16%, 72.75% and 66.68%, respectively. The micro-averaged accuracy, precision and recall were 64.79%, 66.86% and 62.27%, respectively. The following confusion matrices were obtained:

$$\begin{array}{l} \text{Macro Confusion Matrix:} \\ \text{Micro Confusion Matrix:} \end{array} \begin{bmatrix} 0.3682 & 0.1318 \\ 0.1666 & 0.3334 \\ 26773 & 13001 \\ 15008 & 24766 \end{bmatrix}$$

Table 2.7 displays the number of features chosen from each feature group, and Table 2.8 displays the cumulative importance for each feature group. A total of 71 features were selected, with nearly all feature groups being represented for each wavelet transform. The highest ranking feature group in terms of importance was first order, followed by GLCM, GLDM, GLRLM, GLSZM and NGTDM. The wavelet-HL transform was the most important, followed by wavelet-LH and and wavelet-HH transforms.

Table 2.7: Number of features chosen for each feature group for the binary classifier for the wavelet transform group.

Feature Group	Selected Features			Total
	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	9	10	4	23
GLCM	9	5	5	19
GLDM	6	5	1	12
GLRLM	5	2	4	11
GLSZM	2	0	0	2
NGTDM	1	2	1	4
Total	32	24	15	71

Table 2.8: Feature importance considering only the wavelet transform group for binary classification.

Feature Group	Feature-Transform Group Importance			Total
	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	14.28%	16.08%	5.39%	35.76%
GLCM	12.12%	7.04%	7.06%	26.22%
GLDM	8.25%	6.65%	1.38%	16.28%
GLRLM	6.34%	2.78%	5.33%	14.45%
GLSZM	2.31%	0.00%	0.00%	2.31%
NGTDM	1.19%	2.50%	1.30%	4.99%
Total	44.49%	35.05%	20.46%	100.00%

2.3.1.3 Combined Features

Combining the wavelet transform feature groups with the original feature group boosted performance and resulted in a macro-averaged accuracy, precision and recall of 76.02%, 80.88% and 71.82%, respectively. The micro-averaged accuracy, precision and recall were 71.88%, 75.82% and 68.05%, respectively. The following confusion matrices were obtained:

$$\begin{array}{l} \text{Macro Confusion Matrix:} \\ \text{Micro Confusion Matrix:} \end{array} \begin{bmatrix} 0.4011 & 0.0989 \\ 0.1409 & 0.3591 \\ 30114 & 9660 \\ 12706 & 27068 \end{bmatrix}$$

Table 2.9 displays the cumulative importance for each feature group, and Table 2.10 displays the number of features chosen from each feature group. A total of 200 features were selected, with all but one transform feature group being represented for each transform group. The highest ranking feature group in

terms of importance was first order, followed by GLCM, GLRLM, GLDM, GLSZM and NGTDM. The original transform group was the most important, followed by wavelet-LL, wavelet-HL, wavelet-HH and wavelet-LH.

Table 2.9: Feature importance considering both original and wavelet transform groups for binary classification.

	Feature Transform Group Importance					Total
	Original	Wavelet-LL	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	13.75%	7.85%	1.86%	2.96%	3.26%	29.67%
GLCM	4.00%	4.56%	3.51%	5.05%	5.35%	22.47%
GLDM	3.99%	3.48%	4.06%	2.38%	2.70%	16.61%
GLRLM	2.57%	3.58%	3.27%	2.83%	4.71%	16.96%
GLSZM	2.10%	2.67%	0.78%	0.80%	0.00%	6.34%
NGTDM	1.65%	2.18%	1.37%	1.50%	1.26%	7.95%
Total	28.06%	24.32%	14.84%	15.51%	17.26%	100.00%

Table 2.10: Number of features chosen from each feature group for the binary classifier of all transform groups.

	Selected Features					Total
	Original	Wavelet-LL	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	15	10	4	7	9	45
GLCM	10	9	6	12	12	49
GLDM	8	7	8	6	7	36
GLRLM	5	7	7	7	10	36
GLSZM	6	6	2	2	0	16
NGTDM	4	4	3	4	3	18
Total	48	43	30	38	41	200

2.3.2 Grade Classification: Sarcoma Grade

2.3.2.1 Original Features

Table 2.11 displays the results of grade classification using only original features. The table implies a higher difficulty in classifying grades 1 and 2 when compared to grade 3. Overall accuracy shows that the model is able to classify above chance (33%), but precision and recall for grades 1 and 2 could use a lot of improvement.

Table 2.11: Macro- and micro-averaged results for grade classification using original features.

Macro				Micro		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Accuracy	45.54%			46.80%		
Precision	36.84%	31.25%	51.52%	28.39%	44.97%	50.64%
Recall	21.88%	27.78%	66.67%	21.88%	27.78%	66.67%

The following confusion matrices were obtained for the macro- and micro-averaged classification results:

Macro Confusion Matrix:

$$\begin{bmatrix} 7 & 5 & 20 \\ 1 & 5 & 12 \\ 11 & 6 & 34 \end{bmatrix}$$

Micro Confusion Matrix:

$$\begin{bmatrix} 1685 & 768 & 7829 \\ 1407 & 1644 & 7068 \\ 2843 & 1244 & 15286 \end{bmatrix}$$

The number of chosen features and feature importance can be seen in Tables 2.12 and 2.13, respectively. Unsurprisingly, first-order features ranked the highest in importance, followed by GLSZM, NGTDM, GLCM, GLDM and GLRLM.

Table 2.12: Number of features selected when considering the original transform group for grade classification.

Feature Group	Selected Features
First Order	11
GLCM	3
GLDM	3
GLRLM	1
GLSZM	8
NGTDM	5
Total	31

Table 2.13: Feature importance considering purely the original feature groups for grade classification.

Feature Group	Feature Group Importance
First Order	37.81%
GLCM	8.94%
GLDM	8.88%
GLRLM	3.12%
GLSZM	22.82%
NGTDM	18.43%
Total	100.00%

2.3.2.2 Wavelet Features

Table 2.14 displays the results of grade classification, applying only wavelet features. The results are quite similar to the original feature run; there were general improvements to the metrics, with a slight reduction in macro accuracy, grade 3 precision, grade 2 recall and micro grade 2 recall.

Table 2.14: Macro- and micro-averaged results for grade classification using wavelet features.

	Macro			Micro		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Accuracy		43.56%			53.28%	
Precision	42.11%	56.25%	40.91%	34.81%	55.94%	56.59%
Recall	29.63%	26.47%	67.50%	29.63%	26.47%	67.50%

The following macro- and micro-averaged confusion matrices were obtained:

$$\text{Macro Confusion Matrix: } \begin{bmatrix} 8 & 1 & 18 \\ 4 & 9 & 21 \\ 7 & 6 & 27 \end{bmatrix}$$

$$\text{Micro Confusion Matrix: } \begin{bmatrix} 2066 & 749 & 6551 \\ 1736 & 2045 & 6551 \\ 2133 & 862 & 17081 \end{bmatrix}$$

The number of selected features and the feature importance for each wavelet transform type can be seen in Tables 2.15 and 2.16, respectively. The tables indicate that the wavelet-HH transform was most important, followed by wavelet-HL and wavelet-LH. GLCM features ranked as most important, followed by GLRLM, first order, GLDM, NGTDM and GLSZM.

Table 2.15: Number of features selected when considering the wavelet transform group for grade classification.

	Selected Features			Total
	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	0	2	3	5
GLCM	4	4	3	11
GLDM	0	0	4	4
GLRLM	1	1	4	6
GLSZM	0	0	0	0
NGTDM	0	1	0	1
Total	5	8	14	27

Table 2.16: Feature importance considering only wavelet transform feature groups for grade classification.

	Feature Transform Group Importance			Total
	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	0.00%	7.29%	12.36%	19.65%
GLCM	17.78%	18.00%	9.47%	45.25%
GLDM	0.00%	0.00%	11.81%	11.81%
GLRLM	3.34%	3.48%	13.02%	19.84%
GLSZM	0.00%	0.00%	0.00%	0.00%
NGTDM	0.00%	3.44%	0.00%	3.44%
Total	21.12%	32.22%	46.67%	100.00%

2.3.2.3 Combined Features

Using the combined feature set of both original and wavelet features showed some general improvements (see Table 2.17). Macro-averaged accuracy is slightly lower when compared to the pure wavelet transform, although the micro-averaged is at the highest of all runs. Grade 3 macro- and micro-averaged precision and recall were relatively high compared to the other runs, and grade 2 precision was quite high, while the other metrics suffered.

Table 2.17: Macro- and micro-averaged results for grade classification using all transform groups.

	Macro			Micro		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Accuracy	44.55%			57.60%		
Precision	31.58%	31.25%	51.52%	32.60%	60.20%	62.20%
Recall	24.00%	23.81%	61.82%	24.00%	23.81%	61.82%

The resulting confusion matrices are

$$\text{Macro Confusion Matrix: } \begin{bmatrix} 6 & 3 & 16 \\ 0 & 5 & 16 \\ 13 & 8 & 34 \end{bmatrix}$$

$$\text{Micro Confusion Matrix: } \begin{bmatrix} 1935 & 576 & 6577 \\ 1142 & 2201 & 4831 \\ 2858 & 879 & 18775 \end{bmatrix}$$

Tables 2.18 and 2.19 show the number of selected features and the feature importance, respectively. The highest ranking feature group was first order, followed by GLCM, GLRLM, NGTDM, GLDM and GLSZM. Wavelet-HH was the most important transform group, followed by the original, wavelet-LL, wavelet-HL and, finally, wavelet-LH.

Table 2.18: Number of features selected when considering both original and wavelet transform groups for grade classification.

	Selected Features					Total
	Original	Wavelet-LL	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	10	5	0	0	1	16
GLCM	0	1	2	4	3	10
GLDM	0	0	0	0	4	4
GLRLM	1	2	0	3	6	12
GLSZM	1	1	0	0	0	2
NGTDM	2	2	1	0	0	5
Total	14	11	3	7	14	49

Table 2.19: Feature importance considering both the original and wavelet transform feature groups for grade classification.

	Feature Transform Group Importance					Total
	Original	Wavelet-LL	Wavelet-LH	Wavelet-HL	Wavelet-HH	
First Order	19.57%	11.05%	0.00%	0.00%	2.47%	33.08%
GLCM	0.00%	1.62%	5.91%	9.19%	5.93%	22.65%
GLDM	0.00%	0.00%	0.00%	0.00%	7.60%	7.60%
GLRLM	1.64%	2.62%	0.00%	6.16%	11.68%	22.10%
GLSZM	1.35%	1.36%	0.00%	0.00%	0.00%	2.71%
NGTDM	4.57%	4.31%	2.97%	0.00%	0.00%	11.86%
Total	27.13%	20.97%	8.88%	15.34%	27.68%	100.00%

2.4 Discussion

The results of this study highlight the potential of radiomics for sarcoma classification. Tagliafico et al. [62] found that GLCM features were key in distinguishing healthy tissue from local recurrence (LR) in soft tissue sarcoma (STS), a finding consistent with this paper's results. They also identified four features correlated with tumor grade. Yu Zhang et al. [63] used a standardized MRI protocol on 33 patients to successfully differentiate between low- and high-grade STS. Similarly, Corino et al. [64] showed that first-order statistics were the most important features in distinguishing between high- and intermediate-grade sarcomas. While few studies have applied wavelet transforms for sarcoma classification, Shan Hu et al. [65] found that Sym4 and Db4 wavelet transforms combined with GLCM could accurately detect sarcoma. Peng Lin et al. [66] applied 3D wavelet transforms for preoperative osteosarcoma (OS) evaluation with positive results, and Silin Chen et al. [67] combined raw and wavelet features for STS prognosis prediction. Niu et al. [68] applied various image transformations to MRI glioma volumes to predict tumor grade with high accuracy. Radiomics-based predictive modeling has also demonstrated use cases beyond oncology. Ricciardi et al. [69] applied tree-based algorithms (including random forests and gradient boosting) to radio densitometric parameters obtained from mid-thigh CT scans. They achieved high performance for coronary heart disease, cardiovascular disease and chronic heart failure, supporting the relevance of radiomic features for clinical risk stratification.

The feature importance analysis aligns with existing literature [62], [64], confirming the utility of wavelet transforms to improve random forest (RF) model performance [65], [66]. Most studies focus on a single sarcoma subtype with standardized imaging, which improves model generalization. However, there are several limitations in this study: dataset complexity, the use of image transform algo-

rhythms, dataset imbalance and limitations associated with clinical sarcoma grading.

2.4.1 Dataset Complexity

The dataset is complex, as sarcomas are diverse, making generalization difficult [46]. Variations in MRI scan protocols, such as differing contrast, fat suppression and weighting methods, contribute to variability in tumor appearance (Figure 2.3). Considering multiple orientations—axial, coronal and sagittal—adds complexity that the model must handle. Additionally, differences in gray-value ranges and MRI specifications complicate standardization. Some sarcoma types, like OS, are under-represented, making grade classification challenging, as the model may misclassify OS as a lower grade. Including sarcoma subtype information could improve classification.

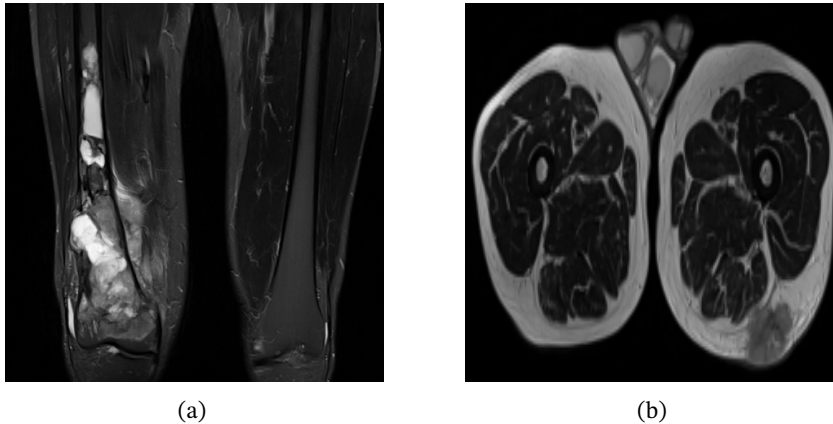


Figure 2.3: MRI slices: **(a)** A case of OS where the machine uses specific settings and contrast media to brighten the tumor region and suppress fat response. **(b)** Another patient (with liposarcoma in this case) where the tumor region is dark in contrast and fat tissue is not suppressed.

Non-pathological tissue was grouped as healthy, which can cause misclassification, as biological tissue may resemble pathological regions (Figure 2.4). Small homogeneous regions in segmented 3D images may confuse the model, especially when similar to healthy tissue. Image quality varied significantly between patients (Figure 2.5), with some images being grainy, blurred or corrupted, which may affect model performance.

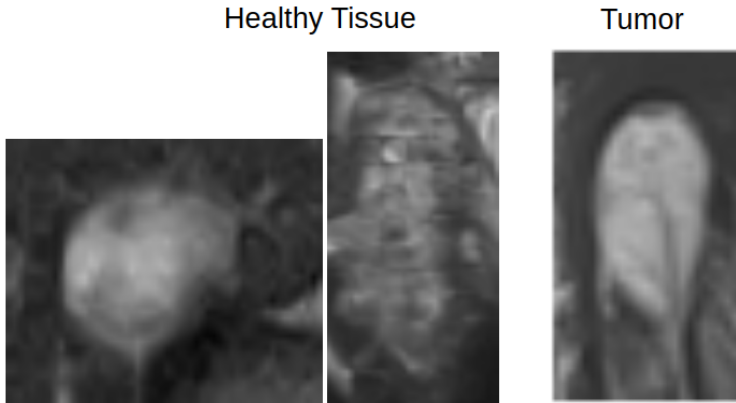


Figure 2.4: An example of how healthy and pathological tissue can overlap in terms of appearance. The left two images show healthy tissue, and the rightmost image shows a tumor region.

To improve generalization, it is crucial to balance the dataset by ensuring adequate representation of all sarcoma types and grades. For grade classification, the recommended number of samples per feature is at least 10 [70], with a maximum of 39 features for three-grade classification based on the dataset's smallest minority-class size when considering each slice of the mask as a data point. Hence, SLIC was used to cluster pixels into super pixels for further analysis.

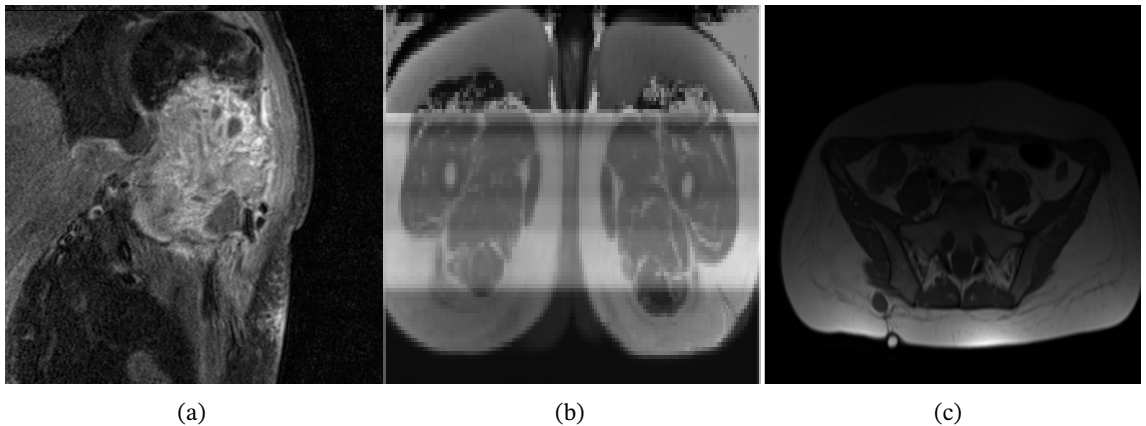


Figure 2.5: Variance in image quality. (a) A grainy slice with relatively high noise. (b) A corrupted image. (c) An example of uneven lighting distribution.

2.4.2 Image Transform Algorithms

Wavelet transforms were used for feature extraction, but other transforms, such as Hessian transform [71], could provide additional relevant information. Applying different transforms, such as Laplacian of Gaussian or gradient transforms [55], may uncover additional textural or topological features that could improve classification performance. However, caution is needed when using higher order transformations, as they may amplify image noise [72].

While combining original and transformed features improved performance for binary classification, there was no single feature that consistently predicted outcomes for all patients, underscoring the complexity of feature interactions and suggesting that optimal feature selection could improve results. Using the combined dataset did not improve performance to a notable degree for grade classification, which was unexpected. This implies that improvements must be made to the feature selection or, perhaps, indicating the need to include more transforms.

2.4.3 Dataset Imbalance

To balance healthy and pathological groups, healthy data were randomly selected per patient to match the number of pathological data points. However, this process could lead to slight variations, especially in patients with small tumors. Over-sampling could address this issue, improving generalization during training. Despite dataset challenges, the classification results were generally successful, highlighting the interdependence of features. The findings indicate that there is considerable potential in identifying sarcoma tissue based on radiomics features.

The under-representation of grade 1 and 2 tumors further complicates the task, as grade 3 tumors, being larger, generate more SPs [73]. Precision and recall were poorer for grades 1 and 2, as seen in Tables 2.11, 2.14 and 2.17.

To address class imbalance, a balanced random forest model was used [61], but other techniques, such as random oversampling or ensemble methods, could also be effective [58], [61]. The results suggest that improvements can be made in grade classification, particularly for grades 1 and 2, though the low number of cases in these categories should be considered.

However, the overall results offer a possible pathway of applying computerized models to accurately diagnose patients. This could decrease the time taken to diagnose patients, as no invasive biopsies are required and another practitioner does not have to analyze the results of the biopsy.

2.4.4 Clinical Sarcoma Grading

Sarcoma grading, according to FNCLCC guidelines, relies on tumor differentiation, the necrosis ratio and the mitotic count. These factors are subject to human

error, making grade classification particularly challenging and leading to potential misdiagnosis of the sarcoma grade.

2.5 Conclusions

The results of training an RF classifier on the IceSG dataset to discriminate healthy versus pathological tissue and classify the sarcoma grade using radiomics features extracted from raw images and image wavelet transforms show considerable promise. The results indicate the following:

- A pathway may exist to high-accuracy identification of sarcomas and their grades.
- Medical imaging shows promise in bypassing invasive biopsies.
- Possible human error in complicated diagnostics can be minimized or avoided.
- Treatment of sarcoma can be started earlier, thereby promoting patient prognosis.

Chapter 3

An Explainable Radiomics-Based Classification Model for Sarcoma Diagnosis

Abstract: Objective: This study introduces an explainable, radiomics-based machine learning framework for the automated classification of sarcoma tumors using MRI. The approach aims to empower clinicians, reducing dependence on subjective image interpretation. Methods: A total of 186 MRI scans from 86 patients diagnosed with bone and soft tissue sarcoma were manually segmented to isolate tumor regions and corresponding healthy tissue. From these segmentations, 851 handcrafted radiomic features were extracted, including wavelet-transformed descriptors. A Random Forest classifier was trained to distinguish between tumor and healthy tissue, with hyperparameter tuning performed through nested cross-validation. To ensure transparency and interpretability, model behavior was explored through Feature Importance analysis and Local Interpretable Model-agnostic Explanations (LIME). Results: The model achieved an F1-score of 0.742, with an accuracy of 0.724 on the test set. LIME analysis revealed that texture and wavelet-based features were the most influential in driving the model's predictions. Conclusions: By enabling accurate and interpretable classification of sarcomas in MRI, the proposed method provides a non-invasive approach to tumor classification, supporting an earlier, more personalized and precision-driven diagnosis. This study highlights the potential of explainable AI to assist in more secure clinical decision-making.

Keywords: Sarcoma diagnosis; Radiomics; Machine Learning; Classification; Explainability

3.1 Introduction

Sarcomas are a rare and diverse group of malignant tumors arising from mesenchymal tissue [38], [74]. In the United States, sarcomas account for 1% of new cancer cases and cancer-related deaths [75]. But this statistic hides a sad truth: among children and adolescents, sarcomas strike with disproportionate force, responsible for 13% of all cancer-related deaths in patients between 0 and 19 years [76]. Several studies have reported a higher incidence of sarcoma in males and it is particularly evident in childhood sarcomas, which show a stronger association with the male sex. However, the biological mechanisms underlying these sex differences remain poorly understood [77].

Clinical signs are often nonspecific, which frequently leads to diagnostic delays and negatively impacts patient outcomes [78]. Moreover, since soft tissue sarcoma is a rare tumor, general practitioners typically diagnose only one case in 24 years of practice. Their histological and molecular heterogeneity, combined with low incidence and aggressive clinical behavior, pose significant challenges for timely and accurate diagnosis and treatment [38]. Early diagnosis improves outcomes, as it can prevent the growth of the mass and decrease the risk of metastases, which are predictors of survival [78]. As a matter of fact, Pisters et al. reported that patients with soft tissue sarcomas larger than 5 cm had a relative risk of death 2.1 times higher than those with tumors smaller than 5 cm [79]. Then, Kolovich et al. reported that each additional centimeter in tumor size leads to a 3-5% reduction in the survival rate.

Radiological imaging plays a crucial role in the diagnostic process. For bone sarcomas, radiography, MRI, and CT scans are the primary diagnostic tools. Instead, ultrasound and MRI are the main modalities used to evaluate soft-tissue lesions [8], [80]. However, their interpretation remains subjective, as it relies on the radiologist's expertise [81].

Although histopathology remains the gold standard for definitive diagnosis, it is still obtained after an invasive procedure and therefore not without limitations. To address these challenges, genetic profiling has emerged as a valuable complementary approach, identifying subtype-specific mutations and molecular signatures that enhance classification accuracy and support targeted therapeutic strategies [43], [82].

Nowadays, Artificial intelligence (AI) offers a helping hand to physicians to address these limitations [83]. Through the extraction and analysis of quantitative features, AI algorithms can identify nuanced patterns imperceptible to human observers, facilitate standardized diagnostic workflows, and potentially enhance diagnostic accuracy [46], [47], [84].

This study evaluates the possibility of using machine learning, in particular Random Forest (RF), to distinguish healthy tissue from sarcomas across a broad spectrum of anatomic sites, MRI protocols, and different acquisition image con-

ditions. Radiomic features were extracted from manually segmented MR images and used to train a RF classifier. Using a heterogeneous, real-world dataset collected since 2004, we assess the robustness and generalizability of the proposed approach. Our broader aim is to demonstrate how AI-driven image analysis can serve as a second opinion alongside traditional diagnostic workflows, helping to reduce inter-observer variability and ultimately contributing to more consistent and effective patient management through a more tailored diagnostic process.

The following section describes the Materials and Methods. Section 3.3 reports the results obtained, followed by Section 3.4, which includes a review of the state of the art. Finally, the Conclusion section summarizes the main findings.

3.2 Materials and Methods

3.2.1 Dataset

3.2.1.1 Sarcoma Cases

IceSG - Medical Information

The IceSG (Icelandic Sarcoma Group) dataset consists of sarcoma cases collected by Landspítali University Hospital. The patient data were approved by the Landspítali Ethics Committee (no. 5), dated 06/02/2024, while the study received approval from Landspítali Scientific Research Committee (No. 93), dated 09/01/2024. The dataset comprises clinical information on patients diagnosed with various bone and soft tissue tumors, with diagnostic examinations performed over a 20-year period, from 2004 to 2023. The average age at diagnosis is approximately 48.78 (± 23.52) years, reflecting a wide age distribution, with patient ages ranging from a minimum of 5 months to a maximum of 93 years, at diagnosis.

The majority of cases are classified as soft tissue sarcomas, accounting for 54 patients. Osteosarcoma is the second most frequent diagnosis, with 14 cases, followed by chondrosarcoma and Ewing sarcoma, each with 7 cases. Less common diagnoses include dermatofibrosarcoma protuberans, with a single case, and giant cell tumor of bone, with two cases.

IceSG - Technical Information

To develop and evaluate the proposed method, MRI scans from the IceSG were used. The dataset includes a total of 86 patients diagnosed with various sarcoma types. Since each patient underwent imaging in multiple anatomical planes, the study comprises 186 MRI exams. Although both T1- and T2-weighted sequences were included in the study to enhance robustness and generalizability, each patient was not necessarily represented by both modalities. These scans also include contrast-enhanced and non-enhanced images, as well as sequences acquired with and without fat suppression, thereby capturing a broad spectrum of real-world imaging conditions. The distribution of MRI exam types, including both contrast-

enhanced and non-contrast sequences, is summarized in Table 3.1. None of the MRI sequence groups contains enough exams to be used individually; therefore, the sequences were aggregated to ensure a meaningful sample size, resulting in a total of 186 MRI exams.

Table 3.1: Number of MRI exams by sequence type, combining contrast-enhanced variants.

Type	# Exams T1	# Exams T2
No Fat Saturation and No Contrast	31	54
With Fat Saturation	13	25
With Fat Saturation and Contrast	29	0
With Contrast	19	0
STIR	1	
Total	95	91

In particular, the STIR case was not excluded since the tumor appearance is visually similar to that observed in T2 with Fat Saturation. Casale et al., in their study on metastasis prediction in soft tissue sarcoma cases using formal methods, employed both STIR and T2 images with fat saturation [85]. Due to the similarity between these sequences, including the STIR exam in our study does not introduce significant variability.

In Figure 3.1 and Figure 3.2 illustrate examples of the different images used.

3.2.1.2 Healthy Cases

Although the dataset does not include scans from healthy subjects, regions of healthy tissue were extracted from the MRI scans of sarcoma patients, as illustrated in Figure 3.3. These regions were manually selected to represent non-pathological anatomy and were used to provide negative samples during model training. This method yielded 175 samples, contributing to a dataset balanced not only in terms of sample count. As a matter of fact, this approach ensures that both tumor and non-tumor tissue characteristics are learned within the same imaging context. The number of healthy tissue samples is moderately lower than that of sarcoma samples, as not all scans contained sufficient healthy tissue for safe segmentation. In such cases, the corresponding healthy images were excluded to prevent volumes from potentially containing tumor traces.

Figure 3.3.A shows the original MRI image with a unilateral soft tissue sarcoma. Figure 3.3.B displays the manual segmentation of the tumor region (blue) on the affected side, while Figure 3.3.C illustrates the corresponding extraction of healthy tissue regions (light blue) from the contralateral limb. These were used as

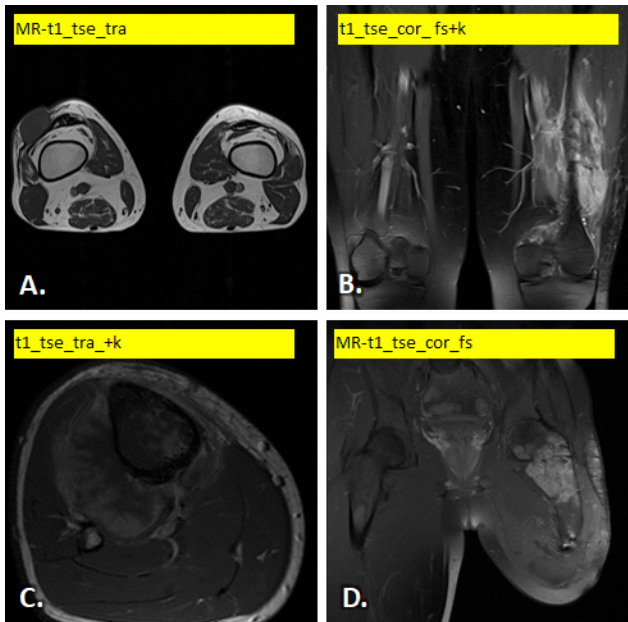


Figure 3.1: Example of T1-weighted MRI scan variants.

negative samples for training, in contrast to the tumor region (blue) used as the positive sample. This strategy enables intra-patient comparison under consistent imaging conditions.

3.2.2 Methods

The workflow started with the segmentation of both sarcoma lesions and healthy tissue volumes from MRI scans. Radiomic features were then extracted from these segmented regions to quantitatively characterize tissue properties. A Random Forest classifier was employed for the classification task [86]. To optimize performance and reduce overfitting, a Nested Cross-Validation (NCV) strategy was adopted [87]. The inner loop was dedicated to hyperparameter tuning, while the outer loop provided an unbiased estimate of generalization performance. After identifying the optimal hyperparameters, the model was retrained on the entire training set. Finally, to improve model interpretability and support clinical decision-making, Feature Importance and Local Interpretable Model-agnostic Explanations (LIME) were applied to identify the most influential radiomic features for each prediction. The complete analysis pipeline is illustrated in Figure 3.4.

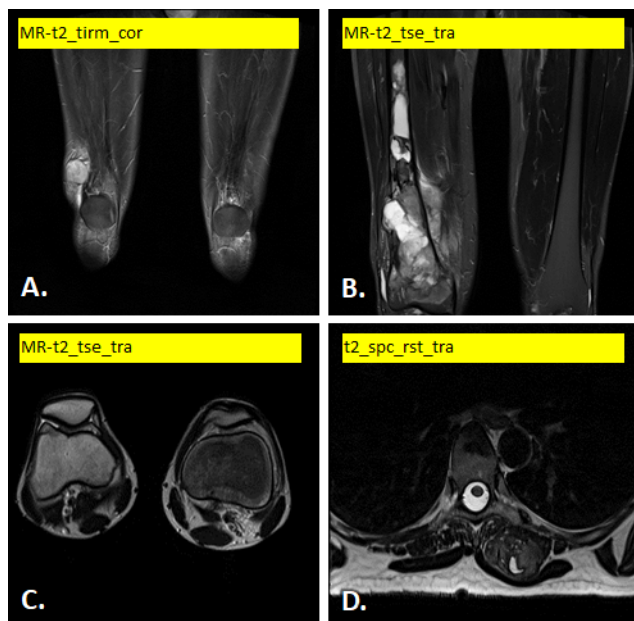


Figure 3.2: Example of T2-weighted MRI scan variants.

3.2.2.1 Segmentation and Features Extraction

In this work, image segmentation was applied to identify and isolate specific regions within medical scans, resulting in the creation of masks. Segmented regions included the sarcoma tumor, labeled as “Sarcoma”, and, when present, healthy tissue, labeled as “Healthy”. As previously mentioned, healthy tissue samples were derived from the same sarcoma patient’s exams. Due to the absence of standardized scan orientation during acquisition, images in axial, coronal, and sagittal planes were included. Each available orientation for a patient was segmented, resulting in up to three image sets per individual. The segmentation process was carried out using Mimics software (Materialize, Leuven, Belgium). A thorough visual inspection of MRI scans was performed to identify suitable regions for segmentation. Areas showing visible abnormalities—such as inflammation, edema, or other suspicious features—were carefully evaluated. For healthy tissue, regions showing signs such as inflammation, edema, or other suspicious features were excluded to ensure that the selected samples were as representative as possible of normal tissue within sarcoma patients. No control volumes were segmented when adequate and enough healthy tissue was not visible. Similarly, tumor regions were selected based on clear visual evidence of the lesion. All segmentations were man-

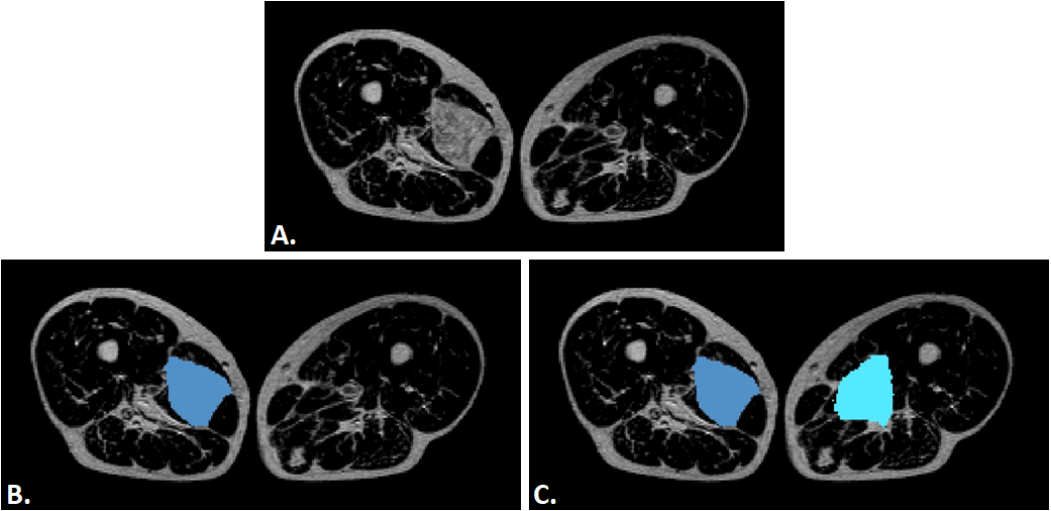


Figure 3.3: Example of tumor and healthy tissue segmentation in axial T1-weighted MRI of the thighs. (A) shows the original MRI with a unilateral soft tissue sarcoma. (B) displays the manual segmentation of the tumor (blue), while (C) shows the corresponding healthy tissue regions (light blue) from the contralateral limb, used as negative samples.

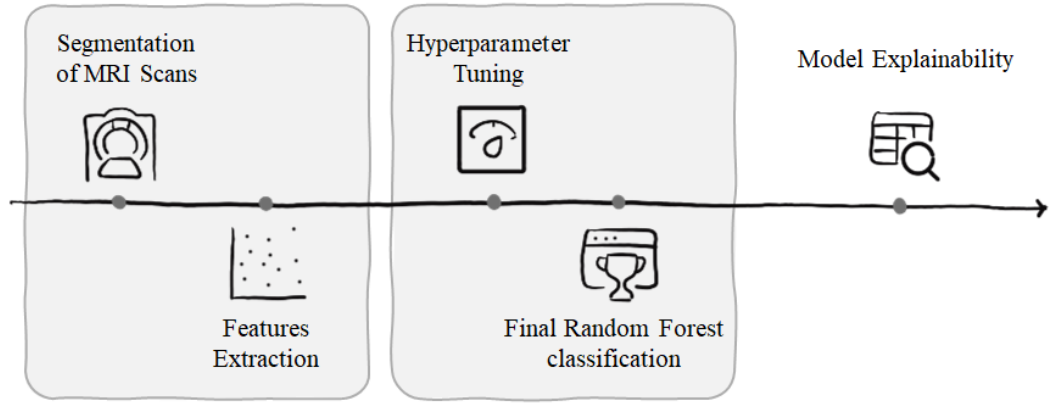


Figure 3.4: Workflow: segmentation of MRI scans, radiomic feature extraction, Random Forest classification after NCV, and model explainability using Feature Importance and LIME.

ually performed and subsequently reviewed by an expert, but no formal inter-rater consistency assessment was conducted. Following segmentation, both MRI images and corresponding masks were exported for radiomic feature extraction and subsequent analysis. Features extraction from segmented volumes was carried out through *PyRadiomics* 3.1.0, a Python library compliant with the Image Biomarker Standardization Initiative (IBSI)[55]. Extracted feature groups span various categories, totaling 107 radiomic features. First-order statistical features describe the overall distribution of voxel intensities within the region of interest (e.g., mean, variance, skewness). Gray Level Co-occurrence Matrix (GLCM) features capture spatial relationships between pairs of pixels by assessing the frequency of specific gray-level combinations. Gray Level Size Zone Matrix (GLSZM) features quantify the size and distribution of homogeneous gray-level zones, extending beyond simple pixel pair comparisons. Gray Level Run Length Matrix (GLRLM) features evaluate the length of consecutive pixels with the same intensity along specific directions. Neighbouring Gray Tone Difference Matrix (NGTDM) features describe local intensity variations by measuring differences between a pixel and its surrounding neighbors, capturing aspects such as contrast and texture coarseness [88]. Shape3D features were included to characterize the geometric properties of the segmented volumes, such as surface area, compactness, sphericity, and elongation, offering insight into the overall morphology of the regions of interest [89].

Furthermore, wavelet transforms were applied to generate higher-order features by decomposing the original image into eight distinct frequency sub-bands, using combinations of low-pass (L) and high-pass (H) filters along the X, Y, and Z axes [90]. These components, listed in Table 3.2, include one approximation image (LLL), capturing low-frequency content and overall structural information, and seven detail images (e.g., LLH, HHL, HHH), which emphasize high-frequency patterns such as edges, fine textures, and directional details.

Table 3.2: 3D Wavelet Transform Components.

Combination	Interpretation
LLL	Low-pass on X, Y, and Z (global structure)
LLH	Low-pass on X and Y, High-pass on Z
LHL	Low-pass on X and Z, High-pass on Y
LHH	Low-pass on X, High-pass on Y and Z
HLL	High-pass on X, Low-pass on Y and Z
HLH	High-pass on X and Z, Low-pass on Y
HHL	High-pass on X and Y, Low-pass on Z
HHH	High-pass on X, Y, and Z (fine details)

From each of these eight wavelet-derived volumes, 93 radiomic features were

extracted, including first-order statistics and texture descriptors. Shape-based features were excluded from the wavelet-derived volumes, as they depend only on the spatial mask and not on gray-level intensities.

The original image contributed 107 features: 93 intensity-based and texture features, and 14 shape-based features. Applying the same 93 intensity-based features to each of the eight wavelet sub-bands yielded an additional 744 features (93×8). Combining all components, the final feature set used for classification comprised 851 features:

$$\text{Total features} = \underbrace{93 + 14}_{\text{original}} + \underbrace{93 \times 8}_{\text{wavelet}} = 851$$

All features were extracted following the PyRadiomics standard pipeline [55].

Manually extracted features were chosen instead of deep features because deep learning approaches typically require large-scale datasets, whereas handcrafted features are more suitable for scenarios with limited data availability, as is the case here [91]–[93].

3.2.2.2 Hyperparameter Tuning and Random Forest classification

Following image segmentation with Mimics Innovation Suite (Materialise NV) and radiomic feature extraction via PyRadiomics, hyperparameter tuning was conducted using Python 3.9. Key libraries included *scikit-learn* v1.5.1 for machine learning, *pandas* v2.2.1 for data manipulation, and *numpy* v1.26.4 for numerical operations commonly employed in AI healthcare applications [94]–[96].

Prior to model training, the dataset was partitioned using a stratified hold-out strategy. A test set of 58 samples was reserved as an independent hold-out subset, used exclusively for the final evaluation of model performance. The remaining samples constituted the development set, employed for model optimization and validation. This stratified partitioning preserved class distribution across both subsets, minimizing sampling bias and enhancing generalizability.

To ensure unbiased performance estimation and robust hyperparameter selection, we adopted a Nested Cross-Validation (NCV) framework. NCV is used for evaluating machine learning models when hyperparameters require tuning, as it effectively prevents information leakage between the model selection and evaluation phases. We implemented a k -fold stratified outer loop to assess generalization capability, coupled with a k -fold stratified inner loop for hyperparameter optimization. Stratification was used in both loops to maintain consistent class distributions across all folds.

Numerous studies report that $k = 10$ and $k = 5$ are the most commonly used values in k -fold cross-validation among machine learning practitioners. However, the literature highlights that there is no definitive rule for choosing the value of k [97]. In our case, we selected $k = 9$ for the outer loop, as the more commonly used

values did not yield a unique or consistent selection of model parameters during validation.

Within each outer loop iteration, a Grid Search was performed over a predefined hyperparameter set for the Random Forest classifier, including the number of estimators (*n_estimators*), maximum tree depth (*max_depth*), and the minimum number of samples required to split an internal node (*min_samples_split*). The parameter grid included:

- *n_estimators*: 100, 200, and 1000,
- *max_depth*: None, 5, and 10,
- *min_samples_split*: 2 and 5.

Tuning *max_depth* and *min_samples_split* helps control and mitigate overfitting by limiting the growth and complexity of individual trees, while adjusting *n_estimators* balances the trade-off between computational cost and model variance. Setting *max_depth* to **None** allows each decision tree in the forest to grow without a predefined limit, enabling the model to capture complex patterns in the data. Setting *min_samples_split* to **2** allows nodes to be split as long as they contain at least two samples, encouraging fine-grained partitioning. Finally, choosing **200** for *n_estimators* offers a practical trade-off between performance and computational efficiency.

For every outer fold and its corresponding best combination of hyperparameters, we evaluated model performance using the metrics: accuracy, precision, recall, F1-score, and the area under the ROC curve (ROC-AUC). Employing a diverse set of evaluation metrics provides a more comprehensive view of classifier behavior.

The final Random Forest model was retrained on the entire development set using the optimal hyperparameters identified through the NCV process. It was subsequently evaluated on the unseen hold-out test set to obtain an unbiased performance estimate. The test set comprised 28 Healthy patients and 30 Sarcoma patients.

3.2.2.3 Model Explainability

To enhance the interpretability of the final Random Forest model, two explainability techniques were employed: Feature Importance and Local Interpretable Model-agnostic Explanations (LIME). These methods offer insights into global and local model behavior and help ensure transparency and trust in the model's decisions, which is important in medical imaging.

Feature Importance:

Global interpretability was achieved by analyzing feature importance via the model's

feature_importances_ attribute. This metric quantifies each feature’s contribution to the model’s predictions, based on the average decrease in impurity across all trees. Features were ranked according to their importance scores to identify the most relevant variables for classification. This analysis was conducted on the final Random Forest model, retrained on the full development set using optimal hyperparameters identified via nested cross-validation. The resulting importance scores were used to determine which radiomic features most influenced the model’s decision-making process.

LIME:

LIME [98] was involved to explain individual predictions by identifying the most influential features and quantifying their impact. Each feature is associated with a value range and an assigned weight. A positive weight indicates that the feature supports the predicted class, while a negative weight suggests influence toward another class. This approach is particularly valuable when applied to complex, non-linear models like Random Forests, which comprise ensembles of decision trees and are inherently difficult to interpret. By approximating the local behavior of the Random Forest with a simpler, interpretable model, LIME enables users to understand which features contributed most to a specific prediction. This enhances transparency and trust, particularly in domains where interpretability is critical, such as medicine. In particular, LIME is introduced for its ability to provide interpretable and accurate explanations for the outputs of any classification model. LIME was implemented using the *lime* Python package, specifically the *lime.lime_tabular.LimeTabularExplainer* class.

3.3 Results

3.3.1 Results of the Hyperparameter Tuning

Table 3.3 presents the optimal hyperparameter combinations and corresponding performance metrics for each outer fold. Despite variations across folds, the most frequently selected configuration was `max_depth = None`, `min_samples_split = 2`, `n_estimators = 200`.

3.3.2 Final Random Forest classification

The resulting confusion matrix, shown in Figure 3.5, illustrates the classification outcomes. In the matrix, **TN** (True Negatives) denotes Healthy samples correctly classified as Healthy, while **FP** (False Positives) represents Healthy samples incorrectly classified as Sarcoma. **FN** (False Negatives) corresponds to Sarcoma cases misclassified as Healthy, while **TP** (True Positives) refers to Sarcoma samples correctly identified.

Table 3.3: Hyperparameter combinations and performance metrics across outer folds.

Outer Fold	Best Combinations	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1	max_depth: None min_samples_split: 2 n_estimators: 200	0.765	0.812	0.722	0.765	0.885
2	max_depth: None min_samples_split: 2 n_estimators: 200	0.794	0.867	0.722	0.788	0.872
3	max_depth: 10 min_samples_split: 2 n_estimators: 200	0.794	0.789	0.833	0.811	0.878
4	max_depth: 5 min_samples_split: 5 n_estimators: 1000	0.765	0.765	0.765	0.765	0.827
5	max_depth: None min_samples_split: 2 n_estimators: 1000	0.853	0.875	0.824	0.848	0.938
6	max_depth: 10 min_samples_split: 5 n_estimators: 200	0.727	0.700	0.824	0.757	0.768
7	max_depth: 5 min_samples_split: 2 n_estimators: 1000	0.818	0.762	0.941	0.842	0.886
8	max_depth: None min_samples_split: 5 n_estimators: 200	0.758	0.714	0.882	0.789	0.831
9	max_depth: 10 min_samples_split: 5 n_estimators: 100	0.758	0.714	0.882	0.789	0.860

TPerformance metrics obtained from the hold-out test set are presented in Table 3.4. This approach ensured that the metrics were not biased by the model selection process.

Table 3.4: Performance metrics from classification on the hold-out test set.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.724	0.719	0.767	0.742	0.871

		Predicted	
		Healthy	Sarcoma
Actual	Healthy	19 TN	9 FP
	Sarcoma	7 FN	23 TP

Figure 3.5: Confusion matrix for classification on the hold-out test set.

3.3.3 Model Explainability

3.3.3.1 Feature Importance Results

As a result of the analysis, 849 out of 851 features contributed to the model’s decisions, even if only marginally. Table 3.5 lists the ten most influential features used in the classification.

Table 3.5: Top 10 most important features ranked by Random Forest.

Feature	Importance
wavelet-LHL_firstorder_Mean	0.024852
wavelet-HLL_firstorder_Median	0.015829
wavelet-HLL_firstorder_Mean	0.015710
original_gldm_LargeDependenceLowGrayLevelEmphasis	0.013242
wavelet-LHL_firstorder_Median	0.011875
wavelet-LHL_gldm_ClusterShade	0.010387
original_gldm_LowGrayLevelEmphasis	0.010175
original_gldm_LowGrayLevelRunEmphasis	0.009607
original_gldm_ShortRunLowGrayLevelEmphasis	0.009075
wavelet-HLL_firstorder_Skewness	0.009016

Feature importance analysis indicated that a broad spectrum of radiomic features contributed to the model’s decision-making process, with the highest-ranked features predominantly derived from wavelet-transformed images, especially from the LHL and HLL decomposition levels. This suggests that texture patterns captured through wavelet filtering play a crucial role in distinguishing between healthy and sarcoma tissues. Many of the most influential features were first-order statisti-

cal descriptors (e.g., mean, median, skewness), indicating that the distribution of voxel intensities within the segmented regions carries significant discriminative information. Additionally, texture-based features derived from GLDM, GLRLM, and GLCM were also prominent. These features reflect spatial relationships and structural complexity within the tissue, such as the presence of fine textures, low gray-level emphasis, and local intensity variations. The integration of intensity-based and texture-based descriptors, especially those extracted from wavelet domains, highlights the model's reliance on both global and localized image characteristics. This diverse feature set aligns with the heterogeneous nature of sarcoma, which often exhibits complex and subtle imaging patterns. Overall, the feature importance analysis reinforces the biological plausibility of the model's predictions and strengthens confidence in its interpretability.

3.3.3.2 Local Interpretability with LIME

To illustrate the behavior of the model across different outcomes, LIME was applied to four representative test cases, covering all combinations of ground truth and predicted labels (where 0 refers to "Healthy" and 1 to "Sarcoma"):

1. true = 0, predicted = 0 (true negative);
2. true = 1, predicted = 1 (true positive);
3. true = 0, predicted = 1 (false positive);
4. true = 1, predicted = 0 (false negative).

It is worth noting that individual feature contributions were relatively small (mostly in the range of ± 0.01 to ± 0.03), which is expected given that the model's decisions are based on a high-dimensional feature space comprising 849 radiomic features. In such high-dimensional settings, the overall decision is distributed across numerous features, with no single attribute dominating the prediction.

Case 1: True = 0, Predicted = 0 (Correct Classification, Figure 3.6)

In this case, all ten top-ranked features exhibited negative contributions, suggesting that they individually lead the prediction toward class 0.

Case 2: True = 1, Predicted = 1 (Correct Classification, Figure 3.7)

This prediction was strongly supported by the feature contributions. Nine out of ten principal features had positive weights, promoting class 1. This indicates that the model made a confident and coherent decision, guided by radiomic features aligned with the true label, even if only one feature showed a small negative contribution.

Case 3: True = 0, Predicted = 1 (Misclassification, Figure 3.8)

In this misclassified case, the feature contributions were mixed. Some features supported the true label (negative weights), while others promoted the incorrect

prediction (positive weights). The overall influence of features that exhibited a positive contribution was sufficient to overturn the decision, highlighting how small variations in key features can lead to incorrect classifications.

Case 4: True = 1, Predicted = 0 (Misclassification, Figure 3.9)

In this example, the majority of influential features had negative contributions pulling the prediction toward class 0. Only three features provided weak positive support for class 1. The model behaved consistently with the observed feature values, but the decision was ultimately incorrect. This outcome may reflect overlapping feature distributions between the classes.

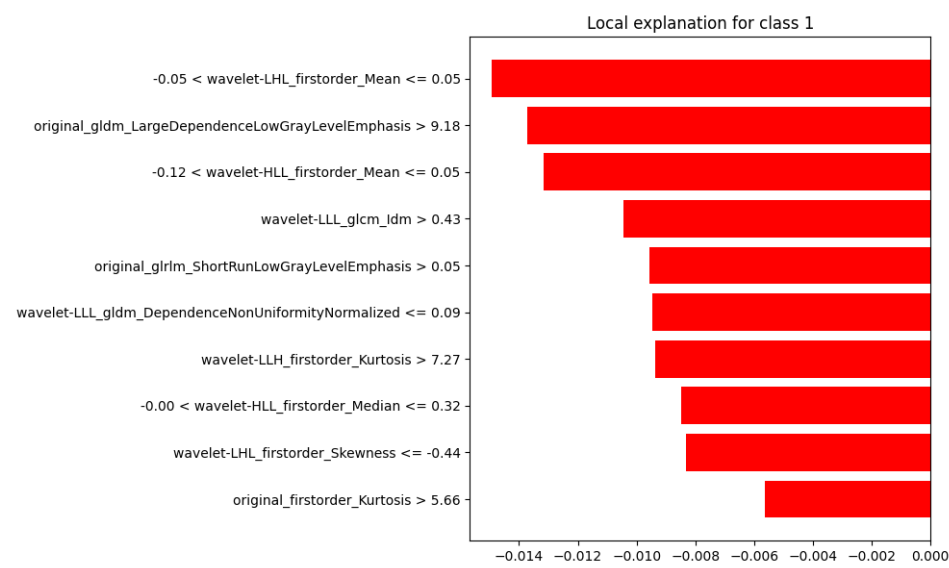


Figure 3.6: LIME Explanation: True = 0, Predicted = 0 (Correct Classification).

3.3.4 Performance comparison

Table 3.6 presents a comparative evaluation of several classification algorithms applied to the same dataset used by the proposed method. Models were evaluated using five standard metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC, in order to benchmark their performance against that of the Random Forest classifier. All classifications were performed using WEKA, a software platform designed for machine learning experimentation and data analysis.

LWL and SVM showed competitive results, especially in Accuracy and F1-score. However, the proposed method based on Random Forest, achieved the

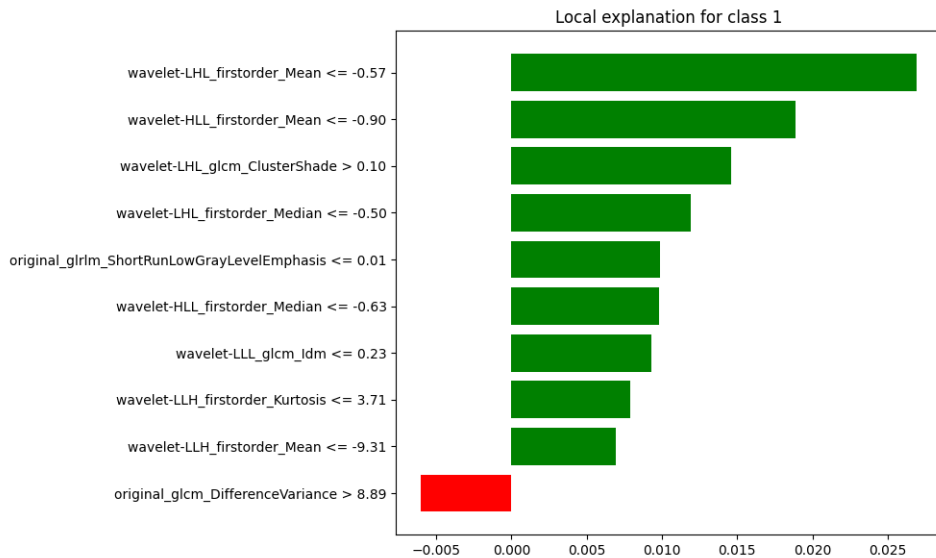


Figure 3.7: LIME Explanation: True = 1, Predicted = 1 (Correct Classification).

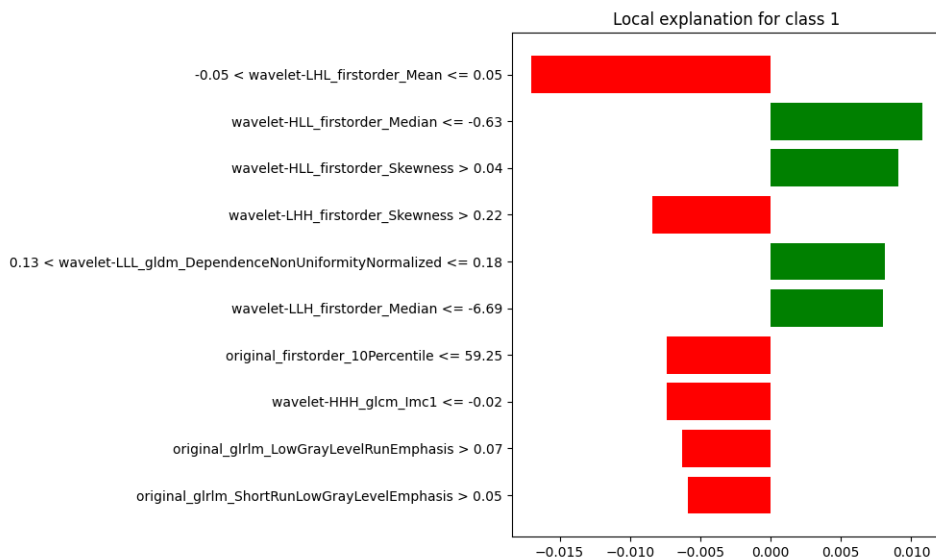


Figure 3.8: LIME Explanation: True = 0, Predicted = 1 (Misclassification).

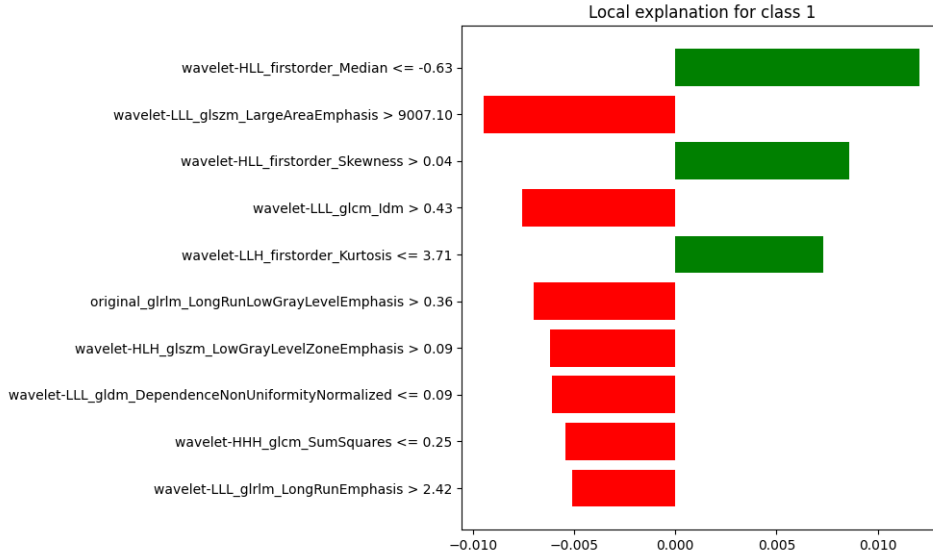


Figure 3.9: LIME Explanation: True = 1, Predicted = 0 (Misclassification).

best overall performance, obtaining the highest F1-score (0.742) and ROC-AUC (0.871), achieving the highest F1-score (0.742) and ROC-AUC (0.871), which reflect strong class discrimination and a well-balanced trade-off between precision and recall. Although IBK achieved a relatively high Recall (0.741), it had lower Precision (0.571). Naive Bayes and BayesNet exhibited a tendency to over-predict the positive class, leading to imbalanced performance across metrics. KStar, in particular, consistently assigned all patients to the negative class, failing to correctly identify any positive cases. Consequently, Precision and F1-score could not be computed, and its ROC-AUC value remained close to chance level (0.500), highlighting its poor discriminatory ability for this classification task.

3.4 Discussion

In this work, we investigated the effectiveness of machine learning techniques for sarcoma diagnosis. Specifically, classification was performed using a Random Forest model on radiomic features extracted from manually segmented volumes containing either sarcoma or healthy tissue.

The real strength of this study lies in the results obtained using heterogeneous MRI images acquired with varying protocols, both with and without contrast, and from multiple anatomical regions. Additionally, the dataset spans a wide temporal

Table 3.6: Comparative performance of classification algorithms on the balanced binary dataset.

Classifier	Accuracy	Precision	Recall	F1-score	ROC-AUC
RandomForest	0.724	0.719	0.767	0.742	0.871
RandomTree	0.706	0.593	0.593	0.653	0.700
IBK	0.529	0.571	0.741	0.645	0.628
LWL	0.741	0.773	0.630	0.694	0.740
KStar	0.534		0.000		0.500
SVM	0.724	0.739	0.630	0.680	0.718
NaiveBayes	0.573	0.475	0.704	0.567	0.512
BayesNet	0.517	0.487	0.704	0.576	0.552
AdaBoostM1	0.672	0.654	0.630	0.642	0.781

range (2004-2023), capturing significant variability in MRI technology and acquisition parameters. The sample size of 86 patients is a limitation and a constraint in terms of the generalizability of the results. For this very reason, examinations of different plans were used for each patient, thus obtaining 186 “Sarcoma” samples for classification. Future work will focus on validating the model on larger datasets to improve robustness and generalizability.

Due to this high heterogeneity, the classification task was particularly challenging and required the use of 849 out of 851 available radiomic features to achieve robust performance. This underscores the generalizability of our approach across diverse imaging conditions, which is a critical requirement for real-world clinical applications. These findings are especially encouraging considering the complexity of the task, which involves differentiating between pathological and non-pathological tissue across a heterogeneous set of MRI acquisitions. The dataset includes various sarcoma types and imaging protocols, introducing significant variability in terms of anatomical location, acquisition parameters, and lesion appearance. Despite these challenges, the model maintained robust and consistent predictive performance, further supporting its potential clinical applicability. Assessing the contribution of individual MRI sequences could potentially enhance classification performance. However, conducting separate analyses for each modality would have resulted in a further reduction of the effective sample size, due to the incomplete availability of all sequences for every patient. Given the already limited dataset, such stratification would have compromised statistical power and model reliability. For this reason, we opted to combine available sequences to maximize the use of data and ensure more stable model training. Furthermore, the application of explainability techniques provided valuable insights into the model’s decision-making process. These explainability analyses highlighted specific ra-

diomic features, particularly wavelet-transformed first-order and texture metrics, as influential predictors, enhancing the model's transparency and potential clinical trustworthiness.

This suggests that texture patterns captured via wavelet filtering play a crucial role in distinguishing between healthy and sarcoma tissues. Several of the most influential features were first-order statistics (such as mean, median, and skewness), indicating that the distribution of voxel intensities within the segmented regions carries significant discriminative information. Additionally, texture-based features derived from GLDM, GLRLM, and GLCM were also prominent. These features reflect spatial relationships and structural complexity within the tissue, such as the presence of fine textures, low gray-level emphasis, and local intensity variations. The combination of intensity-based and texture-based descriptors, especially those extracted from wavelet domains, highlights the model's reliance on both global and localized image characteristics. This diverse feature set aligns with the heterogeneous nature of sarcoma, which often exhibits complex and subtle imaging patterns. Overall, the feature importance analysis supports the biological plausibility of the model's predictions and enhances confidence in its interpretability.

Prioritizing the F1-score is especially justified in medical domain, where both false positives and false negatives can have significant clinical consequences. The F1-score, as the harmonic mean of precision and recall, provides a balanced measure that accounts for both types of errors. In healthcare applications, an overemphasis on either precision or recall alone may lead to suboptimal patient outcomes, such as missing true cases (low recall) or causing unnecessary interventions (low precision). By maximizing the F1-score, we aim to achieve a clinically meaningful balance between sensitivity and specificity, ultimately supporting safer and more effective decision-making.

This approach ensures that performance metrics accurately reflect the model's generalization capability, minimizing the risk of overly optimistic estimates due to selection bias or overfitting.

It is important to note that individual feature contributions were relatively small (mostly in the range of ± 0.01 to ± 0.03), which is expected given that the model's decisions are based on a high-dimensional feature space comprising 849 radiomic features. In such cases, the overall decision is distributed across many features, with no single attribute dominating the prediction.

Nevertheless, recent advances in machine learning, particularly deep learning, have significantly enhanced the analysis of radiological images for the diagnosis and management of sarcomas as well as the quantification of the pathologists' visual pattern for whole slide image diagnosis [99].

Our research group recently explored the use of artificial intelligence for sarcoma diagnosis and grading based on MRI scans and radiomics [100]. In particular, when focused on the binary classification (i.e., diagnosis of sarcoma), they

achieved an accuracy of 76.02%. Although starting from the same dataset, the main difference from this study is that the authors analyzed individual slices, thus obtaining a larger dataset, whereas in the present work, volumes were considered in their entirety. In addition, segmentation of healthy tissue was performed by excluding areas containing sarcoma and edema and including the rest of the image. From a methodological standpoint, although Random Forest was also employed in that study, validation was conducted by leave-one-out cross-validation. By contrast, in the study presented here, a portion of the dataset was isolated from the beginning and used exclusively for final validation. The remaining dataset was used for model tuning, performed through nested k-fold cross-validation, and subsequently for final training, in order to avoid any form of data leakage.

By contrast, several studies have focused on lipomas, which are benign fatty tumors. For example, Gitto et al. analyzed 150 cases to differentiate lipomas from atypical (low-grade sarcomas) lipomatous tumors using three models. Their best model achieved 92% sensitivity but only 33% specificity, suggesting a high false positive rate likely due to dataset imbalance [101]. In contrast, Malinauskaite et al. evaluated 38 cases to distinguish lipomas from liposarcomas using four models. The Support Vector Machine (SVM) performed best, achieving an AUC of 0.926, 88% sensitivity, perfect specificity (100), and 92.7% accuracy [102]. Unlike these studies, our work is based on a larger and balanced dataset comprising 361 patients (186 with sarcoma and 175 healthy), allowing for more robust and generalizable model training and evaluation.

The study by Dai et al. involved 172 patients and evaluated four different CNN architectures for deep feature extraction, combined with a Random Forest classifier, to distinguish uterine sarcomas from atypical leiomyomas [103]. In their approach, the authors based the analysis on only three MRI slices, whereas our method leverages the entire tumor volume, providing a more comprehensive representation of the lesion.

Collectively, these studies demonstrate that machine learning techniques can enhance diagnostic accuracy and patient stratification in soft tissue tumors. However, their performance varies depending on the model used, sample size, and complexity of the diagnostic task. In fact, the main difference between our work and the cited studies is that we distinguish between healthy tissue and sarcomas, encompassing nearly all types rather than focusing on a specific niche.

Although developed in a different domain, the approach proposed by Nan et al. [99] uses expert-derived information in the form of pathologists' gaze patterns to train deep learning models for histopathological diagnosis. In contrast, our model relies on structured clinical data and confirmed diagnoses to produce interpretable predictions at the time of referral. These represent complementary strategies for integrating clinical expertise into decision-support systems.

Researchers are also exploring Generative Adversarial Networks (GANs) to generate synthetic radiological images, addressing the challenge of limited train-

ing data due to the rarity of sarcomas [104], [105]. A promising research direction involves integrating radiological features with other data modalities (such as histopathological, transcriptomic, and genomic data) through multi-modal AI systems [106]–[108].

Despite the encouraging results, limitations including the relatively small dataset size, variability in imaging planes, and manual segmentation processes suggest the need for further validation on larger, multi-center cohorts and the investigation of automated segmentation methods. Moreover, the “black box” nature of deep learning models poses interpretability challenges that may hinder clinical adoption. However, as demonstrated in this study, Explainable AI techniques are being explored to improve model transparency and foster trust in AI-driven decisions.

In conclusion, although technical and logistical barriers persist, machine learning is poised to transform radiological practice in sarcoma care, offering the potential for earlier diagnosis, more accurate prognostication, and truly personalized therapy.

3.5 Conclusions

This study demonstrates the feasibility of using radiomic features from MRI combined with a Random Forest classifier to distinguish sarcoma from healthy tissue:

1. The model, optimized via nested cross-validation, achieved promising performance on an independent test set (F1-score: 0.742), showing a balanced trade-off between sensitivity and specificity.
2. The integration of explainability techniques, including global feature importance and LIME, offered valuable insights into the model’s decision-making process, enhancing its transparency and clinical relevance. These findings underscore the potential of radiomics and machine learning as impactful tools in oncologic imaging.

Future work must focus on integrating multimodal imaging data and advanced deep learning architectures to further enhance classification accuracy and model robustness.

Chapter 4

A Machine Learning Approach to Sarcoma Diagnostics and Staging

4.1 Introduction

In the age of information and computing power improvements, machine learning (ML) has seen drastic developments and has given us the tools needed to improve many processes in modern society. ML has the potential to aid radiologists and oncologists through data driven methodologies to assist with sarcoma diagnosis and staging. Herein the methodologies and theories are defined, which result in a pipeline to develop assistive tools in sarcoma diagnostics and staging.

4.1.1 Medical Imaging

Medical imaging is the process of obtaining detailed visual data of patient anatomy, and is mandatory for the task of sarcoma diagnosis and treatment. Namely, the ability to gather multiple image slices and stacking them to form a three dimensional representation of anatomy has greatly improved diagnostics, treatments, surgical preparation and data-driven techniques in these fields. With techniques such as magnetic resonance imaging (MRI), computed tomography (CT) and positron emission tomography (PET), the acquisition of any type of anatomical data is possible and has become straight forward in healthcare practices.

Segmentation is the process of highlighting specific regions of interest (ROI) in an image. The highlighted regions are referred to as masks and label different regions of anatomy e.g. bones, fat, soft tissue or specific organs. These masks and images can then be used for further analysis in an attempt to find and explain underlying patterns in anatomical structures.

4.1.2 Radiomics

Radiomics is a rapidly evolving field that focuses on extracting quantifiable metrics from medical image data with the aim to characterize varying bodily tissue through differences in visual features and is applied to solve clinical problems programmatically[109]. A key feature in radiomics, is the standardized and automated extraction of quantitative features from medical images that converts visual information into structured numerical descriptors of tissue heterogeneity, intensity, and morphology [110].

This approach differs fundamentally from traditional radiology, which is largely qualitative and dependent on the subjective and often inconsistent interpretation of image data by specialists [111]. Radiomics, however, enable reproducible and objective data analysis, which can be integrated into computational workflows for diagnostic and prognostic decision-making [112]. Radiomics-based approaches have even been shown to be more effective than radiological assessments, in some cases [111].

The motivation behind radiomics lies in its capacity to uncover subtle patterns in image data that may correlate with underlying biology, treatment response, or patient outcomes[109]–[112]. In clinical settings, radiomics has shown promise in capturing anatomical heterogeneity at a much larger scale than biopsies, which sample only localized tissue [109]. Radiomics therefore address a critical limitation in conventional workflows by offering a whole-lesion, non-invasive, and consistent assessment strategy.

4.1.2.1 General Workflow

The typical radiomics workflow consists of a sequential pipeline: image acquisition, segmentation of the ROI, radiomic feature extraction, and post-processing or modeling. Standard imaging modalities such as CT, MRI, or PET are used to acquire images, often under strict protocols to ensure data quality [112]. These images are then segmented using tools such as 3D Slicer [112] or Materialize MIM-ICS[51].

The extracted features, often numbering in the hundreds or thousands, are then structured into feature matrices that can be input into machine learning models or statistical pipelines [110]. This post-processing step may include normalization, dimensionality reduction, and modeling via classifiers such as Random Forests or Support Vector Machines [110], [112]. Dimensionality reduction techniques like LASSO regression are also commonly applied to avoid overfitting in predictive models [109].

4.1.2.2 Feature Classes

Radiomic features are typically classified into first-order statistics (FO), shape-based, and texture-based categories. First-order features quantify intensity distributions within the ROI and include measures such as mean, variance, skewness, and kurtosis [109]. Shape features describe the 3D morphology of lesions, including volume, compactness, and sphericity [110].

Texture features capture spatial patterns and heterogeneity within the tumor. These include metrics derived from the Gray-Level Co-Occurrence Matrix (GLCM), Gray-Level Run-Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Neighbourhood Gray-Tone Difference Matrix (NGTDM), and Gray-Level Dependence Matrix (GLDM) [110], [112]. Each texture class offers a unique method for characterizing anatomical complexity: for instance, GLCM describes pairwise intensity relationships, while GLSZM captures the size of homogeneous regions. These descriptors form the foundation of radiomic analysis by enabling nuanced quantification of anatomy beyond human perception.

4.1.2.3 Image Transforms

Image transforms are sets of mathematical operations applied to images to accentuate or emphasize certain details in an image, often times used in conjunction with radiomics to extract more information from an image than the raw image could produce alone [113]. Transforms have been shown to potentially boost classification performance in radiomics tasks when compared to using raw images alone [114]. With a plethora of available image transforms tested in radiomic tasks, the potential applications increase drastically and offers many opportunities for research. Figure 4.1 displays an example of multiple image transformations applied to a slice of an MRI image of a sarcoma tumor, the figure shows the potential of image transforms' ability to emphasize different details of an image, which can be regarded as new, additional information available for further analysis.

4.1.2.4 Quantitative Capacity and Data Integration

Radiomics transforms unstructured medical images into structured, high-dimensional numerical data, making it inherently compatible with data mining and machine learning techniques [109]. Its quantitative nature allows for seamless integration with complementary clinical data types, including genomics, histopathology, and demographic variables [112]. This integrative capacity is particularly vital in precision medicine, where radiomics can augment traditional biomarkers and contribute to the generation of multi-modal disease signatures.

The value of such integration has been highlighted in studies where radiomic features are combined with clinical parameters to significantly improve model

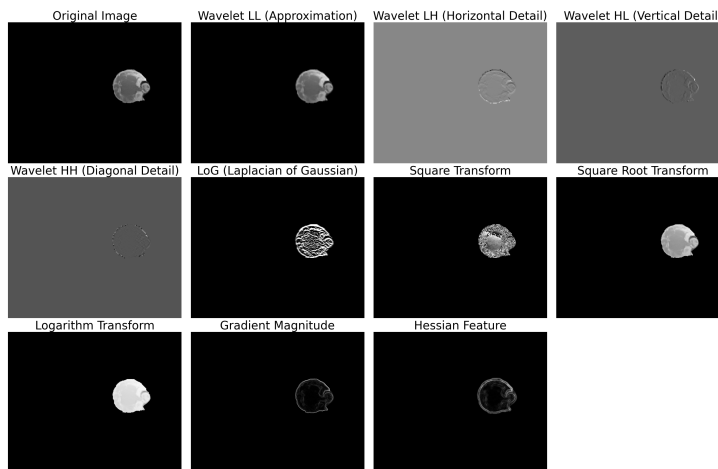


Figure 4.1: An example of image transforms applied to a slice of an MRI scan of a sarcoma tumor.

performance for outcome prediction in head and neck cancers [110], [112]. Furthermore, radiogenomics (linking radiomic descriptors to underlying gene expression) has emerged as a promising subfield demonstrating the biological validity of imaging-derived data [109].

4.1.2.5 Applications in Oncology and Beyond

Radiomics has shown notable utility in various oncologic tasks including tumor classification, treatment response prediction, recurrence risk stratification, and survival estimation. In non-small cell lung cancer, specific PET/CT-derived radiomic features have been associated with gene expression signatures and epidermal growth factor receptor mutation status [109]. Similarly, in gliomas, PET-based NGTDM features have proven capable of distinguishing tumor grades [109].

These applications underscore the potential of radiomics not only as a standalone biomarker strategy but as a synergistic element in multi-omics pipelines [110]. In head and neck cancer, models that combine radiomic and clinical data outperform those based on clinical features alone for predicting locoregional recurrence and distant metastases [110].

4.1.2.6 Challenges and Considerations

Despite its promise, radiomics faces several challenges. One of the most critical is the variability introduced by differences in image acquisition and reconstruction

protocols across institutions and scanners. Factors such as voxel size, contrast agent timing, and motion artifacts can all impact feature reproducibility [109], [112].

Standardization efforts such as the Image Biomarker Standardisation Initiative (IBSI) aim to harmonize feature definitions and extraction protocols [109], but widespread adoption remains limited. Additionally, radiomic models are susceptible to overfitting due to the high dimensionality of the data relative to typical sample sizes, necessitating careful feature selection and validation practices [109]. These issues must be addressed before radiomics can reliably translate into routine clinical practice.

4.1.2.7 Emergence of Machine Learning

Machine learning has become an integral component of radiomics workflows, particularly in tasks such as outcome prediction and risk stratification. Models such as random forests, support vector machines, and deep neural networks are employed to uncover complex, non-linear associations between imaging features and clinical endpoints [110], [112]. The ability to integrate radiomics with ML has shown improved predictive performance in retrospective cohorts, but prospective studies are still needed for clinical validation [112].

4.1.2.8 Empirical Applications of Radiomics

Tomaszewski et al.[115] extracted 164 radiomic features and combined it with histological classification and smoking history to improve selection criteria for clinical trials of a treatment for soft tissue sarcoma; the study shows that the clinical trial SARCO21 could have met the survivability requirement had this type of model been used. Gangil et al.[112] showed that radiomics coupled with clinical data boosted machine learning model performance and generalization when compared to clinical data alone by up to 21% for certain tasks. Vallières et al.[110] demonstrated the large potential for radiomics in risk assessments for head and neck cancer, concluding that it could have an important clinical impact on the personalisation of medical treatments.

4.1.3 Tumor Habitat Analysis

Tumor habitat analysis is a relatively new method for analyzing tumor heterogeneity and has seen applications in cancer diagnostics [116]–[118]. The main idea behind it, is to split up a tumor volume into extremely small volume segments on the order of a few cubic millimeters, and clustering together these volume segments into habitats, based on similarity measures from radiomics feature values. Gener-

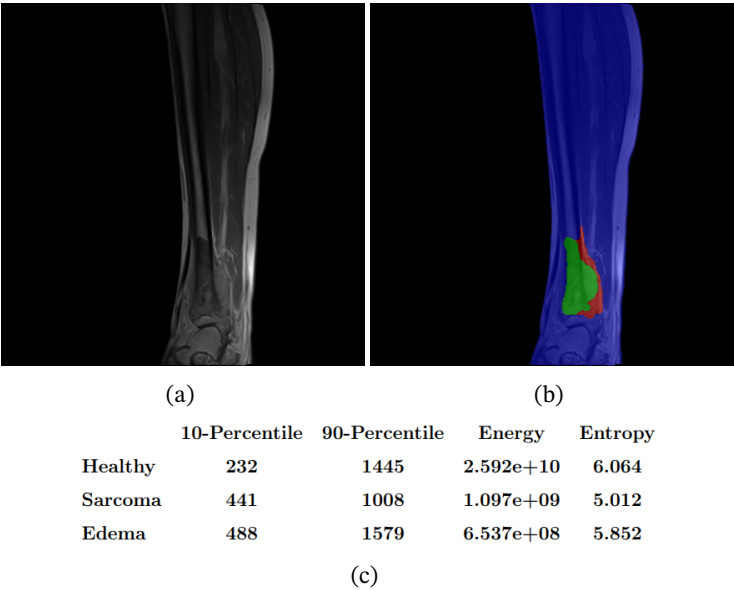


Figure 4.2: The general radiomics feature extraction method. (a) displays the raw image (b) the highlighted areas represent: healthy region in blue, sarcoma in green and the edema in red (c) the numerical results acquired from extracting 4 features from the different highlighted areas.

ally, the number of habitats is quite small, around 3-4, and these habitats can be analyzed in a few different ways.

The behavior of habitats can tell much about the heterogeneity of tumors, if habitats are generally disconnected and spread out, then the tumor is more heterogeneous. Heterogeneity has been correlated with cancer severity, and therefore correlations between tumor heterogeneity and sarcoma grade might exist. Studies have also shown the efficacy of performing radiomics extraction on habitat volumes with success [116]–[118].

4.1.3.1 General Workflow

Starting with medical image scans and masks of the ROI, an unsupervised algorithm such as simple linear iterative clustering (SLIC) is applied on the masked region of the medical image to split the mask into small super pixel (SP) volumes. Specific radiomics features are extracted that correlate with the class trying to be predicted, and the extracted feature values are used to cluster together individual SPs into however many habitats are desired, using a clustering method such as the

K-means clustering algorithm. This forms the baseline habitats that can be used for subsequent analysis.

Habitat interactions can be captured using the multi-regional spatial interaction (MSI) matrix [116], which quantifies heterogeneity by counting the number of pixels in a habitat that border another habitat. The MSI matrix is then a $K \times K$ matrix, where K is the number of habitats, and statistical measures can be extracted from it, e.g. mean, standard deviation, kurtosis, skewness, etc. Each patient scan then provides a single data-point that can be used for machine learning training.

Another approach is to utilize the habitats as masks for radiomics feature extractions [119]. In this case, habitats are treated as masks and radiomics feature extraction is performed directly on it. The extracted data is then used as the input for ML model training.

4.1.3.2 Applications

Wu et al. [116] analyzed intratumoral spatial heterogeneity of breast cancer patients' MRI scans with an aim to predict recurrence-free survival (RFS). They divided tumors into three spatially segregated, phenotypically consistent habitats, based on perfusion MRI parameters. Next, they calculated the MSI matrix for each tumor followed by the extraction of 22 features. The results showed correlation between tumor aggression and a larger volume of specific habitats, as well as interaction between poorly and moderately perfused habitats and surrounding parenchyma.

Wei et al. [117] explored the application of habitats to discriminate between adult-type grade 4 astrocytoma and glioblastoma. They extracted radiomics features from manually segmented masks as well as from generated habitats. The result showed that automatically clustered habitats performed comparably to manually segmented habitats in classification accuracy.

Wu et al. [118] review habitat analysis as an extension of radiomics in the literature. They found evidence in the literature that support the notion that habitats can be applied in certain tasks like tumor grading, differentiation, gene expression predictions and therapy response assessment, especially when coupled with multi-modal imaging and ML approaches.

Xing et al. [120] utilized maps of the apparent diffusion coefficient (ADC) and T_2 relaxation to calculate the diffusion-weighted MRI (dwMRI) of soft tissue sarcoma (STS) patients. By applying calculated parameters to generate 5 distinct habitats, they were able to accurately estimate tumor sub-regions such as necrosis.

The extension of radiomics to include habitat analysis shows much promise in the literature and has a wide potential for aiding in the assessment and diagnostics of sarcoma.

4.1.4 Machine Learning

Machine Learning (ML) is an ever developing field that can be applied in almost all disciplines. Although it has been around since the middle of the 20th century, ML began to blossom in the 1990s and has become an integral part in everyday life in the present[121]. ML employs mathematical learning models that aim to predict outcomes or values based on a set of input data by training on data where the outcome is known. These models are very complicated sets of statistical equations with a multitude of parameters that are tweaked during the training process using computer algorithms, optimizing for decreasing the error of the calculated, predicted output compared to the actual ground truth[122].

The field of medicine has seen a surge in ML applications such as personalized medicine, by leveraging the vast amount patient data available to hospitals and institutions, predictive models can be trained to more confidently draft a personalized diagnostic and treatment path as they can grasp slight variations in patients which may cause them to be phenotypically distinct from the norm[123]. Zhao Ting et al.[124] collected clinical data of children suffering from epilepsy and applied it to predict the steady-state serum concentration of levetiracetam, to better administer the correct dosage of medicine with success. Fisher et al.[125] managed to predict the trajectory of Alzheimer's disease prognosis by using 18-month trajectories of 44 clinical variables for individual patients with success. In the field of bioinformatics, Chatzianastasis et al.[126] managed to identify cancer genes with higher accuracy than conventional models using explainable multi-layer graph neural networks. Dhudum et al.[127] compiled a review of research in the field of drug discovery using ML. Researchers are using ML to, for example, classify cells, creating new compounds, calculating compound attributes and a variety of other tasks.

4.1.4.1 ML Pipeline

The general implementation of ML on a data-set is not too complicated. The first task is to structure the data in such a way that it is compatible with the desired ML model. If working with numerical data: the data must be structured in a consistent format usually in the form of a table with columns representing the different features and the rows individual data points or samples. Computer vision tasks require a dataset of consistently labelled images or masks. It can be argued that the quality of the structured data-set is the most important task, as inconsistencies will result in non-robust models. The second task is model training, where the aim is to generate a model that can generalize to new data. This step typically involves splitting data into training and testing sets, where the model learns the training set and tries to generalize to the unseen test set. The choice of models is vast and ever enlarging, therefore choosing the best model often times requires

simply testing out different models and analysing the results. The final step is optimization, wherein the aim is to slightly tweak the specific parameters of the model to boost performance. Feature selection is applied in this step as some features do not necessarily correlate with task that the model is trying to learn, and therefore an optimal feature set is desired. The mathematics behind ML models is very diverse and therefore require the tuning of different parameters between the applied models and although today's models are good at optimizing for these parameters depending on the input data, small tweaks may offer performance boosts.

4.1.4.2 ML in Sarcoma Research

ML and radiomics have gained increasing attention in sarcoma research, offering tools to support diagnosis, grading, and prognostication through quantitative imaging analysis. Tagliafico et al. [62] investigated the potential of MRI-based radiomics in soft tissue sarcoma (STS) by analyzing scans from 11 adult patients. They identified four features significantly correlated with tumor severity and another four correlated with tumor size, demonstrating that radiomic descriptors could effectively differentiate between healthy and pathological tissue. Their findings supported the feasibility of integrating radiomics into routine diagnostic workflows for STS assessment.

Zhang et al. [63] developed a radiomics-based model to distinguish between low- and high-grade STS using MRI. By selecting five discriminative features, they constructed a classifier that achieved a predictive accuracy of 88% in grade differentiation. This work highlighted the role of targeted feature selection in maximizing classification performance and reinforced the potential of radiomics for non-invasive tumor grading in clinical settings.

Similarly, Corino et al. [64] applied radiomic analysis to apparent diffusion coefficient (ADC) maps derived from diffusion-weighted MRI of 19 patients with STS. Using first-order statistics and texture-based features (GLCM and GLRLM), they identified as few as three features capable of distinguishing intermediate- from high-grade tumors, achieving an accuracy of 88% and high AUC values in both validation and test sets. The study demonstrated that robust grading performance could be achieved using a minimal, carefully selected radiomic feature set.

Gao et al. [128] addressed the challenge of differentiating osteosarcoma (OS) from chondrosarcoma (CS) using MRI sequences, fat-suppressed T2-weighted imaging (T2WI-FS) and contrast-enhanced T1-weighted imaging (CET1). From 530 extracted radiomic features, LASSO regression identified 15 T2WI-FS and 8 CET1 features most relevant for classification. Their logistic regression-based radiomics model achieved AUC values of up to 0.97 in the training set and 0.90 in validation, showing that MRI-based radiomics can provide a non-invasive, accurate alternative to biopsy in distinguishing between OS and CS.

Hu et al. [65] explored computer-aided diagnosis of OS using texture features extracted from bone computed radiography (CR) images. By applying wavelet transforms (Sym4 and Db4) and gray-level co-occurrence matrices (GLCM), followed by support vector machine classification, they achieved recognition accuracies of 93.44% for epiphyseal OS and 96.25% for diaphyseal OS. Their findings demonstrated the advantage of multi-resolution wavelet-based feature extraction over conventional texture descriptors for OS detection.

Chen et al. [67] combined MRI-derived radiomics features with clinical variables to perform a survival analysis in 62 patients with extremity and trunk STS. Their model identified radiomic signatures associated with disease-free survival, showing that imaging biomarkers could serve not only for diagnosis and grading but also for prognostication. This integration of quantitative imaging data into survival modeling illustrated the expanding role of radiomics in personalized sarcoma care.

Xu et al. [129] developed a preoperative MRI-based radiomics and machine learning framework to predict histopathologic grades in soft tissue sarcoma (STS). Using T1-weighted and fat-suppressed T2-weighted MR images from 105 histologically confirmed STS cases, they extracted a comprehensive set of radiomic features and applied 11 feature selection methods in combination with seven classifiers. The most effective approach, applied least absolute shrinkage and selection operator (LASSO) for feature selection coupled with a random forest classifier and achieved an AUC of 0.92 and accuracy of 91.43% in the validation set. This approach addressed the frequent underestimation of tumor grade in biopsy specimens due to intratumoral heterogeneity, demonstrating that noninvasive, image-based grading can improve preoperative decision-making and supplement histopathologic evaluation.

Building on the integration of quantitative imaging and advanced analytics, Zhou et al. [130] applied a deep learning-enhanced radiomics pipeline to distinguish between benign and malignant bone tumors, including osteosarcoma and Ewing sarcoma, on MRI. Their method combined automated tumor segmentation with convolutional neural networks (CNNs) for feature extraction, followed by a gradient boosting classifier to perform final predictions. By leveraging CNN-derived deep features alongside traditional radiomic descriptors, the hybrid model achieved superior diagnostic performance, with AUC values exceeding those of conventional radiomics alone. The results illustrated how combining deep learning with classical machine learning methods can enhance tumor characterization accuracy, paving the way for fully automated, high-throughput diagnostic tools in sarcoma imaging.

4.1.5 Research Aims

The overall goal of this research is to develop a pipeline which can assist health care professionals in diagnosis and grading of sarcoma. This is carried out through the collection of MRI image data from the IceSG database and the subsequent radiomics based data mining to create a feature data table representation of patient data. The table is then used to train multiple ML models to complete the following tasks:

1. Distinguish healthy tissue from pathological.
2. Predict the severity or grade of the sarcoma tumor.
3. Predict the sarcoma sub-group, soft tissue sarcoma (STS) or bone sarcoma (BS).

The underlying aim of this research is to further techniques for non-invasive sarcoma diagnostics using predictive algorithms, as this reduces the number of tests conducted on patients and could reduce the time from the appearance of symptoms to treatment, which improves patient survival rate.

In summary, this research explores multiple techniques of data extraction, from viewing and analyzing sub-segments of tumor morphology, to viewing and analyzing the tumor as a whole. The results of different techniques are subsequently compared and discussed and conclusions are proposed.

4.2 Materials and Methods

This research has a three main objectives: to apply machine learning (ML) to classify between healthy and pathological tissue, classify sarcoma grade (FNCLCC) and classify sarcoma sub-group (STS and BS). The overall, high level workflow is depicted in figure 4.3. The DICOM images of the IceSG database are segmented to masks of specific regions of interest. The masks are then used to extract descriptive features from the distinct regions. This is followed by a step of feature ranking and feature selection. All feature sets are then tested by training different classification models. The models are tested and performance metrics are calculated. The results are then analyzed and discussed.

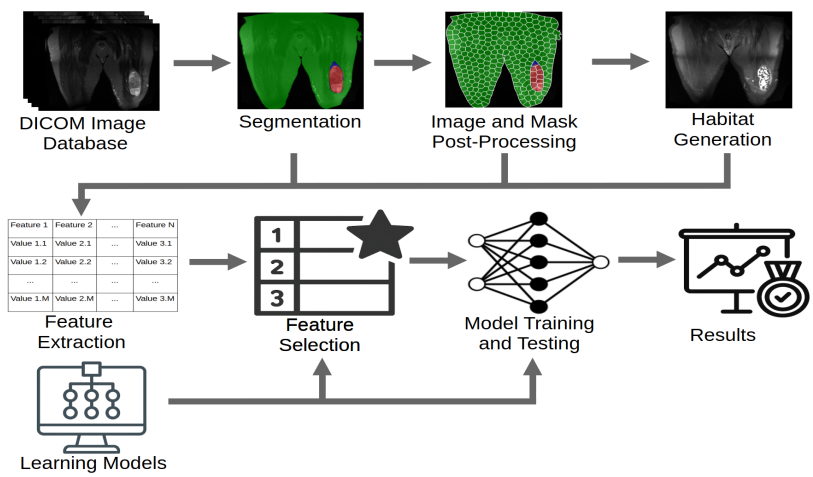


Figure 4.3: The overall workflow methodology in the research, depicting a high-level description.

4.3 DICOM Image Dataset

The database employed for this research is the Icelandic Sarcoma Group (IceSG) database. The data available consists of computed tomography (CT), magnetic resonance imaging (MRI), X-ray images and positron emission tomography (PET) scans from 212 sarcoma patients of all sub-types. These data consist of all sarcoma patients in Iceland from 2004 to 2023. Only MRI image sets are considered in this research as a way of standardizing the imaging modality and for their ability to capture soft tissue details, the literature also has a tendency towards MRI when it

comes to sarcoma [63], [67], [117], [120], [129]. Therefore, out of the 212 patients in the database, only 101 are considered.

It should be noted that there was no standardization in the MRI protocol between patients and scans, therefore all contrast media, weighting and other variables pertaining to the imaging are considered. The database also contains images which have become corrupted over the years, these are also included as a challenge, as it may reflect real world conditions.

The data is obtained through a web portal and is served as a set of digital imaging and communications in medicine (DICOM) files. The files contain raw image slices of the region of interest, often times from multiple perspectives: coronal - front view, sagittal - side view and axial - top down view.

Table 4.1 shows the distribution of the patient cohort in regards to sarcoma sub-type, severity and gender. The most prevalent sub-type is STS making up 65 out of 101 cases. There are 59 males and 42 females in the patient cohort mirroring the trend that males tend to develop sarcoma at a slightly higher rate than females [131]. There are 19 instances of grade 1 sarcoma, 16 of grade 2 and 66 of grade 3, introducing a large imbalance which could manifest difficulties in analysis.

The use of this data has been approved by Landspítali Ethics Committee (5/2024). The research has been approved by the Landspítali Scientific Research Committee (01/2024 - no 93). The World Health Organization (WHO) recognizes

Table 4.1: Dataset demographics. Included sarcoma subtypes are chondrosarcoma (CS), dermatofibrosarcoma protuberans (DFSP), ewing sarcoma (ES), giant cell tumor of bone (GCTB), osteosarcoma (OS) and soft tissue sarcoma (STS).

Age Group	0–19	20–39	40–59	60–79	80+	Total
Subtype						
STS	4	10	19	22	10	65
OS	10	0	3	1	0	14
CS	0	1	6	0	1	8
ES	5	4	1	1	0	11
DFSP	0	1	0	0	0	1
GCTB	1	0	1	0	0	2
Grade						
1	3	0	10	3	3	19
2	1	2	3	8	2	16
3	16	14	17	13	6	66
Sex						
Male	12	8	17	14	8	59
Female	8	8	13	10	3	42

the applicability and need of AI use in healthcare and has set ethical guidelines [132]. They outline 6 core principles for the ethical use of AI in healthcare: protect human autonomy, promote human well-being, safety and public interest, ensure transparency, explainability and intelligibility, foster responsibility and accountability, ensure inclusiveness and equity, and to promote AI that is responsive and sustainable. This research aims to fulfill these principles to the fullest extent through secure anonymous data use, clearly reporting drawbacks and constraints and thereby not using models in clinical practices, until all 6 principles are in full compliance with WHO standards.

4.3.1 Segmentation

Segmentation is the process of marking areas on an image into different labeled sub-regions called masks. Segmented into separate masks are: the sarcoma, surrounding edema, all healthy tissue and tissue of a comparable healthy region, the software used for segmentation is Materialise MIMICS 25 (Materialise NV, Leuven, Belgium). The sarcoma mask contains the boundaries surrounding the sarcoma tumor. Edema is the biproduct of an expanding mass in tissue, as there is more internal pressure on the anatomy surrounding the tumor, it often times creates regions of high response in MRI around the tumor. The visibility of edema is dependent on many factors including MRI protocol, tumor location and size, and is therefore not always visible. The healthy tissue is the remaining anatomy outside of the sarcoma and edema regions. The healthy tissue of a comparable region is manually chosen as a region in the anatomy that resembles the tissue of the tumor region, such as healthy bone for bone sarcoma, or healthy muscle tissue for soft tissue sarcoma. This is the image set used in chapter 3 and consists of 86 patients.

Many patients in the data-set have multiple scans available, and therefore segmentations are carried out for all distinct available perspectives (axial, coronal and sagittal).

After this step, the resulting masks are directly fed into feature extraction (see section 4.3.4). They are also processed further.

4.3.2 Image and Mask Post-Processing

This step includes all of those processes required to standardize imaging data across the whole IceSG database and the steps required for processing images and masks for subsequent applied methodologies.

4.3.2.1 Normalization

To get accurate and comparable values from images and their masks, normalization steps are very important. Depending on the machine and imaging parameters, the pixel values can range significantly. This is a method to standardize pixel intensities across the dataset, which in turn should make all patients more comparable.

Another normalization technique is called resampling. As the height and width of a pixel in millimeters is not standardized across machines and scans, they will result in different scales across the DICOM image dataset. Resampling uses interpolation of pixel values to standardize height, width and depth of a pixel, increasing comparability of size scale differences between patients and tumors.

4.3.2.2 SLIC

Simple linear iterative clustering (SLIC) is an algorithmic function which splits an image into smaller regions called super pixels (SP) [52]. This method is used to split each image slice into multiple sub-regions allowing for a deeper analysis on the inner anatomy of all tissue regions. The SPs are computed separately on the sarcoma region and the healthy region, such that a SP can never contain both sarcoma and healthy tissue, see Appendix B for a more detailed explanation of SLIC.

The parameters of the SLIC algorithm are:

- Compactness: how uniform in shape the SP should be, a higher value gives more square and uniform SPs, a lower value gives more weight to the image colors and intensities when grouping pixels.
- No. pixels per SP: approximately how many pixels are desired for each SP.
- Maximum number of iterations: the maximum amount of iterations while waiting for the algorithm to converge.
- Minimum/Maximum size factor: how much larger or smaller a SP is allowed to be than the defined norm.

The SLIC algorithm is applied twice to the masks. The first pass aims to split masks into 2D segments, independently for each slice, aiming to produce segments of size approximately 400 pixels, compactness was set to 5, maximum number of iterations to 50 and the minimum and maximum size factors set to 0.8 and 1.2 respectively. The 2D SLIC segments are then fed to the feature selection component (see section 4.3.4). The second pass works in 3D, with an aim to split a sarcoma volume into very small volumes of size approximately 40mm^3 , with a compactness of 0.05, maximum number of iterations of 10, and minimum and maximum

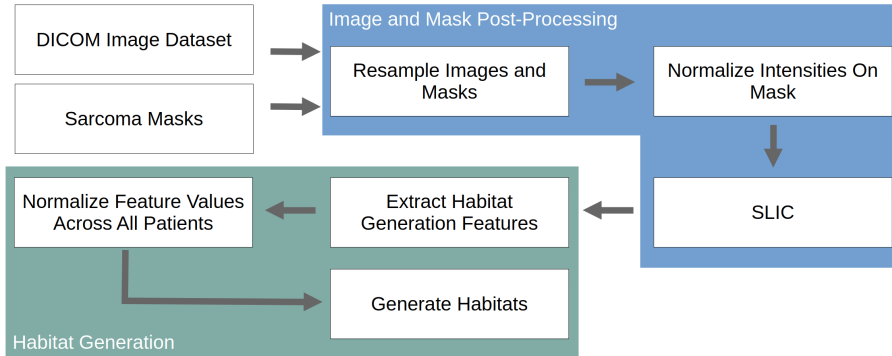


Figure 4.4: The general workflow for generating habitats.

size factors of 0.5 and 3 respectively. These SP segments are processed further into habitats.

4.3.3 Habitat Generation

Another method is also introduced to analyze fine details and tumor sub-region interaction, this method creates what are referred to as habitats. This method groups together SPs into clusters or habitats based on similarity measures and allows for analysis on the interaction the habitats have one another. To calculate the habitat boundaries, the input DICOM images and sarcoma masks are resampled such that each voxel, a pixel in 3D, has the same dimensions of 1 mm^3 . The image intensity is then normalized on the mask region, all pixels above the 99th and below the 1st percentiles are not included in the normalization calculations, as they often times reflect corruption or noise in the image, voxels outside of this range are therefore set to the upper and lower bounds. After normalization, SLIC is applied on the masked region of the image, the aim is to define segment sizes in cubic millimeters, since resampling was done such that each voxel is 1 mm^3 , the number of pixels equals the volume of the segment in cubic millimeters, the target volume was set to 40 mm^3 , which should offer a good trade off of having small enough SPs to create descriptive habitats, but not so many that compilation time is too high. The high level workflow for habitat generation is depicted in figure 4.4.

The next step is to group together newly generated SPs based on a similarity measure, this is done by extracting specific radiomics features known to have responsiveness in tumor biology for each SP. These include the mean, standard deviation, inter-quartile range, 10th percentile value and 90th percentile value of the voxel intensities [133], the mean gradient of the segment, the histogram values of the local binary pattern split into 65 bins, and contrast, homogeneity, angular

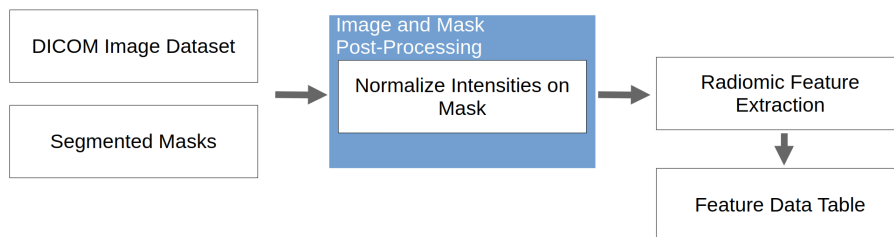


Figure 4.5: The overall radiomic feature extraction protocol.

second momentum and dissimilarity of the gray level co-occurrence matrix [134]. See section 4.3.4 for more details on radiomic features.

Once feature calculations have been completed for all patients, they are normalized globally across all patients, such that the habitats are meaningful and comparable globally. The final step is to cluster together those SPs that are similar. This is done by defining the number of habitats to a constant number K , chosen to be 3 based on other works [116]. Then a K-means clustering algorithm is applied that groups all similar SPs together and labels them to its corresponding habitat [117]. The anatomy surrounding the tumor is also considered, therefore there are 4 effective habitats. It should be noted that the numbering of habitats is arbitrary, meaning habitat 1 of patient 1 is not necessarily comparable to habitat 1 of patient 2. This arbitrary numbering therefore requires metrics that too are independent of habitat numbering, this is addressed in section 4.3.4.

4.3.4 Feature Extraction

Feature extraction is the method of obtaining quantifiable information from some source, that isn't directly available. In the case of this research, quantifiable parameters are extracted from MRI images, affectively reducing an image to a set of numerical features. Information can be obtained in many ways, through simple mathematical operations, complex protocols with multiple layers of operations before obtaining values or even automatically as is done in many artificial neural networks, known as latent features. Feature extraction in this research uses only easily understandable and explainable features, meaning features quantified and calculated through operations verified by the literature, not latent features.

4.3.4.1 Radiomic Feature Extraction

Radiomic feature extraction is the method for extracting numerical information from medical images, masks and SPs on MRI scans in this case, that have correlation with underlying pathologies and prognostic indicators [133]. PyRadiomics

Table 4.2: Types of image transforms applied in the research.

Image Transform	Highlighted Details
Original	The raw image
Wavelet	Horizontal, vertical and diagonal details
Laplacian of Gaussian	Enhances edges
Square	Squares the pixel intensities
Square Root	Takes the square root of image intensities
Logarithm	Calculates the logarithm of each pixel
Exponential	Calculates the exponential of the intensity $e^{\text{intensity}}$
Gradient	Emphasizes intensity changes between pixels
Local Binary Pattern	Captures local texture information

[55] is a package available for the popular programming language Python (Python version 3.9 [135]) which allows for in depth radiomics feature extraction from image masks.

Image transforms are mathematical operations performed on images to accentuate or dampen specific details and often times highlight details not otherwise visible in raw images. The research explores the image transforms outlined in table 4.2, it displays the applied image transformations and the details emphasized by them, these transforms were chosen and tested due to their availability and ease of use in PyRadiomics. Detailed mathematical formulations of the applied image transforms can be found in Appendix A. The wavelet transform has seen much use in ML applications in oncology with success especially in tumors known to be heterogeneous [56], [65], [66], [136]. The LoG transform has been applied somewhat in radiomics analyses in oncology [137], [138], and has been shown to increase predictive performance [139]. The LBP of medical images is widely used in radiomics, due to its ability to capture pixel intensity patterns which are correlated with heterogeneity of tumors [140]. Gradient based features have been applied in radiomic oncology analyses and have shown to improve predictive robustness [141]. Other transforms such as square, square root, logarithm and exponential, although not explicitly explored individually on their impact on predictive models in radiomics, have been used and analyzed in radiomics tasks in oncology and may offer valuable information for predictiveness of sarcoma [142].

There are seven types of feature classes, each with a subset of extractable radiomics features and are outlined in table 4.3. First order (FO) statistics contains 19 features, shape contains 10 2-dimensional and 16 3-dimensional features, Gray Level Co-occurrence Matrix (GLCM) contains 24 features, Gray Level Run Length Matrix (GLRLM) contains 16 features, Gray Level Size Zone Matrix (GLSZM) con-

Table 4.3: Feature classes and their description

Feature Class	Description
FO	Statistical variables that describe the distribution of intensities.
Shape	Descriptive features regarding the shape and outline of the region of interest.
GLCM	Comparisons of pixels with other single pixels in the region.
GLRLM	Features measuring the number of consecutive identical pixels.
GLSZM	A comparison of regions with identical pixel intensities.
NGTDM	Features comparing gray values to neighbours within a distance δ .
GLDM	Quantifies the dependence of a region a distance δ from pixel to said pixel.

tains 16 features, Neighbouring Gray Tone Difference Matrix (NGTDM) contains 5 features and Gray Level Dependence Matrix (GLDM) contains 14 features. Detailed descriptions and mathematical equations regarding the individual features can be found in the PyRadiomics documentation [55]. Each transform produces 93 features to add to the pool, the wavelet transform does four passes of high and low pass filter combinations to extract horizontal, vertical, diagonal and approximation features resulting in a total of 1116 features to analyze. After this step, all calculated feature and SP data are saved in a structured query language (SQL) database, the chosen database type is Python's built in SQLite3 [135]. To speed up loading times, the database is also compiled to a parquet file, this is a file format designed to load quickly, the drawback is that it cannot be queried, best used in those cases where a dataset or subset is frequently loaded. The resulting table containing numerical results of the feature extraction process are referred to as feature data tables.

Shape features are omitted from this research, as SP shape is decided algorithmically not from anatomy and could skew results. To keep the feature space consistent between different processing pipelines, they are also omitted from full mask analyses.

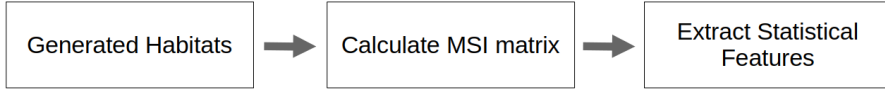


Figure 4.6: Workflow for extracting features from the generated habitats.

4.3.4.2 Habitat Feature Extraction

The feature extraction protocol differs slightly for habitat analysis. Instead of extracting radiomic feature values, the extracted features represent measures of interactions between the formed habitats within the region of interest. The multi-regional spatial interaction (MSI) matrix, $\mathcal{L}$, is defined as the count of how many pixels of habitat i touch or border habitat j .

$$\mathcal{L} = \begin{bmatrix} 0 & C_{1,2} & C_{1,3} & \cdots & C_{1,K} \\ C_{2,1} & 0 & C_{2,3} & \cdots & C_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{K,1} & C_{K,2} & C_{K,3} & \cdots & 0 \end{bmatrix} \quad (4.1)$$

$$\mathcal{L}_{norm} = \frac{\mathcal{L}}{\sum_{i < j} \mathcal{L}_{ij}} \quad (4.2)$$

Where K is the number of habitats, $C_{i,j}$ represent the number of times habitat i touches habitat j , therefore the matrix $\mathcal{L}$ is symmetric. $\mathcal{L}$ represents the absolute count of touching habitats and $\mathcal{L}_{norm}$ the normalized counts, such that the sum of the upper triangle of $\mathcal{L}_{norm}$ is 1, the diagonal is set to 0 as habitat self-touching is not quantifiable. MSI matrices in their raw form can be thought of as a measure of heterogeneity, the more heterogeneous the tumor the more disconnected the individual habitats become, increasing the number of border pixels. In the normalized form, it represents the distribution of habitat contacts.

With 3 tumor habitats and 1 habitat for the surroundings, the resulting MSI matrix is used for extracting statistical features, outlined in table 4.4.

Row strength, entropy and MaxFrac along with rim exposure and RowFrac are habitat dependent and give 1 value per habitat, as habitat numbering is to be arbitrary, the vector containing those features are operated on to make numbering arbitrary by taking the STD, mean, skewness, kurtosis, minimum value and maximum value. The total number of features collected from the MSI matrix is therefore 52. This data table is considered the feature data table for habitat tasks going forward, and follows the same pipeline as the radiomic feature table tasks. The question answered with this method is whether or not heterogeneity measures alone can accurately predict sarcoma severity and sub-group. The parameters of

Feature	Definition	Description
Pair Sum	$\sum_{i<j} \mathcal{L}_{ij}$	Total mixing across all pairs
Balanced STD	$\text{std}\{\mathcal{L}_{ij} : i < j\}$	Variation in pairwise mixing
Edge Entropy	$-\sum_{i<j} p_{ij} \log p_{ij}$	Diversity of pairwise interactions
Dominant Pair	$\arg \max_{i<j} \mathcal{L}_{ij}$	Habitat pair with strongest mixing
Dominant Value	$\max_{i<j} \mathcal{L}_{ij}$	Strength of dominant pair
Rim Total	$\sum_{i \geq 1} \mathcal{L}_{i0}$	Total tumor background mixing
Size Mix Corr	$\rho(s_i, \text{size}_i)$	Correlation of size and mixing
Size Weighted Mix	$\sum_{i<j} \mathcal{L}_{ij} \frac{\text{size}_i + \text{size}_j}{\sum_k \text{size}_k}$	Mixing weighted by habitat sizes
Edge NearZero	$\frac{\#\{i<j : \mathcal{L}_{ij} < 10^{-6}\}}{\binom{K}{2}}$	Fraction of negligible edges
Row Strength	$s_i = \sum_j \mathcal{L}_{ij}$	Total exposure of habitat i
Row Entropy	$-\sum_j p_{ij} \log p_{ij}$	Diversity of i 's neighbor contacts
Row MaxFrac	$\max_j p_{ij}$	Dominant neighbor fraction for i
Rim Exposure	$\mathcal{L}_{i0}$	Habitat background mixing
Rim RowFrac	$\mathcal{L}_{i0}/s_i$	Fraction of exposure at rim

Table 4.4: Summary of features extracted from normalized MSI matrices $\mathcal{L}$.

table 4.4 are extracted both for the MSI matrix $\mathcal{L}$ and $\mathcal{L}_{norm}$, brining the total number of extracted features to 104.

4.3.5 Learning Models

Learning models (LMs) are the computer algorithms trained on input data with an aim to predict some discrete class or continuous value. Each type has multiple varying input parameters and their efficacy depends heavily on data structure and specific use cases. The models applied here are

- Random Forest (RF)
- XGBoost (XGB)
- Light GBM (LGBM)
- Cat Boost (CB)
- Logistic Regression (LR)
- Ridge Classifier (RC)
- SGD Classifier (SGDC)

The models can be obtained through Python's Sci-Kit Learn package [58] or through the python package installer Pip.

Classification refers to the action of labelling a data sample to a specific discrete group or class. Binary classification is the classification of 2 classes and they are often referred to as the positive and negative class. Multi-class classification is the classification of any number of classes (>2).

Model training is the process of fitting a predictive function to the underlying structure of a dataset. The approach differs substantially between linear and tree-based models. Decision trees are trained through a recursive process in which potential splits along feature thresholds are evaluated at each node. These splits are assessed using impurity measures such as information gain for classification or squared error reduction for regression. The branch corresponding to the best split is expanded, and this procedure continues until predefined stopping conditions, such as maximum depth or minimum samples per leaf, are reached. In contrast, linear models are trained by optimizing parameters that minimize a chosen loss function, which quantifies the discrepancy between predictions and ground truths. Depending on the formulation, this optimization may be solved in closed form, as in ordinary least squares, or iteratively using gradient-based methods when dealing with logistic or regularized variants.

4.3.6 Feature Selection

Feature selection is a very important step, as not all radiomics features correlate with the objectives and may therefore introduce noise, reducing model performance [143]. Also, different types of LMs do not handle highly correlated features and low variance features adequately. It is also possible to achieve very positive results with few features, if they correlate well with the classes being predicted. Initially all features with variance less than 0.01 are removed [144] along with one feature of a correlated pair with correlation greater than 0.99 [145]. Many papers use other correlation thresholds e.g. 0.85 [144], however the feature elimination step is also designed to remove redundant features.

4.3.6.1 SHAP

The Shapley Additive Explanations (SHAP) is a method for ranking the importance of features [146], based on a concept from cooperative game theory. The creator, Shapley, desired to answer how to distribute a payout fairly among players, based on their contribution to the game outcome. Given N players, a subset of players $S \subseteq N$, a value function of $v(S)$ which is the total sum of payoffs, the fair payoff for player i is then given by

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (4.3)$$

In modern day ML tasks, the players are features and the payoffs are their importance, a high Shapley value indicates that a feature correlates highly with the positive class, and a negative value correlates more with the negative class. The ranking therefore consists of calculating the Shapley value of each feature, with the aim of selecting the feature set with values furthest from 0. The power of this method comes from its versatility, it can be easily applied to multiple types of models. Due to the inherent mathematical differences between learning models and the fact that feature interactions can be highly complex, especially with large feature sets, feature rankings may vary between selected models. The benefit of SHAP is that it is model dependent, meaning that the ranking algorithm requires an input model in order to complete ranking calculations. SHAP was chosen as the feature ranking algorithm for this reason, as one function can rank features for all models, reducing risks of human error while coding and standardizing the methodology.

4.3.6.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a process of feature selection. Starting with the entire feature list, excluding highly correlated and low variance features, all features are ranked based on a subset of the entire dataset, the radiomics or MSI feature data tables. Once the features are ranked, remove some ratio of the lowest ranked features and repeat feature ranking, this is done iteratively until the desired number of minimum features is reached. Features are ranked on every iteration, because the interaction between features can be very complex and the removal of a feature can change the balance of feature interactions [147]. This process needs to be applied to every model and dataset independently, since models behave differently and utilize features in different ways, therefore the ranking of a feature might vary across different types of models.

It was decided to remove 10% of features on each feature ranking iteration and remove 10 features at a time once 100 features remain. This was deemed as a good balance of having enough samples to see how models behave with a specific amount of features, without compilation times being too long, as SHAP calculations are very difficult and can take a long time based on the number of features and the number data samples. The remaining set of features after each iteration of ranking and elimination are saved for testing, they are referred to as feature subsets. Due to few features (104) extracted from the MSI matrices and the fact that habitats are generated per tumor (not per slice or segment) the number of samples is equivalent to the number of available scans, which is 195. Therefore compilation times are not an issue and one feature is removed at a time after each ranking

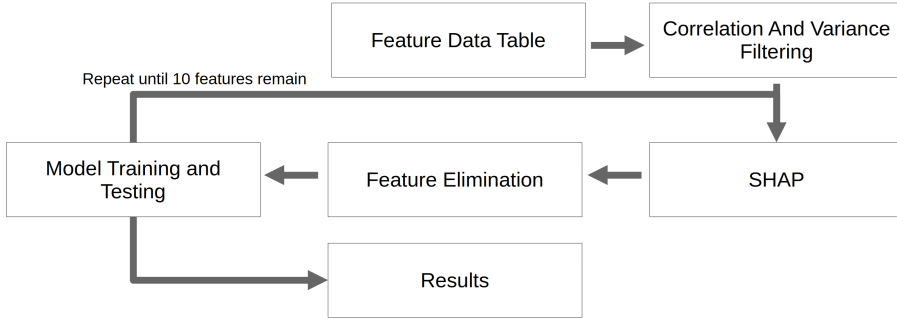


Figure 4.7: The training workflow for all tasks. The only differences between tasks are what the prediction aim is and the initial full feature data table: extracted features from the full sarcoma mask, extracted features from the SLIC segments and statistical features extracted from the MSI matrices of generated habitats. SHAP ranks features, feature elimination removes the bottom ranked features, remaining features are trained and tested on for all models, this is done recursively until few features remain. Result metrics are calculated at each recursion iteration.

iteration. The feature rankings are analyzed to quantify the feature groups and image transforms that have the most positive effect on performance.

To analyze feature selection behaviour the Jaccard index is used, and is defined by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4.4)$$

Where A and B are the two feature subsets being compared, the equation is calculating the number of features shared between the two sets divided by the total number of features in the sets. A value closer to 1 means that the sets are similar, and more dissimilar if the value is close to 0. The Jaccard index is used to compare the similarities of feature sets between models and datasets, datasets are defined in section 4.3.7 and in table 4.6. First, for each dataset, the top 100 selected features are compared between the models, to get an understanding of how feature selection varies between the models. Then for each dataset, the top 100 features of each model are pooled into a super-set and are compared to the other datasets, giving an intuitive measure of how different features are selected depending on the classification and mask post-processing steps.

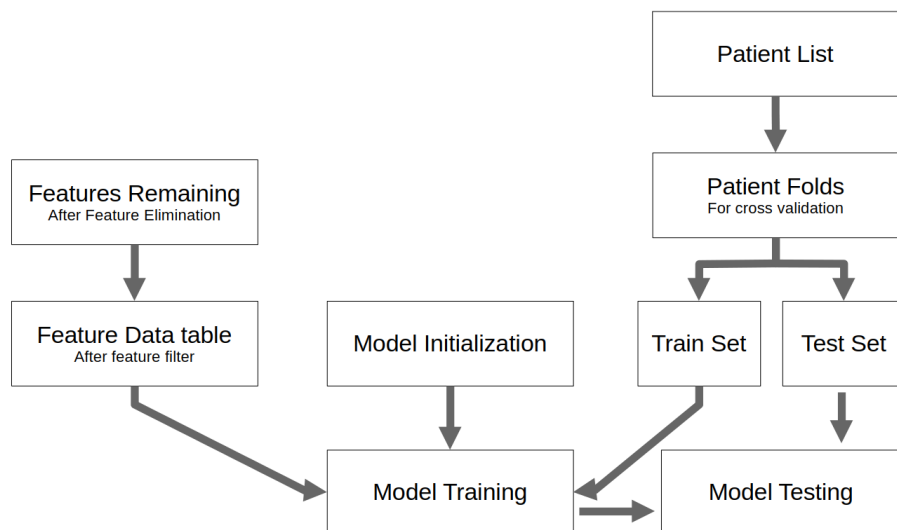


Figure 4.8: A more detailed workflow of the model training and testing component (see figure 4.7 for an overview). Patients are split into folds, based on criteria set by the training task (sarcoma vs. healthy, grade and sub-group classifications)

4.3.7 Model Training and Testing

When talking about model training, it is important to define the training and testing sets. The training set is used for updating model parameters, this is the set of data that the models learn from. The testing set is the subset of data where the model applies its learned information to see how well it can generalize to unseen data. This research employed a cross-validation (CV) approach [148]: the data is split into a number of subsets called folds and each fold is used as the testing set once, while the rest are used as the training set, then the test set is changed until all folds have had a chance to be the test set. This ensures that the results are not affected by a fold randomly being too easy or difficult to train or test on. Each fold contains approximately 10 patients therefore patients cannot be in both the training and testing sets, this ensures no data-leakage between the sets. Data-leakage is when similar data is in both training and test-sets, which often times increases model metrics, but decreases generalization as, in a real-world scenario, a model is not trained on a new patient's data that it is analyzing. The folds are kept identical throughout the experiments such that an accurate result of model performance is achieved across different processing steps. Other benefits of CV include increased robustness, reduced bias, prevention of overfitting and works well with limited data. Although CV increases training times drastically, it is performed for

all models at all stages in the training process.

As has been discussed the MRI image data go through various pipelines and each results in different feature data tables, e.g. masks processed using SPs result in different feature data tables than those processed without. The varying feature data tables are referred to as datasets.

Imbalance is inherent in IceSG database, the distribution of grades is severely skewed towards grade 3, same goes for sarcoma sub-types as most are STS and there is generally orders of magnitude more non-pathological tissue than pathological. Due to this, balancing methods are implemented in the form of calculating an imbalance ratio, defined as the number of samples of the minority class divided by the number of samples in the majority class, for binary classification. For multiclass classification, the imbalance ratio for class i is the total number of samples divided by the product of the number of classes and the number of samples belonging to class i , this provides an imbalance vector. This gives more weight to minority classes during classification to avoid the majority class bias, where models simply predict a single class and report abnormally high accuracy due to large imbalances. The imbalance metrics are then provided to the learning models, and they can then train to avoid the majority bias.

Once the ranking of all features and feature subsets has been achieved, a model is trained multiple times to approximate the optimal number of features that give the highest score. The metric optimized for is the area under the curve (AUC) of the receiving operating characteristics (ROC) curve, abbreviated to AUC-ROC. The AUC-ROC score is a measure of a given models ability to distinguish between classes, meaning how well it can generate a separation of classes. A high AUC-ROC score generally means that the model can generalize better to new unseen data.

The metrics extracted to rate model performance are accuracy, AUC-ROC, F1-score, precision, recall see table 4.5 for metric definitions. The metrics are calculated per fold. Macro and micro averaged metrics are calculated. Micro-averaging computes metrics based on the absolute contribution of all classes, meaning that larger classes have larger weight. Macro-averaging computes metrics independently for each class and takes a simple mean of the results, ensuring that each class has equal weight in the results. Metric calculations as well as metric averaging is handled by functions provided by the Sci-Kit Learn package [58]. The output of model testing are a list of predictions and a list of the probability that it belongs to the predicted class. In the prediction vector, each value is the predicted class of the input sample, the prediction vector is compared to the ground truth vector, if the i^{th} value of the ground truth is identical to the i^{th} value of the predictions vector, then prediction is correct, otherwise it is incorrect. In the probability vector, each value represents the likelihood that the respective sample belongs to a class, for binary classification it is a single value for the probability of the positive class, in multiclass classification each value is a vector with the probability that

the sample belongs to one of the K classes. If $\hat{p}^k(x)$ is the probability of sample x belonging to class k , then the AUC-ROC score is given by

$$ROCAUC_k = \frac{1}{n_k n_{-k}} \sum_{i: y_i=k} \sum_{j: y_j \neq k} \mathbf{P}(\hat{p}^k(x_i) > \hat{p}^k(x_j)) + \frac{1}{2} \mathbf{P}(\hat{p}^k(x_i) = \hat{p}^k(x_j)) \quad (4.5)$$

Where

$$\mathbf{P}(\text{condition}) = \begin{cases} 1, & \text{if the condition is true,} \\ 0, & \text{if the condition is false.} \end{cases} \quad (4.6)$$

In Layman terms, sum over all samples (i) belonging to class k and all samples (j) not in class k if the probability that the i^{th} samples belongs in class k is greater than the probability of sample j belonging to class k then the function $\mathbf{P}$ is 1, if the probabilities are equivalent add 1/2. The total sum is then divided by the product of the total number of samples belonging to class k : n_k , and the total number of samples not belonging in class k : n_{-k} . This is repeated for all K classes and therefore the AUC-ROC score is a vector when the number of classes is greater than 2, for binary classification the AUC-ROC is calculated only for the positive class. To rank a run where the number of classes exceeds 2, the macro averaged AUC-ROC score is used as the main metric for assessment.

Precision measures, for class k , how many predictions of class k are correct:

$$\text{Precision}_k = \frac{TR_k}{TR_k + FR_k} \quad (4.7)$$

Here TR_k and FR_k denote true and false positives for class k , respectively. Recall measures, for class k , how many of the true instances of class k were identified:

$$\text{Recall}_k = \frac{TR_k}{TR_k + FN_k} \quad (4.8)$$

where FN_k denotes false negatives for class k . The F1-score for class k is the harmonic mean of its precision and recall:

$$F1_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \quad (4.9)$$

A high F1-score requires both high precision and high recall.

Overall accuracy represents the fraction of correctly classified samples:

$$\text{Accuracy} = \frac{\sum_{k=1}^K TR_k}{N}, \quad (4.10)$$

where N is the total number of samples and $\sum_k TR_k$ corresponds to the sum of correctly classified samples. To give equal weighting to each class, macro averaging

computes the mean of the per-class metrics. This approach reduces the influence of class imbalance and ensures that minority classes contribute equally to the overall score.

$$\text{Precision}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \text{Precision}_k \quad (4.11)$$

$$\text{Recall}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \text{Recall}_k \quad (4.12)$$

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \text{F1}_k \quad (4.13)$$

The macro recall is another method to measure class balanced accuracy, giving a more robust representation of imbalance independent accuracy. To capture overall performance across the dataset, micro averaging aggregates true positives, false positives, and false negatives across all classes before computing the metric. This approach reflects how the model performs on all predictions, where larger classes contribute proportionally more to the result.

$$\text{Precision}_{\text{micro}} = \frac{\sum_{k=1}^K TP_k}{\sum_{k=1}^K (TP_k + FP_k)} \quad (4.14)$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{k=1}^K TP_k}{\sum_{k=1}^K (TP_k + FN_k)} \quad (4.15)$$

$$\text{F1}_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \quad (4.16)$$

In multi-class classification problems where each sample belongs to exactly one class, the micro-averaged precision, recall, and F1-score all reduce to the same value, which is equivalent to accuracy. Macro metrics highlight performance equally across classes, while micro metrics reflect overall performance weighted by the size of each class. Refer to table 4.5 for definitions on calculated metrics.

Figures 4.7 and 4.8 depict the standardized training workflow for all tests. One of the processed feature tables is fed into the SHAP ranking function and outputs features and their relative ranking, bottom ranked features are eliminated and the remaining features enter the model training and testing block, from which model assessment metrics are extracted. The detailed model training and testing consists of generating patient folds from the list of patients to create a training and test sets, and to perform CV. Once the model has been initialized, the feature data table is filtered based on the current iteration of training patients and current feature subset. The model is trained on the subset feature data table, and then tested on

Table 4.5: Metrics used for the assessment of model performance and their respective definitions.

Metric	Definition
Accuracy	Proportion of all samples that are classified correctly.
Precision	Share of predicted class members that truly belong to that class.
Recall	Share of actual class members that are correctly identified.
F1-score	Balance between precision and recall, expressed as their harmonic mean.
CM	Table showing correct classifications and misclassifications for each class.
AUC-ROC	Measure of how well the model separates classes across decision thresholds.
Macro average	Average across classes, giving each class equal weight.
Micro average	Computed over all predictions, giving larger classes more influence; equal to accuracy in multi-class tasks.

the testing patients (with the same feature subset, of course) from which model assessment metrics are calculated.

Training is carried out using both central processing unit (CPU) and graphics processing unit (GPU) compatible code. Generally, GPU training is faster and more desirable as the processing unit is designed for faster matrix and tensor calculations and higher memory bandwidth, however due to many tree-based models' architecture of creating branches, which introduce irregular memory access patterns, CPUs are more adaptive for these tasks. A range of CPUs and GPUs were used in the processing of calculations in this research, GPUs include: NVidia 5090, NVidia A100 and NVidia 4060TI. CPUs used were: Intel Xeon W-1350, Intel Xeon Gold 6248R (16 and 48 core) and AMD 9950X3D. Workloads were distributed among the array of available CPUs and GPUs, with multithreading enabled to speed up all heavy calculation tasks such as model training and feature extraction. GPU trained models are: CatBoost, LightGBM and XGBoost; CPU trained models are: RandomForest, LogisticRegression, Ridge Classifier and SGD Classifier.

4.3.7.1 Tissue Classification

This classification task aims to discriminate healthy tissue against sarcoma tissue. Figure 4.9 depicts the tests made in tissue classification. First tested, is the ability to classify sarcoma against all healthy tissue (SvAH dataset), this test includes no processing of masks. The next test is the same as the first, but SLIC is applied over

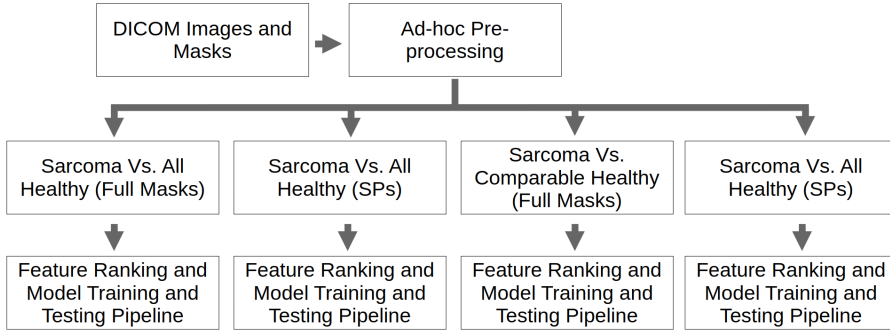


Figure 4.9: A depiction of the tests made for classifying healthy vs. pathological tissue.

sarcoma and healthy masks separately (SvAH-SP dataset), turning each mask into multiple smaller segment masks. Simply put, given a small segment of tissue, is it possible to distinguish between sarcoma and healthy tissue? The next test aims to distinguish sarcoma tissue from tissue in a comparable pathology-free region (SvCH dataset), using the full mask of both regions. The final test aims to do the same as the previous but by applying SLIC to the two masked regions (SvCH-SP dataset).

In an attempt to increase the amount of data, a publicly available dataset of STS [149] was added to the IceSG database. Available are 50 patients, along with masks of sarcoma and the surrounding edema, healthy tissue masks were subsequently segmented manually and added to the SvAH and SvAH-SP datasets. The patients are split into 10 folds for CV.

4.3.7.2 Grade Classification

Grade classification aims to diagnose the grade of the sarcoma of the patient: 1, 2 or 3 according to the FNCLCC guidelines. As has been discussed in section 1.4, the grading of sarcoma can be subjective and prone to human error, therefore it could be of great help to the diagnostics field to have a supporting algorithm to help conceptualize visual features to a sarcoma grade. As there are 19 instances of grade 1, 16 of grade 2 and 66 grade 3 sarcomas, the dataset is imbalanced. This issue is tackled by using a class weights, giving higher weight to the underrepresented classes.

CV is implemented in a similar manner to the tissue classification splitting patients across 10 folds, with the added caveat that each fold has a similar distribution of grades. Again, patients cannot be in both the training and test sets, to avoid data-leakage and to improve model robustness. Grade is predicted using

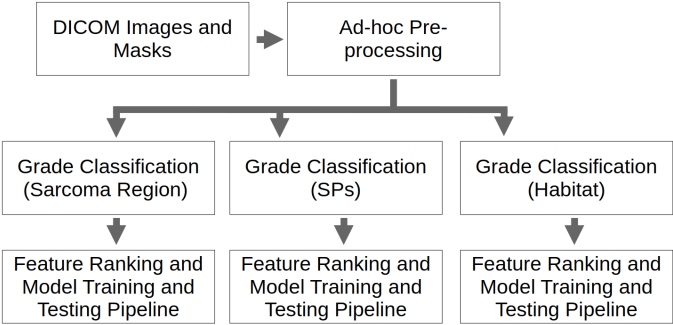


Figure 4.10: A depiction of the tests made for classifying sarcoma grade.

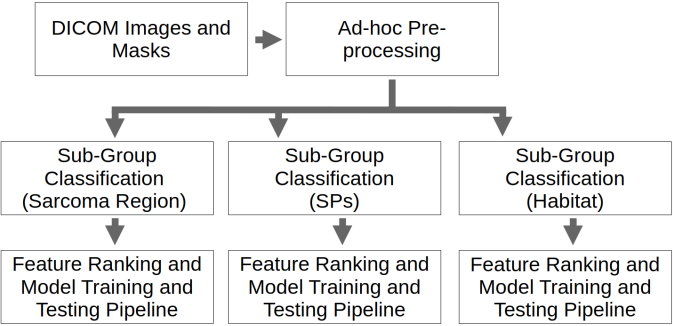


Figure 4.11: A depiction of the tests made for classifying sub-groups.

full masks (Grade dataset), by using SPs (Grade-SP dataset) and through habitat analysis (Grade-H dataset).

4.3.7.3 Sub-Group Classification

This classification aims to predict which sub-group the sarcoma belongs to: soft tissue sarcoma (STS) or bone sarcoma (BS). Due to a lack of patients belonging to each specific sub-type in the dataset and a majority of patients (65) being diagnosed with STS, all patients diagnosed with osteosarcoma, giant cell tumor of bone and ewing sarcoma were pooled together to form the BS sub-group. The total number of patients belonging to the the BS group is 27. The total number of patients considered in this step is therefore 92.

CV is performed by designating patients across 9 folds with approximately 10 patients per fold, ensuring that there is an even distribution of the minority group

in each fold. Sub-group predictions are made using full masks (SG dataset), SPs (SG-SP dataset) and through habitats (SG-H).

Table 4.6: A list of all ML tests applied in the research. Note: each of the 10 tests go through the feature ranking, feature selection, feature elimination and cross-validation independently.

Classification	Dataset	Abbreviation
Tissue	Sarcoma vs. All healthy	(SvAH)
	Sarcoma vs. Comparable healthy	(SvCH)
	Sarcoma vs. All healthy with SPs	(SvAH-SP)
	Sarcoma vs. Comparable healthy with SPs	(SvCH-SP)
Grade	Sarcoma Region Only	(Grade)
	Sarcoma Region Only with SPs	(Grade-SP)
	Sarcoma Habitat	(Grade-H)
Sub-group	Sarcoma Region Only	(SG)
	Sarcoma Region Only with SPs	(SG-SP)
	Sarcoma Habitat	(SG-H)

4.4 Results

The results of feature selection and model training show that radiomics and habitat methods have good correlation with predicting different diagnostics and staging components of sarcoma. Tissue classification of all healthy tissue against sarcoma tumors show that models can easily distinguish between healthy and pathological anatomy (although with caveats). When viewing comparable regions against sarcoma, there is also very high discriminative ability, hinting that radiomics has much potential in sarcoma diagnostics. Even though classifying sarcoma grade didn't obtain great results, they were above chance level showing that there is some correlation between specific radiomics features and sarcoma grade. Predicting sarcoma sub-group, BS or STS, showed very promising results, given the limitations of the dataset. Interestingly, habitat generation and MSI matrix feature extraction was also positively correlated with sarcoma grade and sub-group, showing that heterogeneity could be used as a measure when diagnosing sarcoma. Different models opted for different features, shown through low Jaccard index of feature sets, showing how models may grasp different information and perform similarly even though they use different features.

Although, dataset sizes are rather dynamic, making comparisons a bit difficult, they do capture quantifiable information relating to the classes being predicted. Even with limitations, such as, unstandardized imaging protocols, image corruption, low representation of specific classes, techniques to combat this were effective and gave positive results.

The results are set up to mirror the methodology. Starting with segmentation and considerations, pitfalls and other details relating to it. Mask pre-processing results are stated and discussed, along with parameters used, caveats and limitations. Feature extraction and RFE show the results of the best feature sets of each model and the feature selection behavior between models and datasets. Model training and testing is presented separately for each classification set: tissue, grade, sub-group and habitat classifications.

4.4.1 DICOM Image Dataset and Segmentation

Mask segmentations were carried out over a 6 month period, during this time, the segmentations were approved by a medical professional. Many segmentations were very difficult due to MRI parameters giving little contrast between the sarcoma region and non-pathological surroundings, image corruption, blurry, grainy or noisy images, fused slices and uneven lighting distribution.

Figure 4.12 shows two examples of a corrupt image in the dataset. In these cases, segmentation was carried out to try and avoid including corrupt pixels in the mask, but in some cases corruption was so rampant that it was included. The thought process behind keeping the corrupt pixels was that excluding the corrupt

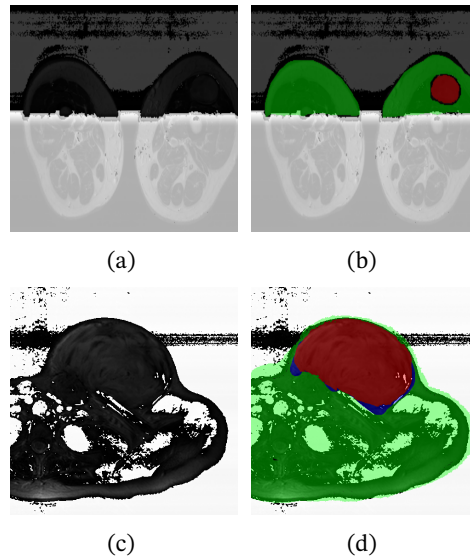


Figure 4.12: (a) An example of a corrupt image in the IceSG dataset (b) the mask of the sarcoma and healthy regions on a corrupt image. (c) Another example of a corrupt image (d) The resulting mask of the second corrupt image.

pixels, could skew the data to be less reflective of real world scenarios, especially if the corrupt pixels are within the tumor. Also, training models on difficult data has been shown to improve model resilience and robustness [150]. Lastly, it may represent real world cases, perhaps computer errors result in a corrupt scan that must be analyzed, which might create a use case for models that can handle extreme cases, although this has become significantly less common.

Figure 4.13 shows an example of segmentations where the contrast between the sarcoma tumor and surrounding anatomy was not great. These are cases where one would speculate great difficulty in classification, as for example, figure 4.13(a) where the sarcoma resembles greatly the surrounding bone, and in figure 4.13(c) where the tumor mimics the surrounding fatty tissue. To the untrained eye, nothing looks too out of the ordinary, but with an understanding of sarcoma and its presentations in MRI scans and general anatomical knowledge, this can be segmented and offers an interesting challenge for the learning models.

Figure 4.14 shows some cases of other difficult segmentations. There were cases of images being blurred together, almost as if 2 slices were fused. There were instances of artefact lines displayed on top of the MR images. Many scans were rather noisy and grainy. Uneven lighting across images also causes difficulties as normalization across the pixel intensities would be skewed and non-reflective of

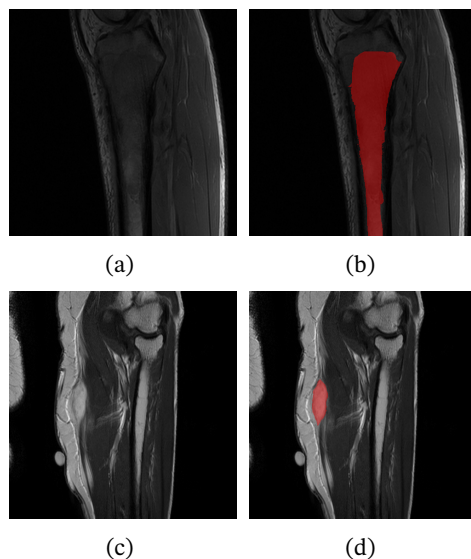


Figure 4.13: (a) An example of a bone sarcoma, where the border of the tumor is very unclear (b) The resulting mask (c) Another example of a sarcoma appearing very similar to surrounding anatomy (d) The resulting mask.

true pixel intensity distributions.

Despite some scans increasing the complexity of the dataset, the majority of scans and images were clean with no visual drawbacks as the ones discussed from figures 4.12 to 4.14. The 101 patients used for training the SvAH, Grade and SG datasets, resulted in 2883 image slices of sarcoma regions and 6003 non pathological image slices. The 86 patients used in the SvCH dataset, resulted in 2184 image slices of sarcoma and healthy regions.

4.4.2 Image and Mask Post-Processing

The image and mask post-processing pipeline turned out to be robust and gave good results. Compilation times varied greatly, especially when SLIC was applied to all healthy anatomy along with sarcoma. It resulted in a total of 1,402,939 samples, 72,749 of which were tumor super pixels (SP) and 1,330,190 non-pathological SPs, for the SvAH-SP, Grade-SP and SG-SP datasets. Compilation times were greatly managed by multithreading calculations, reducing run times by a factor of approximately 5. Figure 4.15 displays an example of the segmentation and SP generation on an input image.

Habitats were also successfully generated after all resampling and normaliza-

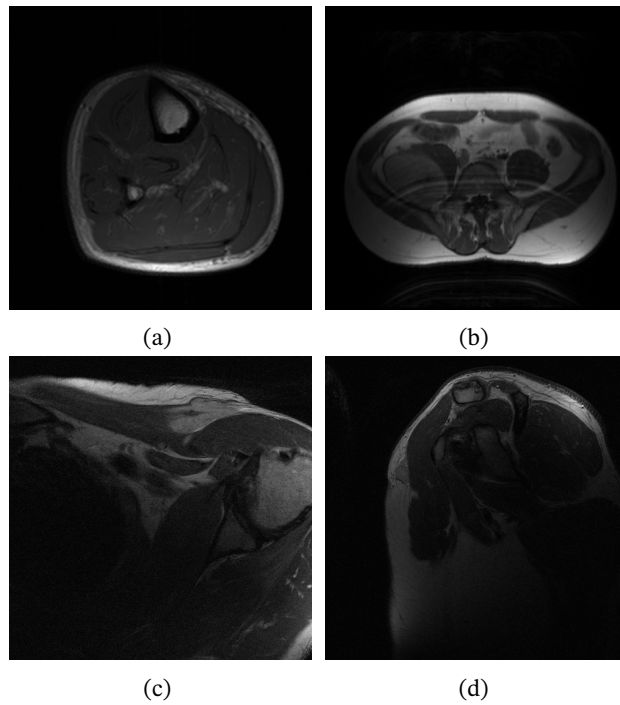


Figure 4.14: Examples of other difficulties in the segmentation process. (a) An MRI slice that seems to be a mix of two slices (b) Visual artefacts in an image slice (c) An example of a relatively noisy and grainy image (d) Uneven lighting or response throughout the image.

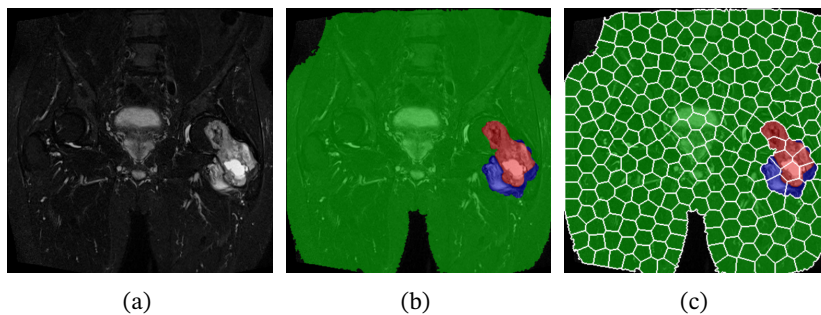


Figure 4.15: An example of a raw image slice and resulting SLIC processed super pixels on masks. (a) displays the original raw image slice (b) displays the masks; green is healthy, red is sarcoma and blue is edema (c) the resulting SPs.

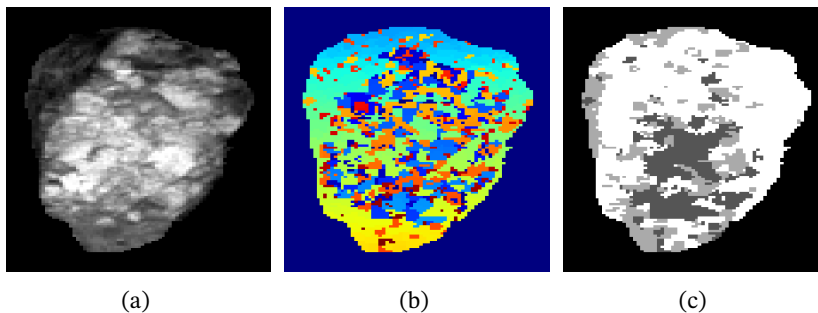


Figure 4.16: An example of a generated habitat (a) the original raw image slice of only the sarcoma (b) the SPs generated from the SLIC process, with an approximate volume of $30mm^3$, displayed in color due to the large number of SPs generated in this step (c) the resulting habitat generated.

tion steps. Figure 4.16 display an example of a generated habitat, displayed is only a single slice, the habitats do however span in all 3 dimensions.

4.4.3 Feature Extraction and RFE

Presented in this section are results from the feature extraction and recursive feature elimination (RFE) steps for radiomics and habitat features.

4.4.3.1 Radiomic Feature Space

The image and mask post-processing steps were slightly different. The first step was to filter out all features with variance less than 0.01 and one of the pairs of features with correlation at or above 0.99, thus the number of features filtered varied slightly between datasets. The SvAH, Grade and SG datasets contained 740 features after the filtering step. The SvAH-SP, Grade-SP and SG-SP dataset had 823 features remaining after filtering. The SvCH dataset was reduced to 796 features and the SvCH-SP dataset had 818 features remaining after the preliminary filtering step.

Due to the number of runs and resulting tables, the results of RFE resulting in the best score is displayed in tables 4.7 to 4.13. The "Total" columns in the tables show the total number of features selected from each type of image transform and total number of features selected from each feature group. Interestingly, nearly, all image transforms have representation in the optimal feature sets for all runs, the same goes for feature groups. See section 4.4.4 for figures of scores for each model and respective number of features tested.

Table 4.7: Selected features for highest scoring set of features for binary classification for all healthy against sarcoma masks. Best model was LogisticRegression with 100 features and a AUC-ROC score of 0.999.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	5	3	2	2	1	0	13
Exponential	2	5	2	6	1	1	17
Gradient	1	1	1	3	1	2	9
LBP-2D	3	4	1	0	1	0	9
LoG	5	0	0	0	3	0	8
Logarithm	2	3	0	0	0	1	6
Square	2	1	0	3	1	0	7
Square Root	1	2	0	1	2	1	7
Wavelet-HH	0	1	3	0	2	0	6
Wavelet-HL	1	1	1	2	1	0	6
Wavelet-LH	0	0	0	2	0	0	2
Wavelet-LL	1	0	3	1	3	2	10
Total	23	21	13	20	16	7	100

Table 4.8: Selected features for highest scoring set of features for binary classification for all healthy against sarcoma masks processed with SPs. Best model was CatBoost with 90 features and a AUC-ROC score of 0.885.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	1	0	0	1	1	0	3
Exponential	0	1	4	5	0	1	11
Gradient	2	3	1	0	0	0	6
LBP-2D	0	2	0	5	2	0	9
LoG	2	2	1	2	0	0	7
Logarithm	5	2	0	1	1	0	9
Square	2	0	0	0	0	0	2
Square Root	3	2	0	0	2	0	7
Wavelet-HH	2	1	5	2	2	0	12
Wavelet-HL	3	3	2	0	0	0	8
Wavelet-LH	3	2	3	1	0	0	9
Wavelet-LL	1	2	0	1	2	1	7
Total	24	20	16	18	10	2	90

Table 4.9: Selected features for highest scoring set of features for binary classification for comparable healthy against sarcoma masks. Best model was RidgeClassifier with 247 features and a AUC-ROC score of 0.974.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	5	6	4	5	1	0	21
Exponential	9	6	2	3	1	1	22
Gradient	7	3	4	5	1	0	20
LBP-2D	6	11	4	4	2	0	27
LoG	4	9	1	4	2	0	20
Logarithm	7	11	1	1	3	1	24
Square	5	9	5	8	1	1	29
Square Root	4	6	4	5	3	0	22
Wavelet-HH	1	1	0	1	0	0	3
Wavelet-HL	5	3	5	6	1	0	20
Wavelet-LH	7	6	2	6	0	0	21
Wavelet-LL	1	7	3	4	3	0	18
Total	61	78	35	52	18	3	247

Table 4.10: Selected features for highest scoring set of features for binary classification for comparable healthy against sarcoma masks processed with SPs. Best model was LogisticRegression with 100 features and a AUC-ROC score of 0.865.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	6	4	0	1	0	0	11
Exponential	3	3	3	1	0	0	10
Gradient	3	3	3	3	0	0	12
LBP-2D	1	2	3	3	2	0	11
LoG	0	2	0	1	0	0	3
Logarithm	1	2	1	0	1	0	5
Square	2	1	1	1	0	0	5
Square Root	2	3	1	4	0	1	11
Wavelet-HH	0	0	1	1	0	0	2
Wavelet-HL	0	4	0	1	0	0	5
Wavelet-LH	3	2	1	3	0	0	9
Wavelet-LL	0	6	2	7	1	0	16
Total	21	32	16	26	4	1	100

Table 4.11: Selected features for highest scoring set of features for grade classification for whole masks. Best model was SGDClassifier with 60 features and a AUC-ROC score of 0.690.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	2	1	1	1	1	0	6
Exponential	4	2	3	2	0	0	11
Gradient	0	0	0	1	1	0	2
LBP-2D	0	0	1	1	0	1	3
LoG	0	1	0	1	0	0	2
Logarithm	0	0	0	2	1	1	4
Square	4	2	1	3	0	0	10
Square Root	1	2	1	1	0	1	6
Wavelet-HH	0	0	0	1	0	0	1
Wavelet-HL	1	1	1	3	0	0	6
Wavelet-LH	0	1	1	2	0	0	4
Wavelet-LL	0	1	1	2	1	0	5
Total	12	11	10	20	4	3	60

Table 4.12: Selected features for highest scoring set of features for grade classification for masks processed with SPs. Best model was RidgeClassifier with 10 features and a AUC-ROC score of 0.670.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	1	0	0	0	0	0	1
Exponential	0	0	0	0	0	0	0
Gradient	0	1	0	0	0	0	1
LBP-2D	0	0	0	0	0	0	0
LoG	1	0	0	0	0	0	1
Logarithm	0	0	0	0	0	0	0
Square	1	0	0	0	0	0	1
Square Root	0	0	0	0	0	0	0
Wavelet-HH	0	0	0	0	0	0	0
Wavelet-HL	0	0	1	1	0	0	2
Wavelet-LH	0	0	1	1	0	0	2
Wavelet-LL	0	2	0	0	0	0	2
Total	3	3	2	2	0	0	10

Table 4.13: Selected features for highest scoring set of features for sub-group classification for whole masks. Best model was RidgeClassifier with 149 features and a AUC-ROC score of 0.863.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	3	3	1	4	1	1	13
Exponential	3	5	0	0	1	2	11
Gradient	4	5	3	2	1	0	15
LBP-2D	2	7	0	1	1	1	12
LoG	2	4	0	3	2	0	11
Logarithm	3	5	0	5	1	0	14
Square	6	4	1	4	0	0	15
Square Root	2	6	0	3	1	0	12
Wavelet-HH	2	2	1	1	1	0	7
Wavelet-HL	4	2	0	3	0	0	9
Wavelet-LH	5	5	1	4	2	0	17
Wavelet-LL	1	6	0	3	2	1	13
Total	37	54	7	33	13	5	149

Table 4.14: Selected features for highest scoring set of features for sub-group classification for masks processed with SPs. Best model was SGDClassifier with 100 features and a AUC-ROC score of 0.802.

	FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM	Total
Original	2	1	0	0	0	0	3
Exponential	3	5	1	5	1	0	15
Gradient	3	3	2	2	0	0	10
LBP-2D	5	2	1	1	1	0	10
LoG	4	0	1	0	0	0	5
Logarithm	3	3	1	1	1	1	10
Square	3	3	1	0	0	0	7
Square Root	0	2	0	0	1	0	3
Wavelet-HH	1	1	2	3	0	0	7
Wavelet-HL	1	1	3	2	0	0	7
Wavelet-LH	3	6	1	7	0	0	17
Wavelet-LL	1	5	0	0	0	0	6
Total	29	32	13	21	4	1	100

Table 4.15: The Jaccard index of the top 100 for all models of the SvAH dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.538	0.198	0.053	0.504	0.130	0.351
XGB	0.538	1.000	0.170	0.058	0.408	0.111	0.290
SGDC	0.198	0.170	1.000	0.105	0.205	0.242	0.190
RC	0.053	0.058	0.105	1.000	0.058	0.124	0.064
CB	0.504	0.408	0.205	0.058	1.000	0.111	0.493
LR	0.130	0.111	0.242	0.124	0.111	1.000	0.105
RF	0.351	0.290	0.190	0.064	0.493	0.105	1.000

Table 4.16: The Jaccard index of the top 100 for all models of the SvAH-SP dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.653	0.093	0.036	0.613	0.081	0.316
XGB	0.653	1.000	0.105	0.053	0.587	0.105	0.290
SGDC	0.093	0.105	1.000	0.290	0.105	0.333	0.117
RC	0.036	0.053	0.290	1.000	0.064	0.307	0.117
CB	0.613	0.587	0.105	0.064	1.000	0.081	0.307
LR	0.081	0.105	0.333	0.307	0.081	1.000	0.124
RF	0.316	0.290	0.117	0.117	0.307	0.124	1.000

The Jaccard index was also analyzed to see how feature selection behavior varied between models, datasets and classifications. Table 4.15 shows the similarity of the top 100 features, between models for tissue classification using the SvAH dataset. Table 4.16 displays the Jaccard index of the top 100 features between varying models of the SvAH-SP dataset. Table 4.17 displays the Jaccard index of the top 100 features between varying models of the SvCH dataset. Table 4.18 displays the Jaccard index of the top 100 features between varying models of the SvCH-SP dataset.

The Jaccard index for grade classification using the Grade and Grade-SP datasets is presented in tables 4.19 and 4.20, respectively. It is similarly shown for subgroup classification for the SG and SG-SP datasets in tables 4.21 and 4.22, respectively. Finally, the similarity behavior of the top 100 features pooled together for each model for each classification and dataset is displayed in table 4.23, comparing how different classification tasks and different mask post-processing affect feature selection behavior.

Table 4.17: The Jaccard index of the top 100 for all models of the SvCH dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.626	0.124	0.036	0.626	0.087	0.399
XGB	0.626	1.000	0.124	0.053	0.562	0.093	0.379
SGDC	0.124	0.124	1.000	0.183	0.124	0.575	0.099
RC	0.036	0.053	0.183	1.000	0.047	0.183	0.070
CB	0.626	0.562	0.124	0.047	1.000	0.099	0.460
LR	0.087	0.093	0.575	0.183	0.099	1.000	0.099
RF	0.399	0.379	0.099	0.070	0.460	0.099	1.000

Table 4.18: The Jaccard index of the top 100 for all models of the SvCH-SP dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.613	0.070	0.036	0.575	0.053	0.429
XGB	0.613	1.000	0.053	0.036	0.626	0.047	0.351
SGDC	0.070	0.053	1.000	0.290	0.053	0.460	0.075
RC	0.036	0.036	0.290	1.000	0.036	0.399	0.064
CB	0.575	0.626	0.053	0.036	1.000	0.047	0.379
LR	0.053	0.047	0.460	0.399	0.047	1.000	0.053
RF	0.429	0.351	0.075	0.064	0.379	0.053	1.000

Table 4.19: The Jaccard index of the top 100 for all models of the G dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.504	0.047	0.042	0.379	0.047	0.242
XGB	0.504	1.000	0.020	0.036	0.389	0.020	0.170
SGDC	0.047	0.020	1.000	0.250	0.042	0.439	0.058
RC	0.042	0.036	0.250	1.000	0.026	0.266	0.075
CB	0.379	0.389	0.042	0.026	1.000	0.047	0.242
LR	0.047	0.020	0.439	0.266	0.047	1.000	0.064
RF	0.242	0.170	0.058	0.075	0.242	0.064	1.000

Table 4.20: The Jaccard index of the top 100 for all models of the G-SP dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.600	0.053	0.058	0.429	0.058	0.351
XGB	0.600	1.000	0.047	0.047	0.429	0.036	0.299
SGDC	0.053	0.047	1.000	0.290	0.042	0.429	0.099
RC	0.058	0.047	0.290	1.000	0.026	0.449	0.087
CB	0.429	0.429	0.042	0.026	1.000	0.036	0.258
LR	0.058	0.036	0.429	0.449	0.036	1.000	0.087
RF	0.351	0.299	0.099	0.087	0.258	0.087	1.000

Table 4.21: The Jaccard index of the top 100 for all models of the SG dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.515	0.075	0.070	0.460	0.105	0.258
XGB	0.515	1.000	0.070	0.070	0.429	0.075	0.266
SGDC	0.075	0.070	1.000	0.235	0.099	0.493	0.081
RC	0.070	0.070	0.235	1.000	0.064	0.235	0.081
CB	0.460	0.429	0.099	0.064	1.000	0.105	0.351
LR	0.105	0.075	0.493	0.235	0.105	1.000	0.093
RF	0.258	0.266	0.081	0.081	0.351	0.093	1.000

Table 4.22: The Jaccard index of the top 100 for all models of the SG-SP dataset.

	LGBM	XGB	SGDC	RC	CB	LR	RF
LGBM	1.000	0.575	0.111	0.058	0.587	0.087	0.379
XGB	0.575	1.000	0.111	0.053	0.504	0.070	0.316
SGDC	0.111	0.111	1.000	0.266	0.111	0.399	0.149
RC	0.058	0.053	0.266	1.000	0.075	0.389	0.130
CB	0.587	0.504	0.111	0.075	1.000	0.081	0.408
LR	0.087	0.070	0.399	0.389	0.081	1.000	0.130
RF	0.379	0.316	0.149	0.130	0.408	0.130	1.000

Table 4.23: The Jaccard index from the top 100 features pooled from each model of a dataset compared amongst datasets.

Dataset	SvAH	SvAH-SP	SvCH	SvCH-SP	Grade	Grade-SP	SG	SG-SP
SvAH	1.000	0.272	0.273	0.239	0.350	0.277	0.310	0.260
SvAH-SP	0.272	1.000	0.312	0.460	0.277	0.414	0.275	0.391
SvCH	0.273	0.312	1.000	0.338	0.288	0.300	0.317	0.275
SvCH-SP	0.239	0.460	0.338	1.000	0.288	0.382	0.258	0.353
Grade	0.350	0.277	0.288	0.288	1.000	0.316	0.335	0.283
Grade-SP	0.277	0.414	0.300	0.382	0.316	1.000	0.265	0.431
SG	0.310	0.275	0.317	0.258	0.335	0.265	1.000	0.273
SG-SP	0.260	0.391	0.275	0.353	0.283	0.431	0.273	1.000

Table 4.24: Jaccard index when comparing top 10 ranking habitat features for all models when classifying sarcoma grade.

	LR	SGDC	RC	RF	CB	LGBM	XGB
LR	1.000	0.538	1.000	0.176	0.111	0.176	0.176
SGDC	0.538	1.000	0.538	0.250	0.176	0.176	0.176
RC	1.000	0.538	1.000	0.176	0.111	0.176	0.176
RF	0.176	0.250	0.176	1.000	0.538	0.538	0.538
CB	0.111	0.176	0.111	0.538	1.000	0.538	0.667
LGBM	0.176	0.176	0.176	0.538	0.538	1.000	0.667
XGB	0.176	0.176	0.176	0.538	0.667	0.667	1.000

4.4.3.2 Habitat Feature Space

Extracted from the 4x4 habitat matrices were a total of 104 statistical features. For grade classification, filtering out highly correlated features (≥ 0.99) and low variance (≤ 0.01) features resulted in only 28 remaining features, showing that most extracted variables are useless. 7 features remained from the unnormalized MSI matrix, and 21 remain from the normalized MSI matrix. Table 4.24 displays the Jaccard index of the top 10 ranking features for grade classification using habitats.

Similarly for sub-group classification using habitats, 29 out of the 104 features remain after correlation and variance filtering. The difference in feature count after filtering differs slightly from grade classification as the patient sets are slightly different. 21 features remain from the normalized MSI matrix, and 8 features from the raw MSI matrix. Table 4.25

Table 4.25: Jaccard index when comparing top 10 ranking habitat features for all models when classifying sarcoma sub-group.

	LR	SGDC	RC	RF	CB	LGBM	XGB
LR	1.000	0.667	0.818	0.176	0.176	0.176	0.111
SGDC	0.667	1.000	0.538	0.250	0.250	0.250	0.176
RC	0.818	0.538	1.000	0.250	0.176	0.111	0.111
RF	0.176	0.250	0.250	1.000	0.538	0.429	0.538
CB	0.176	0.250	0.176	0.538	1.000	0.667	0.818
LGBM	0.176	0.250	0.111	0.429	0.667	1.000	0.818
XGB	0.111	0.176	0.111	0.538	0.818	0.818	1.000

4.4.4 Model Training and Testing

The main results of model training and testing are presented in this section. More detailed tables and graphs of results and metrics are provided in Appendix C. The results are simply stated and discussed in more detail in the discussion, section 4.5.6.

4.4.4.1 Tissue Classification

The resulting AUC-ROC score for each run of the tissue classification for the 4 slightly different processing pipelines and the ROC curve of the best performing model are displayed in figures 4.17 to 4.23. Further results on model metrics can be found in the figures and tables of Appendix C. Figure 4.17 shows the AUC-ROC score for tissue classification using the SvAH dataset, results are extremely promising. Although it should be noted that this classification is quite straight forward as healthy anatomy is orders of magnitude larger than pathological, and sizes can be indirectly inferred through GLCM features and LBP transform. This is mitigated when processing with SPs, and noticeable drops in AUC-ROC score can be seen in figure 4.19, where the AUC-ROC score is presented for the SvAH-SP dataset. The SvCH dataset's AUC-ROC score can be seen in figure 4.21 where the models use comparable regions as healthy samples instead of the entire anatomy. The results of the AUC-ROC score of the SvCH dataset show that classifying sarcoma against similar or comparable healthy regions provide very good results, and in this case differences in input mask sizes is not an issue, as the healthy and pathological masks are nearly identical, only in different regions. Similar performance drops are noticeable for the SvCH-SP dataset when viewing the AUC-ROC scores in figure 4.23.

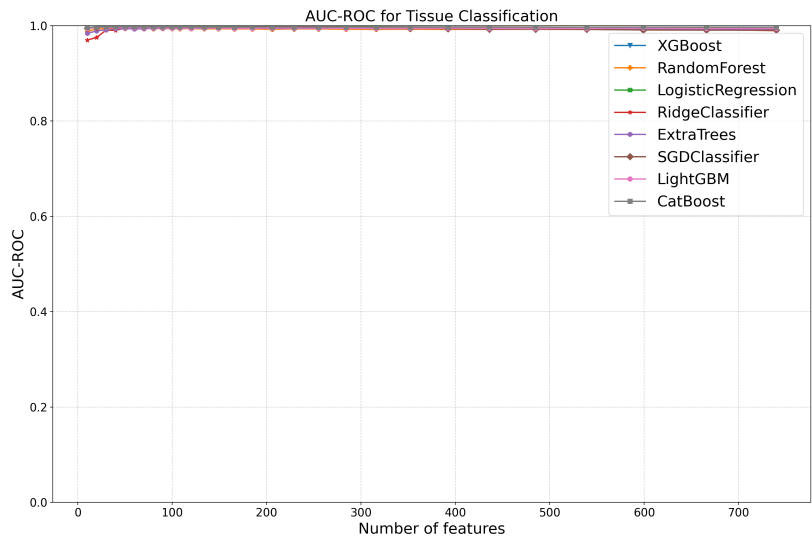


Figure 4.17: AUC-ROC score against the number of feature applied for each model using the SvAH dataset.

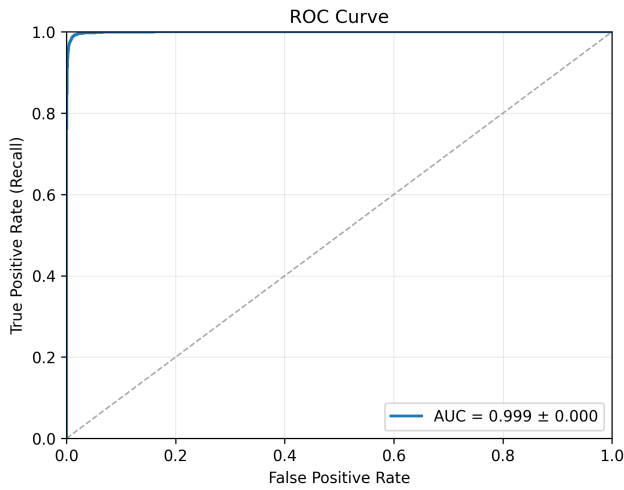


Figure 4.18: ROC curve of the highest scoring run: logistic regression at 100 features for tissue classification using the SvAH dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

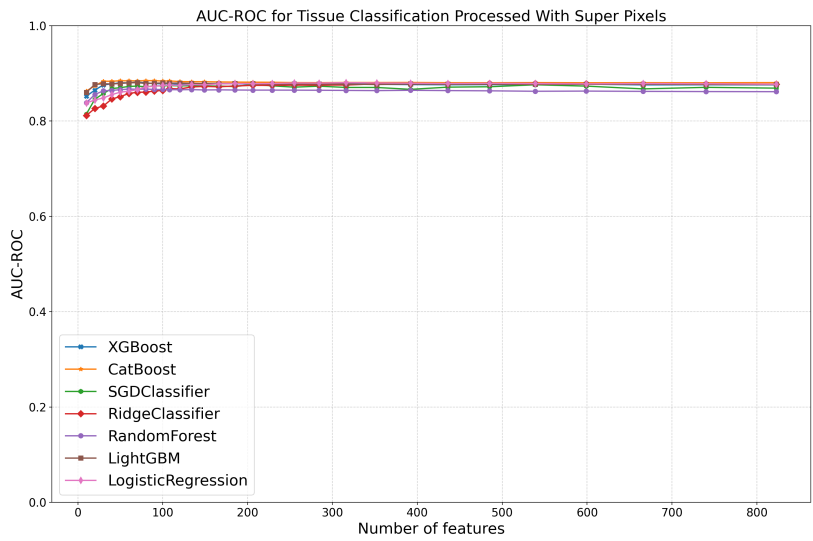


Figure 4.19: AUC-ROC score against the number of feature applied for each model using the SvAH-SP dataset.

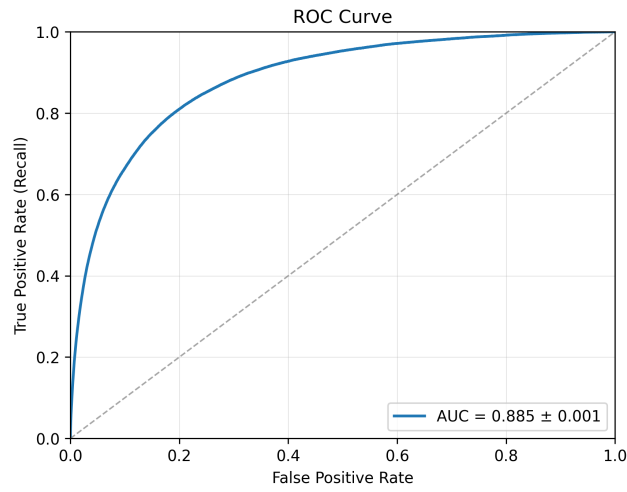


Figure 4.20: ROC curve of the highest scoring run: CatBoost at 90 features for tissue classification using the SvAH-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

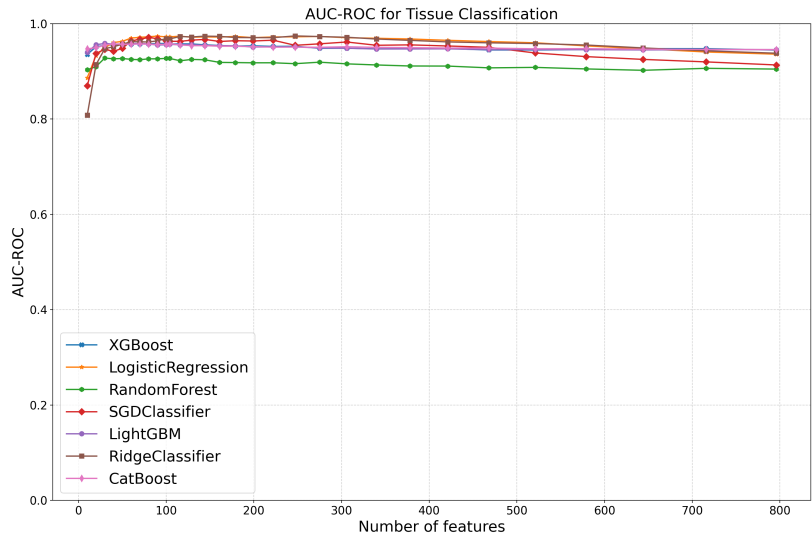


Figure 4.21: AUC-ROC score against the number of feature applied for each model using the SvCH dataset.

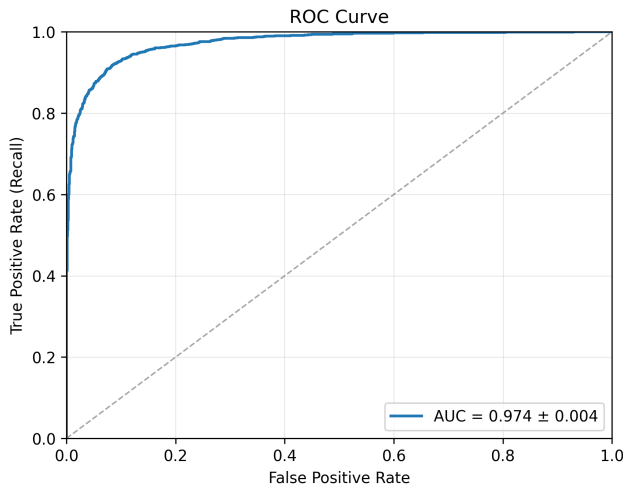


Figure 4.22: ROC curve of the highest scoring run: logistic regression at 90 features for tissue classification using the SvCH dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

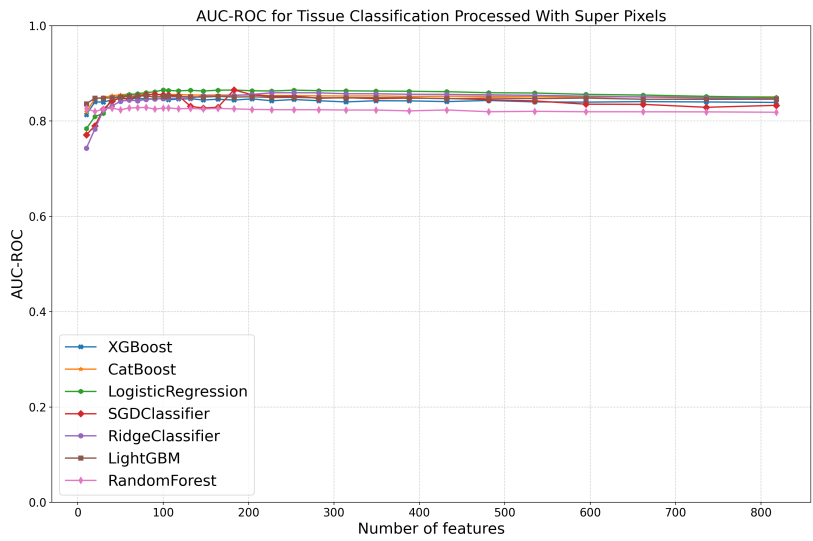


Figure 4.23: AUC-ROC score against the number of feature applied for each model using the SvCH-SP dataset.

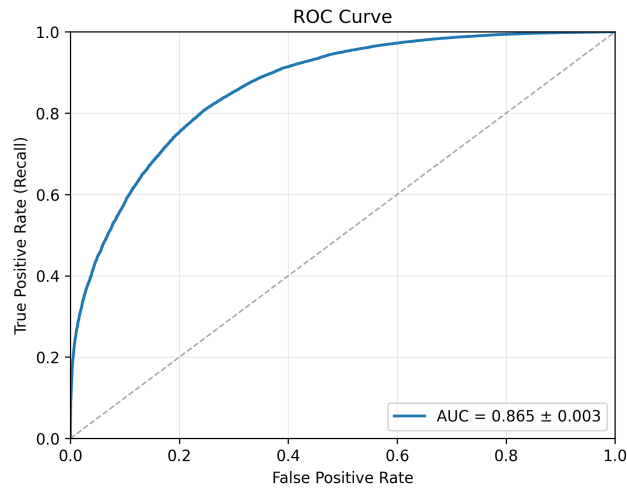


Figure 4.24: ROC curve of the highest scoring run: logistic regression at 100 features for tissue classification using the SvCH-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

4.4.4.2 Grade Classification

The resulting AUC-ROC score for each run of grade classification, processed with and without SPs are presented in figures 4.25 and 4.27. The ROC curve for the best runs of the two tasks can be found in figures 4.26 and 4.28. Further results of metrics can be found in Appendix C. Figure 4.25 shows the results for grade classification using whole masks, and performance is noticeably better than when the classifying use SPs (figure 4.27), although by a small margin. The AUC-ROC results show some extent of discriminative capability between the 3 grades, although it is not high enough to use reliably in a clinical setting.

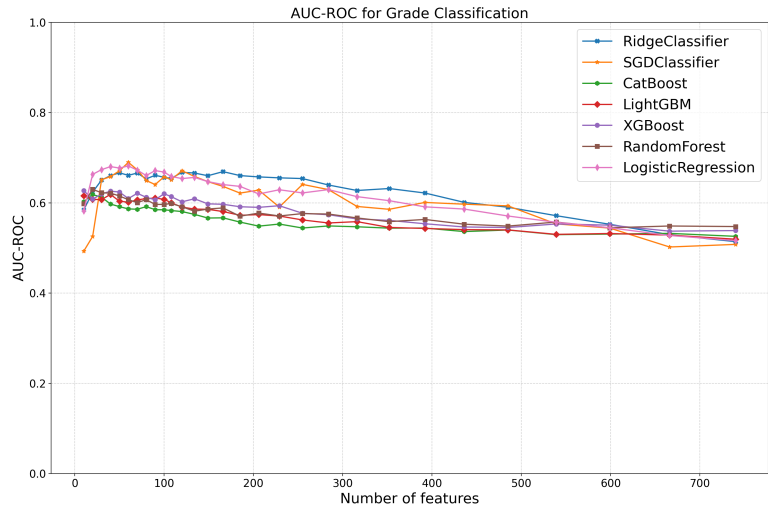


Figure 4.25: AUC-ROC score against the number of feature applied for each model for grade classification.

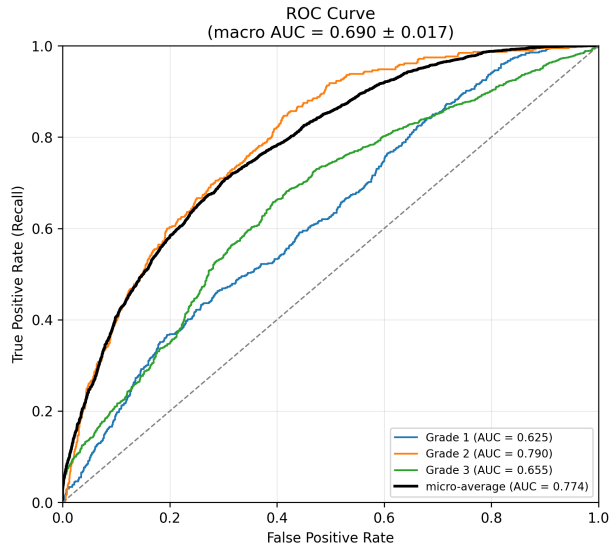


Figure 4.26: ROC curve of the highest scoring run: SGD Classifier at 60 features for grade classification using the grade dataset. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title.

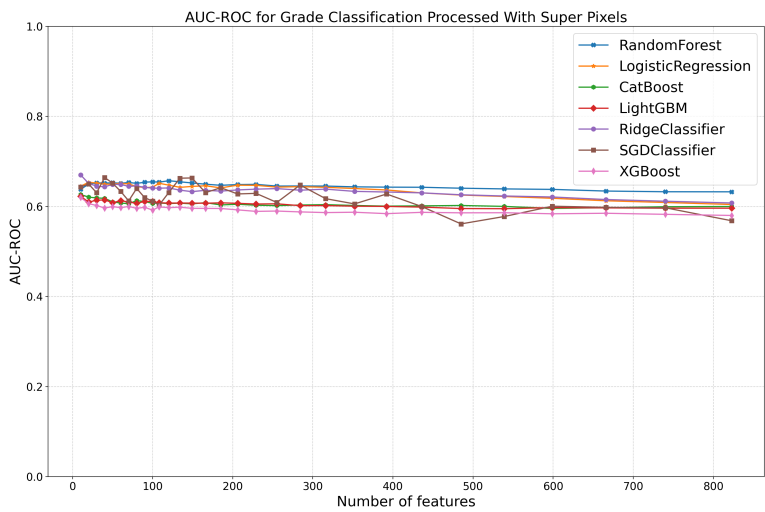


Figure 4.27: AUC-ROC score against the number of feature applied for each model for grade classification using SPs.

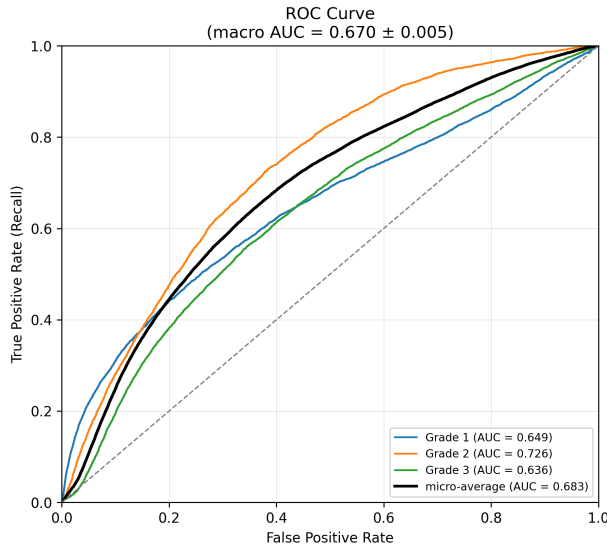


Figure 4.28: ROC curve of the highest scoring run: ridge classifier at 10 features for grade classification using the grade-SP dataset. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title.

4.4.4.3 Sub-Group Classification

The resulting AUC-ROC score for each run of sub-group classification, processed with and without SPs is displayed in figures 4.29 and 4.31. The ROC graphs of the best runs for sub-group classification processed without and with SPs can be found in figures 4.30 and 4.32, respectively. Further metrics can be found in Appendix C. Figure 4.29 shows that classifying sub-group using full masks has relatively high discriminative abilities, although there is much variance between the models, possibly due to the ratio of number of features to samples, which is not inherent when classifying using SPs, figure 4.31. There is however a drop in discriminative ability when processing using SPs, which is to be expected as there is less information contained within each sample. The stability between models can be partially explained by the much higher ratio of samples to number of features compared to sub-group classification using only masks.

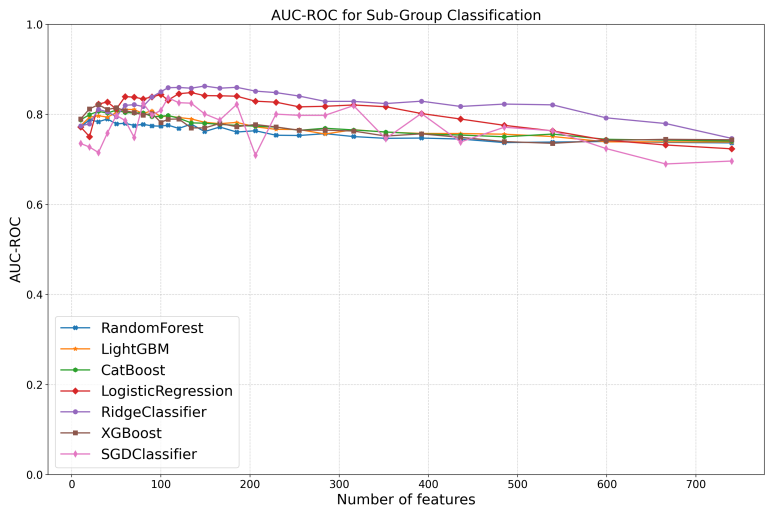


Figure 4.29: AUC-ROC score against the number of feature applied for each model for sub-group classification.

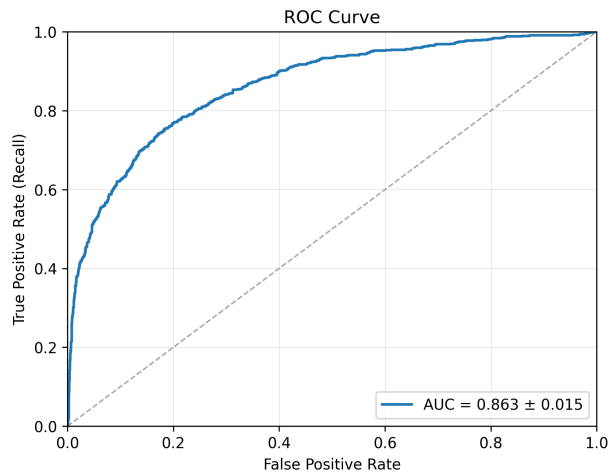


Figure 4.30: ROC curve of the highest scoring run: ridge classifier at 149 features for sub-group classification using the SG dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

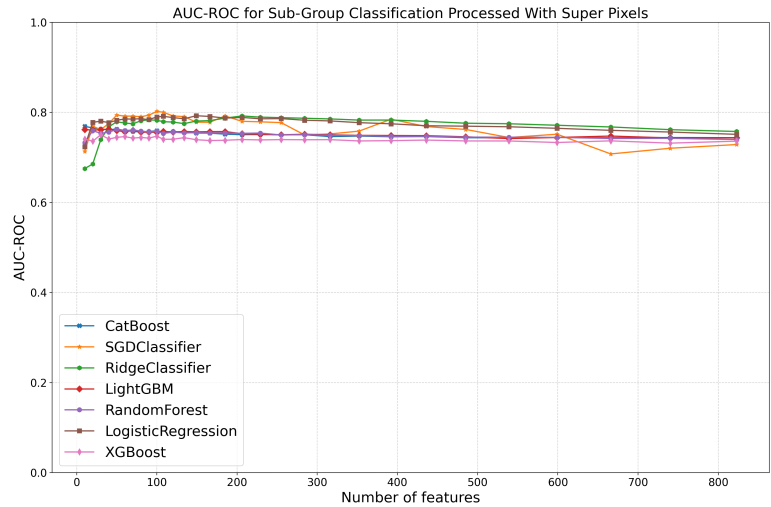


Figure 4.31: AUC-ROC score against the number of feature applied for each model for sub-group classification using SPs.

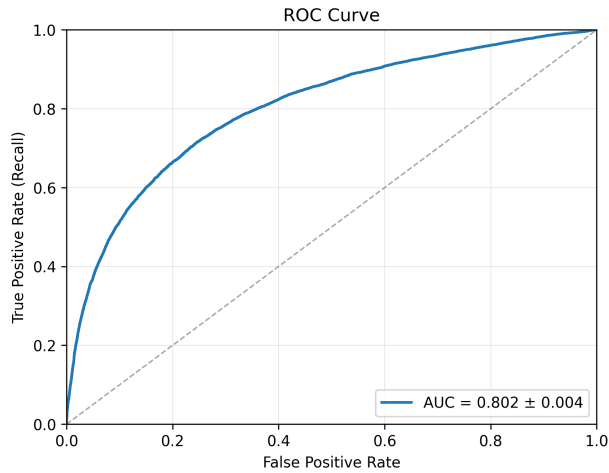


Figure 4.32: ROC curve of the highest scoring run: SGD classifier at 100 features for sub-group classification using the SG-SP dataset. The 95% confidence interval of the AUC-ROC score is shown in the legend.

4.4.4.4 Habitat Classification

The resulting AUC-ROC score for each run of grade and sub-group classification using habitats can be found in figures 4.33 and 4.35 respectively. Further metrics can be found in Appendix C. Grade classification using habitats had slightly worse AUC-ROC than when processing using radiomic feature data, although there were some instances of relatively good AUC-ROC scores with few features, see XGBoost at 4 features and LightGBM at 5 features in figure 4.33. This shows that few features can have large potential in sarcoma grade staging. Sub-group classification using the habitats was quite erratic between the models, see figure 4.35, but do show some discriminative ability between the two classes, soft tissue sarcoma (STS) and bone sarcoma (BS). This shows that there is some potential in using heterogeneity measures to classify grade and sub-group.

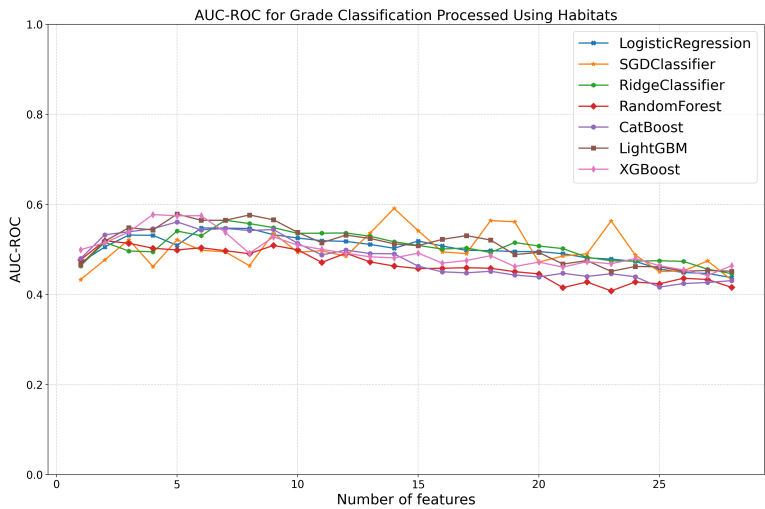


Figure 4.33: AUC-ROC score against the number of feature applied for each model for grade classification using habitats.

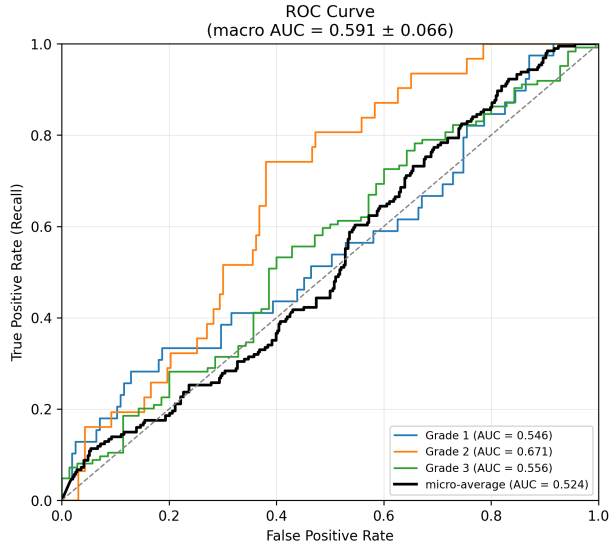


Figure 4.34: ROC curve of the highest scoring run: SGD classifier at 14 features for grade classification using habitats. The 95% confidence interval of the macro averaged AUC-ROC score is shown in the title.

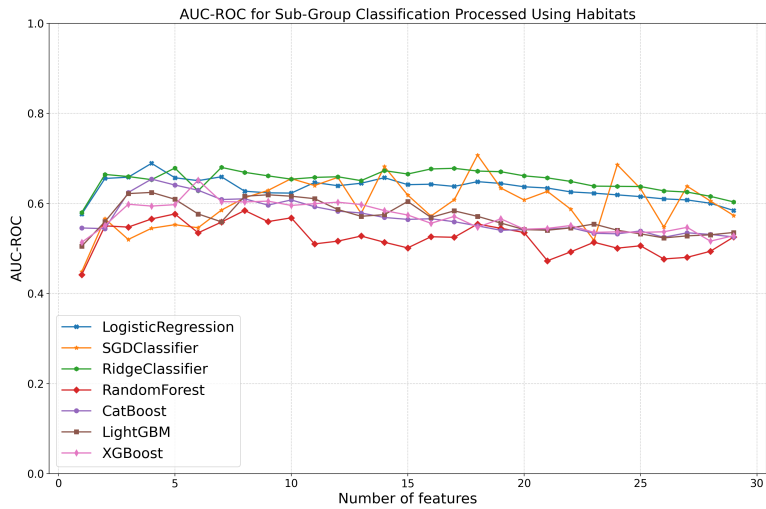


Figure 4.35: AUC-ROC score against the number of feature applied for each model for sub-group classification using habitats.

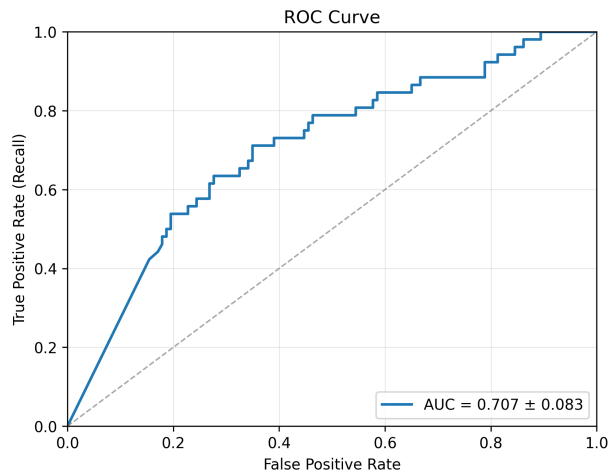


Figure 4.36: ROC curve of the highest scoring run: SGD classifier at 18 features for sub-group classification using habitats. The 95% confidence interval of the AUC-ROC score is shown in the legend.

4.5 Discussion

4.5.1 DICOM Image Database and Segmentation

The database used in the classification is a very interesting one and should be discussed. Generally, anatomy is considered to be quite complex and heterogeneous in nature, making classification tasks more difficult. On top of that, sarcomas are considered to be a heterogeneous, diverse group of malignancies with multiple sub-types and presentations [46]. Given this, it should be emphasized that the patient cohort may not provide enough instances of sarcoma to allow for effective generalization of sarcoma tissue, or for model performance to extend reliably beyond the results presented here, especially when MRI protocols are not standardized throughout all scans. Differences in MR weighting, contrast, and fat suppression protocols can have a profound effect on the appearance of the region of interest [151]. As seen in figure 4.37, even scans of the same anatomy can look very different depending on the imaging protocol. Scans of all orientations: axial, coronal, and sagittal, were included which increases both the variance and the complexity that the learning model must overcome. Tumors sometimes appear homogeneous and at other times heterogeneous, and varying the weighting and use of contrast media affects whether the tumor appears brighter or darker in the images, this can be seen in figures 4.37 and 4.38. MRI protocols can also influence whether surrounding edema is visible or not [151]. One way to mitigate these effects could be to include protocol-specific feature columns in the processed data table, which might help the model generalize better to the different appearances of sarcoma tumors. Another important consideration is that gray-value ranges differ drastically between patients, and standardization of intensity values to a representative scale depends on multiple metadata factors and machine specifications [152]. This, in turn, adds yet another level of complexity for learning models tasked with classification.

Another important limitation is the uneven representation of sarcoma sub-types in the dataset. Subtle differences exist between different sarcoma types, but when only a handful of cases are available for certain sub-types, these nuances are easily lost. For example, there is only a single case of dermatofibrosarcoma protuberans (DFSP) in the dataset. With no comparable examples to learn from, it is not realistic to expect a model to accurately grade that one DFSP case. In contrast, the majority of cases are soft tissue sarcomas (STS), meaning the models primarily learn to distinguish and grade STS. When asked to classify or grade an underrepresented type such as DFSP, the model inevitably tries to interpret it through the lens of STS, which raises questions about whether it is even possible to meaningfully grade these rare cases with such a limited dataset.

During the segmentation phase, all non-pathological tissue was grouped together as healthy tissue. However, biological tissue in the human body is highly

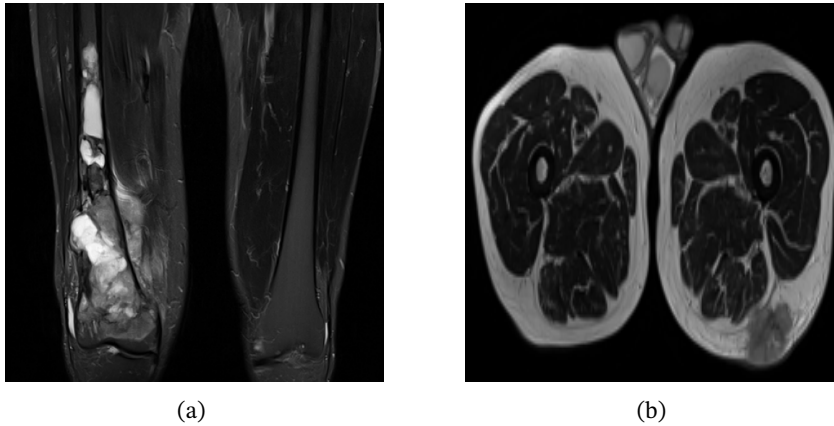


Figure 4.37: MRI slices: (a) shows a case of OS where the machine uses specific settings and contrast media to brighten the tumor region and suppress fat response. (b) displays another patient, liposarcoma in this case, where the tumor region is dark in contrast and fat tissue is not suppressed.

diverse in both appearance and behavior, which can pose difficulties when certain healthy tissues resemble pathological regions. In general, training a model on a greater number of tissue classes can improve classification of the pathological region, as it reduces the likelihood of false positives. By grouping all non-pathological tissue into a single class, the distribution of feature values in the “healthy” space can become so broad that it overlaps with the pathological feature space, ultimately making sarcoma classification more difficult. This effect is exemplified in figure 4.38.

The sarcoma and edema regions were fully segmented in 3D segmentation software. As a result, in some slices, small remnants of pathological tissue remain visible, often appearing as very small and homogeneous regions. Training a model on such small areas can be problematic, as they closely resemble healthy fat tissue and, without broader spatial context, may easily lead to misclassification.

Image quality was not standardized across patients, as shown in figure 4.39. Some MR scans appear grainy, slightly out of focus, or with variable pixel density. Such inconsistencies increase the variance of the dataset, but without necessarily improving generalization. While larger and more diverse datasets often benefit from higher variance by making models more robust [150], [153], in a small dataset this additional variance can instead harm performance, as the model struggles to distinguish noise from meaningful variation. Compounding this issue, many images were corrupted (although not a significant amount), with entire regions of slices rendered unusable. This is particularly detrimental if the corrupted areas

overlap with the pathological region of interest, as crucial diagnostic information is lost.

4.5.2 Image And Mask Post-Processing

The post-processing of images and masks require a multitude of parameters, where the range of viable values can be large. For example when deciding the size of SPs, one might speculate which size is the best, small SPs might lack enough pixel intensity variation to be meaningful, and too large results in spatial averaging which

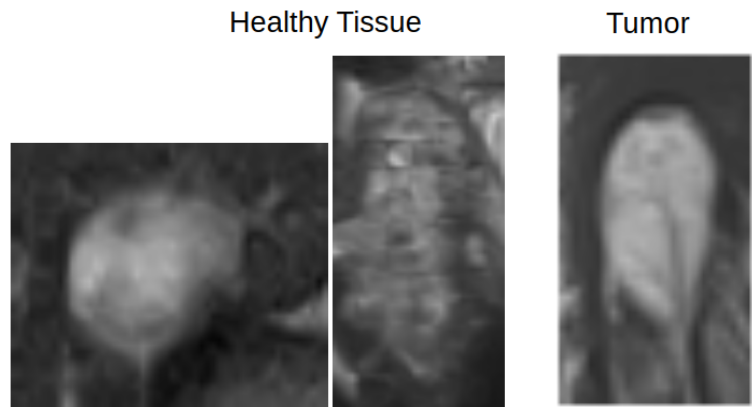


Figure 4.38: An example of how healthy and pathological tissue can overlap in terms of appearance. Left two images show healthy tissue, right most image shows a tumor region.

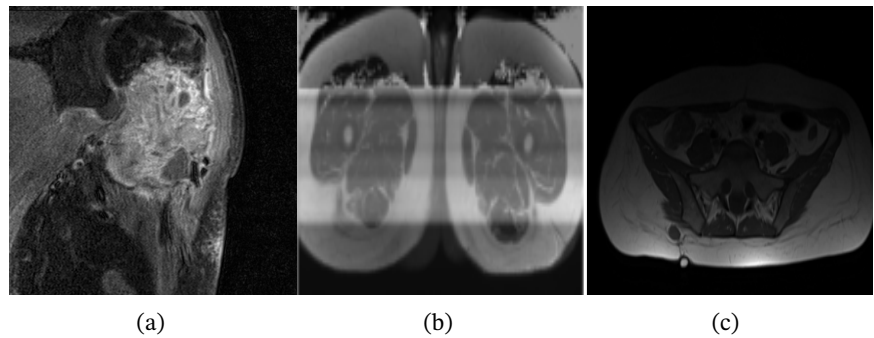


Figure 4.39: Variance in image quality: (a) Shows a grainy slice, with relatively high noise (b) Displays a corrupted image (c) An example of uneven lighting distribution.

dampen fine grained details. Also, due to the drawbacks that the IceSG database presents, normalization could be heavily affected by noise, image artefacts and corruption, even though measures such as normalizing only between the 1st to 99th percentiles were taken to dampen these effects. Nevertheless, decisions had to be taken on values on all parameters, and slightly different values will have a trickle-down effect all the way down to classification results. Proving that the parameter values chosen are the most optimal ones is a very difficult process, as it cannot be confirmed until results are acquired, which is unfortunately not feasible in the presented pipelines due to computational restraints.

Although, finding and testing the optimal values for parameters is next to impossible, different values were tested such as the size of SPs and the value chosen was deemed a good trade off of speed, size and overall results. The compactness variable of the SLIC function was also tested slightly, a lower compactness gives image gradients a larger weight when deciding borders, which seems perfect for heterogeneous tumors such as sarcoma, but the drawback is that the SPs become extremely irregular, causing more problems further down the pipeline, most notably was that the differences in sizes had much more variance, making SPs less comparable. This could partly be tackled by increasing the parameter for the maximum number of iterations of the SLIC algorithm, but it is quickly seen how only 3 parameters can have a large affect on the overall results, as multiple combinations of parameters would require testing.

The SLIC clustering method is not perfect, and iterations have been made to improve the algorithm such as the differentiable SLIC algorithm [154] which solves the issue of the non-differentiability of the original SLIC method. This allows for easier integration with classical neural network models. These newer SLIC algorithms have shown improved performance [154], [155], and could increase performance if integrated into the pipeline. It should be mentioned that these studies base SLIC performance on its segmentation abilities, not on classification of specific SPs.

Therefore, the parameter values chosen can be improved, probably with diminishing returns, but the results of model performance was adequate enough to be able to draw some conclusions.

4.5.3 Habitat Generation

Habitat analysis is a relatively new method, with the earliest found application emerging in a 2010 article [156], and not many articles exist that detail the workflow. There is much room for optimizations in this component, such as which features create the best habitat clusters, that best represent heterogeneity of sarcoma in the current database. Generally, in the literature, authors used standardized MR protocols, since pixel intensities of different anatomy vary under different MRI protocols [109], [112]. Also, different features aid in discriminating sub-groups

and grades, meaning that a feature set used for generating habitats for grade classification may not perform as well for sub-group classification. This research used the same features and generated habitats for classifying sarcoma grade and sub-groups.

It could be possible to follow the grade classification pipeline to obtain the feature sets that best correlate with sarcoma grade, and use these features to cluster SPs based on similarity in the K-means clustering step. This process by itself is also time consuming, as it requires much testing to find the best clustering features to predict grade or sub-group.

4.5.4 Feature Extraction

The radiomic feature extraction was the most straight forward as it was mainly handled by PyRadiomics, which has been tested and approved by the community. The thought process behind the radiomics extractions was to simply extract everything possible, even though many extracted values were irrelevant. It was decided to go into the feature extraction in an unbiased manner, even though many clues exist in the literature of which features might perform the best. It needs to be confirmed whether the same features are applicable in the IceSG database. It was rather decided to trust the feature selection for filtering out unnecessary features. This has its drawbacks however, Jiarpakdee et al. [157] tested how metrics were selected at different correlation and variance filtering thresholds, and found, using their feature ranking algorithm (not SHAP), that the best feature sets varied with different thresholds for feature filtering. Therefore, different thresholds for correlation and variance filtering should be tested to see which gives the best results.

Habitat feature extractions were also straight forward, and followed the same philosophy as the radiomic feature extraction: extract everything possible and let the feature selection filter out redundant features. This inturn ensures a less biased feature set, ultimately used for classification tasks.

4.5.5 Feature Selection

Feature selection is one of the most important steps in the workflow, as it has the responsibility for creating the best possible model-dependent feature subsets and removing redundant features. Out of the total 1116 radiomic features extracted, after the correlation and variance filtering, generally, around 300 features were immediately removed. The best performing models, tables 4.7 to 4.14, also showed that the best models performed on 60 to 183 features, even less in some cases. Although, many models did perform well on few features such as most models with only 10 features in tissue classification for all datasets.

What was extremely interesting, were the results of the Jaccard Index of the top 100 performing features pooled across models for datasets. When comparing

the pooled feature space of SvCH and SvAH, the Jaccard index was only 0.273, meaning that the two datasets only share 27.3% of their features. This is to be somewhat expected as the SvAH dataset contains much more anatomy with much more variance in radiomic feature data, explaining to some degree the low similarity of chosen features. Although, this did increase to 46% identical features when processed with super pixels. The Jaccard index does not take into account the correlations of features, meaning that although the features might have been different, correlation between selected features might be high or quantify similar information.

It was also interesting to look at how features were selected between models in the same dataset. For tissue classification, and the top 100 features, of the SvAH dataset no features were selected by all models, 11 features were selected by 6 out of the 7 models, 13 by 5, 29 by 4, 35 by 3, 76 by 2 and 182 selected by a single model. This trend is quite similar for all datasets and classification tasks, where few features are chosen by many models and most features are shared by 1-3 models. This trend shows the complexity of how models apply their chosen features to predict the classes. Also, when viewing the selected features of the top 20 features between models, there was an even larger gap between the chosen features, for all datasets and classifications, no single feature was ever shared by all models and most features chosen were unique to a single model. Meaning there was no convergence to an optimal feature set, but model results at 20 features were very good in many cases, showing that models utilize different details and information to create associations to the predicted class. This has powerful implications. The fact that models utilize different features and arrive at similar results, means that there is a potential for decision pooling where different models process information and cast a vote as to their predictions, another abstract form of saying ensemble methods can be applied to pool together varying results.

4.5.6 Model Training and Testing

Overall discussion of the results of the model training and testing step are presented in this section. Model performance in specific cases is displayed and discussed in section 4.5.7, highlighting where performance was good and where it wasn't.

4.5.6.1 Tissue Classification

For the SvAH dataset, the best results were obtained with the SGD Classifier at 50 features, achieving an AUC-ROC of 0.999 with similarly high accuracy. Even with as few as 20 features, XGBoost reached a score of 0.998, which demonstrates the overall strength of this classification task. This outcome is not unexpected, as size information can be indirectly captured through LBP and GLCM features, even

though explicit shape features were not included. Metrics shown in figures C.1 to C.7 further confirm that performance across all models was consistently very high, both in macro and micro measures.

When moving to the SvAH-SP dataset, performance dropped compared to SvAH, although scores remained strong. The maximum AUC-ROC values across models ranged from 0.867 to 0.885, with CatBoost being the top performer, at 90 features. Interestingly, CatBoost also achieved a score of 0.863 with only 10 features, suggesting that in some cases it may be more efficient to use fewer features since inference requires fewer computational resources. As seen in figures C.8 to C.15, accuracy and balanced accuracy remained fairly stable across models and feature counts, hovering around 80%. Precision, however, likely suffered due to class imbalance, as more healthy samples were misclassified as tumor than the number of tumor samples in the database. Nonetheless, the ratio of correctly to incorrectly classified samples was favorable, and overall performance remained very good. These results highlight that it is possible to classify even very small segments of masks as healthy or pathological, which opens the possibility of scanning through MRI volumes to quickly locate regions most likely to contain tumors.

For the SvCH dataset, results were slightly worse than those of SvAH. This is partly because size is no longer implicitly captured, since the masks for the two classes are of equal size. The dataset is also somewhat biased, as masks were manually translated to a healthy region, which limits realism. Despite these limitations, all metrics pointed to strong generalization and robustness, with overall accuracy and balanced accuracy consistently above 80%, and some models exceeding 90%. AUC-ROC scores, table C.3, are extremely promising, the maximum score for each model ranged from 0.928 to 0.974, with CatBoost scoring 0.947 with only 10 features. These findings suggest that discriminating healthy tissue from pathological tissue is achievable with acceptable accuracy. Metrics are visualized in figures C.16 to C.23.

Finally, the SvCH-SP dataset showed performance patterns similar to those observed when comparing SvAH and SvAH-SP. Here, precision did not decline as much, likely because this dataset is balanced with equal numbers of samples from each class. Performance was generally stable across models and feature counts, with differences usually within 10% across all metrics. Accuracy and balanced accuracy were slightly lower compared to the SvAH-SP dataset, but this may be explained by the fact that SvAH-SP contains more patients, and thus more training data, which aids generalization. Metrics are visualized in figures C.24 to C.31.

Taken together, these results demonstrate that sarcoma can be distinguished from healthy tissue at multiple scales, ranging from very small segments many times smaller than the tumor itself to full tumor masks spanning an entire image slice.

4.5.6.2 Grade Classification

For the Grade dataset, the highest AUC-ROC score was achieved by the SGDClassifier at 0.690 using 60 features. These scores are not nearly as high as those obtained for tissue classification, this is not surprising given the greater difficulty of the task. The amount of training data is much smaller: tissue classification included at least twice as many samples across all cases. Class imbalance further complicates learning, with only 19 grade 1 cases and 16 grade 2 cases compared to 66 grade 3 cases. With so few examples of grades 1 and 2, the models may not be able to capture or explain the variance between classes. The dataset structure also adds difficulty, as data are split slice-wise; at the edges of the tumor volume, slices may contain little or no useful information for grading. Importantly, FNCLCC grading is strongly dependent on biopsy results, which are not included here and cannot be directly inferred from image transforms or radiomic features. Nevertheless, there is some evidence of correlation: model accuracy and balanced accuracy consistently exceeded chance levels, as shown in figures C.32 and C.33. Macro metrics (figures C.34 to C.36) also show that F1-score, precision, and recall were often above the baseline of 0.33. Still, the overall performance is not sufficient to draw strong conclusions beyond noting that extracted radiomic features do carry some information related to FNCLCC grade that can be partially explained and used by different models for classification.

For the Grade-SP dataset, the highest AUC-ROC score was again achieved by the SGDClassifier, reaching 0.670 at 10 features. While this score is modest, it suggests that models can still explain some variance between grades using radiomic features alone. Interestingly, most models performed best at 10 features. Performance varied considerably across models, as shown in figures C.40 to C.47, with accuracy differences of up to 20% between the best and worst performing approaches. The highest balanced accuracy was obtained by RidgeClassifier at 10 features, reaching 51.5%, though its overall accuracy was similar to other models. Conversely, LightGBM reached the best overall accuracy of 70.5% with 10 features, but its balanced accuracy dropped to 37.6%, suggesting a clear bias toward the majority class.

Overall grade classification has proven to be very difficult given the input data and the task. The extracted radiomics features from MRI scans of sarcoma tumors are not sufficient for diagnosing grade reliably. There are possible explanations for this. FNCLCC grade is heavily dependent on biopsy results, that information is not inferred or deduced from image transform and radiomic features, using the described pipeline of this research. The diagnosed grade is not necessarily 100% accurate, as has been discussed [35]–[37], grading is very subjective and prone to error, errors in grading would propagate down the pipeline and could be detrimental to classification metrics.

The studies discussed in section 4.1.4.2 found better correlation with severity

and grade than these results. Possible explanations include the general use of standardized imaging protocols, relatively small patients cohorts, and fact that most research focuses on a single sarcoma type. As different sarcoma types have different radiomic signatures, as is evident from the sub-group classifications, constraining the pipelines to specific types could result in different optimal feature sets to performing grading. The potential benefits are that trained models would not need to generalize and explain variance of all sarcoma, but could focus on explaining differences between phenotypically similar sarcoma. The literature shows that often times fewer than 10 features were needed to accurately and robustly discriminate patient grade, not evident in the presented results, as often times model metrics suffered at 10 features.

4.5.6.3 Sup-Group Classification

For sub-group classification using whole masks (SG dataset), performance was notably stronger than for grade classification, though still below the very high levels seen in tissue classification. Across models, AUC-ROC scores ranged between 0.790 and 0.863 depending on the number of features (table C.7). Logistic Regression and Ridge Classifier achieved the top results, with logistic regression peaking at 0.848 with 134 features and Ridge Classifier reaching the highest score overall at 0.863 with 149 features. Other models, including XGBoost and CatBoost, consistently performed above 0.80, while Random Forest generally trailed slightly lower.

Accuracy results (figure C.48) were stable across models, typically around 70-80%. Balanced accuracy (figure C.49) highlighted class imbalance effects, with Ridge Classifier maintaining values near or above 75% while most other models hovered closer to 70%. Macro-averaged F1-scores and precision/recall values (figures C.50-C.52) support this pattern, showing that models differed in how well they captured minority versus majority sub-groups.

For sub-group classification using masks processed with super pixels (SG-SP dataset), performance was slightly lower but broadly consistent across models. The best AUC-ROC values for each model (table C.8) ranged from 0.750 to 0.802, with the strongest results coming from SGD classifier (0.802 at 100 features), Ridge Classifier (0.792 at 206 features), and logistic regression (peaking at 0.793 at 149 features). Accuracy (figure C.56) and balanced accuracy (figure C.57) both stabilized around 70-72%, with less variation between models than in the SG dataset. Macro-averaged metrics (figures C.58-C.60) showed that the models handled class imbalance relatively evenly, though without the higher peaks observed for Ridge Classifier in the whole-mask setting. As in the SG dataset, several models performed well, even with relatively few features, reinforcing the idea that subgroup classification depends more on a small set of discriminative radiomic features than on the full feature space.

Overall, these results suggest that sub-group classification is feasible but remains challenging. Whole-mask classification using the SG dataset achieved stronger performance, with AUC-ROC values peaking above 0.85 and accuracy levels near 80%. The approach using the SG-SP dataset produced slightly lower results, with AUC-ROC values mostly below 0.80 and overall accuracy closer to 70-72%. While both approaches demonstrate that radiomic features contain meaningful information for distinguishing sarcoma sub-groups, the difference highlights a trade-off: whole-mask features capture more global characteristics useful for classification, whereas super-pixel features offer finer granularity but may dilute the global signals needed for robust subgroup prediction.

4.5.6.4 Habitat Classification

Grade predictions using habitat features performed slightly worse overall compared to the standard radiomics-based approaches. The highest AUC-ROC score was achieved by SGD classifier with 14 features at 0.591 (table C.9). LightGBM also performed competitively, peaking at 0.579 with 5 features. Although these scores remain modest, they consistently exceeded random baseline, indicating that habitat-derived features do capture some signal related to tumor grade.

The evaluation metrics in figures C.66-C.68 highlight considerable variation between models. XGBoost, in particular, often achieved stronger micro-averaged performance at low feature counts, suggesting that it was able to capture more of the dominant class signal compared to other models. LightGBM, random forest and CatBoost also showed relatively stable micro F1-scores and recall values across different feature subsets.

The large spread in performance across models and feature counts points to instability in predictions and emphasizes the difficulty of the task. Unlike tissue classification or subgroup prediction, grade is inherently tied to histological properties (e.g., mitotic index, necrosis, cellularity) that are not directly captured by imaging features alone. Still, the fact that several models consistently performed above chance demonstrates that habitat-level information encodes biologically relevant signals related to grade, even if insufficient for clinically reliable predictions in the present dataset.

Sub-group classification showed somewhat better results using habitats compared to grade classification. The best performing model was the SGD Classifier using 18 features, achieving an AUC-ROC score of 0.707. Other models performed comparatively at few features, random forest displayed the worst performance. Reviewing the metrics in Figures C.72 to C.79, it is clear that XGBoost and LightGBM generally outperformed the other models across most metrics, particularly in micro-averaged F1-score and precision, where they consistently remained above baseline. Balanced accuracy peaked at 64.0%, which is still relatively low for clinical applicability and lower than radiomic sub-group classification results. These

results highlight that habitats capture meaningful patterns relevant to sub-group differentiation, though the signal remains weak.

Overall, habitats do show potential for application in sarcoma diagnostics, but their effectiveness appears sensitive to the choice of features and dataset characteristics. In this work, the features used to generate habitats were selected based on prior literature, which often builds on datasets collected under standardized imaging protocols and single sarcoma type. These conditions can emphasize certain radiomic properties that may not generalize directly to the IceSG database, where acquisition protocols and patient variability introduce additional variance. As a result, the chosen features may not represent the optimal set for constructing habitats in this context. Incorporating feature sets identified through systematic feature selection analyses could therefore improve habitat grade classification performance. Similarly, sub-group habitats were derived using the same feature set as for grade classification, even though the underlying biological distinctions between sub-groups may rely on different imaging characteristics. A dedicated feature selection workflow tailored specifically to sub-group prediction could yield more informative habitats and lead to stronger model performance.

4.5.6.5 Final Take

There are several limitations inherent to the IceSG database that likely contributed to the challenges observed in classification. The most pressing is the small sample size, which results in poor representation of FNCLCC grades 1 and 2. Grade 3 tumors are also generally larger, meaning that on average they contribute more slices and super pixels per patient. This exacerbates the imbalance, as the models are exposed to disproportionately more grade 3 data compared to the lower grades. A similar issue is present in sub-group classification, where there were 65 soft tissue sarcoma (STS) cases but only 27 bone sarcoma (BS) cases, again biasing the dataset toward one class.

Another factor is the lack of standardized imaging protocols. Variability in acquisition introduces differences in texture, intensity, and resolution that are unrelated to grade or subtype, and have been shown to create issues in comparing radiomics feature values [109], [112]. This additional variance reduces the ability of classifiers to generalize. Image corruption, scanner noise, motion artefacts, and other technical drawbacks further degrade robustness by obscuring or distorting relevant tumor information.

It must also be noted that sarcoma behavior and severity are heavily influenced by gene expression, which is not captured by radiomic features alone. For example, one patient in the IceSG dataset was initially diagnosed as FNCLCC grade 1 but experienced multiple relapses; the grade was later manually revised to grade 3 to reflect its clinical severity. This case highlights both the imperfection of the FNCLCC system and the risk of label noise in small datasets, where just a few misclas-

sified patients can strongly impact model training and evaluation. Incorporating genetic data could address this issue, as several studies have demonstrated associations between gene signatures and sarcoma aggressiveness [43]–[45]. Similarly, image-derived features from biopsy samples could complement radiomic features by providing the histological information that FNCLCC explicitly relies on.

Given these limitations, it may be worth questioning whether FNCLCC grade itself is the optimal prediction target. One alternative could be the development of a severity index derived directly from clinical outcomes, such as overall survival, recurrence-free survival, and cause-specific mortality. Such an index could serve as a more robust endpoint for training predictive models and may yield better generalizability than grading alone. Beyond grading, radiomic features could also be leveraged for treatment guidance. By combining imaging features with demographic data and records of past treatment responses, machine learning models could potentially support personalized therapy recommendations, aligning predictive modeling more directly with clinical decision-making.

With the introduction of more models than simply random forest [100], [158], improvements were hard to miss. For tissue classification in the SvCH dataset, all models had a higher AUC-ROC than random forest. There was no one model that outperformed the rest for all tasks, emphasizing the complexity of the datasets and their inherent variance. This behaviour hints that specific models may be more suited than others for performing specific tasks, although the degree of variance between model performance in the same classification task is often times very minimal and not significantly large enough to draw any concrete conclusions.

Choosing optimal models can be a very subjective task and depends on many variables, for simplicity one might choose a model that performs slightly worse, but has fewer features making processing easier. The power of the robustness of decision-tree models with few features is the low computational load required, and can be run on nearly any CPU. Therefore with smaller models, the ease of access can be a big selling point. Fewer features also make analyzing the respective models a much simpler task, which tackles the general public's issues with large black boxes of machine learning models, where the inner workings are difficult to explain in human terms.

4.5.7 Sarcoma Manager

The easiest way to quantify model performance is visually, as there are too many patients, features and data to go through manually. Therefore, a small program was designed in order to view data, the Sarcoma Manager. The program allows the user to browse through all patient scans, displaying images, masks and super pixels displayed after needs, as well as model classification results. Using this software, it is much easier to see where models perform well and where they do not. Figure 4.40 displays an open case, at the top the user can select the patient, per-

spective and ML model that they want. The top left image displays the raw image slice, to the right of the image, the user can toggle the visibility of different masks and SPs. Patient information is displayed beneath the mask and SP toggle buttons with predicted values displayed beneath the patient information. The prediction section is set up to show how different models and methods classify the patient. SP means that the predictions were made using super pixels, otherwise they are processed using masks. Pooled means that predictions were aggregated for all slices and the mean was taken. Majority vote means that each prediction casts a vote to the likeliest class and the most voted class is displayed. Slice means that information was taken on a per slice basis, so only based on the image displayed.

$$Grade(SP, pooled) = \frac{1}{N} \sum_{slices} \sum_{SP} P(Grade\ 1) + 2P(Grade\ 2) + 3P(Grade\ 3) \quad (4.17)$$

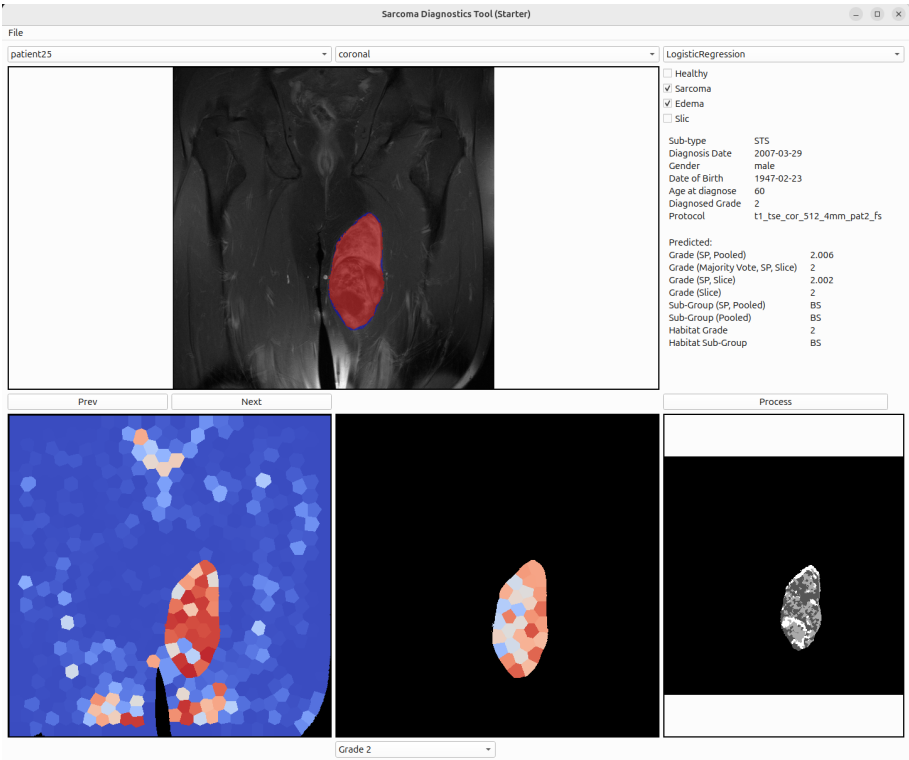


Figure 4.40: Sarcoma Manager.

So, sum the probabilities of each grade for every SP in every slice and divide by the number of samples.

$$Grade(SP, slice) = \frac{1}{N} \sum_{SP} P(\text{Grade 1}) + 2P(\text{Grade 2}) + 3P(\text{Grade 3}) \quad (4.18)$$

Sum the probabilities of each grade for every SP on the current slice and divide by the number of samples. These 2 metrics can be descriptive, as they show the general distribution of sample classification, but due to lacking accuracy, are not the best metrics to view for diagnosing grade.

Below the main image are 3 more images, the left most image shows a heatmap of the chosen tissue classifier. Blue shows low response to sarcoma tissue, and red shows high response to sarcoma tissue. Viewing the classification in this manner helps in seeing where the model performs well and where not, and has the added benefit of showing the likelihood of a sarcoma tumor within regions. The middle image shows the probabilities of the selected sarcoma grades in SPs and the right image shows the generated habitat. After selecting a patient and perspective, pressing the "Process" button infers the different classifiers of the selected model.

The software makes comparing different patients, perspectives, models, and classifiers much easier. Figures 4.41 to 4.46 show examples of patients displayed in the Sarcoma Manager program. Example 1 in figure 4.41 shows that the tissue classifier often mistakes the outer layer of fat as sarcoma, as well as some of the bone SPs, and there is very little response from the actual tumor region, the raw image itself does not offer much contrast, and little heterogeneity is observed in the tumor region. The sub-group classification was correct, except for the habitat sub-group classification. The grade was accurately predicted using a majority vote and using the whole sarcoma mask on the visible slice, but habitats and other grade measures did not classify correctly.

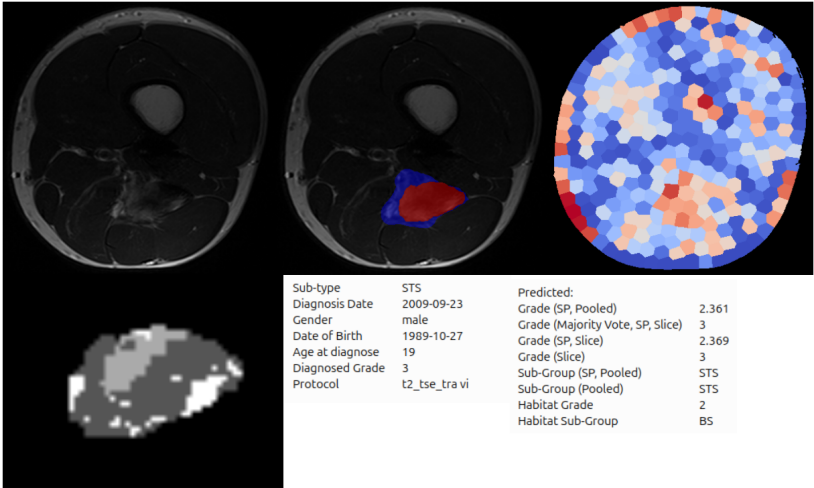


Figure 4.41: Example 1 of a patient viewed in the Sarcoma Manager program.

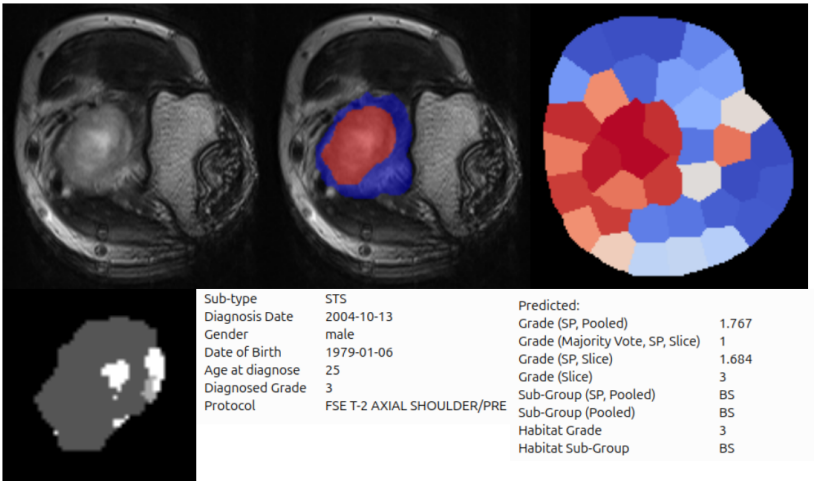


Figure 4.42: Example 2 of a patient viewed in the Sarcoma Manager program.

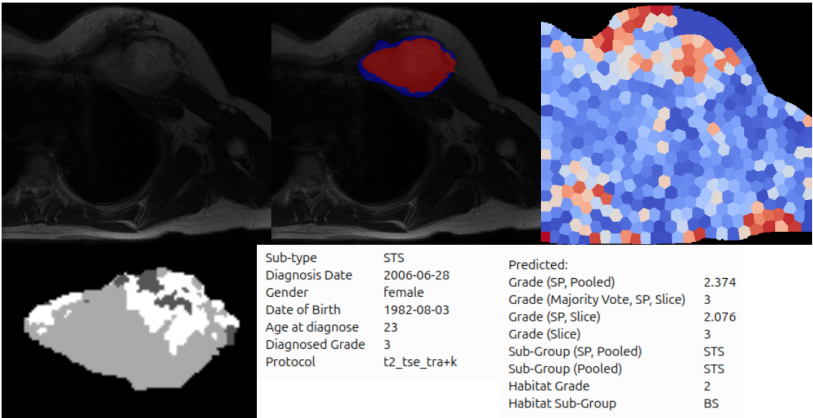


Figure 4.43: Example 3 of a patient viewed in the Sarcoma Manager program.

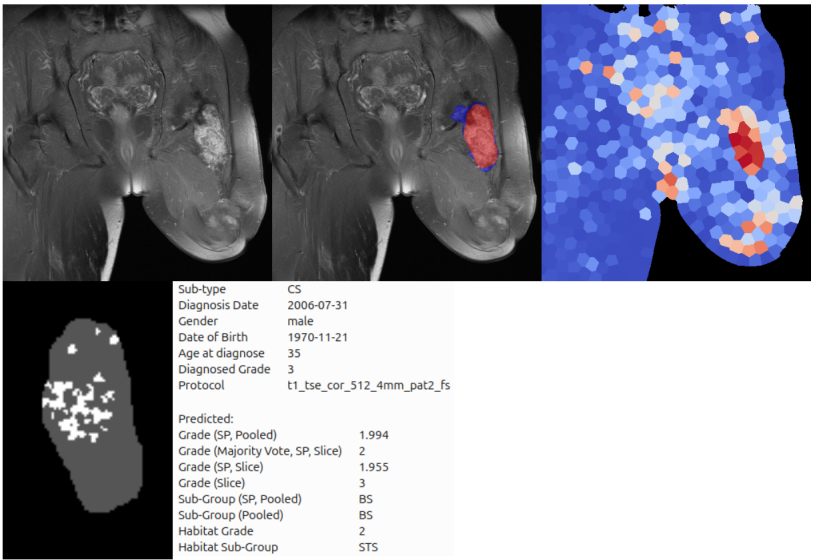


Figure 4.44: Example 4 of a patient viewed in the Sarcoma Manager program.

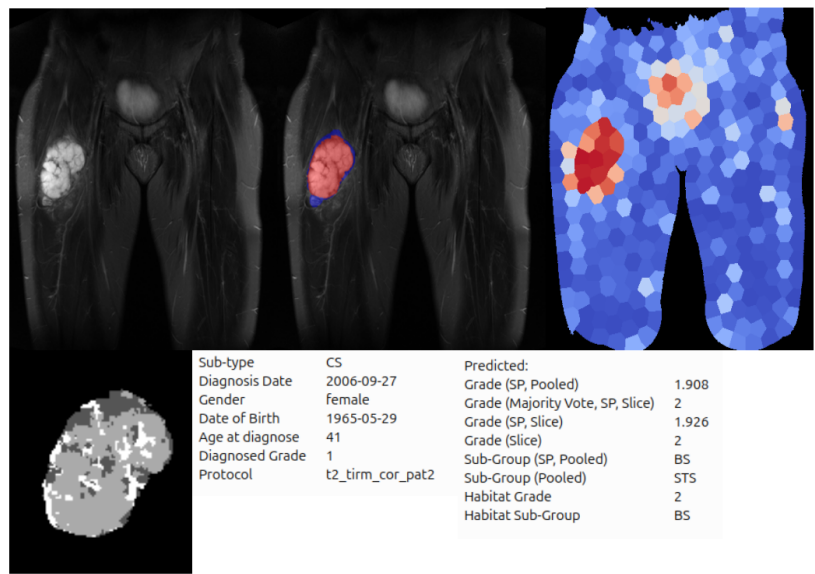


Figure 4.45: Example 5 of a patient viewed in the Sarcoma Manager program.

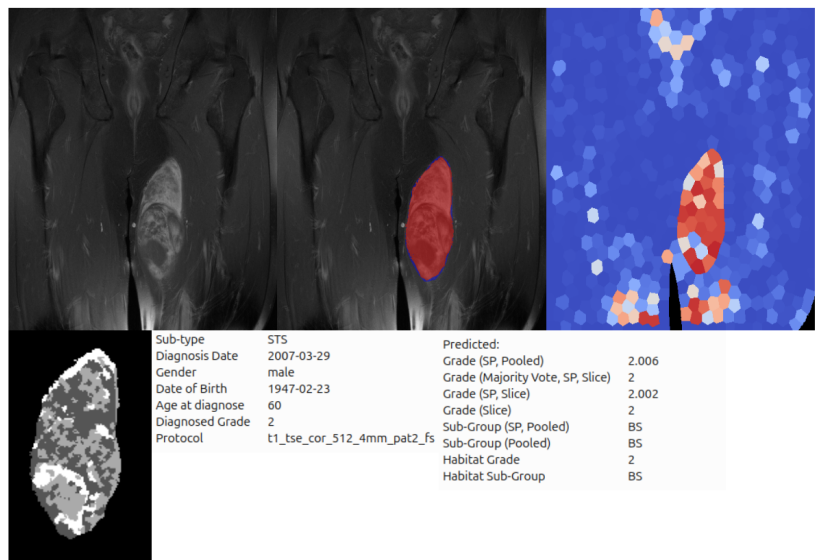


Figure 4.46: Example 6 of a patient viewed in the Sarcoma Manager program.

Figure 4.42 shows a slightly better heat map, in this case the sarcoma region is slightly more red than the surroundings, although there was again high response with the outer fatty layers and sarcoma. The sub-group was incorrectly classified in all cases and grade was correctly classified only using the current slice mask and with habitats. Visually, the raw image slice is a bit blurred and the region surrounding the tumor is quite visually similar to edema, possibly explaining the high resonance to sarcoma in that region.

Figure 4.43 displays a case of grade 3 STS, the raw image has quite bad lighting and provides little visual heterogeneity in the tumor. The heatmap shows very little resonance to sarcoma on the tumor region, although grade was correctly classified by the majority vote, and the slice grade. Sub-group was also correctly classified for by the pooled SPs of the whole tumor and the pooled masks of the whole tumor. Habitats did not correctly classify this patient's scan.

Figure 4.44 displays a patient suffering from chondrosarcoma, a type of bone sarcoma and a diagnosed grade of 3. Grade was correctly classified only by the displayed slice's mask and sub-group was correctly classified using SPs and masks. None of the other classifications were correct.

Figure 4.45 displays another case of chondrosarcoma, this time of grade 1. Visually, the raw image offers very much contrast between the region of interest and surrounding tissue, although tumor sub-regions are relatively homogenous in pixel intensity. The image has fat suppressed, and it can be seen that fat SPs show little probability of being classified as sarcoma. The heatmap is very red in the sarcoma region, showing very distinctly where the tumor lies. The grade was not correctly predicted, all leaning towards grade 2. The sub-group was correctly classified by the pooled SPs and by habitats.

Figure 4.46 shows a patient diagnosed with grade 2 STS. The heatmap shows clear distinction from the surrounding tissue, although a few SPs with relatively high intensity show high probability with being sarcoma SPs, likely explained by their resemblance to edema. The grade was correctly classified in all cases, but sub-group was misclassified in all cases.

These are a few examples of how the different classifiers for tissue, grade and sub-group behaved given a single slice on 6 patients. It can be difficult to draw any concrete conclusions from these images alone, but some patterns do emerge. Fat suppression seems to help the model correctly classify outer fatty layers as healthy anatomy, image quality seems to affect classification results. T1 weighted seems to outperform T2 images, possibly explained by the difference in contrast and sub-region quality of these examples. Contrast media usage cannot be displayed, as this information must be manually fetched from images in the database. It is displayed only in some JPEG images of scans, but not consistently. Therefore, drawing conclusion on contrast media usage is difficult, but generally it does add more contrast and emphasis on the tumor region. The results of the classifiers hint that if a study is designed with the intent to use machine learning to diagnose patients,

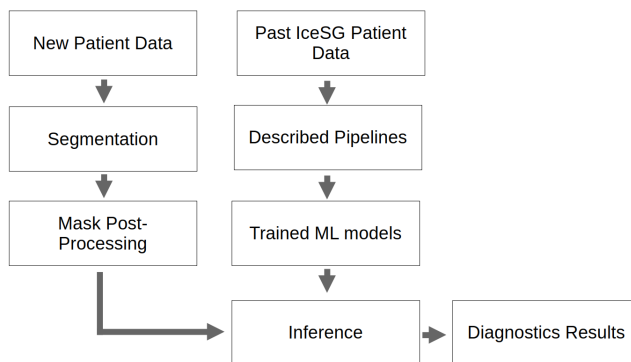


Figure 4.47: The roadmap towards deployment.

then the best imaging protocol for gathering patient data would be T1 weighted with fat suppression on for tissue classification. For grade and sub-group classification, it appears that imaging protocols that emphasize tumor heterogeneity and sub-regions fare better, at least when compared to those cases where little heterogeneity and sub-region contrast is visible.

4.5.8 Roadmap to Deployment

The current framework takes as input DICOM image files across multiple patients and scans, coupled with associated patient metadata regarding their sarcoma diagnosis and staging. These data are processed into feature value tables through the described pipelines, from which machine learning (ML) models are trained to predict an output class: tissue, grade, or sub-group. Figure 4.47 outlines the conceptual roadmap for deploying the trained models in a clinical setting.

By pushing all patients in the IceSG database through the standardized pipelines, a collection of trained ML models is obtained. When a new patient requires sarcoma assessment, a trained professional would first produce the segmentation masks for that patient’s scan. The raw DICOM images and corresponding masks would then be passed through the post-processing steps described earlier to generate the same feature tables used during model training. These tables would then be forwarded to the appropriate ML models for inference, producing diagnostic predictions that can assist in clinical decision making.

In practice, a new clinical protocol would be established for incoming patient data. MRI acquisition parameters should follow the configuration that provided the most consistent radiomic information during development, specifically, T1-weighted sequences with fat suppression. Ensuring protocol consistency is essential for reliable feature extraction and cross-site reproducibility. As model perfor-

mance improves, segmentation could transition from manual to semi-automatic or fully automatic workflows. Once a robust tissue classifier is trained to clinical standards, operator interaction could be limited to reviewing or approving the automatically generated segmentation, while the remainder of the process runs without user input.

For deployment in a clinical environment, the pipeline could be integrated directly with the PACS system via a DICOM interface. A local inference server or containerized application could listen for new studies, automatically trigger processing, and return results as secondary capture DICOM images or structured reports viewable within existing radiology software. This type of setup would allow clinicians to request a “sarcoma analysis” directly from the PACS viewer without manually exporting data.

To assess readiness for clinical translation, inference speed should also be evaluated. The benefit of the many of the models such as XGBoost and logistic regression classifiers is that they are very lightweight and can be run on nearly any CPU, this ensures that model inference is fast and accessible. Optimizations such as pre-caching image metadata, GPU-accelerated feature computation, or lightweight ensemble models could further reduce latency.

Deployment, however, should only proceed once each ML model performs at or above clinical competence for its respective classification task and has been validated on independent, prospective cohorts. Even then, such tools should be used cautiously, as decision-support systems rather than autonomous diagnostic tools, since predictive errors can have clinical consequences. Continuous monitoring, feedback loops, and periodic retraining with new patient data would be essential parts of the deployment roadmap to maintain accuracy and reliability over time. ML models should also be trained on more diverse data to reduce biases, as the data used in this work is gathered only in Iceland it may not reflect international demographics and may therefore be indirectly biased in classification. To ensure unbiased and ethical practices, models must be trained on data with multiple international points of origin.

4.5.9 Improvements

Many improvements can be made to all the pipelines and workflows presented in this research. Including features for MRI protocol could help models grasp the variance between tumor presentation in MR images, for example, models often times misclassify fatty tissue as tumor instead of healthy anatomy when fat suppression is turned off. Including a boolean feature for fat suppression could aid models, as they would then expect certain response from fatty tissue which could be more accurately classified. Also, contrast varies depending on the protocol, including features for contrast would help models as they would expect a certain type of contrast and therefore response of certain features. The tree based algo-

rithms (XGBoost, LightGBM, CatBoost, RandomForest), could make splits early on in the decision trees based on protocol features. Other meta data features such as perspective could also aid models, as the appearance of anatomy is very different depending on perspective, at least when comparing axial against coronal or sagittal, currently the models must account for the differences using only numerical features, this could decrease the load on the models.

Resampling the raw input images and respective masks for the radiomics components could lead to better results. Resampling has been shown to significantly improve model performance in radiomics [159], implementing resampling for the radiomics based datasets could therefore improve model classification results of these works and should be tested.

There are certainly many improvements that can be made in the habitat generation pipeline. A round of feature selection to find the best features that correlate with grade could be included, by using the defined feature selection in the radiomics section, but applied in 3D. Although, this could prove difficult as size scales highlight different features as has been shown when comparing the selected features with and without SPs. Using different feature sets could also potentially improve sub-group classification. Overall, habitats show much potential in sarcoma diagnostics, but need much work to improve model robustness.

None of the applied workflows took context into account, or what the surroundings look like. All models only have information contained within the masks or habitats. If the neighborhood of the data were taken into account then models could be improved. For example, layers of fat often times appear on the outer layers, closest to the skin, if models could see how far a specific SP is from the edge of anatomy, it could classify with higher accuracy that the tissue is healthy. Also, as different sarcoma sub-types develop in different regions of anatomy, having knowledge of the surrounding region could help in classifying sub-groups, e.g. bones have different radiomic values than muscle, if the tumor originates in bone, then the neighbors should have high resonance with bone features, making classification easier.

Introducing more classes of anatomy could also improve model performance. Currently, there is only healthy and pathological tissue masks, and healthy tissue can often times visually resemble heterogeneous sarcoma tumors. That means that the range in variance of healthy tissue is very high and may encapsulate the radiomic feature ranges of sarcoma tissue in many cases. Including masks for other anatomy can help models grasp the subtle differences in varying anatomical structures and improve classification performance.

Combining radiomics features with habitat features could also improve performance. From classification results, it can be seen that habitat and radiomic classifiers are not always in agreement, hinting that they are grasping different information. Although, habitats work in 3D and radiomics in 2D, so whether or not there is different information in habitats than in radiomics is difficult to confirm.

The current workflows are also designed to work in only 2D or 3D, and therefore would need to be reformulated in order to combine radiomics and habitats.

The use of other clinical data could improve model performance. The biopsy images that are used for the FNCLCC grading guidelines contains vital information to tumor grade, including these images and extracting radiomic features has the potential to boost grade classification. Genomic data, has also been shown to improve classification of sarcoma, and it would be interesting to see if its inclusion could boost classification accuracy. Unfortunately, neither biopsy or genomic data is available for the patients included in the IceSG database.

The radiomic workflows should also be tested in 3D. In the presented works, all radiomic tasks were performed in 2D and on a slice-wise basis, it stands to reason, that a single slice does not contain enough information to accurately diagnose grade, but that isn't necessarily true for the 3D case. Features would be extracted from the tumor volume and go through a similar pipeline as described. It would also make it easier to combine habitat data as the input images and masks are identical and on the same scale. The difficulties in performing the method this way is that there are only 195 scans available and more than a thousand features to choose from, having such a low sample to feature ratio could introduce a lot of difficulties in feature selection as models generally require orders of magnitude more samples than features.

Model hyperparameter fine-tuning [160] was not implemented in this this pipeline. Some meta explorations in fine-tuning models were made and only produced incremental improvements to model metrics, $< 2\%$ and was therefore not justified, as tuning requires multiple training cycles and some classifiers such as random forest train very slowly on many of the datasets. If model training time is reduced and higher metrics are achieved through other improvements, then fine-tuning becomes justifiable, especially if deployed in a clinical setting. In this case, a grid search of ML model initialization parameters can be done in an attempt to boost metrics. If models are used in clinical practice then even slight improvements can only benefit the patients down the line and therefore fine tuning becomes imperative.

Splitting the pipelines based on sarcoma types might also offer benefits. In this case a model could focus on explaining the variance of phenotypically similar cases, instead of having to explain differences on the whole range of sarcoma. Different feature sets might be selected depending on sarcoma type, resulting in more specialized predictive capabilities.

4.5.10 Notable Attempts

Many attempts at different methods were tested, and not included in the works. Combining grades 1 and 2 into a set and classifying grade against class 3 did not have any noticeable improvements to overall metrics. There was an attempt to ex-

tract habitat features in 2 dimensions which gave bad results, using the extracted 2D habitat features with radiomic features also gave bad results. Habitats worked best in 3 dimensions, and therefore radiomic features should also be extracted in 3D and tested. There were early attempts at using artificial neural networks applied to the radiomic feature data, but those results were always outperformed by tree methods. U-nets, neural network models designed for segmentation tasks, were also attempted to see how well they could segment and classify healthy from pathological tissue, but unfortunately lacked enough data to perform well at all and required too much computing power to be deemed reasonable for use.

Chapter 5

Conclusion and Future Directions

These works show that novel computer algorithms and technology has the potential to support clinical diagnosis and staging of varying sarcoma. The results indicate that it is possible to teach ML models to distinguish between healthy and pathological tissue at different size scales, opening up the possibility of automated pathways to anatomical searches for sarcoma. The results of FNCLCC grading using ML models also hints at much potential, trained models performed adequately enough to conclude that correlations exist between input radiomics data and grade. The performance of models when predicting sarcoma sub-group also shows potential, even with very few cases of the minority class, performance metrics were relatively high.

Overall, ML displays a large potential in the medical field. With the right quality and magnitude of data, coupled with a robust pre-processing and training pipeline, many factors can be predicted through non-invasive imaging techniques. Although the performance may not be adequate for clinical implementation, development in this field may soon lead to computer algorithms that can accurately diagnose patients to an acceptable degree.

While the works presented establish a foundation for understanding diagnostics, characterization and grading of sarcoma using AI, several opportunities exist for further future development. These include:

- The addition of meta-data features such as the used MRI protocol for each scan can help models account for visual differences caused by the MRI machines.
- Increasing the number of segmentation mask types to include other anatomy such as fat, organs and bone.
- Incorporating more patients in minority classes such as grade 1 and 2, as well as instances of bone sarcoma.

- A deeper analysis on habitat building radiomic features.
- Developing a method for analyzing patients and tumors using both radiomic features and habitat features simultaneously.
- Improving and testing parameter choices for models and functions.
- Implementing spatial context into the learning models.
- Applying the pipelines to specific sub-groups independently.

Other Publications

Assessment of Femoral Cartilage Morphological and Topological Features Using Machine Learning

Gunnarsson, A. E., Recenti, M., Jacob, D., Aubonnet, R., Burgunder, B., Mengual i Escalona, I., Ciliberti, F. K., Forni, R., Donisi, L., Petersen, H., & Gargiulo, P. (2022). In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, Rome, Italy, pp. 277–282. DOI: 10.1109/MetroXRaine54828.2022.9967659

CT- and MRI-Based 3D Reconstruction of Knee Joint to Assess Cartilage and Bone

Ciliberti, F. K., Guerrini, L., Gunnarsson, A. E., Recenti, M., Jacob, D., Cangiano, V., Tesfahunegn, Y. A., Islind, A. S., Tortorella, F., Tsirilaki, M., Jónsson, H. Jr., Gargiulo, P., & Aubonnet, R. (2022). **Diagnostics**, 12(2), 279. DOI: 10.3390/diagnostics12020279

The role of bone mineral density and cartilage volume to predict knee cartilage degeneration

Ciliberti, F. K., Cesarelli, G., Guerrini, L., Gunnarsson, A. E., Forni, R., Aubonnet, R., Recenti, M., Jacob, D., Jónsson, H. Jr., Cangiano, V., Islind, A. S., Gambacorta, M., & Gargiulo, P. (2022). **European Journal of Translational Myology**, 32(2), 10678. DOI: 10.4081/ejtm.2022.10678

Predicting lifestyle using BioVRSea multi-biometric paradigms

Recenti, M., Jacob, D., Aubonnet, R., Burgunder, B., Mengual i Escalona, I., Gunnarsson, A. E., Ciliberti, F. K., Forni, R., Donisi, L., Petersen, H., & Gargiulo, P. (2022). In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, Rome, Italy, pp. 329–334. DOI: 10.1109/MetroXRaine54828.2022.9967685

Bibliography

- [1] P. F. M. Choong, Ed., *Sarcoma: A Practical Guide to Multidisciplinary Management*. Springer Singapore, 2021, ISBN: 978-981-15-9414-4. DOI: 10.1007/978-981-15-9414-4. [Online]. Available: <https://doi.org/10.1007/978-981-15-9414-4>.
- [2] J. Yang, Z. Ren, X. Du, M. Hao, and W. Zhou, "The role of mesenchymal stem/progenitor cells in sarcoma: Update and dispute," *Stem Cell Investigation*, vol. 1, p. 18, 2014, eCollection 2014, ISSN: 2306-9759. DOI: 10.3978/j.issn.2306-9759.2014.10.01. [Online]. Available: <https://doi.org/10.3978/j.issn.2306-9759.2014.10.01>.
- [3] N. Frazzette and G. Jour, "Novel molecular methods in soft tissue sarcomas: From diagnostics to theragnostics," *Cancers*, vol. 17, no. 7, p. 1215, 2025, Published 3 April 2025. DOI: 10.3390/cancers17071215. [Online]. Available: <https://doi.org/10.3390/cancers17071215>.
- [4] L. F. Peltier, "Historical note on bone and soft tissue sarcoma," *Journal of Surgical Oncology*, vol. 30, no. 4, pp. 201–205, 1985, Dec 1985. DOI: 10.1002/jso.2930300403.
- [5] E. J. Odes, P. S. Randolph-Quinney, M. Steyn, *et al.*, "Earliest hominin cancer: 1.7-million-year-old osteosarcoma from swartkrans cave, south africa," *South African Journal of Science*, vol. 112, no. 7–8, pp. 1–5, 2016. DOI: 10.17159/sajs.2016/20150471. [Online]. Available: <https://doi.org/10.17159/sajs.2016/20150471>.
- [6] S. Ekhtiari, K. Chiba, S. Popovic, *et al.*, "First case of osteosarcoma in a dinosaur: A multimodal diagnosis," *The Lancet Oncology*, vol. 21, no. 8, pp. 1021–1022, 2020. DOI: 10.1016/S1470-2045(20)30171-6. [Online]. Available: [https://doi.org/10.1016/S1470-2045\(20\)30171-6](https://doi.org/10.1016/S1470-2045(20)30171-6).
- [7] N. F. Gilbert, C. P. Cannon, P. P. Lin, and V. O. Lewis, "Soft-tissue sarcoma," *Journal of the American Academy of Orthopaedic Surgeons*, vol. 17, no. 1, pp. 40–47, Jan. 2009, ISSN 1067-151X.

- [8] J. N. Cormier and R. E. Pollock, "Soft tissue sarcomas," *CA: A Cancer Journal for Clinicians*, vol. 54, no. 2, pp. 94–109, 2004, First published: March/April 2004. DOI: 10.3322/canjclin.54.2.94. [Online]. Available: <https://doi.org/10.3322/canjclin.54.2.94>.
- [9] National Cancer Institute, *Surveillance, Epidemiology, and End Results (SEER) Program SEER*Stat Database: Incidence - SEER Research Data*, <https://seer.cancer.gov/>, SEER 9 Registries, Nov 2023 Sub (1975–2021) <Katrina/Rita Population Adjustment>. National Cancer Institute, DC-CPS, Surveillance Research Program. Released April 2024, based on the November 2023 submission, 2024.
- [10] B. Serban, B. Cretu, A. Cursaru, C. Nitipir, C. Orlov-Slavu, and C. Cirstoiu, "Local recurrence management of extremity soft tissue sarcoma," *EFORT Open Reviews*, vol. 8, no. 8, pp. 606–614, 2023, Published August 1, 2023. DOI: 10.1530/EOR-23-0095. [Online]. Available: <https://doi.org/10.1530/EOR-23-0095>.
- [11] J. Rytlewski, Q. R. Brockman, R. D. Dodd, M. Milhem, and V. Monga, "Epigenetic modulation in sensitizing metastatic sarcomas to therapies and overcoming resistance," *Cancer Drug Resistance*, vol. 5, no. 1, pp. 25–35, 2022, Published January 4, 2022; eCollection 2022. DOI: 10.20517/cdr.2021.88. [Online]. Available: <https://doi.org/10.20517/cdr.2021.88>.
- [12] A. Italiano, S. Mathoulin-Pelissier, A. L. Cesne, *et al.*, "Trends in survival for patients with metastatic soft-tissue sarcoma," *Cancer*, vol. 117, no. 5, pp. 1049–1054, 2011. DOI: <https://doi.org/10.1002/cncr.25538>. eprint: <https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.25538>. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.25538>.
- [13] A. Greenwood, R. Arora, and H. Shaikh, "Osteosarcoma (osteogenic sarcoma)," *StatPearls [Internet]*, 2025, Updated 2024 Dec 11. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK563177/>.
- [14] S. Weeden, R. Grimer, S. Cannon, A. Taminiau, and B. Uscinska, "The effect of local recurrence on survival in resected osteosarcoma," *European Journal of Cancer*, vol. 37, no. 1, pp. 39–46, 2001, ISSN: 0959-8049. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959804900003622>.
- [15] S. S. Bielack, B. Kempf-Bielack, D. Branscheid, *et al.*, "Second and subsequent recurrences of osteosarcoma: Presentation, treatment, and outcomes of 249 consecutive cooperative osteosarcoma study group patients," *Journal of Clinical Oncology*, vol. 27, no. 4, pp. 557–565, 2009, PMID: 19075282. DOI: 10.1200/JCO.2008.16.2305. eprint: <https://ascopubs.org/>

- doi/pdf/10.1200/JCO.2008.16.2305. [Online]. Available: <https://ascopubs.org/doi/abs/10.1200/JCO.2008.16.2305>.
- [16] L. Twenhafel, D. Moreno, T. Punt, M. Kinney, and R. Ryznar, "Epigenetic changes associated with osteosarcoma: A comprehensive review," *Cells*, vol. 12, no. 12, 2023. DOI: 10.3390/cells12121595. [Online]. Available: <https://www.mdpi.com/2073-4409/12/12/1595>.
- [17] Y. Jha and K. Chaudhary, "Giant cell tumour of bone: A comprehensive review of pathogenesis, diagnosis, and treatment," *Cureus*, vol. 15, no. 10, e46945, 2023. DOI: 10.7759/cureus.46945. [Online]. Available: <https://doi.org/10.7759/cureus.46945>.
- [18] S. Venneker, K. Szuhai, P. C. W. Hogendoorn, and J. V. M. G. Bovée, "Mutation-driven epigenetic alterations as a defining hallmark of central cartilaginous tumours, giant cell tumour of bone and chondroblastoma," *Virchows Archiv*, vol. 476, no. 1, pp. 135–146, 2020, ISSN: 1432-2307. DOI: 10.1007/s00428-019-02699-2. [Online]. Available: <https://doi.org/10.1007/s00428-019-02699-2>.
- [19] E. Palmerini, P. Picci, P. Reichardt, and G. Downey, "Malignancy in giant cell tumor of bone: A review of the literature," *Technology in Cancer Research & Treatment*, vol. 18, p. 1533033819840000, 2019. DOI: 10.1177/1533033819840000.
- [20] F. Limaïem, D. Davis, and K. Sticco, "Chondrosarcoma," *StatPearls [Internet]*, 2025, Updated 2023 Aug 14. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK538132/>.
- [21] W. A. Chow, "Chondrosarcoma: Biology, genetics, and epigenetics," *F1000Research*, vol. 7, F1000 Faculty Rev–1826, 2018, Published November 20, 2018; eCollection 2018. DOI: 10.12688/f1000research.15953.1. [Online]. Available: <https://doi.org/10.12688/f1000research.15953.1>.
- [22] J. Fromm, A. Klein, A. Baur-Melnyk, *et al.*, "Survival and prognostic factors in conventional central chondrosarcoma," *BMC Cancer*, vol. 18, no. 1, p. 849, 2018, Published August 24, 2018, ISSN: 1471-2407. DOI: 10.1186/s12885-018-4741-7. [Online]. Available: <https://doi.org/10.1186/s12885-018-4741-7>.
- [23] O. G. Bloch, B. J. Jian, I. Yang, *et al.*, "Cranial chondrosarcoma and recurrence," *Skull Base*, vol. 20, no. 3, pp. 149–156, 2010, Published April 12, 2010, ISSN: 1531-5010. DOI: 10.1055/s-0029-1246218. [Online]. Available: <http://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0029-1246218>.

- [24] M. E. Pring, K. L. Weber, K. K. Unni, and F. H. Sim, "Chondrosarcoma of the pelvis: A review of sixty-four cases," *The Journal of Bone and Joint Surgery*, vol. 83, no. 11, pp. 1630–1642, 2001, ISSN: 0021-9355. [Online]. Available: https://journals.lww.com/jbjsjournal/fulltext/2001/11000/chondrosarcoma_of_the_pelvis__a_review_of_3.aspx.
- [25] S. Durer, D. Gasalberti, and H. Shaikh, "Ewing sarcoma," *StatPearls [Internet]*, 2025, Updated 2024 Jan 8. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK559183>.
- [26] W. M. Mendenhall, R. A. Zlotecki, and M. T. Scarborough, "Dermatofibrosarcoma protuberans," *Cancer*, vol. 101, no. 11, pp. 2503–2508, 2004. DOI: <https://doi.org/10.1002/cncr.20678>. eprint: <https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.20678>. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.20678>.
- [27] A. Perumal, P. Venkatakrishnan, and N. Janakiraman, "Histopathology image classification for soft tissue sarcoma in limbs using artificial neural networks," Feb. 2024.
- [28] *Osteosarcoma*, <https://en.wikipedia.org/wiki/Osteosarcoma>, Accessed: 2025-08-07, 2025. [Online]. Available: <https://en.wikipedia.org/wiki/Osteosarcoma>.
- [29] *Giant-cell tumor of bone*, https://en.wikipedia.org/wiki/Giant-cell_tumor_of_bone, Accessed: 2025-08-07, 2025.
- [30] *Chondrosarcoma*, <https://en.wikipedia.org/wiki/Chondrosarcoma>, Accessed: 2025-08-07, 2025.
- [31] *Ewing sarcoma*, https://en.wikipedia.org/wiki/Ewing_sarcoma, Accessed: 2025-08-07, 2025.
- [32] *Dermatofibrosarcoma protuberans*, https://en.wikipedia.org/wiki/Dermatofibrosarcoma_protuberans, Accessed: 2025-08-07, 2025.
- [33] G. Porrello, R. Cannella, A. Randazzo, G. Badalamenti, G. Brancatelli, and F. Vernuccio, "Ct and mr imaging of retroperitoneal sarcomas: A practical guide for the radiologist," *Cancers*, vol. 15, no. 11, p. 2985, May 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15112985.
- [34] M. Trojani, G. Contesso, J. M. Coindre, *et al.*, "Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system," *International Journal of Cancer*, vol. 33, no. 1, pp. 37–42, 1984. DOI: 10.1002/ijc.2910330108. [Online]. Available: <https://doi.org/10.1002/ijc.2910330108>.

- [35] X. Lin, S. Davion, E. Bertsch, I. Omar, R. Nayar, and W. Laskin, "Federation nationale des centers de lutte contre le cancer grading of soft tissue sarcomas on needle core biopsies using surrogate markers," *Human Pathology*, vol. 56, pp. 147–154, Oct. 2016, Epub 2016 Jun 23. DOI: 10.1016/j.humpath.2016.06.008. [Online]. Available: <https://doi.org/10.1016/j.humpath.2016.06.008>.
- [36] S. Lin, C. Tran, E. Bandari, *et al.*, "The 1000 mitoses project: A consensus-based international collaborative study on mitotic figures classification," *International Journal of Surgical Pathology*, vol. 32, no. 8, pp. 1449–1458, Dec. 2024, Epub 2024 Apr 16. DOI: 10.1177/10668969241234321. [Online]. Available: <https://doi.org/10.1177/10668969241234321>.
- [37] A. Morisi, T. Rai, N. J. Bacon, *et al.*, "Detection of necrosis in digitised whole-slide images for better grading of canine soft-tissue sarcomas using machine-learning," *Veterinary Sciences*, vol. 10, no. 1, 2023, ISSN: 2306-7381. DOI: 10.3390/vetsci10010045. [Online]. Available: <https://www.mdpi.com/2306-7381/10/1/45>.
- [38] A. R. Dancsok, K. Asleh-Aburaya, and T. O. Nielsen, "Advances in sarcoma diagnostics and treatment," *Oncotarget*, 2017.
- [39] Johns Hopkins Medicine, *Sarcoma*, <https://www.hopkinsmedicine.org/health/conditions-and-diseases/sarcoma>, Accessed: 2025-06-05.
- [40] J. Fayette and J.-Y. Blay, "Adjuvant chemotherapy in the treatment of sarcomas," *Bull Cancer*, 2006.
- [41] Z. Burningham, M. Hashibe, L. Spector, and J. D. Schiffman, "The epidemiology of sarcoma," *Clin Sarcoma Res*, 2012.
- [42] M. Sbaraglia, E. Bellan, and A. P. D. Tos, "The 2020 who classification of soft tissue tumours: News and perspectives," *Pathologica*, 2021.
- [43] M. Farid and J. Ngeow, "Sarcomas associated with genetic cancer predisposition syndromes: A review," *Oncologist*, 2016.
- [44] I.-M. Schaefer, G. M. Cote, and J. L. Hornick, "Contemporary sarcoma diagnosis, genetics, and genomics," *Journal of Clinical Oncology*, 2017.
- [45] L. Wang, H. Nie, Q. Wang, *et al.*, "Use of magnetic resonance imaging combined with gene analysis for the diagnosis of fetal congenital heart disease," *BMC Med Imaging*, 2019.
- [46] F. Al-Obeidat, A. Rashid, W. Hafez, *et al.*, "The accuracy of artificial intelligence in the diagnosis of soft tissue sarcoma: A systematic review and meta-analysis," *Current Problems in Surgery*, p. 101743, 2025. DOI: 10.1016/j.cpsurg.2025.101743. [Online]. Available: <https://doi.org/10.1016/j.cpsurg.2025.101743>.

- [47] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," *Cell*, 2023.
- [48] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "Dream: Diabetic retinopathy analysis using machine learning," *IEEE Journal of Biomedical and Health Informatics*, 2014.
- [49] M. Tanveer, B. Richhariya, R. U. Khan, *et al.*, "Machine learning techniques for the diagnosis of alzheimer's disease: A review," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2020.
- [50] F. Obeidin, "Soft tissue general grading," *Pathology Outlines*, 2022. [Online]. Available: <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [51] Materialise NV, *Mimics Innovation Suite*, <https://www.materialise.com/en/medical/software/mimics-innovation-suite>, Accessed: 2025-05-28, 2024.
- [52] K. Hammoudi, A. Cabani, B. Slika, H. Benhabiles, F. Dornaika, and M. Melkemi, "Superpixelgridmasks data augmentation: Application to precision health and other real-world data," *Journal of Healthcare Informatics Research*, 2023.
- [53] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. DOI: 10.1109/TPAMI.2012.120.
- [54] Z. Khatun, H. Jónsson, M. Tsirilaki, *et al.*, "Beyond pixel: Superpixel-based mri segmentation through traditional machine learning and graph convolutional network," *Computer Methods and Programs in Biomedicine*, 2024.
- [55] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, e104–e107, 2017. DOI: 10.1158/0008-5472.CAN-17-0339. [Online]. Available: <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [56] V. H. Tang, S. T. M. Duong, C. D. T. Nguyen, *et al.*, "Wavelet radiomics features from multiphase ct images for screening hepatocellular carcinoma: Analysis and comparison," *Nature*, 2023.
- [57] B. Sudharani, M. Hemalatha, and B. Deepa, "Wavelet transform for a fuzzy based image fusion," *International Journal of Advances in Electrical and Engineering*, 2015.
- [58] L. Buitinck, G. Louppe, M. Blondel, *et al.*, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

- [59] C. Bühnemann, S. Li, H. Yu, *et al.*, “Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis,” *PLOS ONE*, vol. 9, no. 9, e107105, Sep. 2014. DOI: 10.1371/journal.pone.0107105.
- [60] W. Chen, C. Zhou, Z. Yan, *et al.*, “Using machine learning techniques predicts prognosis of patients with ewing sarcoma,” *Journal of Orthopaedic Research*, vol. 39, no. 11, pp. 2519–2527, 2021. DOI: <https://doi.org/10.1002/jor.24991>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jor.24991>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jor.24991>.
- [61] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: <http://jmlr.org/papers/v18/16-365.html>.
- [62] A. Tagliafico, B. Bignotti, F. Rossi, F. Valdora, and C. Martinoli, “Local recurrence of soft tissue sarcoma: A radiomic analysis,” *Radiology and Oncology*, vol. 53, p. 300, 2019. DOI: 10.2478/raon-2019-0034. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6765164/>.
- [63] Y. Zhang, Y. Zhu, X. Shi, *et al.*, “Soft tissue sarcomas: Preoperative predictive histopathological grading based on radiomics of mri,” *Academic Radiology*, vol. 26, pp. 1262–1268, 2019. DOI: 10.1016/j.acra.2018.10.014. [Online]. Available: <https://www.ajronline.org/doi/full/10.2214/AJR.19.22147>.
- [64] V. Corino, E. Montin, A. Messina, *et al.*, “Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions,” *Journal of Magnetic Resonance Imaging*, vol. 47, pp. 829–840, 2017. DOI: 10.1002/jmri.25788. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/28653477/>.
- [65] S. Hu, C. Guan, W. Tang, and Y. Liu, “Texture feature extraction based on wavelet transform and gray-level co-occurrence matrices applied to osteosarcoma diagnosis,” *Bio-Medical Materials and Engineering*, vol. 24, pp. 129–143, 2014. DOI: 10.3233/BME-130793. [Online]. Available: <https://doi.org/10.3233/BME-130793>.
- [66] P. Lin, P.-F. Yang, S. Chen, *et al.*, “A delta-radiomics model for preoperative evaluation of neoadjuvant chemotherapy response in high-grade osteosarcoma,” *Cancer Imaging*, 2020.

- [67] S. Chen, N. Li, Y. Tang, *et al.*, “Radiomics analysis of fat-saturated t2-weighted mri sequences for the prediction of prognosis in soft tissue sarcoma of the extremities and trunk treated with neoadjuvant radiotherapy,” *Frontiers in Oncology*, vol. 11, p. 710 649, 2021, Epub 2021 Sep 17. DOI: 10 . 3389 / f onc . 2021 . 710649. [Online]. Available: <https://doi.org/10.3389/fonc.2021.710649>.
- [68] J. Niu, Q. Tan, X. Zou, and S. Jin, “Accurate prediction of glioma grades from radiomics using a multi-filter and multi-objective-based method,” *AIMS Press: Mathematical Biosciences and Engineering*, 2022.
- [69] C. Ricciardi, K. J. Edmunds, M. Recenti, *et al.*, “Assessing cardiovascular risks from a mid-thigh CT image: A tree-based machine learning approach using radiodensitometric distributions,” *Scientific Reports*, vol. 10, no. 1, p. 2863, 2020. DOI: 10 . 1038 / s41598 - 020 - 59873 - 9.
- [70] V. Lakshmanan, S. Robinson, and M. Munn, *Machine Learning Design Patterns*. Sebastopol, CA: O’Reilly Media, 2020.
- [71] Q. C. Le, H. Arimura, K. Ninomiya, and Y. Kabata, “Radiomic features based on hessian index for prediction of prognosis in head-and-neck cancer patients,” *Nature*, 2020.
- [72] R. C. Gonzalez and R. E. Woods, *Digital Image Processin 3rd Edition*. Pearson, 2008.
- [73] Z. Tian, L. Liu, Z. Zhang, and B. Fei, “Superpixel-based segmentation for 3d prostate mr images,” *IEEE Transactions on Medical Imaging*, 2015. DOI: 10 . 1109 / TMI . 2015 . 2408362. [Online]. Available: <https://doi.org/10.1109/TMI.2015.2408362>.
- [74] S. Pillozzi, A. Bernini, I. Palchetti, *et al.*, “Soft tissue sarcoma: An insight on biomarkers at molecular, metabolic and cellular level,” *Cancers*, vol. 13, no. 12, p. 3044, 2021.
- [75] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.
- [76] P. Hingorani, K. Janeway, B. D. Crompton, *et al.*, “Current state of pediatric sarcoma biology and opportunities for future discovery: A report from the sarcoma translational research workshop,” *Cancer genetics*, vol. 209, no. 5, pp. 182–194, 2016.
- [77] I. Cosci, P. Del Fiore, S. Mocellin, and A. Ferlin, “Gender differences in soft tissue and bone sarcoma: A narrative review,” *Cancers*, vol. 16, no. 1, p. 201, 2023.
- [78] J. Á. Fernández, B. Gómez, D. Díaz-Gómez, *et al.*, “Diagnostic delay in soft tissue sarcomas: A review,” *Cancers*, vol. 17, no. 11, p. 1861, 2025.

- [79] P. Pisters, D. Leung, J. Woodruff, W. Shi, and M. F. Brennan, "Analysis of prognostic factors in 1,041 patients with localized soft tissue sarcomas of the extremities.," *Journal of clinical oncology*, vol. 14, no. 5, pp. 1679–1689, 1996.
- [80] A. B. Meyers and M. Epelman, "Ultrasound versus magnetic resonance imaging of soft-tissue lesions: Competitive or complementary?" *Pediatric Radiology*, vol. 52, no. 9, pp. 1639–1647, 2022.
- [81] A. M. Schmid, D. L. Raunig, C. G. Miller, *et al.*, "Radiologists and clinical trials: Part 1 the truth about reader disagreements," *Therapeutic innovation & regulatory science*, vol. 55, pp. 1111–1121, 2021.
- [82] E. Merry, K. Thway, R. L. Jones, and P. H. Huang, "Predictive and prognostic transcriptomic biomarkers in soft tissue sarcomas," *npj Precision Oncology*, vol. 5, no. 1, p. 17, 2021.
- [83] Y. Y. Aung, D. C. Wong, and D. S. Ting, "The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare," *British medical bulletin*, vol. 139, no. 1, pp. 4–15, 2021.
- [84] S. Elia, E. Pompeo, A. Santone, *et al.*, "Radiomics and artificial intelligence can predict malignancy of solitary pulmonary nodules in the elderly," *Diagnostics*, vol. 13, no. 3, p. 384, 2023.
- [85] R. Casale, G. Varriano, A. Santone, *et al.*, "Predicting risk of metastases and recurrence in soft-tissue sarcomas via radiomics and formal methods," *JAMIA open*, vol. 6, no. 2, ooad025, 2023.
- [86] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001. DOI: 10.1023/A:1010950718922.
- [87] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, pp. 1–8, 2006.
- [88] G. Varriano, V. Nardone, S. Correra, F. Mercaldo, and A. Santone, "An automatic radiomic-based approach for disease localization: A pilot study on covid-19," *Computerized Medical Imaging and Graphics*, vol. 116, p. 102411, 2024.
- [89] S. Correra, G. Varriano, V. Nardone, and A. Santone, "Identifying ocular diseases through image processing: Pattern matching on single fundus images," in *2024 32nd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, IEEE, 2024, pp. 128–133.
- [90] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

- [91] P. A. M. Arias and J. A. Q. Sepulveda, "Deep learned vs. hand-crafted features for action classification," in *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE, 2018, pp. 170–171.
- [92] S. Alinsaif, "Covid-19 image classification: A comparative performance analysis of hand-crafted vs. deep features," *Computation*, vol. 12, no. 4, p. 66, 2024.
- [93] L. Rundo and C. Militello, "Image biomarkers and explainable ai: Hand-crafted features versus deep learned features," *European Radiology Experimental*, vol. 8, no. 1, p. 130, 2024.
- [94] J. Jovel and R. Greiner, "An introduction to machine learning approaches for biomedical research," *Frontiers in Medicine*, vol. 8, p. 771 607, 2021.
- [95] M. Recenti, C. Ricciardi, R. Aubonnet, *et al.*, "Toward predicting motion sickness using virtual reality and a moving platform assessing brain, muscles, and heart signals," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 635 661, 2021.
- [96] M. Recenti, P. Gargiulo, M. Chang, S. B. Ko, T. J. Kim, and S. U. Ko, "Predicting stroke, neurological and movement disorders using single and dual-task gait in korean older population," *Gait & posture*, vol. 105, pp. 92–98, 2023.
- [97] I. K. Nti, O. Nyarko-Boateng, J. Aning, *et al.*, "Performance of machine learning algorithms with different k values in k-fold cross-validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, 2021.
- [98] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [99] T. Nan, S. Zheng, S. Qiao, *et al.*, "Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis," *Nature Communications*, vol. 16, no. 1, p. 5493, 2025.
- [100] A. E. Gunnarsson, S. Correra, C. T. Sánchez, M. Recenti, H. Jónsson Jr, and P. Gargiulo, "Toward new assessment in sarcoma identification and grading using artificial intelligence techniques," *Diagnostics*, vol. 15, no. 13, p. 1694, 2025.
- [101] S. Gitto, M. Interlenghi, R. Cuocolo, *et al.*, "Mri radiomics-based machine learning for classification of deep-seated lipoma and atypical lipomatous tumor of the extremities," *La radiologia medica*, vol. 128, no. 8, pp. 989–998, 2023.

- [102] I. Malinauskaite, J. Hofmeister, S. Burgermeister, *et al.*, “Radiomics and machine learning differentiate soft-tissue lipoma and liposarcoma better than musculoskeletal radiologists,” *Sarcoma*, vol. 2020, no. 1, p. 7 163 453, 2020.
- [103] M. Dai, Y. Liu, Y. Hu, *et al.*, “Combining multiparametric mri features-based transfer learning and clinical parameters: Application of machine learning for the differentiation of uterine sarcomas from atypical leiomyomas,” *European Radiology*, vol. 32, no. 11, pp. 7988–7997, 2022.
- [104] Y. Gao, V. Ghodrati, A. Kalbasi, *et al.*, “Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion mri and a deep neural network with generative adversarial network-based data augmentation,” *Medical physics*, vol. 48, no. 6, pp. 3262–3372, 2021.
- [105] R. Vadisetty and H. Suyal, “Smart data augmentation using generative adversarial networks for rare oncological disease classification,” *International Journal on Smart & Sustainable Intelligent Computing*, vol. 2, no. 2, pp. 65–79, 2025.
- [106] R. Kumar, K. Sporn, A. Khanna, *et al.*, “Integrating radiogenomics and machine learning in musculoskeletal oncology care,” *Diagnostics*, vol. 15, no. 11, p. 1377, 2025.
- [107] A. Crombé, C. Lucchesi, F. Bertolo, *et al.*, “Integration of pre-treatment computational radiomics, deep radiomics, and transcriptomics enhances soft-tissue sarcoma patient prognosis,” *NPJ Precision Oncology*, vol. 8, no. 1, p. 129, 2024.
- [108] W. Kang, X. Qiu, Y. Luo, *et al.*, “Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis,” *Journal of Translational Medicine*, vol. 21, no. 1, p. 598, 2023.
- [109] M. E. Mayerhoefer, A. Materka, G. Langs, *et al.*, “Introduction to radiomics,” *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020, Epub 2020 Feb 14. DOI: 10.2967/jnumed.118.222893. [Online]. Available: <https://doi.org/10.2967/jnumed.118.222893>.
- [110] M. Vallières, E. Kay-Rivest, L. J. Perrin, *et al.*, “Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer,” *arXiv preprint arXiv:1703.08516*, 2017, Submitted March 24, 2017. DOI: 10.48550/arXiv.1703.08516. [Online]. Available: <https://arxiv.org/abs/1703.08516>.

- [111] G. Chiacchio, D. Castellani, C. Nedbal, *et al.*, “Radiomics vs radiologist in prostate cancer. results from a systematic review,” *World Journal of Urology*, vol. 41, no. 3, pp. 709–724, 2023, Published March 1, 2023, ISSN: 1433-8726. DOI: 10.1007/s00345-023-04305-2. [Online]. Available: <https://doi.org/10.1007/s00345-023-04305-2>.
- [112] T. Gangil, K. Sharan, B. D. Rao, K. Palanisamy, B. Chakrabarti, and R. Kadavigere, “Utility of adding radiomics to clinical features in predicting the outcomes of radiotherapy for head and neck cancer using machine learning,” *PLOS ONE*, vol. 17, no. 12, e0277168, Dec. 2022, Published December 15, 2022. DOI: 10.1371/journal.pone.0277168. [Online]. Available: <https://doi.org/10.1371/journal.pone.0277168>.
- [113] S. Rizzo, F. Botta, S. Raimondi, *et al.*, “Radiomics: The facts and the challenges of image analysis,” *European Radiology Experimental*, vol. 2, no. 1, p. 36, 2018, Published 14 November 2018. DOI: 10.1186/s41747-018-0068-z. [Online]. Available: <https://doi.org/10.1186/s41747-018-0068-z>.
- [114] A. Demircioğlu, “The effect of preprocessing filters on predictive performance in radiomics,” *European Radiology Experimental*, vol. 6, no. 1, p. 40, 2022, Published September 1, 2022, ISSN: 2509-9280. DOI: 10.1186/s41747-022-00294-w. [Online]. Available: <https://doi.org/10.1186/s41747-022-00294-w>.
- [115] M. Tomaszewski, S. Fan, A. Garcia, *et al.*, “Ai-radiomics can improve inclusion criteria and clinical trial performance,” *Tomography*, vol. 8, no. 1, pp. 341–355, Feb. 2022, Published 2022 Feb 2. DOI: 10.3390/tomography8010028. [Online]. Available: <https://doi.org/10.3390/tomography8010028>.
- [116] J. Wu, G. Cao, X. Sun, *et al.*, “Intratumoral spatial heterogeneity at perfusion mr imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy,” *Radiology*, vol. 288, no. 1, pp. 26–35, Jul. 2018. DOI: 10.1148/radiol.2018172462.
- [117] R. Wei, S. Lu, S. Lai, *et al.*, “A subregion-based radiofusionomics model discriminates between grade 4 astrocytoma and glioblastoma on multisequence mri,” *Journal of Cancer Research and Clinical Oncology*, vol. 150, no. 2, pp. 73–83, 2024, ISSN: 1432-1335. DOI: 10.1007/s00432-023-05603-3. [Online]. Available: <https://doi.org/10.1007/s00432-023-05603-3>.
- [118] L. X. Wu, N. Ding, Y. D. Ji, *et al.*, “Habitat analysis in tumor imaging: Advancing precision medicine through radiomic subregion segmentation,” *Cancer Management and Research*, vol. 17, pp. 731–741, Apr. 2025, ISSN: 1179-1322. DOI: 10.2147/CMAR.S511796.

- [119] H. Huang, H. Chen, D. Zheng, *et al.*, “Habitat-based radiomics analysis for evaluating immediate response in colorectal cancer lung metastases treated by radiofrequency ablation,” *Cancer Imaging*, vol. 24, no. 1, p. 44, Mar. 2024, ISSN: 1470-7330. DOI: 10.1186/s40644-024-00692-w.
- [120] S. Xing, C. R. Freeman, S. Jung, R. Turcotte, and I. R. Levesque, “Probabilistic classification of tumour habitats in soft tissue sarcoma,” *NMR in Biomedicine*, vol. 31, no. 11, e4000, 2018, e4000 NBM-18-0087.R1. DOI: <https://doi.org/10.1002/nbm.4000>. eprint: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.4000>. [Online]. Available: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/nbm.4000>.
- [121] J. H. George, *A brief history of machine learning*, <https://www.dataversity.net/a-brief-history-of-machine-learning/>, Accessed: 2025-07-22, 2021.
- [122] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. [Online]. Available: <http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning>.
- [123] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, M. J. Lee, and H. Asadi, “Edoctor: Machine learning and the future of medicine,” *Journal of Internal Medicine*, vol. 284, no. 6, pp. 603–619, 2018. DOI: <https://doi.org/10.1111/joim.12822>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/joim.12822>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/joim.12822>.
- [124] T. Zhao, H. Li, L. Zhang, *et al.*, “Prediction of levetiracetam serum concentration in uygur epilepsy children based on artificial neural network model,” *Chinese Modern Applied Pharmacy*, vol. 38, no. 22, pp. 2875–2880, 2021.
- [125] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1807.03876*, 2018. [Online]. Available: <https://arxiv.org/abs/1807.03876>.
- [126] M. Chatzianastasis, M. Vazirgiannis, and Z. Zhang, “Explainable multi-layer graph neural network for cancer gene prediction,” *arXiv preprint arXiv:2301.08831*, 2023, Submitted January 20 2023; revised May 3 2023. DOI: 10.48550/arXiv.2301.08831. [Online]. Available: <https://arxiv.org/abs/2301.08831>.
- [127] R. Dhudum, A. Ganeshpurkar, and A. Pawar, “Revolutionizing drug discovery: A comprehensive review of ai applications,” *Drugs and Drug Candidates*, vol. 3, no. 1, pp. 148–171, 2024, ISSN: 2813-2998. [Online]. Available: <https://www.mdpi.com/2813-2998/3/1/9>.

- [128] Z. Gao, Z. Dai, Z. Ouyang, *et al.*, “Radiomics analysis in differentiating osteosarcoma and chondrosarcoma based on t2-weighted imaging and contrast-enhanced t1-weighted imaging,” *Scientific Reports*, vol. 14, no. 1, p. 26 594, 2024, Epub 2024 Nov 4. DOI: 10 . 1038/s41598-024-78245-1. [Online]. Available: <https://doi.org/10.1038/s41598-024-78245-1>.
- [129] W. Xu, D. Hao, F. Hou, D. Zhang, and H. Wang, “Soft tissue sarcoma: Preoperative mri-based radiomics and machine learning may be accurate predictors of histopathologic grade,” *American Journal of Roentgenology*, vol. 215, no. 4, pp. 963–969, 2020, Epub ahead of print July 29, 2020. DOI: 10 . 2214/AJR. 19 . 22147.
- [130] L. Tian, D. Zhang, S. Bao, *et al.*, “Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas,” *Clinical Radiology*, vol. 76, no. 2, 158.e19–158.e25, 2021, Epub December 5, 2020. DOI: 10.1016/j . crad. 2020 . 08 . 038.
- [131] I. Cosci, P. Del Fiore, S. Mocellin, and A. Ferlin, “Gender differences in soft tissue and bone sarcoma: A narrative review,” *Cancers*, vol. 16, no. 1, p. 201, Dec. 2023, ISSN: 2072-6694. DOI: 10 . 3390/cancers16010201.
- [132] W. H. Organization, “Ethics and governance of artificial intelligence for health,” World Health Organization, Geneva, Switzerland, Tech. Rep., 2021. [Online]. Available: <https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content>.
- [133] W. Zhang, Y. Guo, and Q. Jin, “Radiomics and its feature selection: A review,” *Symmetry*, vol. 15, no. 10, 2023, ISSN: 2073-8994. DOI: 10 . 3390/sym15101834. [Online]. Available: <https://www.mdpi.com/2073-8994/15/10/1834>.
- [134] J. Xu, P. Qi, X. Ou, *et al.*, “Bi-modal ultrasound radiomics and habitat analysis enhanced the pre-operative prediction of axillary lymph node burden in patients with early-stage breast cancer,” *Frontiers in Oncology*, vol. 15, p. 1 607 442, May 2025, ISSN: 2234-943X. DOI: 10 . 3389 / fonc . 2025 . 1607442.
- [135] P. S. Foundation, *Python 3.9.23 documentation*, <https://docs.python.org/3.9/>, Python Software Foundation, 2024.
- [136] A. Chaddad, P. Daniel, and T. Niazi, “Radiomics evaluation of histological heterogeneity using multiscale textures derived from 3d wavelet transformation of multispectral images,” *Frontiers in Oncology*, vol. 8, p. 96, Apr. 2018. DOI: 10 . 3389/fonc . 2018 . 00096.

- [137] H. Bagher-Ebadian, F. Siddiqui, C. Liu, B. Movsas, and I. J. Chetty, "On the impact of smoothing and noise on robustness of ct and cbct radiomics features for patients with head and neck cancers," *Medical Physics*, vol. 44, no. 5, pp. 1755–1770, 2017. DOI: <https://doi.org/10.1002/mp.12188>. eprint: <https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12188>. [Online]. Available: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12188>.
- [138] Z. Ma, J. Zhang, X. Liu, *et al.*, "Comparative analysis of repeatability in ct radiomics and dosiomics features under image perturbation: A study in cervical cancer patients," *Cancers*, vol. 16, no. 16, 2024, ISSN: 2072-6694. DOI: 10.3390/cancers16162872. [Online]. Available: <https://www.mdpi.com/2072-6694/16/16/2872>.
- [139] A. Demircioğlu, "The effect of preprocessing filters on predictive performance in radiomics," *European Radiology Experimental*, vol. 6, no. 1, p. 40, Sep. 2022, ISSN: 2509-9280. DOI: 10.1186/s41747-022-00294-w. [Online]. Available: <https://doi.org/10.1186/s41747-022-00294-w>.
- [140] H. Erfankhah, M. Yazdi, M. Babaie, and H. R. Tizhoosh, "Heterogeneity-aware local binary patterns for retrieval of histopathology images," *IEEE Access*, vol. 7, pp. 18 354–18 367, 2019. DOI: 10.1109/ACCESS.2019.2897281.
- [141] H. Zhang, J. Liu, M. Dohopolski, Z. Wardak, R. Timmerman, and H. Peng, "Gradient-based radiomics for outcome prediction and decision-making in pulsar: A preliminary study," *International Journal of Particle Therapy*, vol. 15, p. 100 739, 2025, ISSN: 2331-5180. DOI: <https://doi.org/10.1016/j.ijpt.2025.100739>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2331518025000022>.
- [142] S. Volpe, L. J. Isaksson, M. Zaffaroni, *et al.*, "Impact of image filtering and assessment of volume-confounding effects on ct radiomic features and derived survival models in non-small cell lung cancer," *Translational Lung Cancer Research*, vol. 11, no. 12, pp. 2452–2463, Dec. 2022, ISSN: 2218-6751. DOI: 10.21037/tlcr-22-248.
- [143] H. A. Güvenir, "A classification learning algorithm robust to irrelevant features," Nov. 1998. DOI: 10.1007/BFb0057452.
- [144] E. Mylona, D. I. Zaridis, C. N. Kalantzopoulos, *et al.*, "Optimizing radiomics for prostate cancer diagnosis: Feature selection strategies, machine learning classifiers, and mri sequences," *Insights into Imaging*, vol. 15, no. 1, p. 265, Nov. 2024, ISSN: 1869-4101. DOI: 10.1186/s13244-024-01783-9.

- [145] I. Shiri, Y. Salimi, A. Saberi, *et al.*, “Diagnosis of covid-19 using ct image radiomics features: A comprehensive machine learning study involving 26,307 patients,” *medRxiv*, 2021. DOI: 10.1101/2021.12.07.21267367. eprint: <https://www.medrxiv.org/content/early/2021/12/08/2021.12.07.21267367.full.pdf>. [Online]. Available: <https://www.medrxiv.org/content/early/2021/12/08/2021.12.07.21267367>.
- [146] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/1705.07874>.
- [147] B. F. Darst, K. C. Malecki, and C. D. Engelman, “Using recursive feature elimination in random forest to account for correlated variables in high dimensional data,” *BMC Genetics*, vol. 19, no. 1, p. 65, Sep. 2018, ISSN: 1471-2156. DOI: 10.1186/s12863-018-0633-8. [Online]. Available: <https://doi.org/10.1186/s12863-018-0633-8>.
- [148] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, “A guide to cross-validation for artificial intelligence in medical imaging,” *Radiology: Artificial Intelligence*, vol. 5, no. 4, e220232, May 2023. DOI: 10.1148/ryai.220232.
- [149] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, *A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities (soft-tissue-sarcoma)[dataset]*, The Cancer Imaging Archive, 2015. DOI: 10.7937/K9/TCIA.2015.7G02GSKS. [Online]. Available: <https://doi.org/10.7937/K9/TCIA.2015.7G02GSKS>.
- [150] E. Rusak, L. Schott, R. S. Zimmermann, *et al.*, “A simple way to make neural networks robust against diverse image corruptions,” in *European Conference on Computer Vision (ECCV)*, 2020, pp. 1–17. [Online]. Available: https://www.ecva.net/papers/eccv%5C_2020/papers%5C_ECCV/papers/123480052.pdf.
- [151] R. H. Hashemi, W. G. Bradley, and C. J. Lisanti, *MRI: The Basics*, 4th ed. Lippincott Williams & Wilkins, 2022, ISBN: 9781975181702.
- [152] L. G. Nyúl, J. K. Udupa, and X. Zhang, “New variants of a method of mri scale standardization,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, Feb. 2000, ISSN: 0278-0062. DOI: 10.1109/42.836373.
- [153] G. Siedel, W. Shao, S. Vock, and A. Morozov, “Investigating the corruption robustness of image classifiers with random p-norm corruptions,” *arXiv preprint arXiv:2305.05400*, 2023.

- [154] Z. Chen, Z. Zhong, X. Pan, and X. Xi, "A novel improved slic superpixel segmentation algorithm," in *Proceedings of the IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 2022.
- [155] F. He, M. P. Mahmud, A. Z. Kouzani, A. Anwar, F. Jiang, and S. H. Ling, "An improved slic algorithm for segmentation of microscopic cell images," *Biomedical Signal Processing and Control*, 2022. DOI: 10.1016/j.bspc.2022.103458. [Online]. Available: <https://doi.org/10.1016/j.bspc.2022.103458>.
- [156] E. G. Troost, J. Bussink, A. L. Hoffmann, O. C. Boerman, W. J. Oyen, and J. H. Kaanders, "18f-flt pet/ct for early response monitoring and dose escalation in oropharyngeal tumors," *Journal of Nuclear Medicine*, vol. 51, no. 6, pp. 866–874, 2010, ISSN: 0161-5505. DOI: 10.2967/jnumed.109.069310. eprint: <https://jnm.snmjournals.org/content/51/6/866.full.pdf>. [Online]. Available: <https://jnm.snmjournals.org/content/51/6/866>.
- [157] J. Jiarpakdee, C. Tantithamthavorn, and C. Treude, "The impact of automated feature selection techniques on the interpretation of defect models," *Empirical Software Engineering*, vol. 25, no. 3, pp. 2156–2194, 2020. DOI: 10.1007/s10664-019-09782-5. [Online]. Available: <https://ctreude.ca/wp-content/uploads/2020/06/emse20a.pdf>.
- [158] S. Correria, A. E. Gunnarsson, M. Recenti, *et al.*, "An explainable radiomics-based classification model for sarcoma diagnosis," *Diagnostics*, vol. 15, no. 16, 2025, ISSN: 2075-4418. DOI: 10.3390/diagnostics15162098. [Online]. Available: <https://www.mdpi.com/2075-4418/15/16/2098>.
- [159] J. Bleker, C. Roest, D. Yakar, H. Huisman, and T. C. Kwee, "The effect of image resampling on the performance of radiomics-based artificial intelligence in multicenter prostate mri," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 5, pp. 1800–1806, 2024. DOI: <https://doi.org/10.1002/jmri.28935>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.28935>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.28935>.
- [160] S. M. Malakouti, M. B. Menhaj, A. A. Suratgar, *et al.*, "Applying grid search, random search, bayesian optimization, genetic algorithm, and particle swarm optimization to fine-tune the hyperparameters of the ensemble of ml models enhances its predictive accuracy for mud loss," *Research Square*, Dec. 2024, Preprint, Version 1. DOI: 10.21203/rs.3.rs-5187887/v1. [Online]. Available: <https://doi.org/10.21203/rs.3.rs-5187887/v1>.
- [161] Commission Internationale de l'Éclairage (CIE), *Colorimetry* (CIE Publication No. 15), 1st. Vienna, Austria: CIE, 1976.

Appendix A

Image Transforms

Image transforms are the backbone of the feature generation step.

Wavelet Transform

The general wavelet transform is defined by equation A.1.

$$F(a, b) = \int_{-\inf}^{\inf} f(x)\psi^*(x)dx \quad (\text{A.1})$$

In equation A.1, * represents the complex conjugate and ψ is some function that follows certain rules. The specific wavelet function used is the discrete Haar wavelet transform which is given by a sequence of equations. First the high- and low-pass filters are defined as

Low-pass filter

$$h = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right] \quad (\text{A.2})$$

High-pass filter

$$g = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right] \quad (\text{A.3})$$

Example: low-pass filter applied to an arbitrary row

$$h_{row,i} = \frac{f(row, 2(i-1)) + f(row, 2i)}{\sqrt{2}} \quad (\text{A.4})$$

Where $f(i,j)$ represent the pixel value in position (i, j) and $h_{row,i}$ represents the i^{th} value of the low-pass filter vector if an arbitrary row. Similarly for the high-pass

filter

$$g_{row,i} = \frac{f(row, 2(i-1)) - f(row, 2i)}{\sqrt{2}} \quad (A.5)$$

The results from the first pass vertical pass of the filter is then applied horizontally through the columns. Therefore, the resulting transformed image contains 1/4 of the total pixels of the original image. The combination of high- and low-passes therefore give 4 different quarter size images: high horizontal - high vertical, low horizontal - low vertical, high horizontal - low vertical and low -horizontal - high vertical, when considering 2-dimensional images.

Laplacian of Gaussian (LoG) Transform

The LoG transform is applied on an image stack, therefore given a pixel located in position (x,y,z) of the stack and a smoothing parameter σ the LoG transform is defined as

$$\nabla^2 G(x, y, z; \sigma) = \left(\frac{x^2 + y^2 + z^2 - 3\sigma^2}{\sigma^4} \right) \cdot G(x, y, z; \sigma) \quad (A.6)$$

where the Gaussian kernel $G(x, y, z; \sigma)$ is given by

$$G(x, y, z; \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^3} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}} \quad (A.7)$$

Square Transform

The square transform of a pixel x is given by

$$f(x) = (cx)^2 \quad (A.8)$$

Where

$$c = \frac{1}{\sqrt{\max(|x|)}} \quad (A.9)$$

Here $\max(|x|)$ denotes the value of the pixel with the highest intensity of the image.

Square Root transform

The square root transform of a pixel with intensity x

$$f(x) = \begin{cases} \sqrt{cx} & \text{for } x \geq 0 \\ -\sqrt{-cx} & \text{for } x < 0 \end{cases} \quad (A.10)$$

Where

$$c = \max(|x|) \quad (A.11)$$

Here $\max(|x|)$ denotes the maximum intensity of the transforming image.

Logarithm Transform

Given a pixel with intensity x of an image with highest intensity $\max(|x|)$ the logarithm transform is defined by

$$f(x) = \begin{cases} c \log(x + 1) & \text{for } x \geq 0 \\ -c \log(-x + 1) & \text{for } x < 0 \end{cases}, \text{ where } c = \frac{\max(|x|)}{\log(\max(|x|) + 1)} \quad (\text{A.12})$$

Exponential Transform

Given a pixel with intensity x of an image with highest intensity $\max(|x|)$ the exponential transform is defined by

$$f(x) = e^{cx}, \text{ where } c = \frac{\log(\max(|x|))}{\max(|x|)} \quad (\text{A.13})$$

Gradient Transform

Given a pixel in position (x, y, z) and its intensity $f(x, y, z)$, then the gradient transform of an image is given by

$$|\nabla f(x, y, z)| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2} \quad (\text{A.14})$$

Where in the discrete format the partials are defined as

$$\frac{\partial f}{\partial x} \approx \frac{f(x + 1, y, z) - f(x - 1, y, z)}{2} \quad (\text{A.15})$$

and is similarly defined for y and z .

Local Binary Pattern (LBP2D) Transform

The local binary pattern (LBP) transform is a method for encoding local texture of an image through comparison of intensities in a fixed window, typically a 3×3 window.

For a central pixel with intensity I_c , and its surrounding P neighbors I_p , the binary comparison is defined as:

$$LBP(x, y) = \sum_{p=0}^{P-1} s(I_p - I_c) \cdot 2^p \quad (\text{A.16})$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (\text{A.17})$$

Each comparison yields a binary digit, and the pattern formed is converted to a decimal number by weighting each binary value with 2^p . The result is a local texture descriptor value assigned to the central pixel.

For example, consider a 3×3 window:

$$\begin{bmatrix} 50 & 55 & 52 \\ 49 & \mathbf{53} & 54 \\ 51 & 50 & 48 \end{bmatrix}$$

Comparing each neighbor to the central pixel (53), the resulting binary pattern is:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & - & 1 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow 01010000_2 = 80_{10}$$

Where the subscripts denote the base system of the number: 2 for binary, 10 for a base 10 system. The pattern of collecting the binary numbers is as follows:

$$\begin{bmatrix} P_0 & P_1 & P_2 \\ P_7 & C & P_3 \\ P_6 & P_5 & P_4 \end{bmatrix} \Rightarrow P_0, P_1, P_2, P_3, P_4, P_5, P_6, P_7$$

Where P_i denotes a 1 or 0 determined by the intensity comparison. This LBP is calculated for every pixel except border pixels, as a center pixel on a border will not fulfill the required window size needed for the calculations. The resulting LBP image contains the LBP code for each non-border pixel and serves as a compact texture representation. LBP2D is especially useful for capturing micro-texture patterns in medical images and is invariant to monotonic grayscale changes.

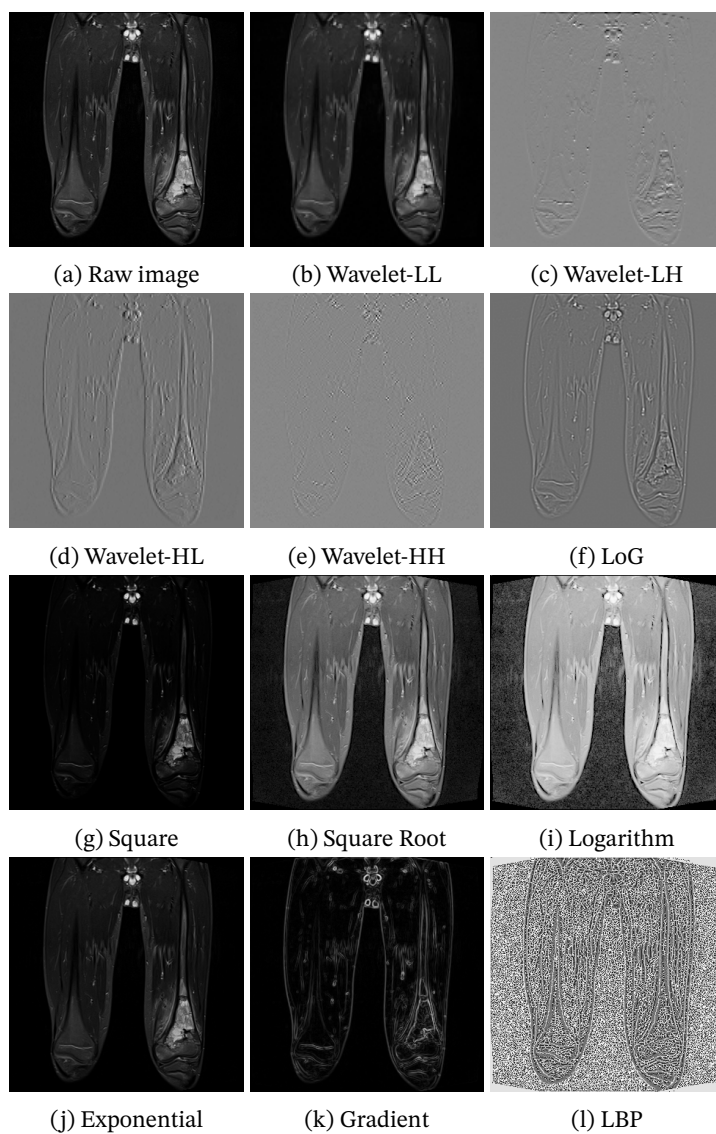


Figure A.1: An example of all transforms applied to a single medical image slice of sarcoma.

Appendix B

Simple Linear Iterative Clustering (SLIC)

SLIC is an algorithm used for segmenting an image or a mask on an image into multiple sub regions based on five dimensions: 3 color channels, and 2 spatial dimensions. The cluster search is limited by parameters to keep its search local, making it a very efficient algorithm for generating super pixels (SP). Originally developed by Achanta et al. in 2012 [53] it has seen much use and success in the field of computer vision.

The algorithm takes in as an input a value K which represents the number of desired SPs from the image. With an image with a total of N pixels, this gives each SP a size of approximately N/K pixels and also defines an approximate center of each SP at $S = \sqrt{N/K}$ intervals in both spatial dimensions for the image. The 5-dimensional vector representing each pixel is given by

$$C_i = [l_i, a_i, b_i, x_i, y_i]$$

Where the $[l, a, b]$ coordinates are a way to represent colors and intensity using 3 coordinates [161], and $[x, y]$ are the positions of each pixel. The SLIC algorithm defines a 5D color-spatial distance D_s as

$$\begin{aligned} d_{lab} &= \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \\ d_{xy} &= \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \\ D_s &= d_{lab} + \frac{m}{S} d_{xy} \end{aligned}$$

Where the parameter m controls the weight given the spatial component the distance measure, a low m will group pixels of similar color together, whereas a

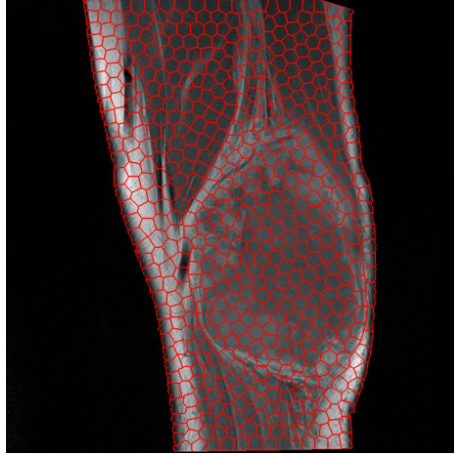


Figure B.1: Example of the product of the SLIC algorithm applied to a medical image

higher m will create more uniform (square) super pixels, this parameter is referred to as the compactness factor. The subscripts i in the equation represents an arbitrary pixel in the region of the cluster K .

The algorithm begins by initializing cluster centers $C_k = [l_k, a_k, b_k, x_k, y_k]$ at regular intervals (S). Since, it is favorable to have intense edges in the image as the boundary of SPs, the relative color gradients are calculated in a small $n \times n$ region around the center C_k , where $n \approx \text{few pixels}$, and the spatial center is moved to the position of the lowest gradient. Then for each cluster center, calculate the best matching pixels in a $2S \times 2S$ distance around the cluster center using D_s and recompute the new cluster centers and the residual error given by the L1 distance between current and previous centers. This is repeated until a stable center is found, or when the residual error is less than a defined threshold. Connectivity is always enforced, meaning that each cluster produces only a single region. Figure B.1 displays the end result of the slic algorithm applied to a medical image used in these works, the red boundaries represents the boundaries of each SP.

Appendix C

Results

C.1 Figures of Results

C.1.1 Tissue Classification

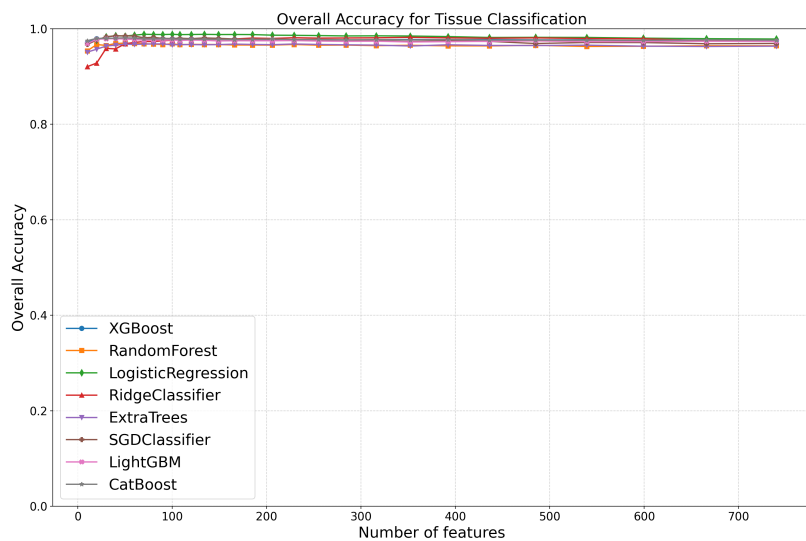


Figure C.1: Tissue classification accuracy: SvAH dataset

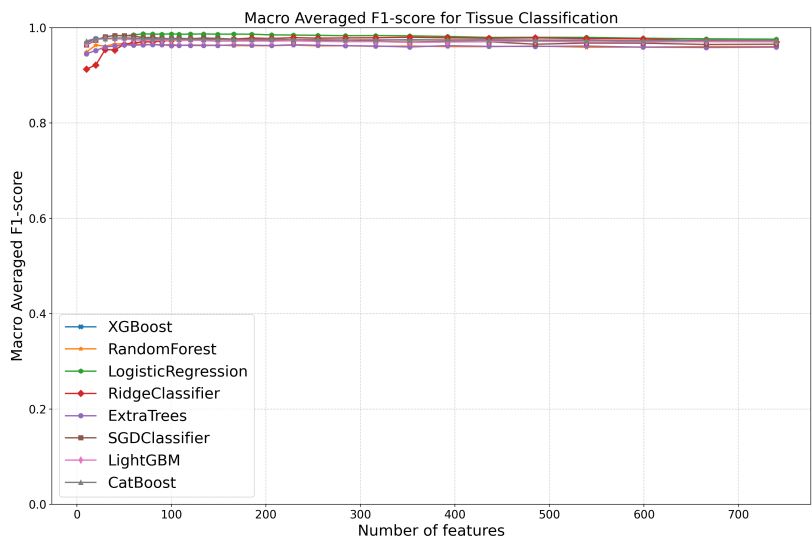


Figure C.2: Tissue classification macro F1-score: SvAH dataset

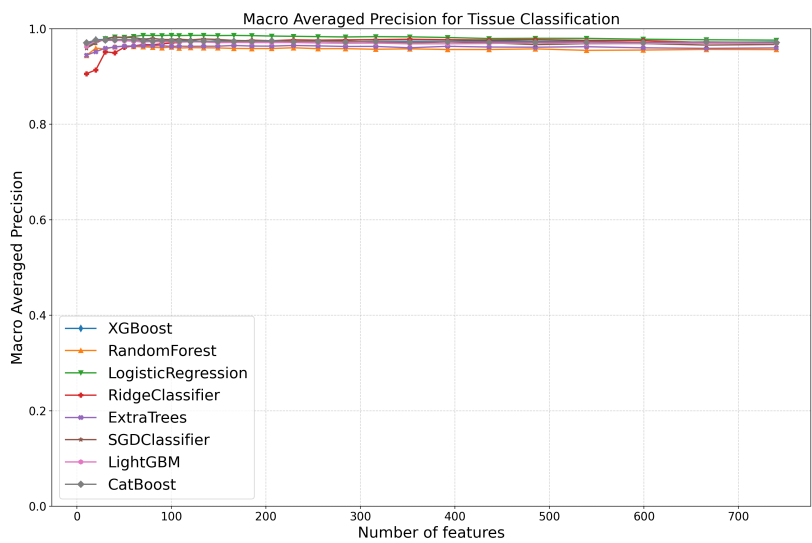


Figure C.3: Tissue classification macro precision: SvAH dataset

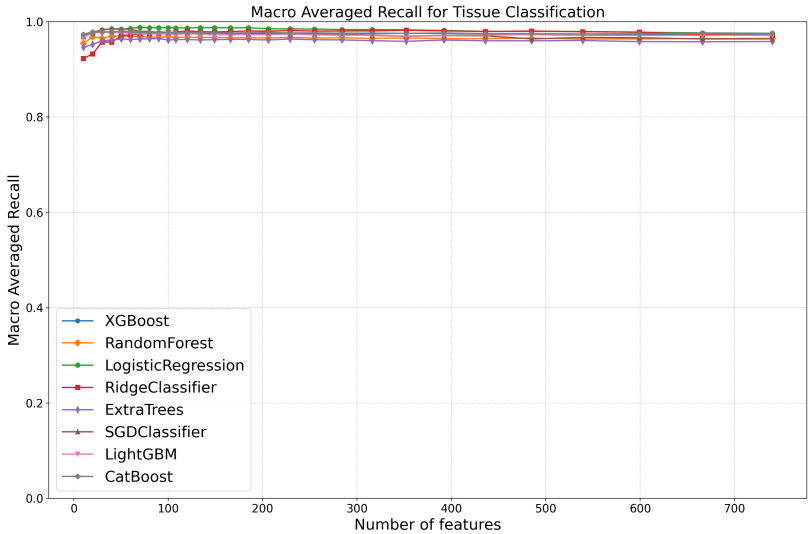


Figure C.4: Tissue classification macro recall: SvAH dataset

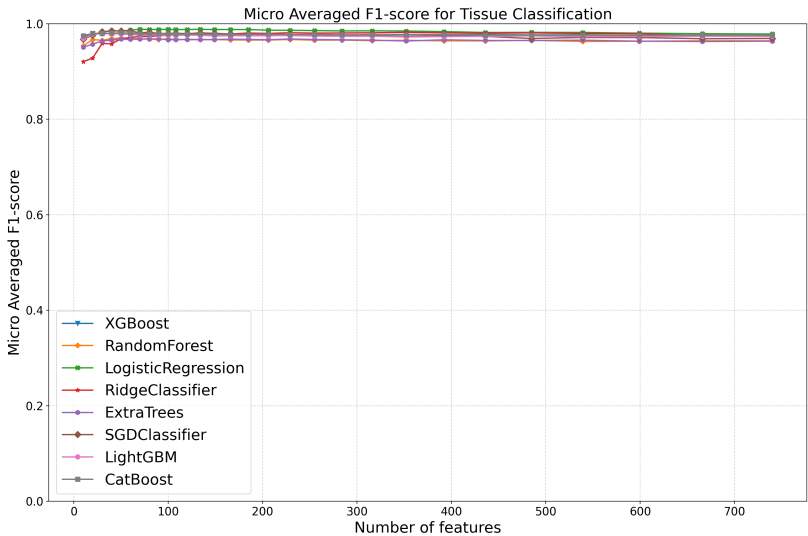


Figure C.5: Tissue classification micro F1-score: SvAH dataset

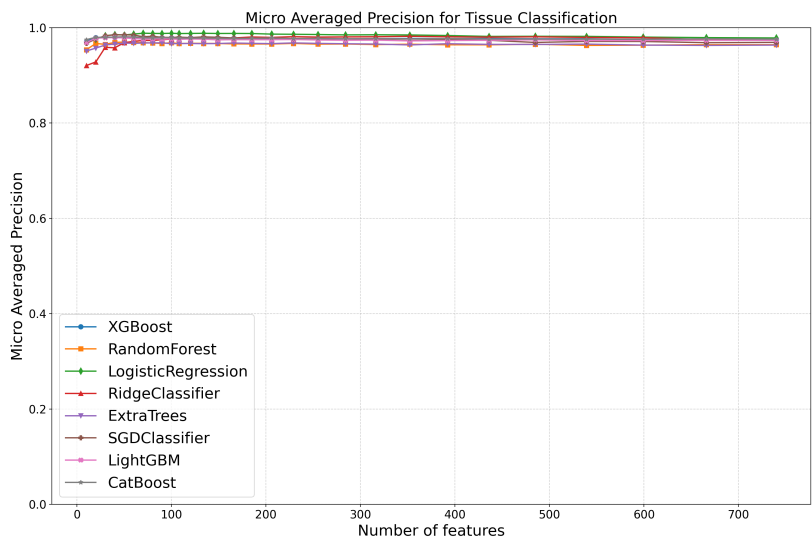


Figure C.6: Tissue classification micro precision: SvAH dataset

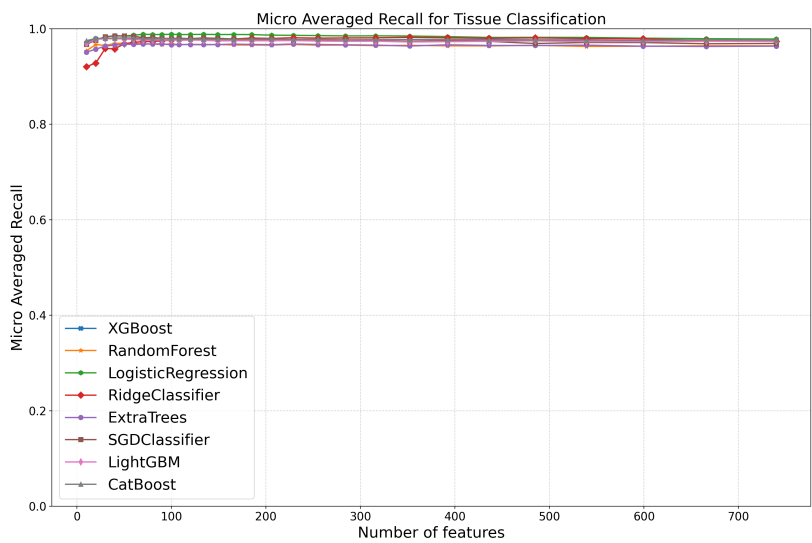


Figure C.7: Tissue classification micro recall: SvAH dataset

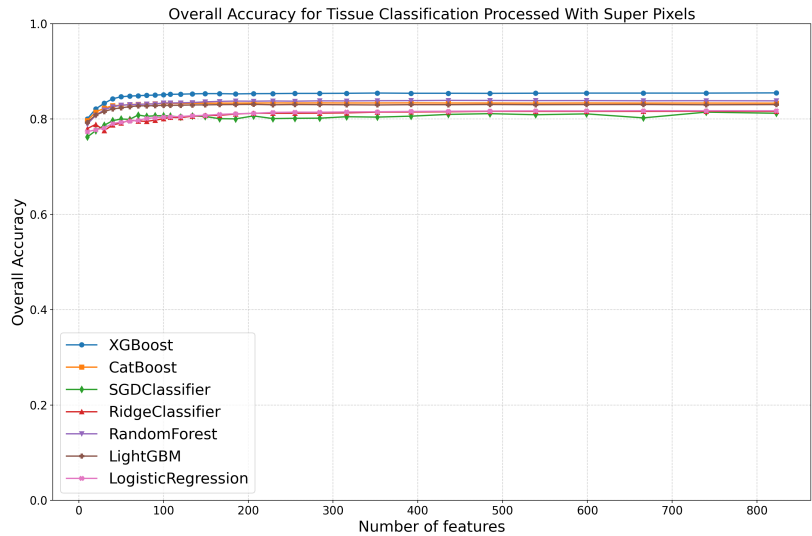


Figure C.8: Tissue classification accuracy: SvAH-SP dataset

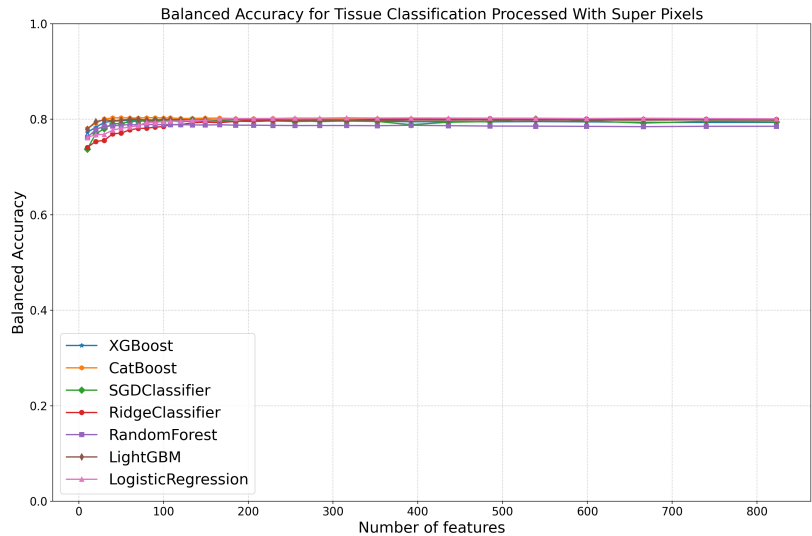


Figure C.9: Tissue classification balanced accuracy: SvAH-SP dataset

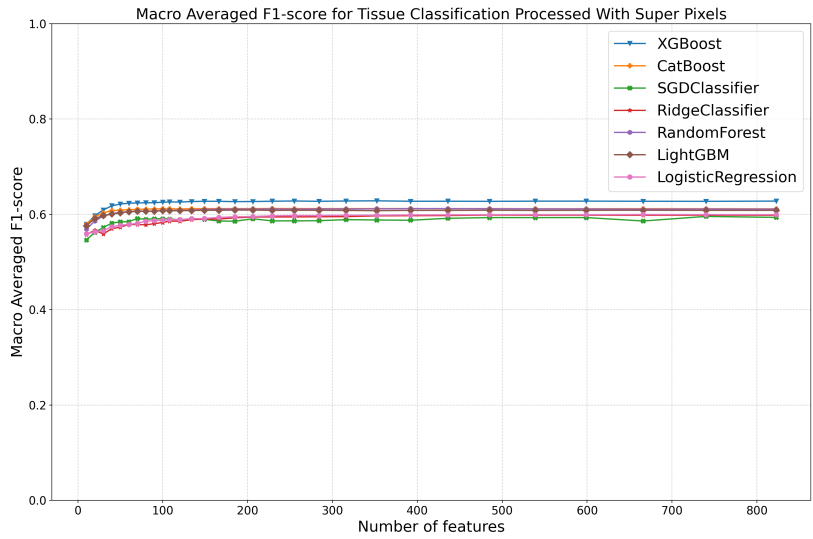


Figure C.10: Tissue classification macro F1-score: SvAH-SP dataset

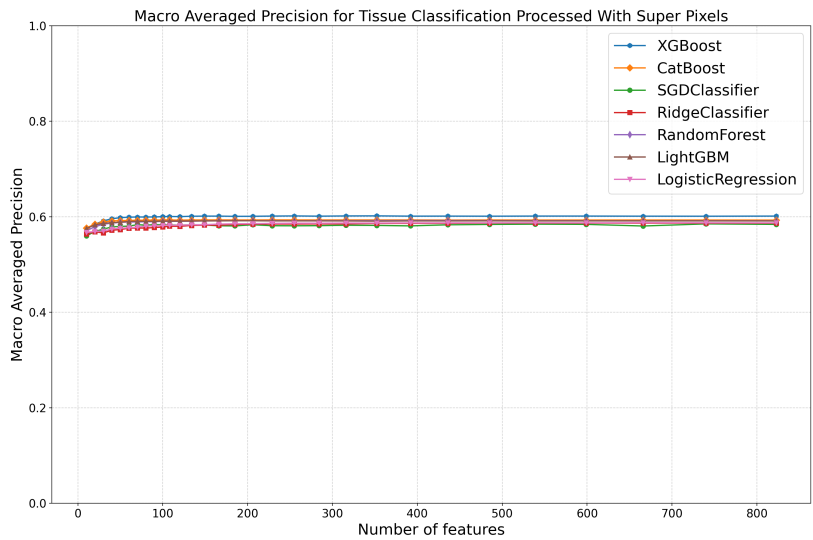


Figure C.11: Tissue classification macro precision: SvAH-SP dataset

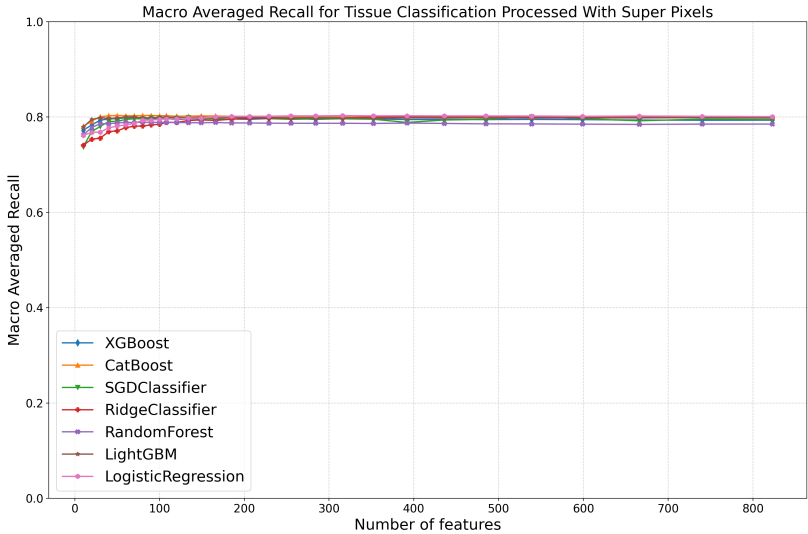


Figure C.12: Tissue classification macro recall: SvAH-SP dataset

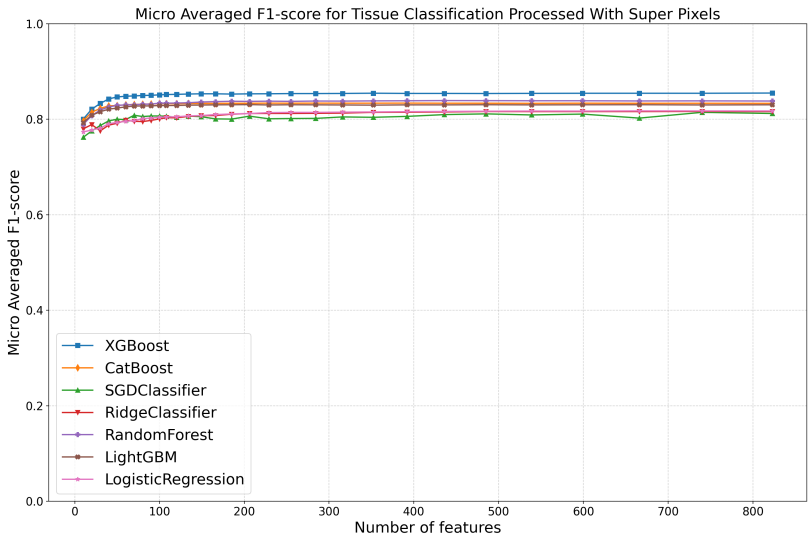


Figure C.13: Tissue classification micro F1-score: SvAH-SP dataset

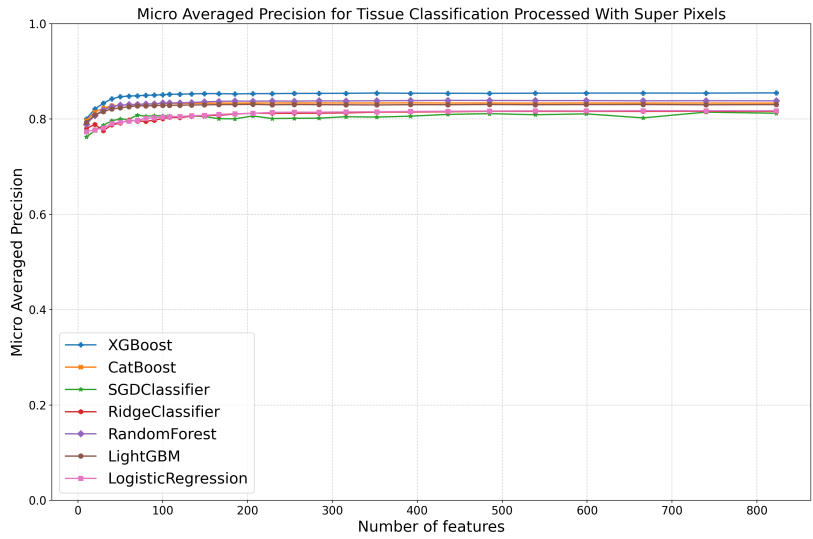


Figure C.14: Tissue classification micro precision: SvAH-SP dataset

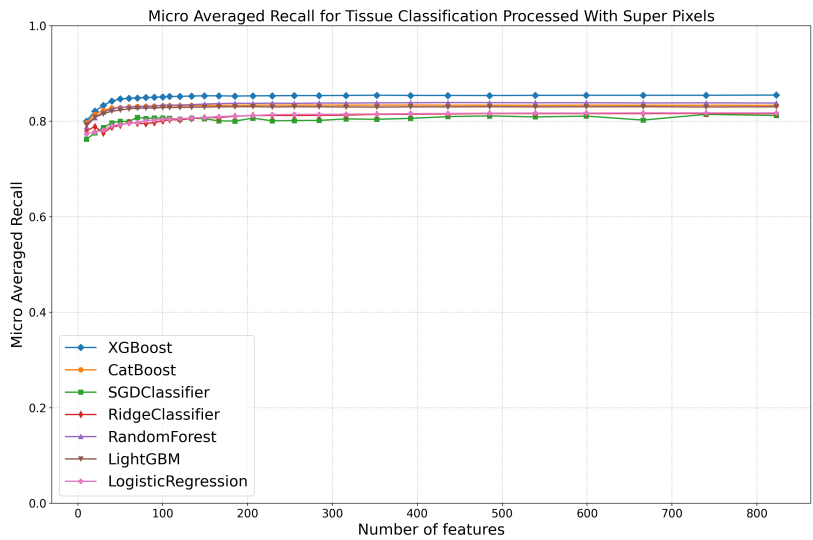


Figure C.15: Tissue classification micro recall: SvAH-SP dataset

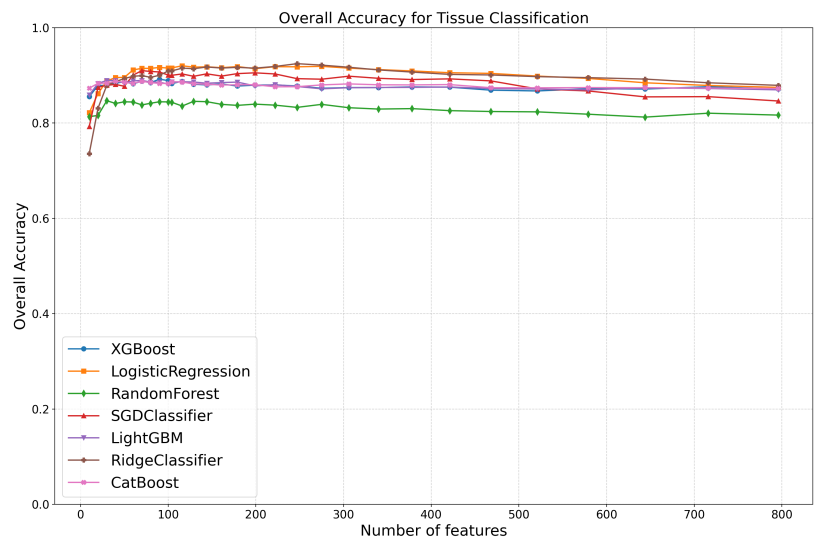


Figure C.16: Tissue classification accuracy: SvCH dataset

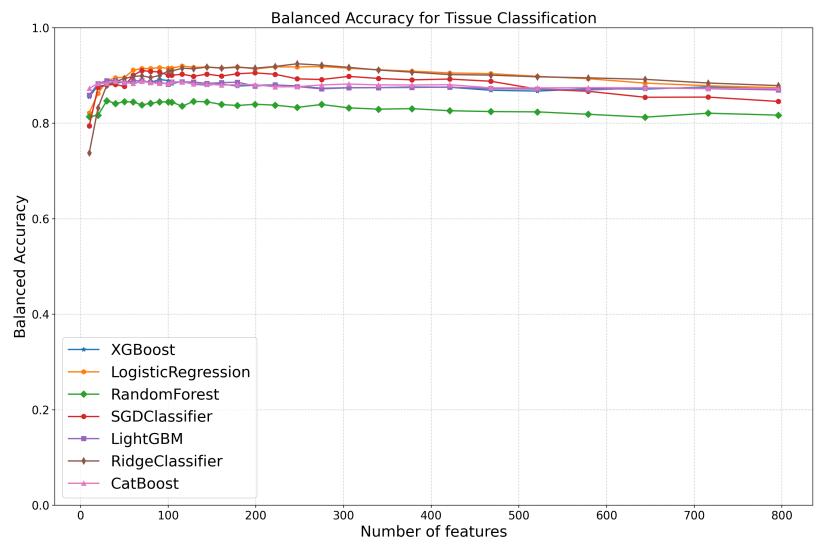


Figure C.17: Tissue classification balanced accuracy: SvCH dataset

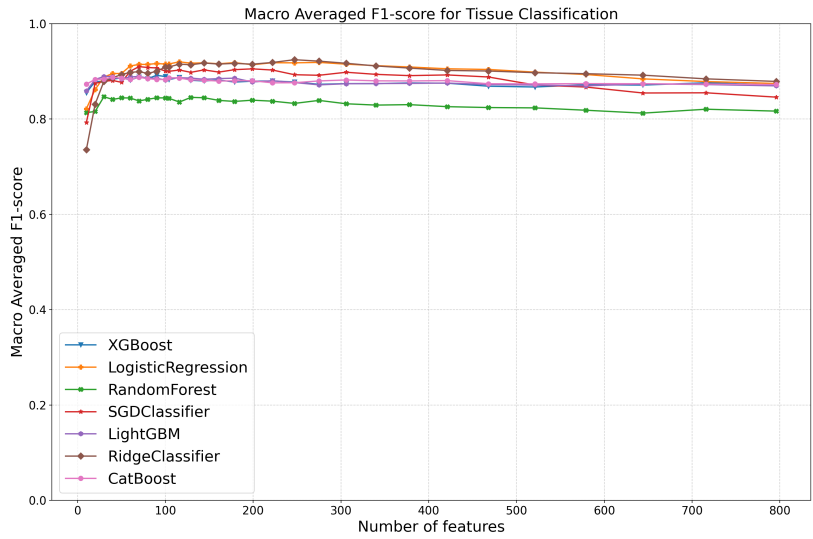


Figure C.18: Tissue classification macro F1-score: SvCH dataset

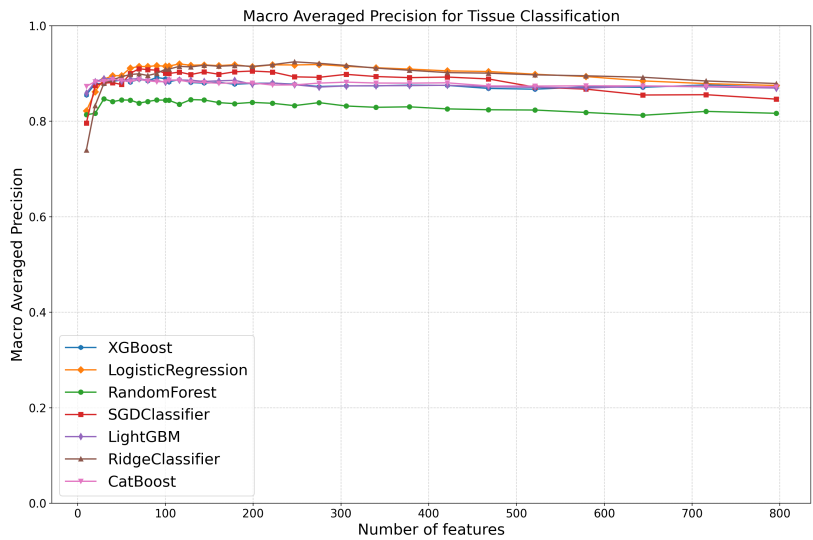


Figure C.19: Tissue classification macro precision: SvCH dataset

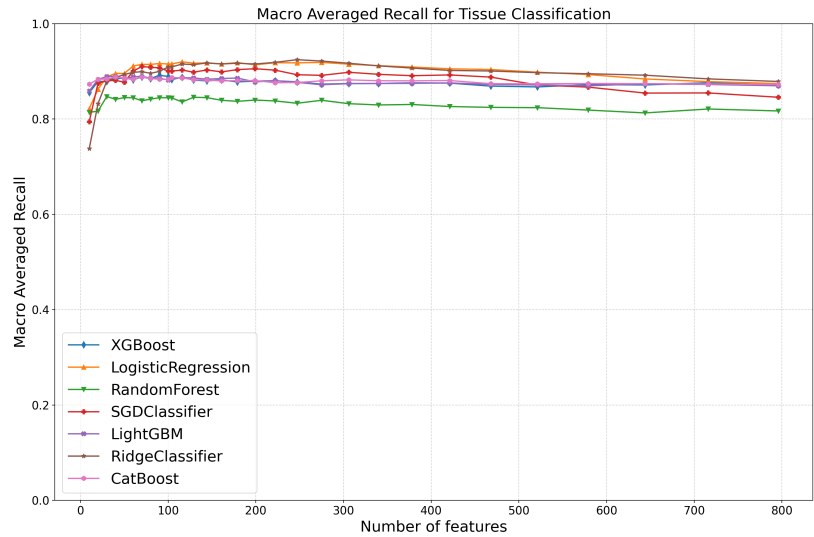


Figure C.20: Tissue classification macro recall: SvCH dataset

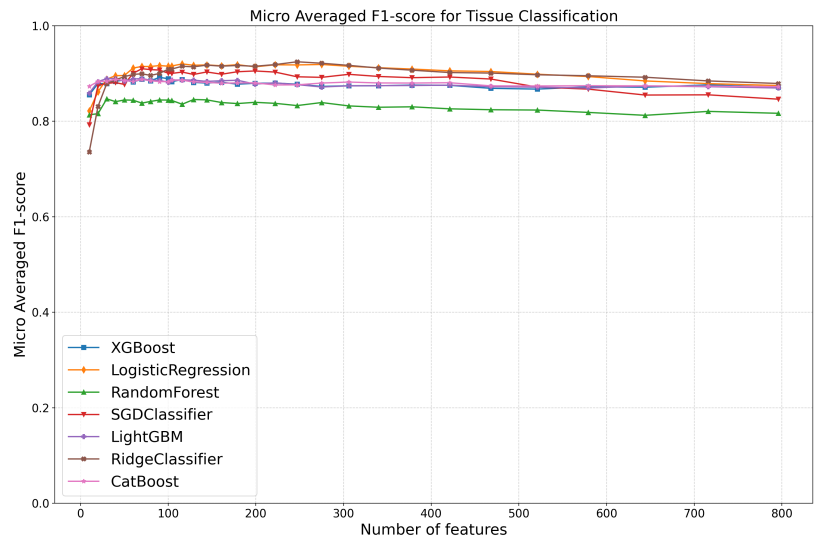


Figure C.21: Tissue classification micro F1-score: SvCH dataset

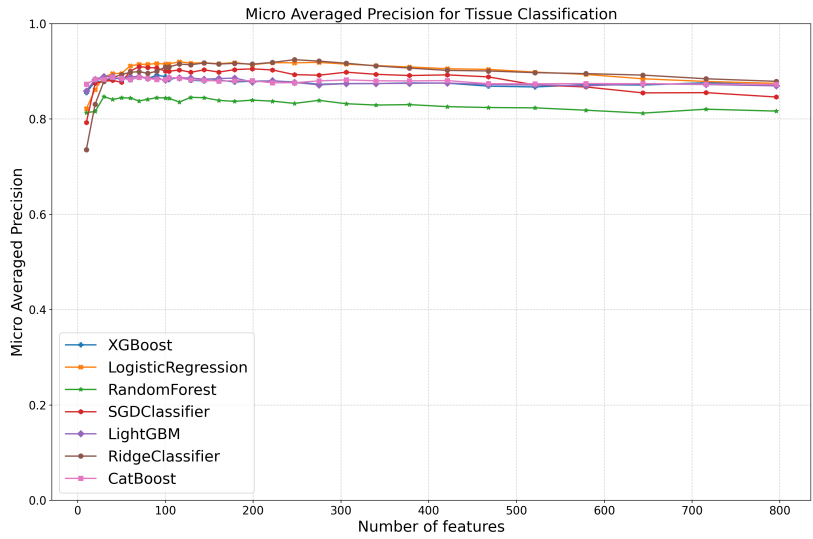


Figure C.22: Tissue classification micro precision: SvCH dataset

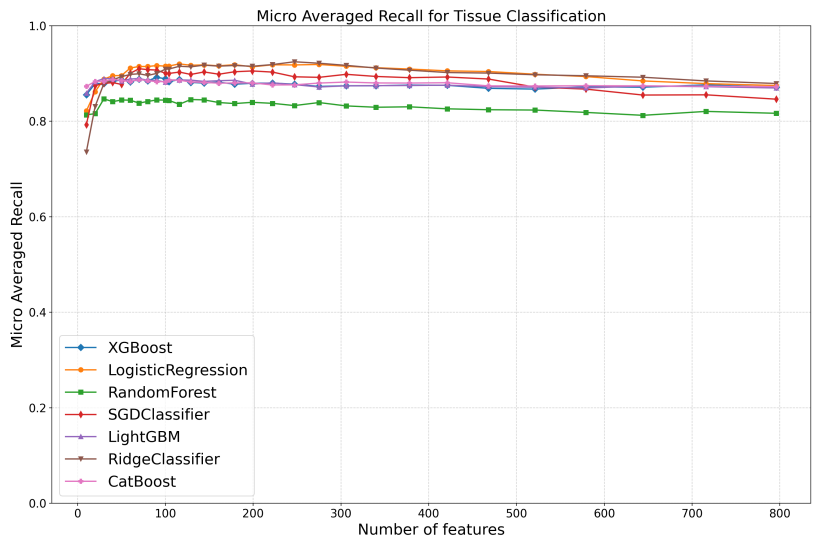


Figure C.23: Tissue classification micro recall: SvCH dataset

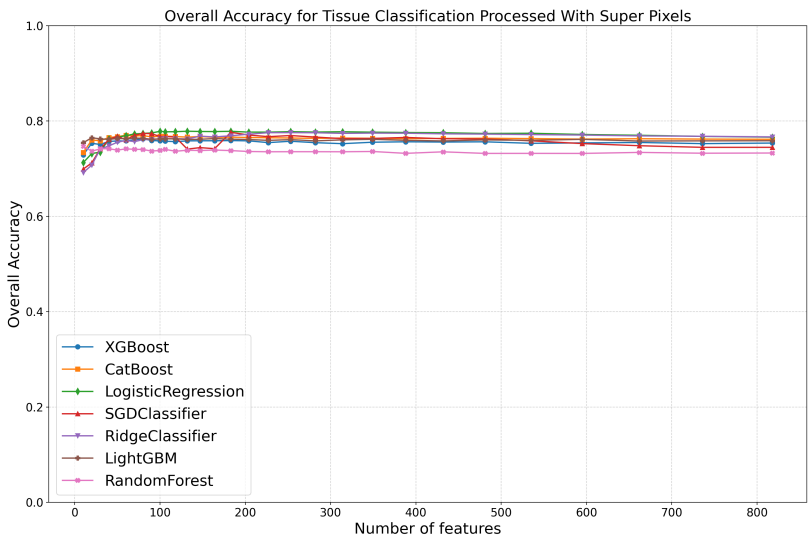


Figure C.24: Tissue classification accuracy: SvCH-SP dataset

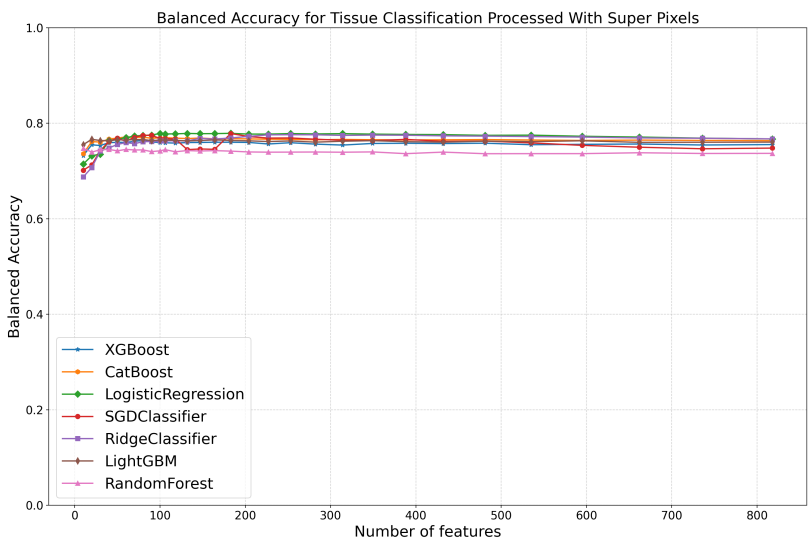


Figure C.25: Tissue classification balanced accuracy: SvCH-SP dataset

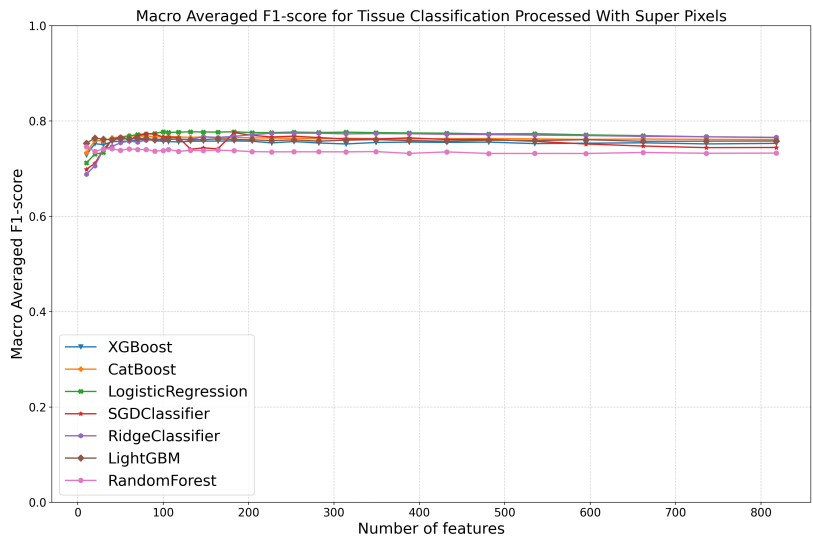


Figure C.26: Tissue classification macro F1-score: SvCH-SP dataset

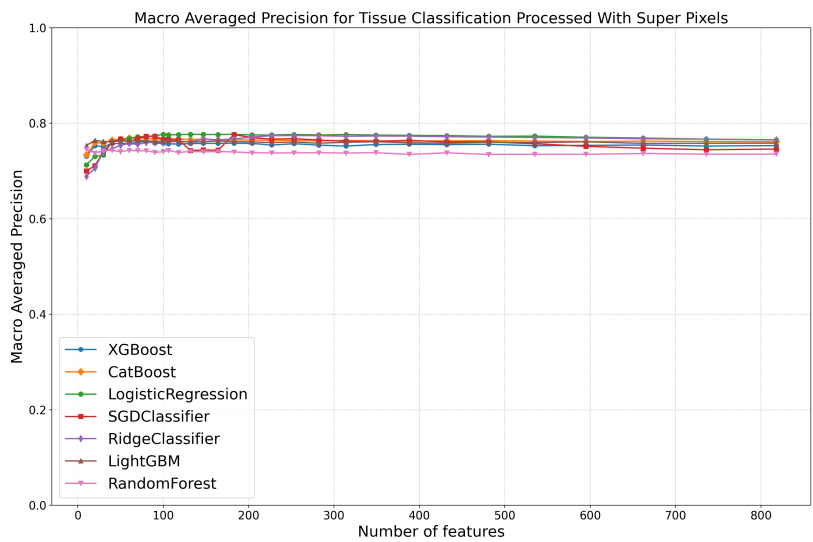


Figure C.27: Tissue classification macro precision: SvCH-SP dataset

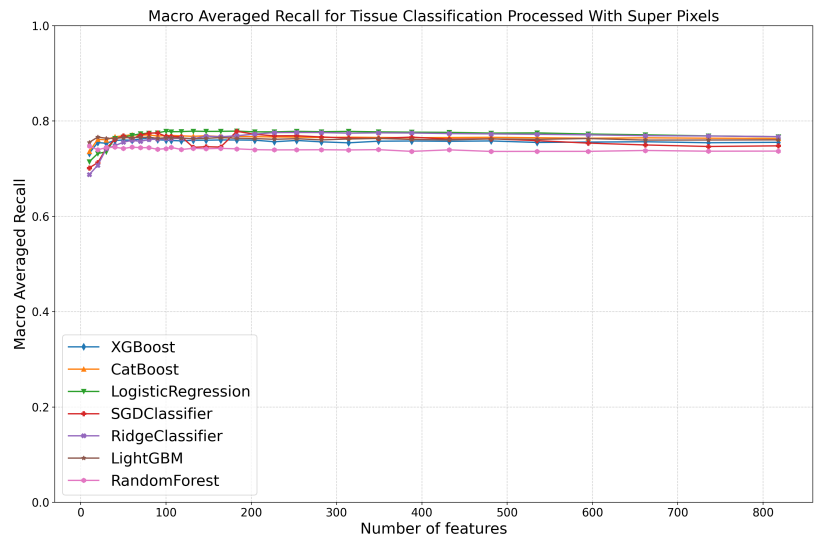


Figure C.28: Tissue classification macro recall: SvCH-SP dataset

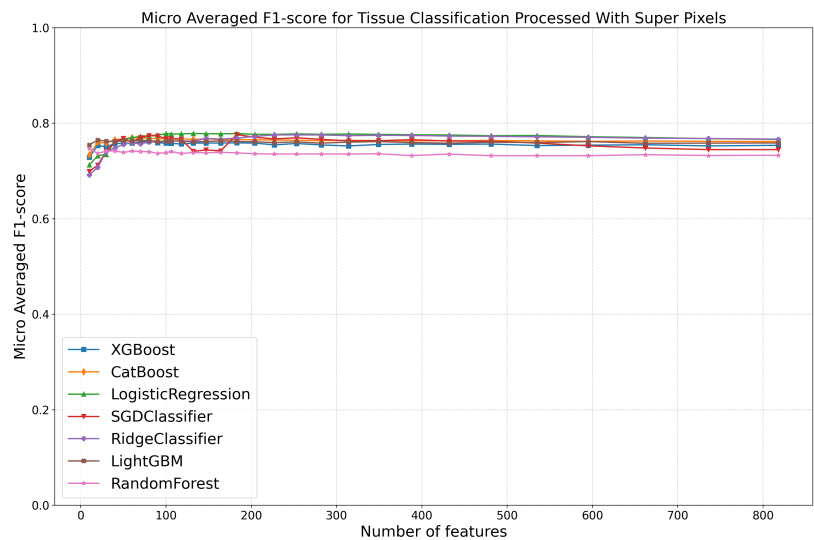


Figure C.29: Tissue classification micro F1-score: SvCH-SP dataset

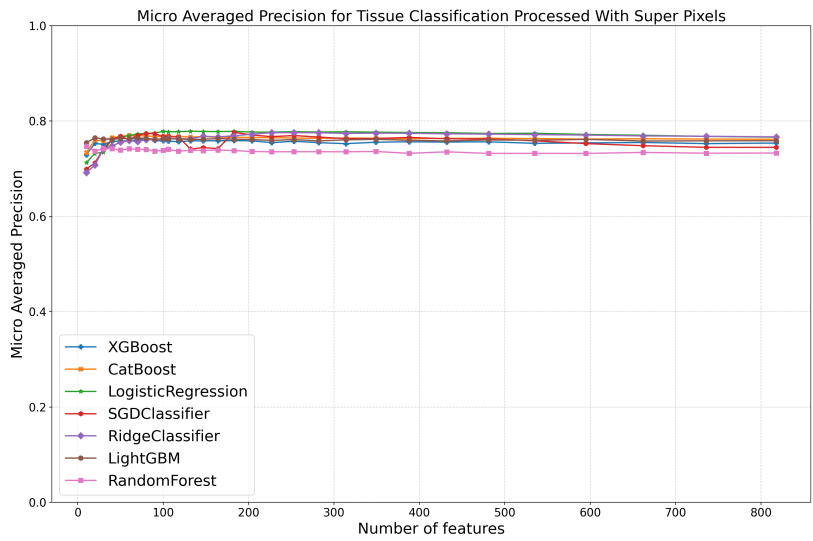


Figure C.30: Tissue classification micro precision: SvCH-SP dataset

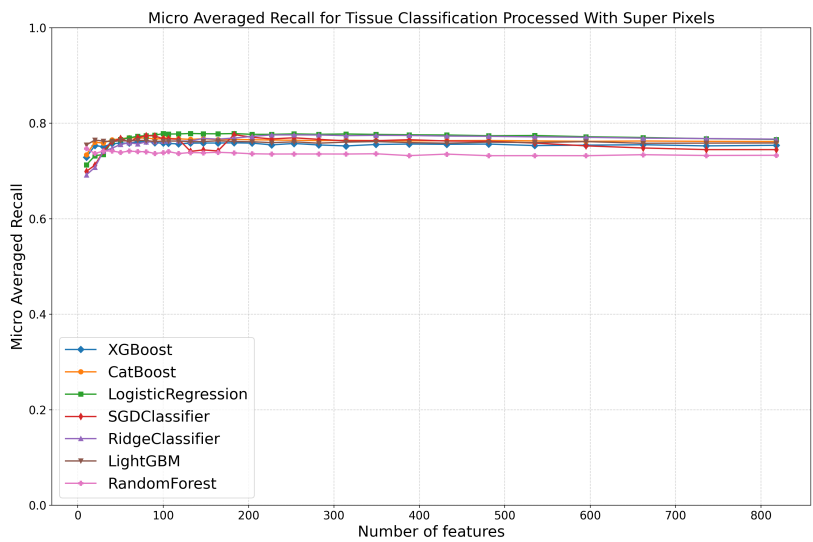


Figure C.31: Tissue classification micro recall: SvCH-SP dataset

C.1.2 Grade Classification

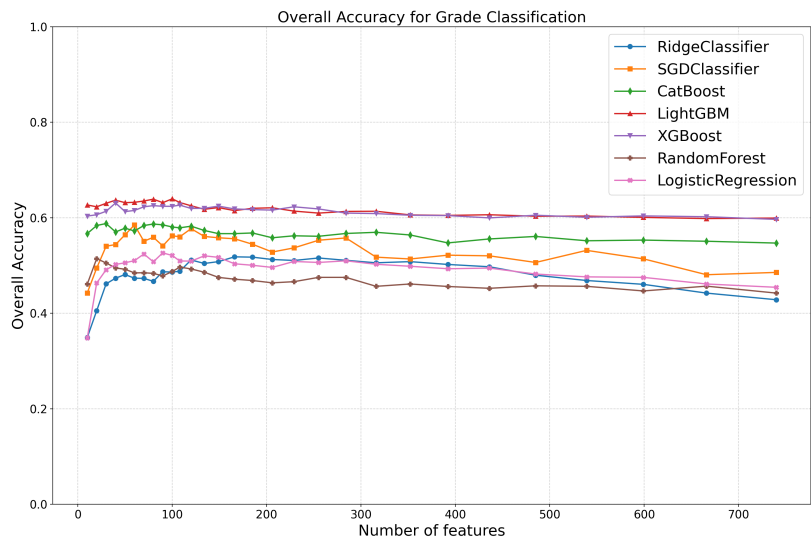


Figure C.32: Grade classification accuracy: G dataset

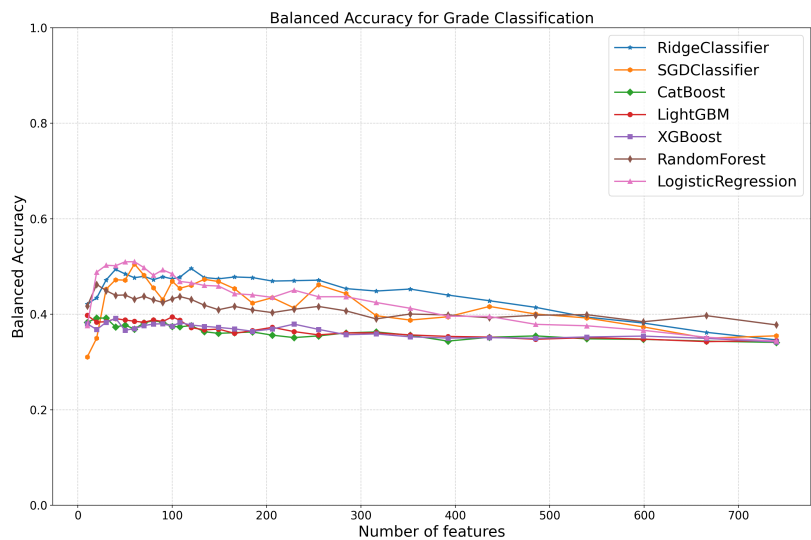


Figure C.33: Grade classification balanced accuracy: G dataset

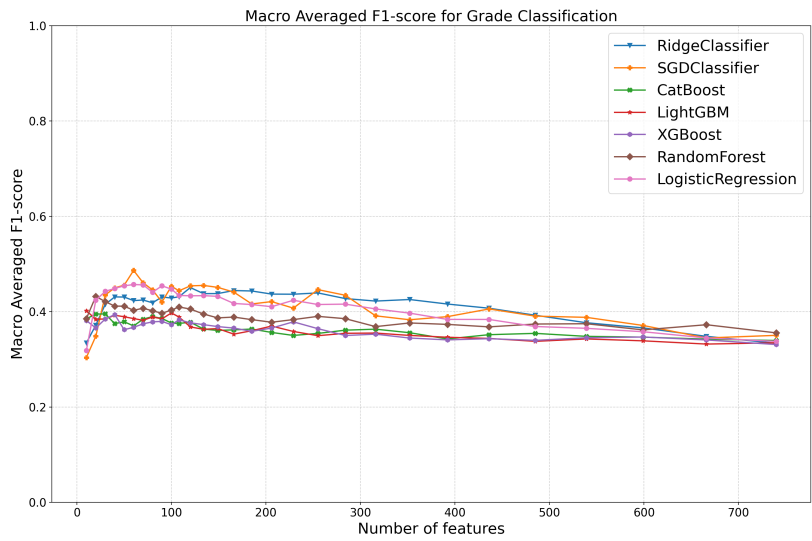


Figure C.34: Grade classification macro F1-score: G dataset

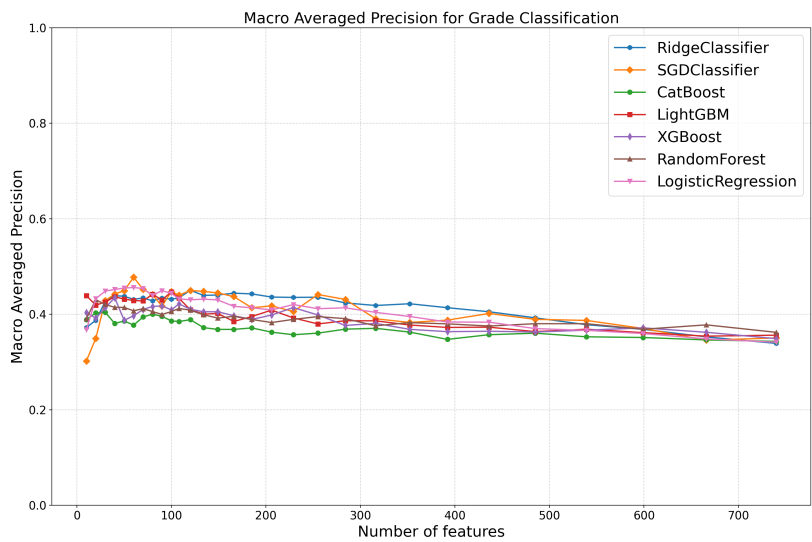


Figure C.35: Grade classification macro precision: G dataset

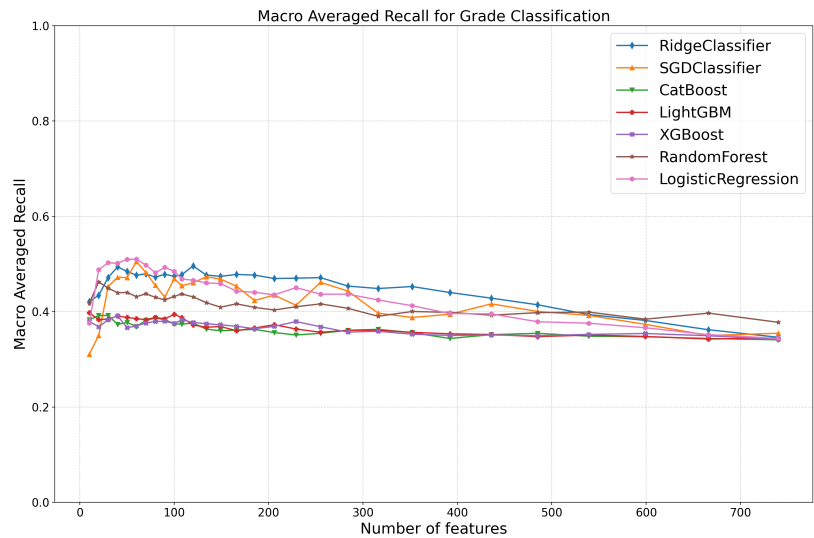


Figure C.36: Grade classification macro recall: G dataset

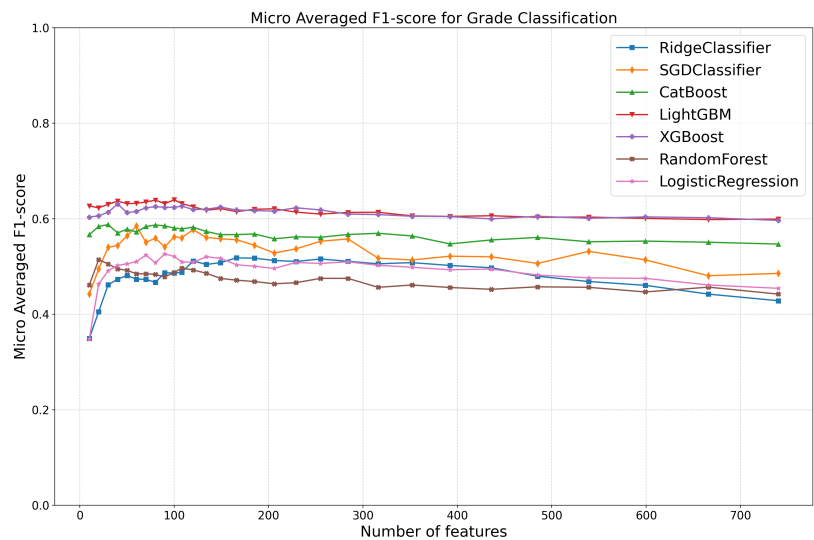


Figure C.37: Grade classification micro F1-score: G dataset

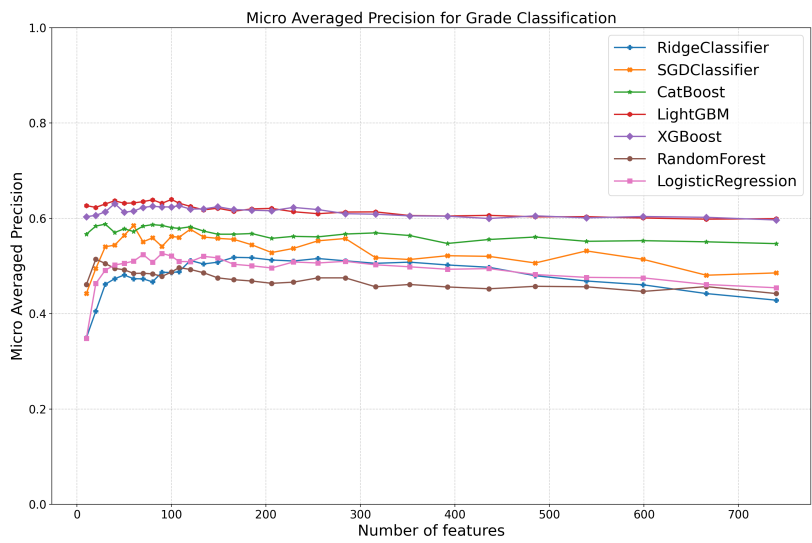


Figure C.38: Grade classification micro precision: G dataset

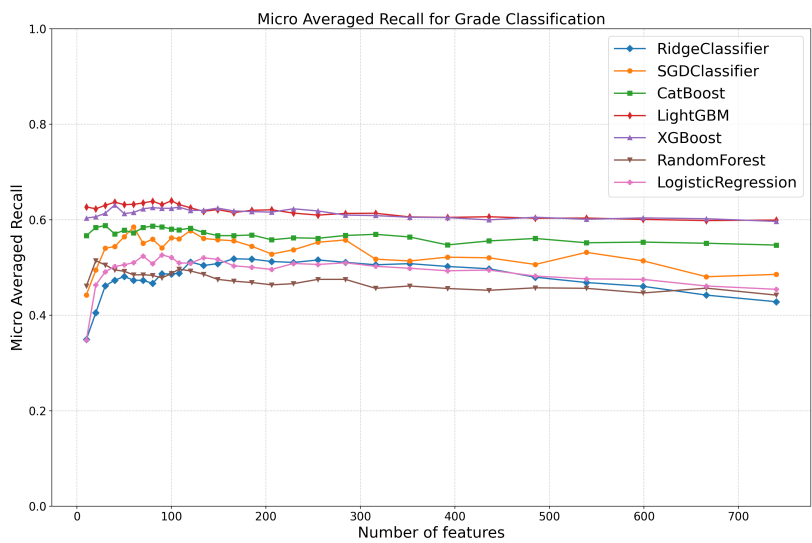


Figure C.39: Grade classification micro recall: G dataset

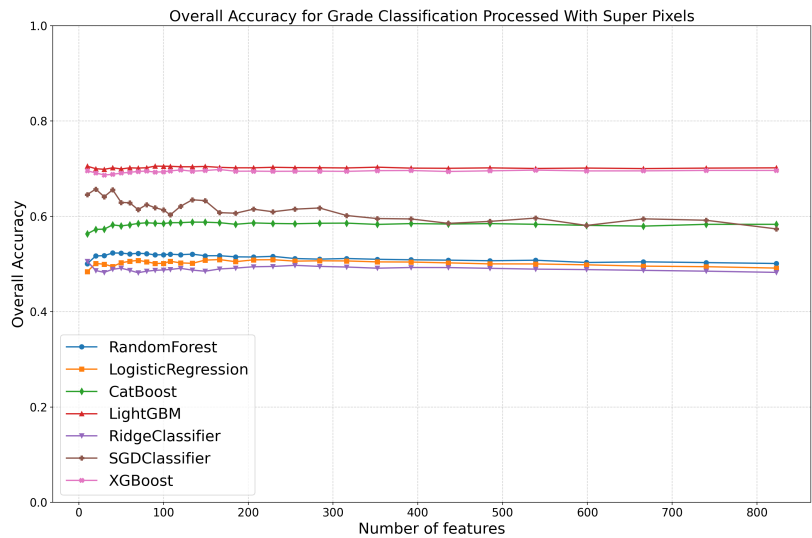


Figure C.40: Grade classification accuracy: G-SP dataset

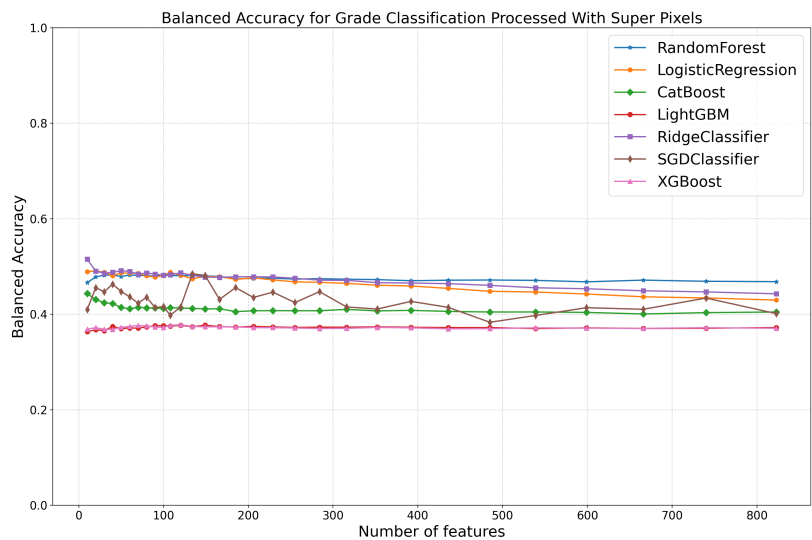


Figure C.41: Grade classification balanced accuracy: G-SP dataset

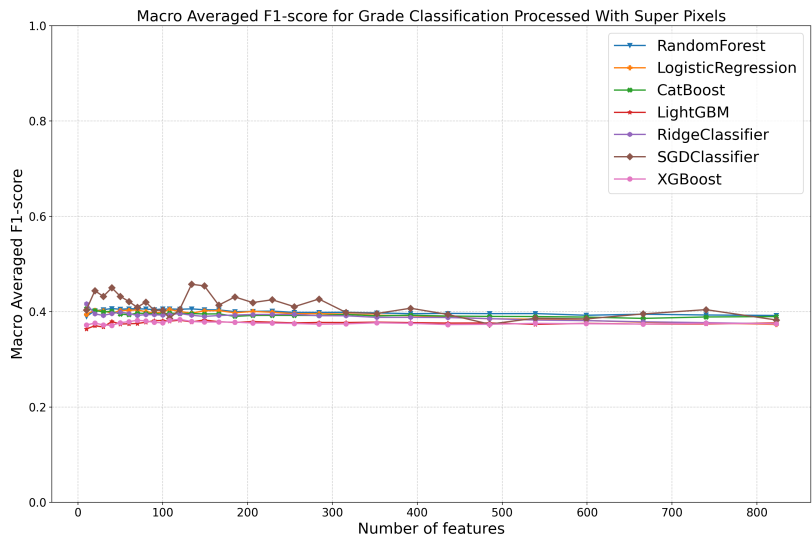


Figure C.42: Grade classification macro F1-score: G-SP dataset

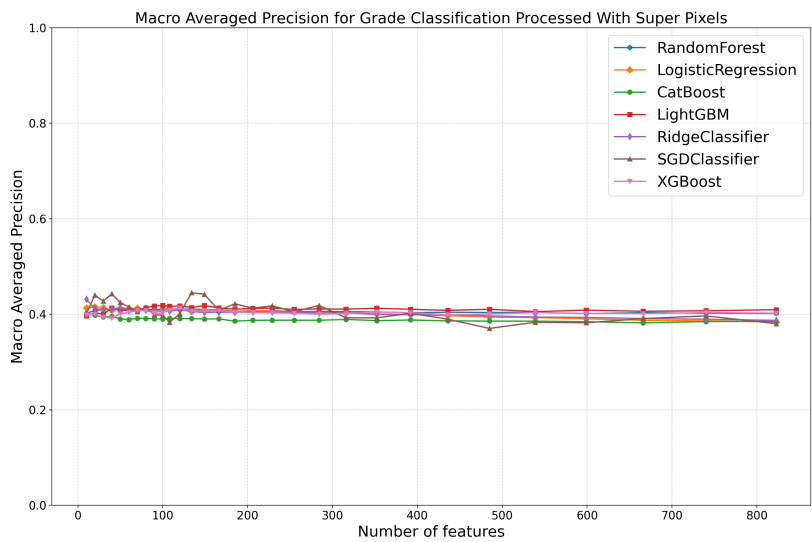


Figure C.43: Grade classification macro precision: G-SP dataset

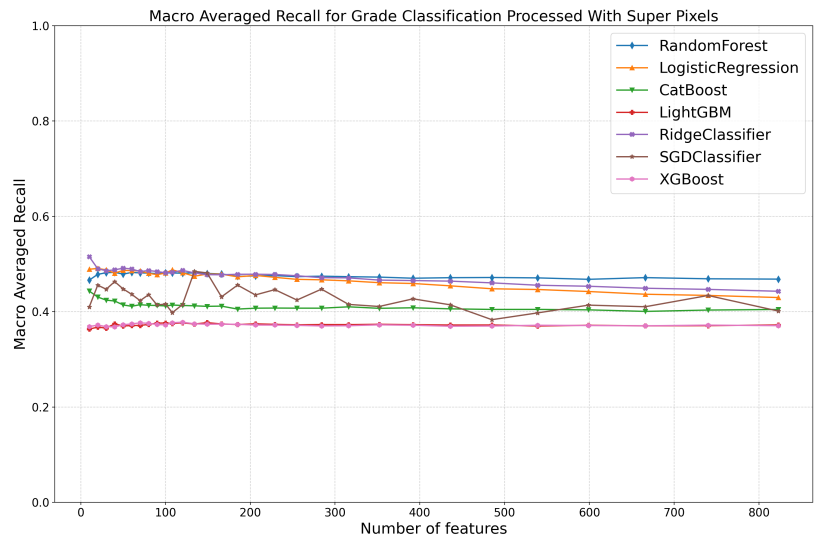


Figure C.44: Grade classification macro recall: G-SP dataset

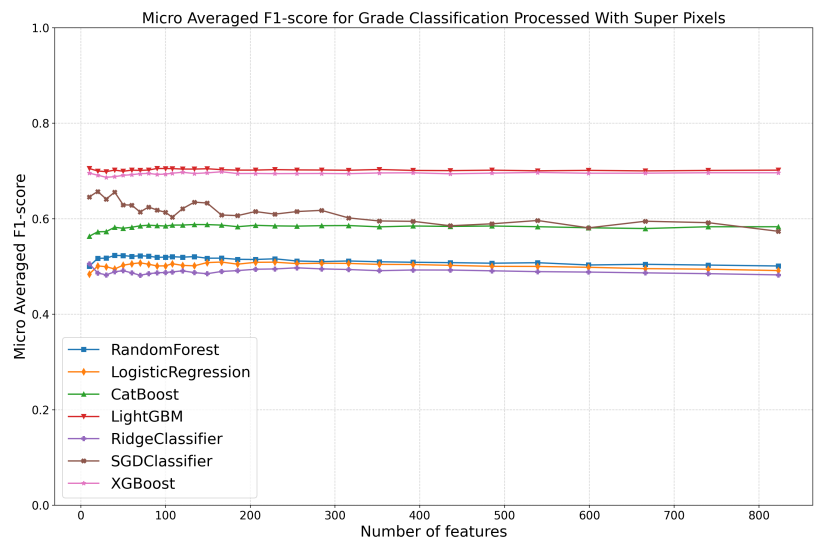


Figure C.45: Grade classification micro F1-score: G-SP dataset

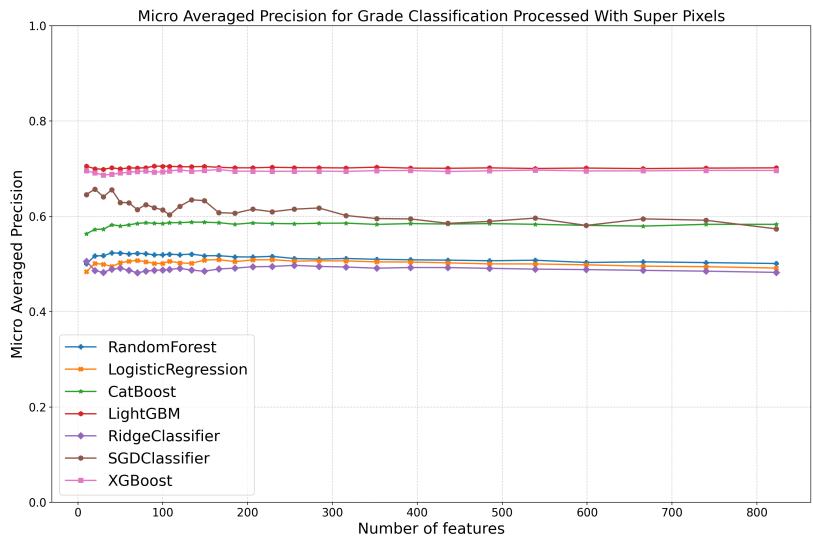


Figure C.46: Grade classification micro precision: G-SP dataset

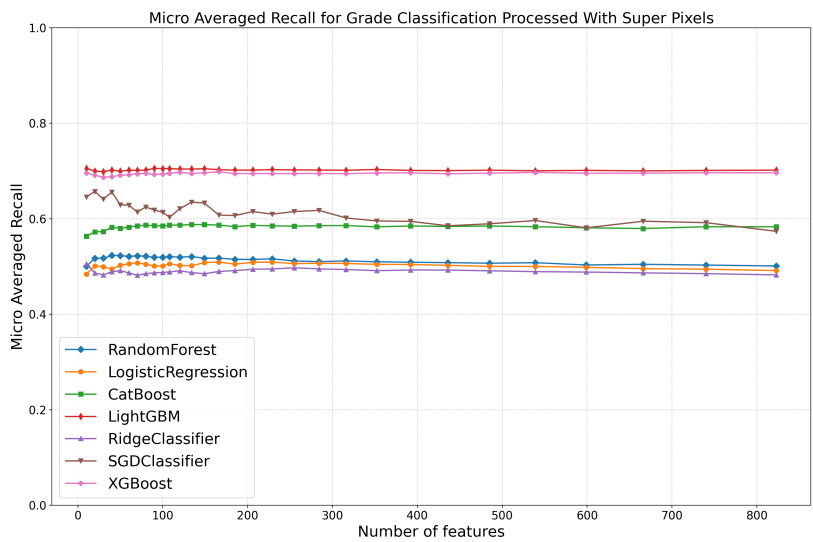


Figure C.47: Grade classification micro recall: G-SP dataset

C.1.3 Sub-Group Classification

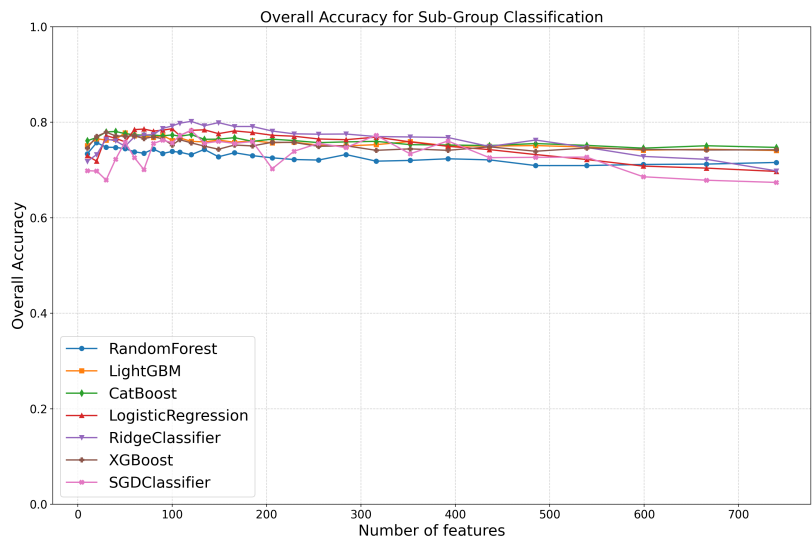


Figure C.48: Sub-group classification accuracy: SG dataset

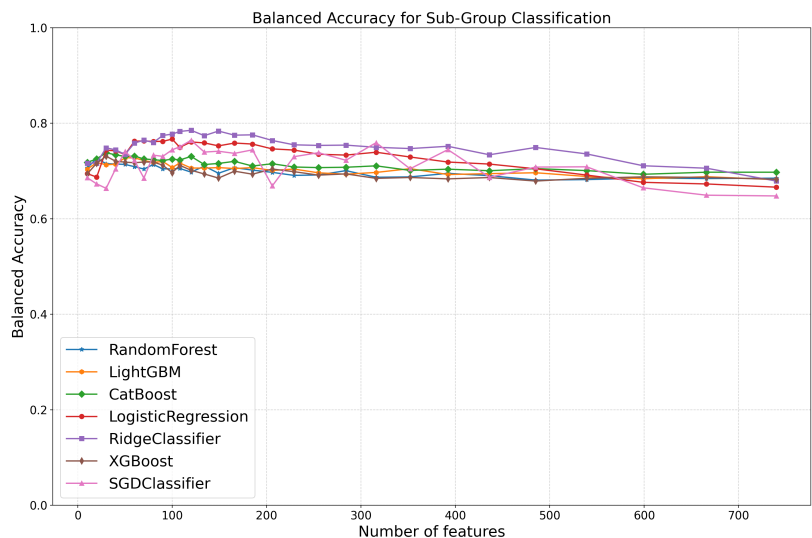


Figure C.49: Sub-group classification balanced accuracy: SG dataset

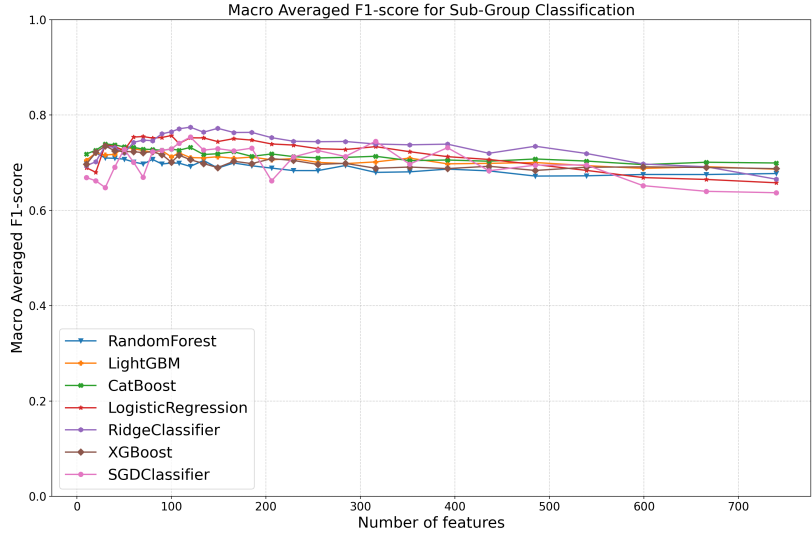


Figure C.50: Sub-group classification macro F1-score: SG dataset

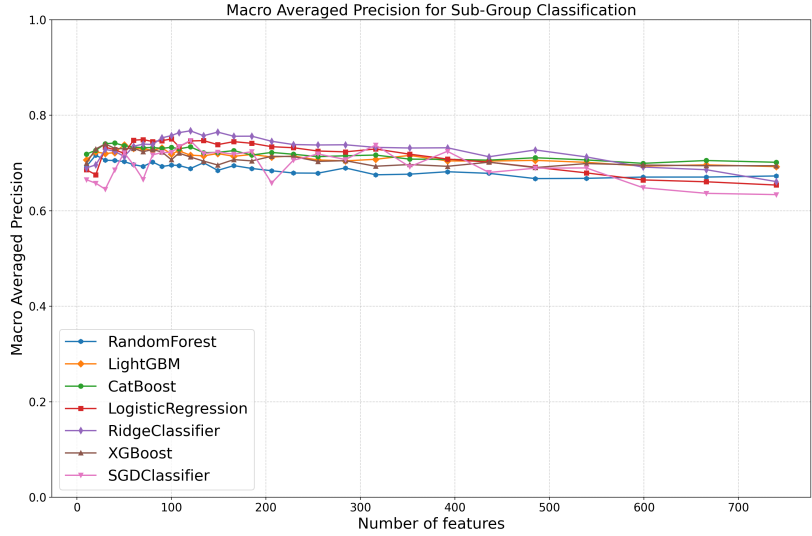


Figure C.51: Sub-group classification macro precision: SG dataset

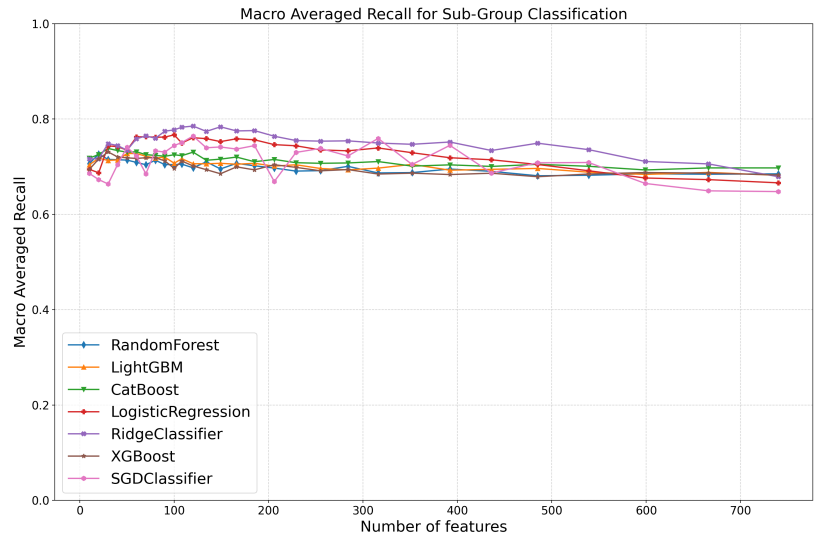


Figure C.52: Sub-group classification macro recall: SG dataset

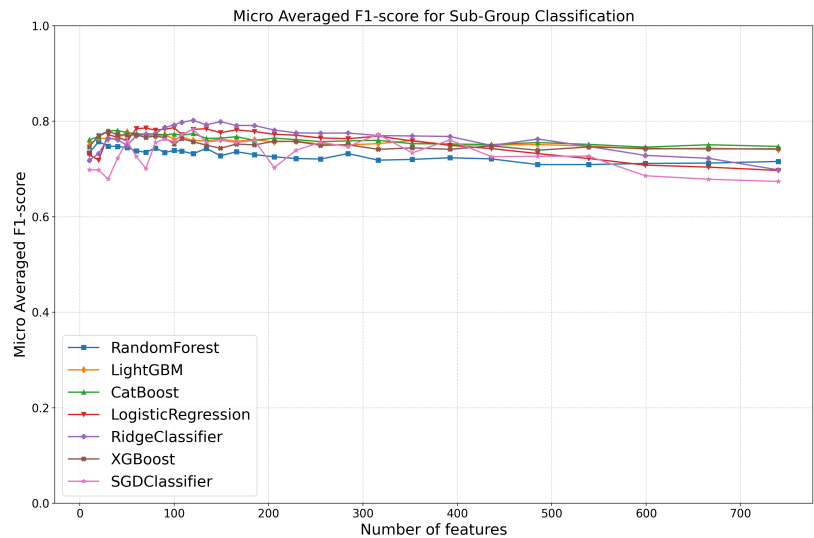


Figure C.53: Sub-group classification micro F1-score: SG dataset

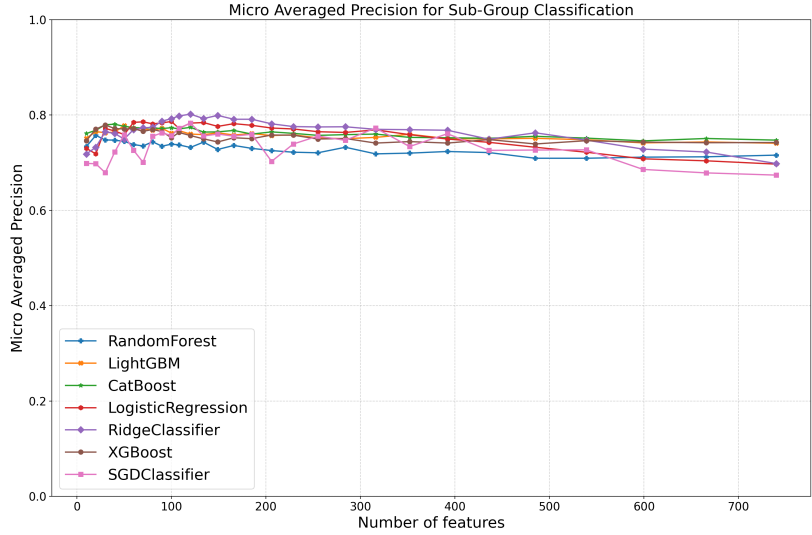


Figure C.54: Sub-group classification micro precision: SG dataset

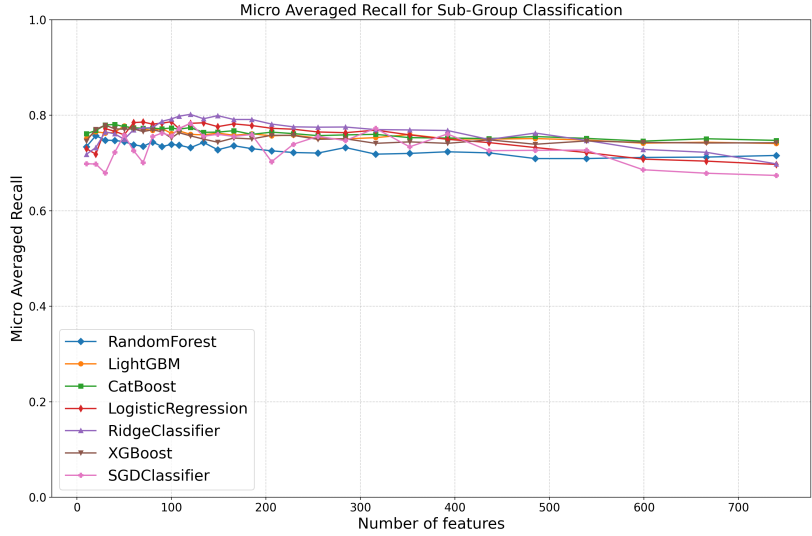


Figure C.55: Sub-group classification micro recall: SG dataset

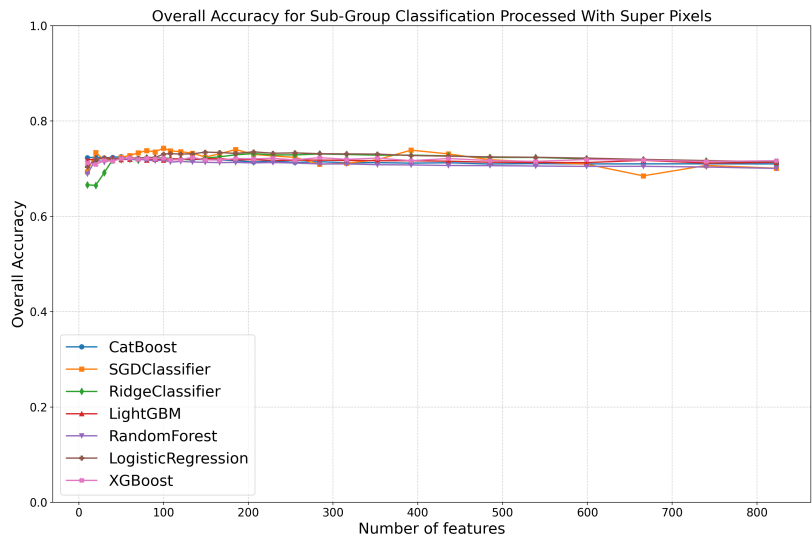


Figure C.56: Sub-group classification accuracy: SG-SP dataset

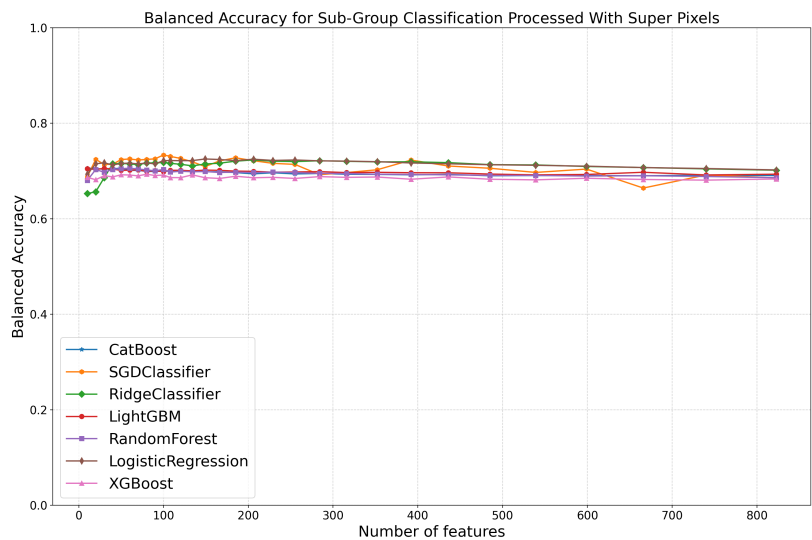


Figure C.57: Sub-group classification accuracy: SG-SP dataset

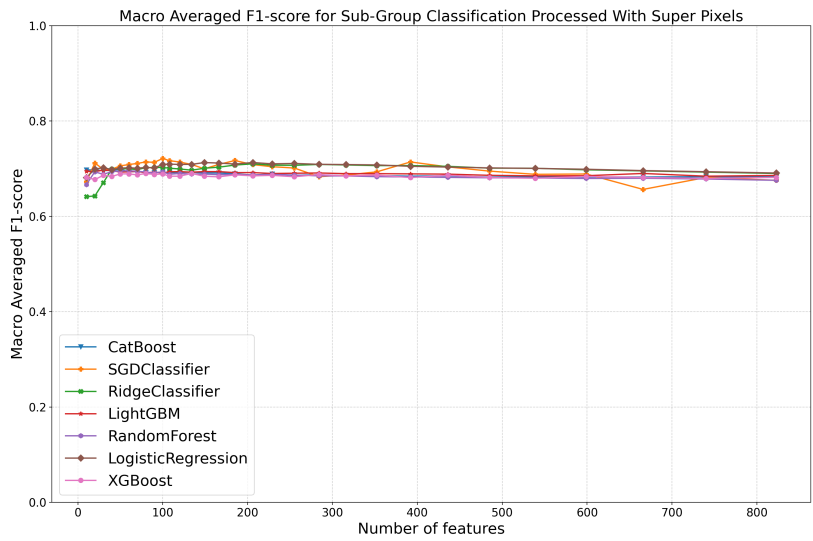


Figure C.58: Sub-group classification macro F1-score: SG-SP dataset

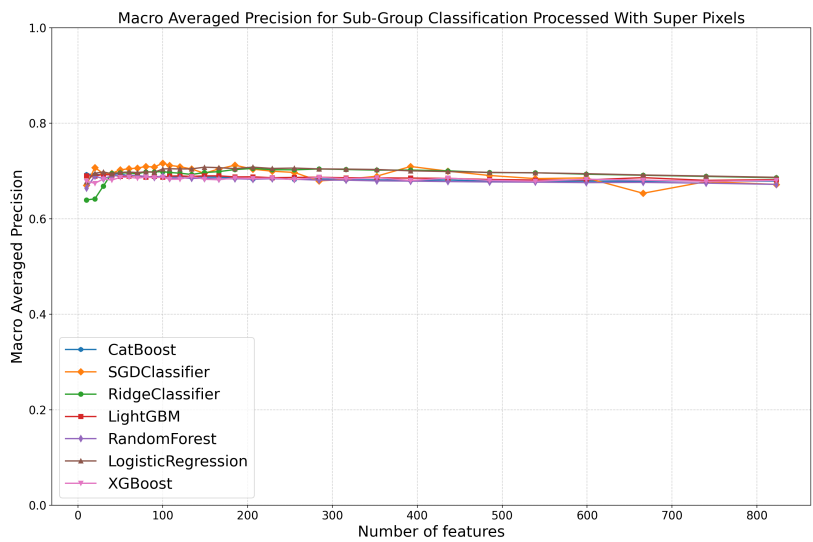


Figure C.59: Sub-group classification macro precision: SG-SP dataset

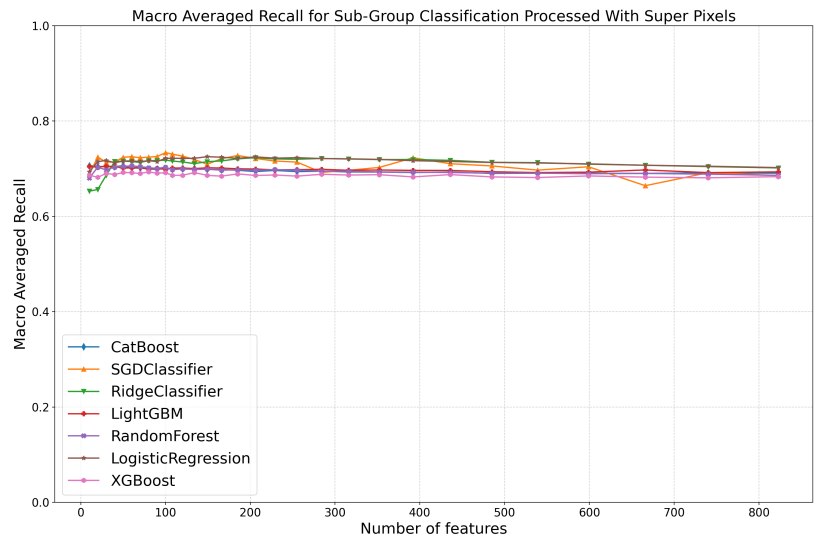


Figure C.60: Sub-group classification macro recall: SG-SP dataset

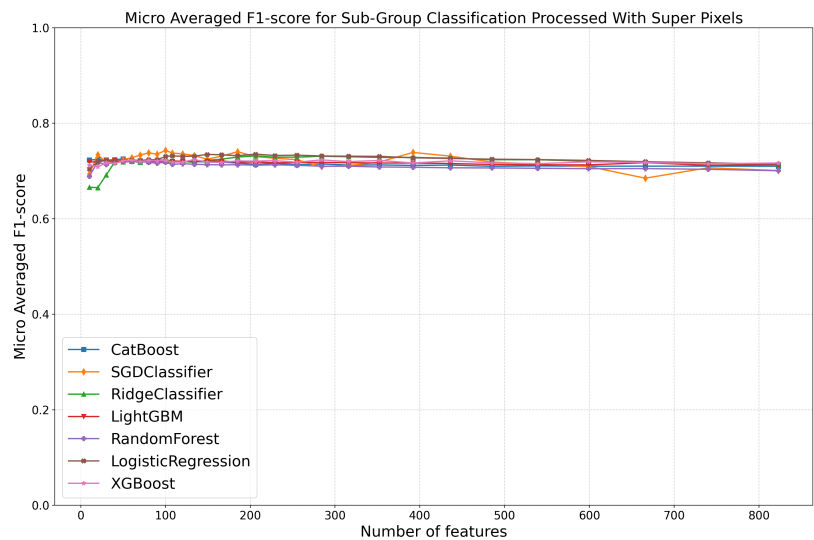


Figure C.61: Sub-group classification micro F1-score: SG-SP dataset

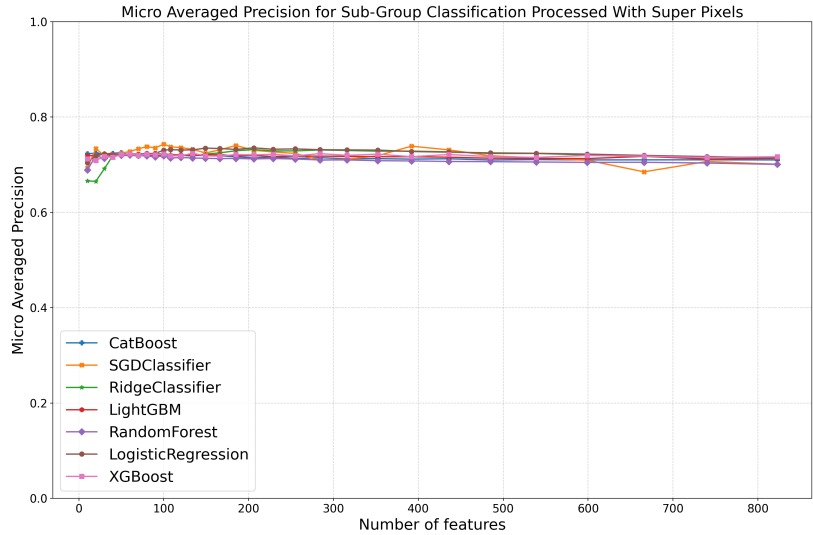


Figure C.62: Sub-group classification micro precision: SG-SP dataset

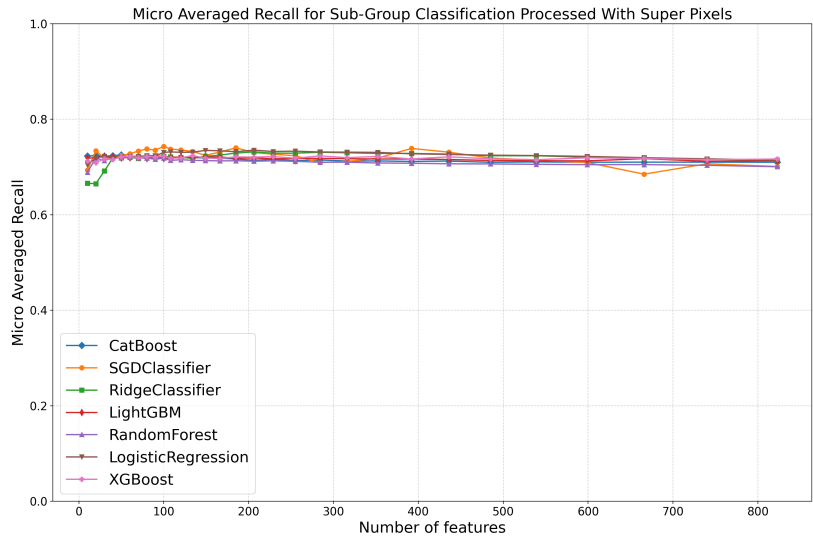


Figure C.63: Sub-group classification micro recall: SG-SP dataset

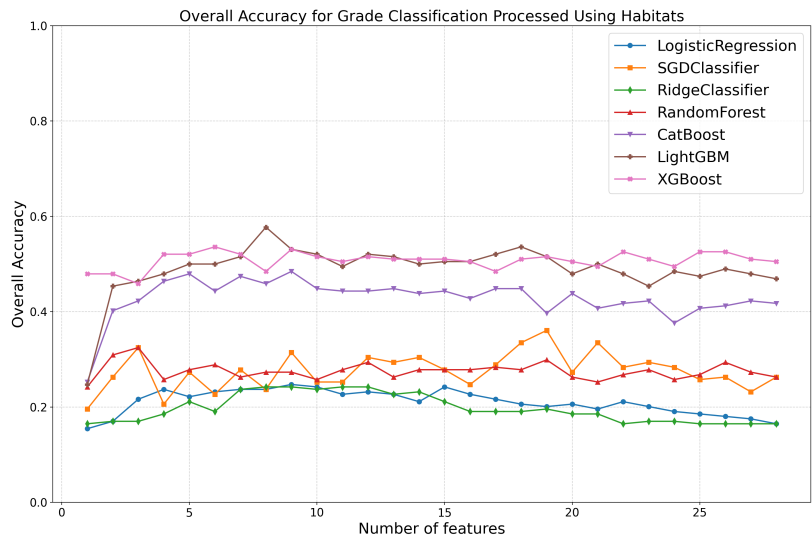


Figure C.64: Grade classification accuracy using habitats

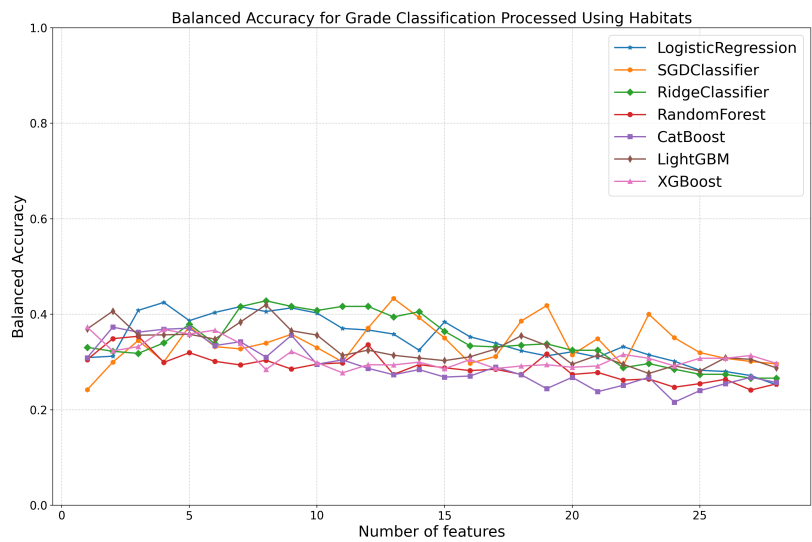


Figure C.65: Grade classification balanced accuracy using habitats

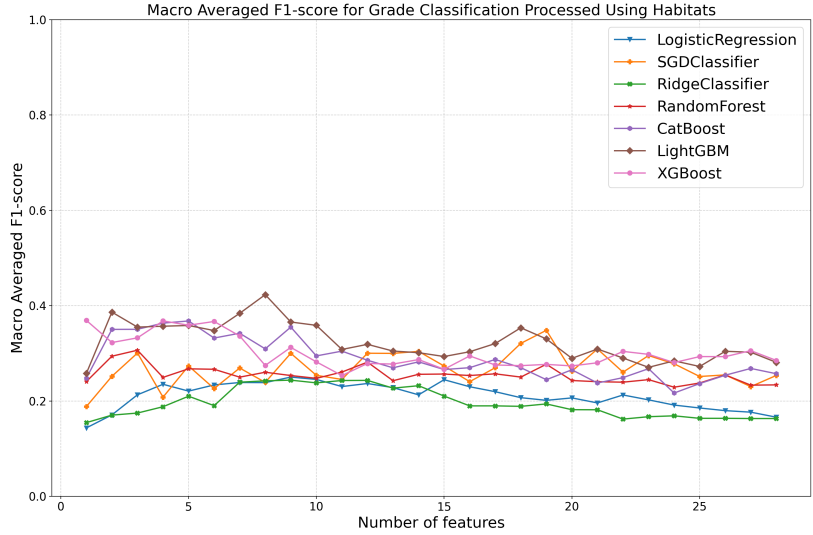


Figure C.66: Grade classification macro F1-score using habitats

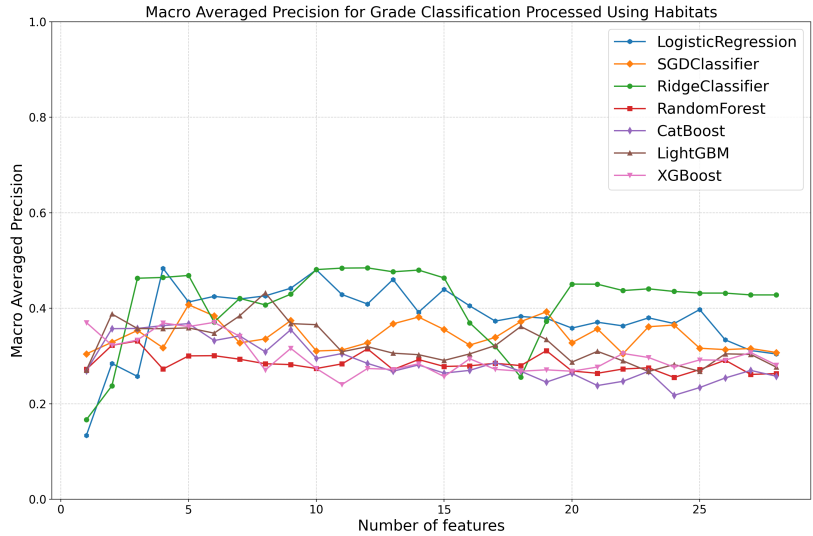


Figure C.67: Grade classification macro precision using habitats

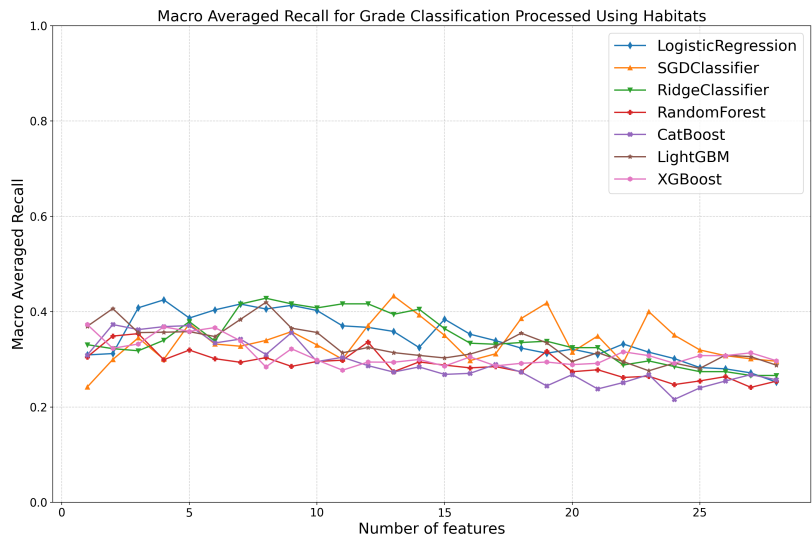


Figure C.68: Grade classification macro recall using habitats

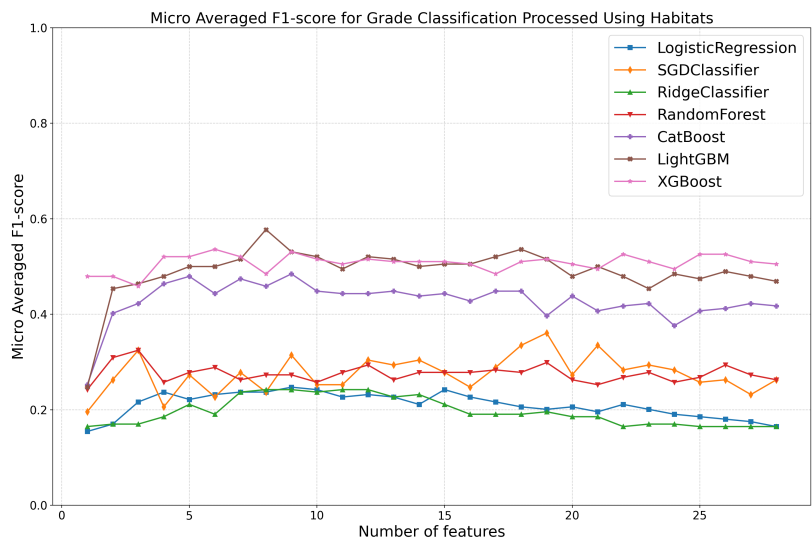


Figure C.69: Grade classification micro F1-score using habitats

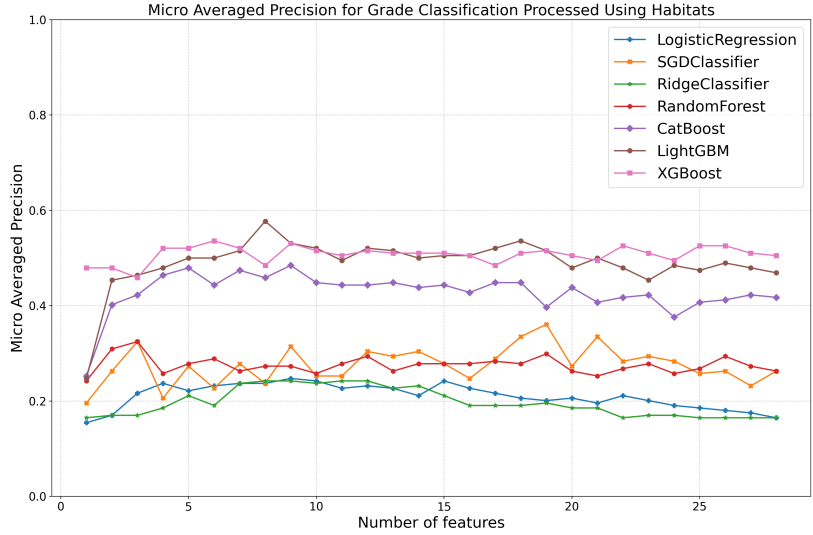


Figure C.70: Grade classification micro precision using habitats

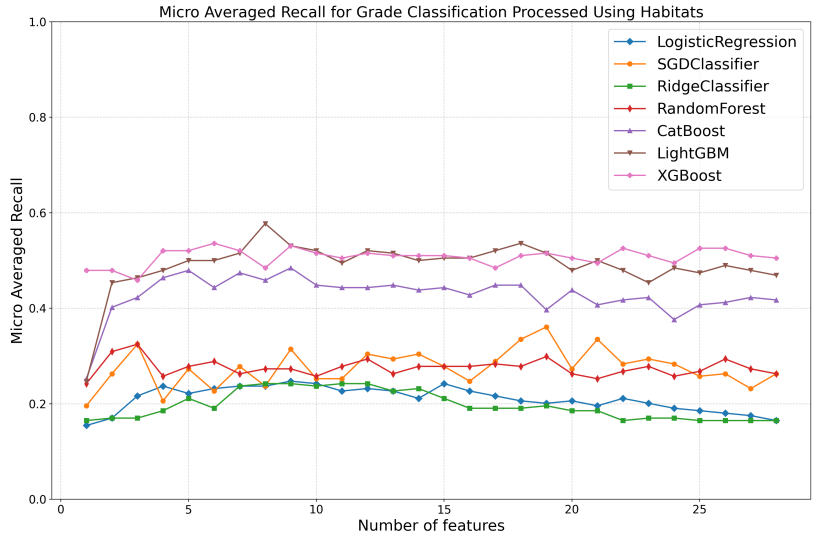


Figure C.71: Grade classification micro recall using habitats

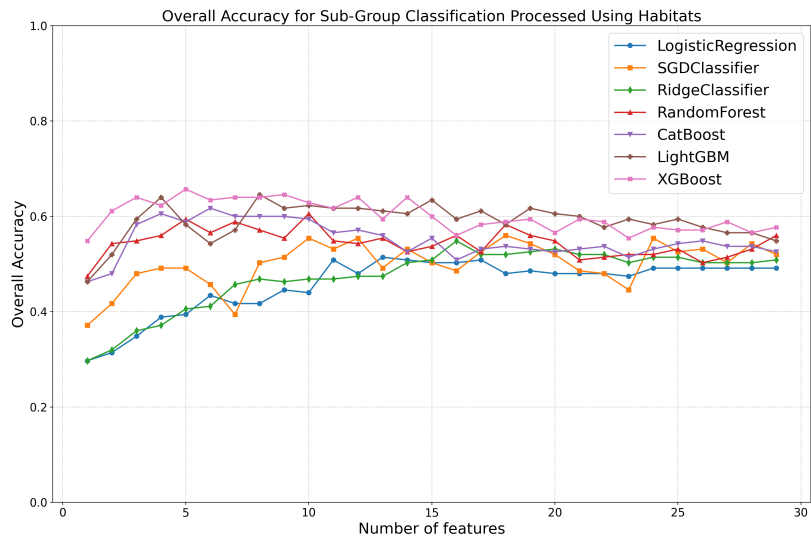


Figure C.72: Sub-group classification accuracy using habitats

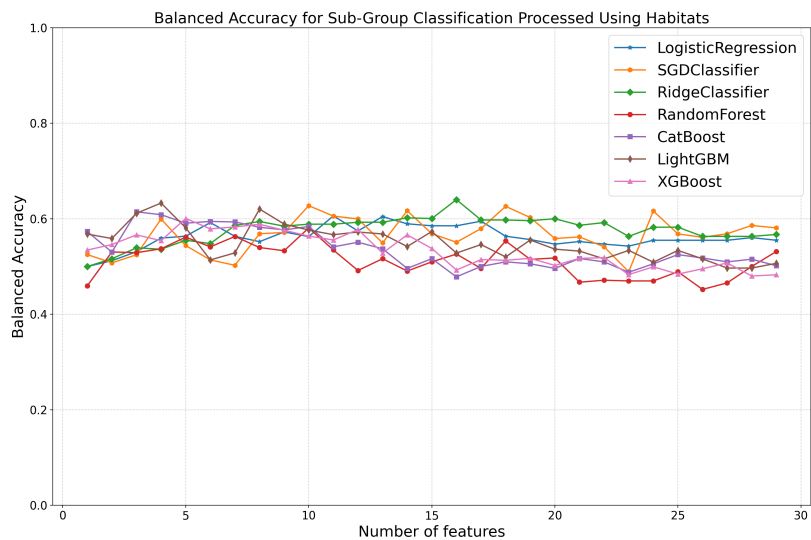


Figure C.73: Sub-group classification balanced accuracy using habitats

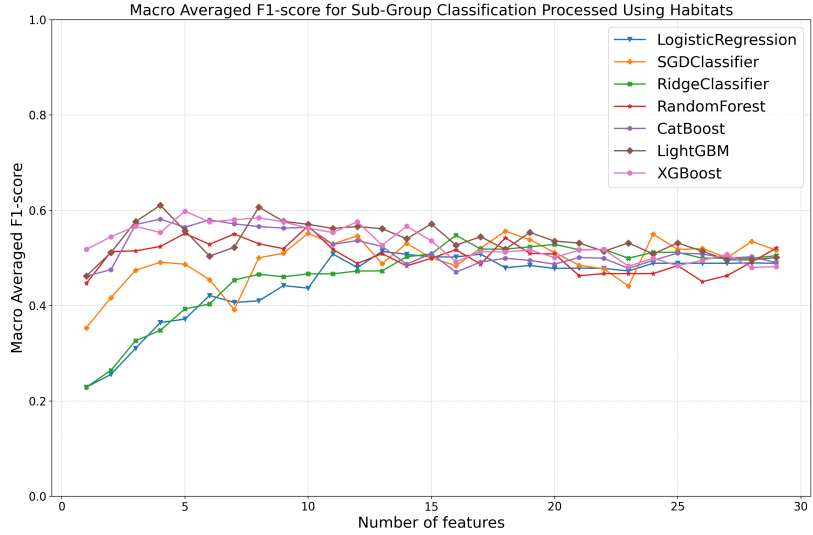


Figure C.74: Sub-group classification macro F1-score using habitats

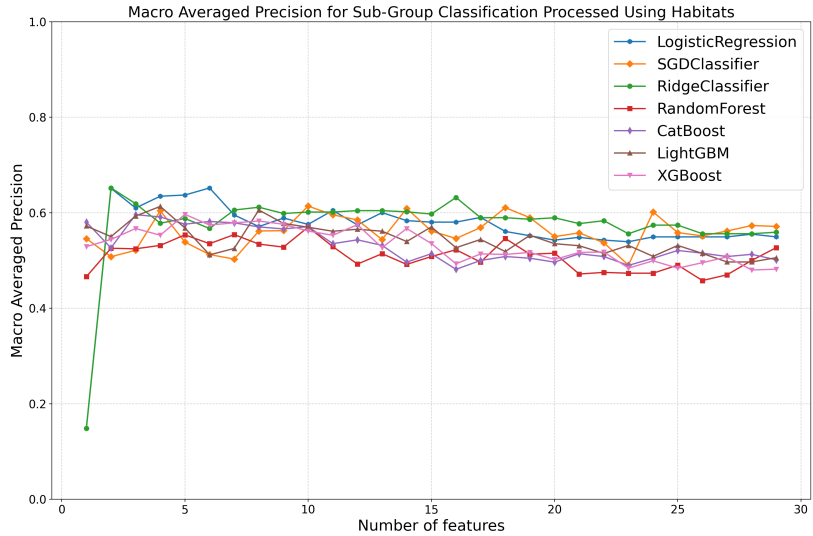


Figure C.75: Sub-group classification macro precision using habitats

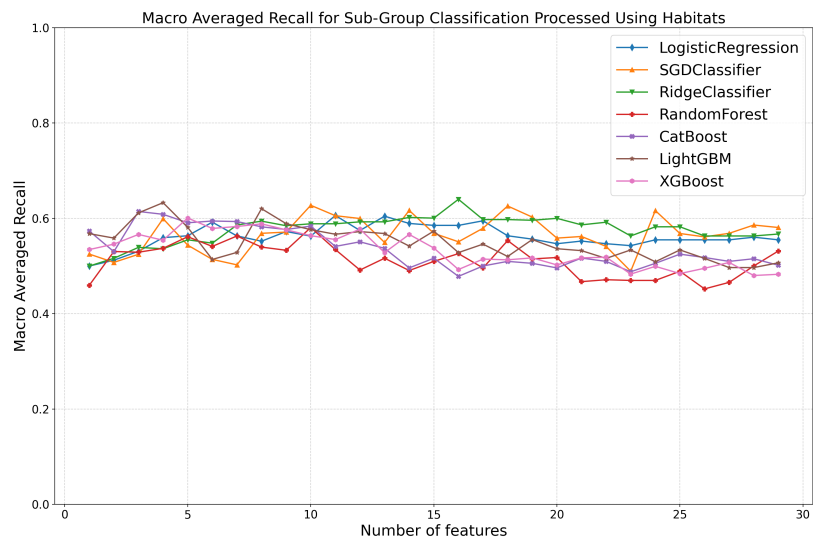


Figure C.76: Sub-group classification macro recall using habitats

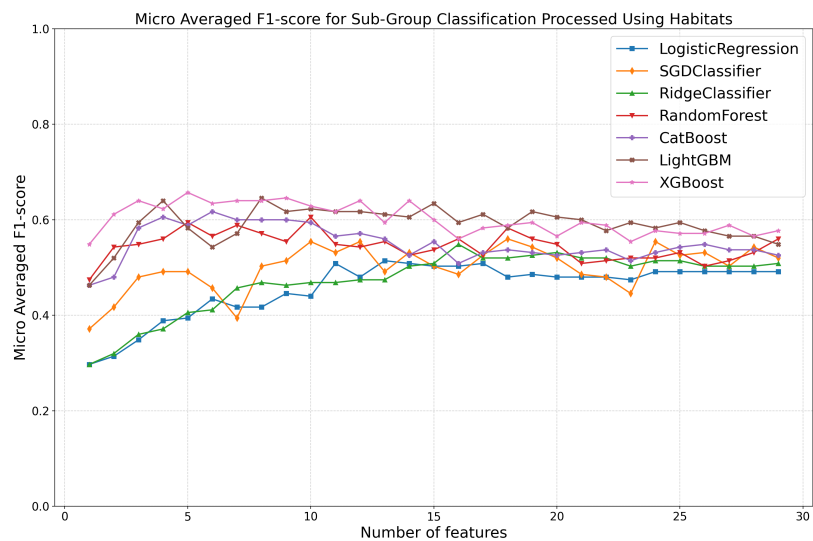


Figure C.77: Sub-group classification micro F1-score using habitats

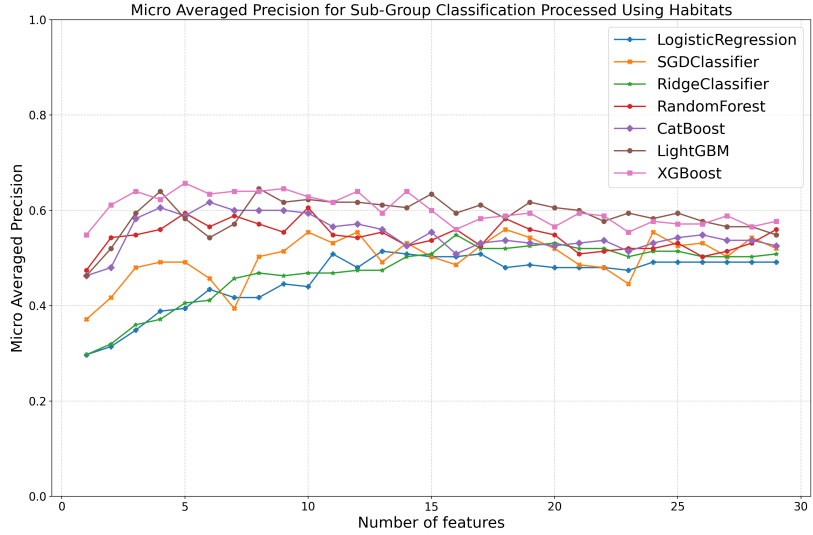


Figure C.78: Sub-group classification micro precision using habitats

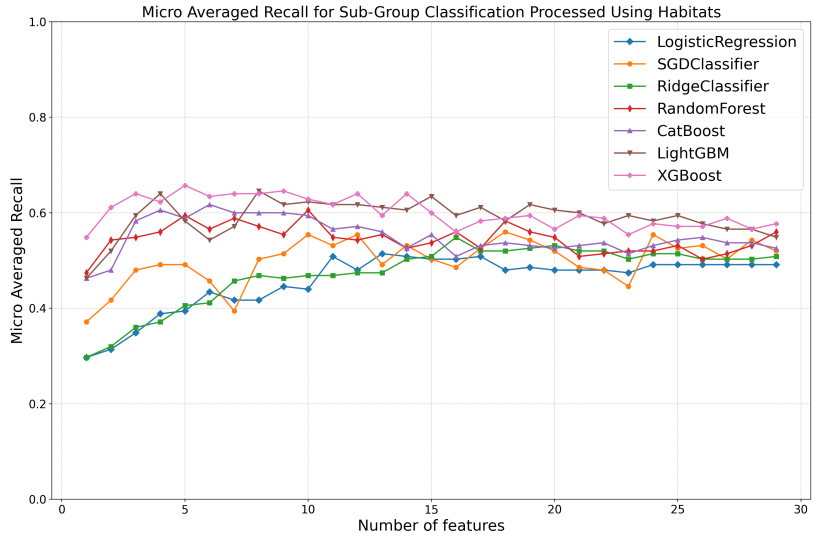


Figure C.79: Sub-group classification micro recall using habitats

C.2 Tables of Results

C.2.1 Tissue Classification

Table C.1: AUC-ROC score for all models and tested number of features for tissue classification with all healthy against sarcoma masks. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.996	0.994	0.995	0.987	0.970	0.995	0.996
20	0.997	0.997	0.997	0.994	0.975	0.996	0.998
30	0.997	0.996	0.998	0.993	0.991	0.998	0.997
40	0.998	0.996	0.999	0.994	0.990	0.998	0.997
50	0.998	0.996	0.999	0.994	0.995	0.998	0.997
60	0.998	0.996	0.999	0.994	0.995	0.998	0.997
70	0.998	0.996	0.999	0.994	0.996	0.997	0.997
80	0.998	0.996	0.999	0.994	0.996	0.997	0.997
90	0.998	0.995	0.999	0.994	0.996	0.997	0.997
100	0.997	0.995	0.999	0.994	0.996	0.997	0.997
108	0.998	0.995	0.999	0.994	0.997	0.996	0.997
120	0.997	0.994	0.999	0.994	0.997	0.996	0.997
134	0.997	0.996	0.999	0.993	0.997	0.996	0.997
149	0.997	0.995	0.999	0.993	0.997	0.996	0.997
166	0.997	0.994	0.999	0.993	0.997	0.996	0.997
185	0.997	0.994	0.999	0.993	0.998	0.996	0.997
206	0.997	0.995	0.998	0.992	0.997	0.996	0.997
229	0.997	0.995	0.998	0.993	0.998	0.996	0.997
255	0.997	0.995	0.998	0.993	0.998	0.995	0.996
284	0.997	0.994	0.998	0.992	0.997	0.996	0.996
316	0.997	0.995	0.998	0.992	0.997	0.995	0.996
352	0.997	0.994	0.997	0.992	0.997	0.994	0.996
392	0.997	0.995	0.997	0.992	0.997	0.994	0.996
436	0.997	0.994	0.997	0.992	0.997	0.993	0.996
485	0.997	0.994	0.997	0.992	0.997	0.993	0.996
539	0.997	0.994	0.996	0.992	0.996	0.992	0.996
599	0.997	0.995	0.995	0.992	0.995	0.991	0.996
666	0.997	0.994	0.994	0.992	0.992	0.991	0.996
740	0.997	0.994	0.993	0.991	0.990	0.990	0.996

Table C.2: AUC-ROC score for all models and tested number of features for tissue classification with all healthy against sarcoma masks processed with SPs. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.863	0.860	0.838	0.838	0.811	0.815	0.852
20	0.874	0.877	0.844	0.856	0.826	0.846	0.865
30	0.883	0.878	0.848	0.863	0.831	0.858	0.876
40	0.883	0.878	0.855	0.864	0.846	0.868	0.878
50	0.884	0.879	0.861	0.866	0.851	0.870	0.880
60	0.884	0.880	0.862	0.867	0.858	0.872	0.880
70	0.884	0.880	0.866	0.867	0.860	0.873	0.882
80	0.885	0.879	0.872	0.867	0.861	0.872	0.880
90	0.885	0.879	0.873	0.866	0.863	0.874	0.879
100	0.884	0.880	0.875	0.866	0.864	0.874	0.878
108	0.884	0.880	0.875	0.866	0.868	0.874	0.879
120	0.883	0.880	0.875	0.866	0.867	0.874	0.877
134	0.883	0.879	0.877	0.866	0.871	0.874	0.878
149	0.882	0.878	0.876	0.865	0.873	0.874	0.879
166	0.882	0.878	0.878	0.865	0.872	0.873	0.879
185	0.882	0.879	0.879	0.865	0.874	0.872	0.878
206	0.881	0.878	0.878	0.865	0.875	0.876	0.877
229	0.881	0.877	0.879	0.865	0.875	0.874	0.878
255	0.881	0.878	0.880	0.865	0.875	0.871	0.879
284	0.880	0.878	0.880	0.864	0.875	0.873	0.877
316	0.881	0.877	0.881	0.864	0.876	0.870	0.878
352	0.881	0.877	0.880	0.864	0.878	0.870	0.877
392	0.881	0.878	0.879	0.864	0.878	0.866	0.876
436	0.880	0.877	0.879	0.864	0.878	0.871	0.876
485	0.880	0.877	0.879	0.863	0.879	0.872	0.876
539	0.881	0.877	0.879	0.862	0.879	0.876	0.876
599	0.880	0.877	0.878	0.863	0.878	0.873	0.877
666	0.880	0.877	0.879	0.862	0.878	0.867	0.876
740	0.880	0.877	0.878	0.862	0.877	0.871	0.876
823	0.881	0.876	0.877	0.862	0.877	0.869	0.876

Table C.3: AUC-ROC score for all models and tested number of features for tissue classification with comparable healthy against sarcoma masks. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.947	0.940	0.886	0.903	0.808	0.870	0.935
20	0.949	0.956	0.933	0.910	0.914	0.938	0.951
30	0.955	0.958	0.950	0.928	0.947	0.946	0.955
40	0.958	0.957	0.960	0.926	0.952	0.942	0.956
50	0.957	0.958	0.963	0.927	0.957	0.948	0.957
60	0.957	0.958	0.970	0.925	0.962	0.964	0.957
70	0.957	0.959	0.971	0.924	0.962	0.967	0.957
80	0.957	0.957	0.973	0.926	0.961	0.971	0.957
90	0.957	0.955	0.974	0.926	0.964	0.969	0.959
100	0.957	0.956	0.972	0.927	0.969	0.964	0.958
104	0.958	0.957	0.972	0.927	0.968	0.963	0.957
116	0.955	0.955	0.974	0.922	0.973	0.963	0.958
129	0.954	0.956	0.972	0.925	0.972	0.965	0.957
144	0.954	0.954	0.972	0.924	0.974	0.967	0.956
161	0.953	0.954	0.972	0.919	0.973	0.963	0.954
179	0.953	0.954	0.973	0.918	0.971	0.964	0.953
199	0.952	0.951	0.971	0.918	0.971	0.964	0.954
222	0.951	0.951	0.972	0.918	0.971	0.965	0.952
247	0.951	0.951	0.972	0.916	0.974	0.955	0.953
275	0.950	0.949	0.973	0.919	0.973	0.958	0.948
306	0.952	0.948	0.971	0.916	0.971	0.962	0.948
340	0.950	0.947	0.969	0.913	0.968	0.955	0.947
378	0.950	0.947	0.968	0.911	0.965	0.956	0.947
421	0.949	0.947	0.965	0.911	0.962	0.953	0.948
468	0.949	0.946	0.962	0.907	0.960	0.950	0.945
521	0.947	0.946	0.960	0.908	0.958	0.939	0.945
579	0.948	0.945	0.953	0.905	0.955	0.931	0.946
644	0.946	0.945	0.947	0.902	0.949	0.925	0.948
716	0.946	0.945	0.941	0.906	0.943	0.920	0.948
796	0.945	0.946	0.936	0.905	0.938	0.913	0.945

Table C.4: AUC-ROC score for all models and tested number of features for tissue classification with comparable healthy against sarcoma masks processed with SPs. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.820	0.836	0.784	0.826	0.743	0.771	0.814
20	0.847	0.848	0.809	0.820	0.783	0.790	0.840
30	0.849	0.848	0.816	0.825	0.823	0.825	0.840
40	0.853	0.849	0.849	0.827	0.832	0.842	0.845
50	0.855	0.850	0.852	0.823	0.841	0.849	0.847
60	0.857	0.852	0.855	0.827	0.844	0.846	0.847
70	0.855	0.853	0.857	0.828	0.842	0.852	0.847
80	0.857	0.852	0.860	0.829	0.845	0.855	0.846
90	0.855	0.850	0.861	0.825	0.846	0.856	0.847
100	0.856	0.851	0.865	0.827	0.848	0.855	0.847
106	0.856	0.852	0.864	0.827	0.848	0.854	0.845
118	0.856	0.851	0.863	0.826	0.848	0.853	0.846
132	0.855	0.851	0.864	0.826	0.848	0.831	0.847
147	0.854	0.850	0.863	0.825	0.853	0.827	0.844
164	0.855	0.852	0.865	0.826	0.852	0.829	0.846
183	0.854	0.850	0.865	0.826	0.854	0.865	0.844
204	0.854	0.851	0.863	0.824	0.856	0.855	0.846
227	0.853	0.849	0.863	0.824	0.859	0.852	0.842
253	0.853	0.849	0.865	0.824	0.859	0.852	0.845
282	0.853	0.849	0.864	0.824	0.859	0.848	0.843
314	0.852	0.849	0.863	0.823	0.857	0.849	0.840
349	0.853	0.849	0.863	0.823	0.857	0.847	0.843
388	0.852	0.849	0.862	0.821	0.856	0.848	0.842
432	0.852	0.847	0.861	0.823	0.856	0.847	0.841
481	0.852	0.848	0.859	0.820	0.855	0.845	0.843
535	0.851	0.847	0.859	0.820	0.854	0.842	0.840
595	0.850	0.848	0.856	0.819	0.852	0.835	0.840
662	0.851	0.846	0.854	0.819	0.850	0.835	0.841
736	0.851	0.846	0.852	0.819	0.848	0.829	0.840
818	0.851	0.846	0.849	0.818	0.847	0.833	0.839

C.2.2 Grade Classification

Table C.5: AUC-ROC score for all models and tested number of features for grade classification with whole masks. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.603	0.616	0.582	0.598	0.584	0.493	0.627
20	0.618	0.608	0.663	0.630	0.624	0.526	0.608
30	0.613	0.607	0.674	0.622	0.651	0.651	0.620
40	0.597	0.619	0.680	0.622	0.660	0.658	0.626
50	0.592	0.604	0.677	0.615	0.667	0.671	0.624
60	0.587	0.602	0.682	0.607	0.661	0.690	0.610
70	0.586	0.607	0.673	0.601	0.666	0.673	0.621
80	0.592	0.610	0.661	0.607	0.653	0.649	0.612
90	0.585	0.611	0.672	0.597	0.661	0.641	0.607
100	0.585	0.608	0.668	0.597	0.657	0.658	0.620
108	0.583	0.600	0.658	0.599	0.654	0.651	0.614
120	0.581	0.591	0.654	0.591	0.668	0.671	0.602
134	0.575	0.586	0.656	0.582	0.666	0.659	0.609
149	0.566	0.586	0.647	0.586	0.660	0.647	0.598
166	0.567	0.581	0.641	0.589	0.669	0.637	0.597
185	0.557	0.572	0.637	0.570	0.660	0.622	0.591
206	0.548	0.574	0.620	0.577	0.658	0.628	0.590
229	0.553	0.571	0.629	0.571	0.655	0.591	0.594
255	0.544	0.562	0.622	0.576	0.654	0.641	0.577
284	0.549	0.555	0.629	0.575	0.640	0.629	0.574
316	0.547	0.558	0.614	0.567	0.627	0.592	0.564
352	0.544	0.546	0.605	0.558	0.632	0.586	0.561
392	0.544	0.543	0.591	0.563	0.622	0.601	0.554
436	0.536	0.541	0.586	0.553	0.601	0.597	0.547
485	0.540	0.540	0.571	0.549	0.590	0.593	0.545
539	0.530	0.530	0.557	0.558	0.572	0.553	0.553
599	0.531	0.532	0.545	0.545	0.553	0.544	0.551
666	0.532	0.529	0.528	0.549	0.530	0.502	0.537
740	0.526	0.520	0.516	0.548	0.514	0.508	0.539

Table C.6: AUC-ROC score for all models and tested number of features for grade classification with masks processed with SPs. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.627	0.623	0.645	0.638	0.670	0.643	0.620
20	0.621	0.610	0.654	0.650	0.651	0.650	0.606
30	0.619	0.614	0.652	0.653	0.645	0.631	0.602
40	0.618	0.615	0.649	0.652	0.644	0.664	0.597
50	0.610	0.609	0.650	0.653	0.649	0.652	0.600
60	0.608	0.613	0.651	0.651	0.649	0.634	0.597
70	0.606	0.610	0.650	0.653	0.645	0.613	0.600
80	0.613	0.607	0.644	0.651	0.646	0.640	0.596
90	0.611	0.611	0.642	0.654	0.643	0.620	0.598
100	0.607	0.611	0.643	0.655	0.641	0.612	0.592
108	0.608	0.608	0.652	0.654	0.641	0.603	0.599
120	0.607	0.608	0.647	0.657	0.641	0.631	0.597
134	0.607	0.608	0.642	0.655	0.636	0.663	0.598
149	0.606	0.608	0.645	0.652	0.633	0.663	0.596
166	0.608	0.607	0.646	0.649	0.636	0.631	0.596
185	0.603	0.608	0.641	0.647	0.634	0.642	0.596
206	0.605	0.607	0.647	0.649	0.637	0.628	0.593
229	0.603	0.605	0.647	0.649	0.638	0.629	0.589
255	0.602	0.606	0.643	0.645	0.640	0.609	0.590
284	0.603	0.602	0.644	0.646	0.636	0.648	0.588
316	0.604	0.602	0.643	0.645	0.639	0.617	0.586
352	0.602	0.601	0.640	0.644	0.634	0.606	0.587
392	0.601	0.600	0.637	0.643	0.632	0.628	0.584
436	0.601	0.599	0.630	0.643	0.630	0.600	0.587
485	0.602	0.595	0.625	0.641	0.626	0.561	0.586
539	0.600	0.595	0.622	0.639	0.623	0.578	0.586
599	0.596	0.598	0.618	0.638	0.621	0.601	0.584
666	0.597	0.597	0.613	0.634	0.615	0.598	0.585
740	0.599	0.596	0.609	0.633	0.612	0.597	0.583
823	0.600	0.596	0.604	0.633	0.608	0.568	0.580

C.2.3 Sub-Group Classification

Table C.7: AUC-ROC score for all models and tested number of features for sub-group classification with whole masks. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.788	0.775	0.772	0.772	0.775	0.735	0.790
20	0.800	0.794	0.751	0.788	0.779	0.728	0.812
30	0.807	0.797	0.823	0.784	0.811	0.715	0.822
40	0.803	0.794	0.828	0.790	0.806	0.759	0.811
50	0.809	0.804	0.812	0.779	0.795	0.798	0.815
60	0.804	0.811	0.840	0.780	0.820	0.786	0.808
70	0.803	0.811	0.838	0.775	0.822	0.749	0.804
80	0.803	0.802	0.834	0.778	0.818	0.825	0.798
90	0.795	0.808	0.839	0.775	0.839	0.799	0.802
100	0.796	0.795	0.844	0.774	0.850	0.809	0.782
108	0.798	0.796	0.832	0.776	0.859	0.837	0.789
120	0.792	0.793	0.846	0.769	0.860	0.826	0.790
134	0.781	0.790	0.848	0.778	0.858	0.825	0.770
149	0.780	0.782	0.842	0.762	0.863	0.801	0.770
166	0.778	0.779	0.841	0.772	0.858	0.787	0.780
185	0.776	0.782	0.840	0.761	0.860	0.822	0.773
206	0.774	0.772	0.829	0.764	0.852	0.710	0.778
229	0.772	0.767	0.827	0.754	0.848	0.801	0.772
255	0.765	0.766	0.817	0.753	0.841	0.798	0.765
284	0.769	0.756	0.818	0.757	0.829	0.798	0.765
316	0.766	0.766	0.821	0.751	0.829	0.820	0.763
352	0.761	0.761	0.817	0.747	0.824	0.746	0.752
392	0.757	0.758	0.802	0.747	0.829	0.802	0.757
436	0.754	0.758	0.790	0.745	0.818	0.739	0.749
485	0.750	0.756	0.776	0.737	0.823	0.772	0.740
539	0.756	0.751	0.763	0.739	0.821	0.764	0.736
599	0.745	0.739	0.743	0.740	0.792	0.724	0.742
666	0.743	0.739	0.732	0.738	0.780	0.690	0.745
740	0.741	0.739	0.724	0.737	0.747	0.696	0.744

Table C.8: AUC-ROC score for all models and tested number of features for subgroup classification with masks processed with SPs. The highest score for each model is in bold.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
10	0.769	0.762	0.724	0.733	0.675	0.713	0.741
20	0.766	0.760	0.778	0.759	0.685	0.769	0.736
30	0.763	0.762	0.781	0.754	0.739	0.762	0.750
40	0.761	0.763	0.777	0.756	0.768	0.775	0.741
50	0.763	0.760	0.782	0.762	0.779	0.794	0.745
60	0.760	0.757	0.786	0.759	0.776	0.792	0.746
70	0.758	0.760	0.785	0.761	0.775	0.792	0.743
80	0.756	0.756	0.786	0.759	0.781	0.790	0.744
90	0.757	0.756	0.784	0.757	0.784	0.794	0.743
100	0.760	0.756	0.790	0.758	0.782	0.802	0.747
108	0.756	0.758	0.791	0.753	0.779	0.800	0.740
120	0.758	0.756	0.788	0.756	0.779	0.792	0.740
134	0.756	0.758	0.786	0.754	0.775	0.791	0.744
149	0.754	0.757	0.793	0.755	0.781	0.778	0.739
166	0.754	0.757	0.791	0.754	0.782	0.778	0.737
185	0.752	0.757	0.787	0.754	0.787	0.792	0.738
206	0.751	0.753	0.789	0.753	0.792	0.780	0.740
229	0.751	0.751	0.786	0.754	0.789	0.779	0.739
255	0.750	0.750	0.786	0.750	0.788	0.777	0.740
284	0.750	0.752	0.783	0.751	0.787	0.750	0.739
316	0.746	0.750	0.781	0.749	0.785	0.752	0.739
352	0.747	0.749	0.777	0.749	0.783	0.758	0.737
392	0.746	0.749	0.775	0.747	0.783	0.785	0.737
436	0.746	0.748	0.770	0.747	0.780	0.769	0.739
485	0.744	0.746	0.769	0.745	0.776	0.762	0.737
539	0.744	0.742	0.768	0.745	0.775	0.744	0.737
599	0.744	0.744	0.765	0.745	0.771	0.752	0.733
666	0.744	0.747	0.760	0.742	0.768	0.708	0.737
740	0.744	0.743	0.756	0.742	0.762	0.720	0.732
823	0.744	0.744	0.751	0.740	0.758	0.729	0.736

C.2.4 Habitat Classification

Table C.9: AUC-ROC scores obtained from habitat analysis when predicting grade.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
1	0.480	0.466	0.471	0.477	0.464	0.433	0.500
2	0.533	0.518	0.506	0.520	0.516	0.477	0.515
3	0.539	0.548	0.532	0.514	0.497	0.522	0.539
4	0.546	0.543	0.532	0.503	0.495	0.462	0.578
5	0.561	0.579	0.509	0.499	0.541	0.522	0.575
6	0.544	0.565	0.548	0.504	0.531	0.499	0.575
7	0.547	0.565	0.548	0.497	0.565	0.495	0.538
8	0.542	0.577	0.547	0.491	0.558	0.464	0.492
9	0.544	0.566	0.533	0.509	0.549	0.536	0.528
10	0.514	0.538	0.526	0.499	0.536	0.494	0.510
11	0.488	0.515	0.520	0.471	0.536	0.497	0.500
12	0.499	0.532	0.518	0.492	0.537	0.485	0.492
13	0.491	0.525	0.511	0.473	0.529	0.536	0.484
14	0.490	0.513	0.503	0.463	0.517	0.591	0.481
15	0.463	0.509	0.519	0.458	0.509	0.542	0.492
16	0.450	0.523	0.508	0.458	0.501	0.495	0.470
17	0.448	0.531	0.498	0.460	0.503	0.491	0.476
18	0.452	0.521	0.498	0.458	0.493	0.564	0.486
19	0.443	0.489	0.495	0.451	0.516	0.562	0.462
20	0.439	0.494	0.496	0.446	0.508	0.472	0.472
21	0.447	0.468	0.491	0.415	0.502	0.486	0.461
22	0.440	0.475	0.482	0.428	0.483	0.491	0.473
23	0.446	0.452	0.479	0.408	0.475	0.564	0.468
24	0.440	0.462	0.474	0.428	0.474	0.488	0.480
25	0.417	0.462	0.457	0.424	0.475	0.451	0.464
26	0.425	0.452	0.449	0.436	0.474	0.453	0.455
27	0.427	0.454	0.447	0.433	0.456	0.475	0.443
28	0.431	0.452	0.438	0.416	0.447	0.432	0.464

Table C.10: AUC-ROC scores obtained from habitat analysis when predicting sub-group.

No. Features	CB	LGBM	LR	RF	RC	SGDC	XGB
1	0.546	0.505	0.576	0.442	0.580	0.449	0.514
2	0.544	0.560	0.656	0.550	0.664	0.566	0.551
3	0.624	0.622	0.658	0.548	0.660	0.520	0.598
4	0.654	0.625	0.689	0.566	0.653	0.545	0.594
5	0.641	0.609	0.657	0.577	0.679	0.553	0.597
6	0.629	0.577	0.651	0.535	0.628	0.546	0.652
7	0.609	0.559	0.659	0.559	0.680	0.585	0.604
8	0.610	0.616	0.627	0.585	0.669	0.613	0.604
9	0.596	0.619	0.624	0.560	0.661	0.629	0.605
10	0.608	0.616	0.623	0.568	0.654	0.655	0.596
11	0.593	0.611	0.647	0.510	0.658	0.640	0.600
12	0.583	0.587	0.639	0.516	0.659	0.658	0.603
13	0.579	0.572	0.645	0.528	0.651	0.578	0.598
14	0.569	0.575	0.657	0.514	0.673	0.682	0.584
15	0.565	0.604	0.642	0.501	0.665	0.619	0.574
16	0.566	0.569	0.643	0.526	0.677	0.572	0.556
17	0.560	0.584	0.638	0.525	0.678	0.608	0.571
18	0.550	0.571	0.649	0.554	0.672	0.707	0.547
19	0.540	0.557	0.644	0.545	0.671	0.634	0.566
20	0.542	0.542	0.637	0.536	0.661	0.608	0.543
21	0.543	0.541	0.634	0.473	0.657	0.627	0.545
22	0.546	0.546	0.626	0.493	0.649	0.587	0.551
23	0.534	0.554	0.623	0.514	0.639	0.518	0.536
24	0.533	0.541	0.619	0.501	0.638	0.686	0.537
25	0.539	0.533	0.615	0.506	0.638	0.633	0.536
26	0.525	0.524	0.610	0.477	0.628	0.548	0.537
27	0.535	0.528	0.608	0.480	0.625	0.638	0.547
28	0.531	0.530	0.601	0.494	0.616	0.606	0.516
29	0.526	0.536	0.584	0.526	0.604	0.573	0.529