



Advancing Knee Osteoarthritis Assessment Using Novel Quantitative Imaging Features and Machine Learning

Federica Kiyomi Ciliberti

Dissertation submitted to the Department of Engineering at Reykjavík University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

November 17, 2025

Thesis Committee:

Paolo Gargiulo, Supervisor
Professor, Reykjavik University, Iceland

Halldór Jónsson jr, Co-supervisor
Professor Emeritus, Reykjavik University, Iceland

Marco Cristani, Examiner
Professor, University of Verona, Italy

Arsenio Cutolo, Examiner
Associate Professor, University of Naples Federico II, Italy

ISBN 978-9935-539-97-7 Electronic version
ISBN 978-9935-539-96-0 Print version
ORCID Federica Kiyomi Ciliberti 0000-0002-8378-9478
<https://orcid.org/0000-0002-8378-9478>
Copyright © 2025 Federica Kiyomi Ciliberti 

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.



Advancing Knee Osteoarthritis Assessment Using Novel Quantitative Imaging Features and Machine Learning

Federica Kiyomi Ciliberti

November 17, 2025

Abstract

Knee osteoarthritis (KOA) is a leading cause of disability worldwide, characterized by progressive degeneration of articular cartilage and associated musculoskeletal structures. Despite its prevalence, early diagnosis and patient-specific assessment remain challenging due to the multifactorial nature of the disease and the limitations of conventional imaging and grading systems. This thesis advances KOA evaluation by integrating multi-modal quantitative imaging features and machine learning methodologies to improve diagnostic accuracy, reproducibility, and early detection.

Magnetic resonance imaging (MRI) and computed tomography (CT) scans were acquired and analyzed within the framework of the EU SINPAIN project, which aims to develop innovative and personalized therapies for KOA. The imaging data were processed through a standardized pipeline for segmentation, registration, and feature extraction of bones, cartilages, and muscles. Morphological, density, and radiomic features were combined to train and validate supervised and semi-supervised learning models and identify the most informative imaging biomarkers of KOA.

The study demonstrates that specific imaging-derived characteristics, such as alterations in cartilage morphology and texture, subchondral bone density distribution, and the balance between muscle and intramuscular adipose tissue, serve as sensitive indicators of degenerative changes in the joint. Cross-tissue analyses revealed consistent interdependence between cartilage degradation and surrounding musculoskeletal adaptations, supporting a holistic interpretation of joint health. Furthermore, robustness and stability assessments established the reliability of the extracted quantitative features under realistic segmentation variations.

By uniting advanced imaging analytics with artificial intelligence, this work contributes a reproducible framework for objective and personalized assessment of KOA. The developed methodology enhances the understanding of tissue-level interactions in joint degeneration and supports the transition toward precision diagnostics and targeted therapeutic strategies in osteoarthritis care.

Keywords: knee OA, medical imaging, CT, MRI, machine learning, early diagnosis of KOA, precision medicine

Ágrip

Slitgigt í hné (e. Knee OsteoArthritis, KOA) er ein helsta orsök bæklunar um allan heim og einkennist af stigvaxandi hrörnun liðbrjósks og tengdra stoðkerfisþátta. Þrátt fyrir víðtæka útbreiðslu slitgigtarinnar er snemmbúin greining og einstaklingstengt mat krefjandi vegna fjölpátta eðlis sjúkdómsins og takmarkana á hefðbundnum myndgreiningar- og flokkunarkerfum. Þessi ritgerð setur fram nýja magngreinalega matsaðferð á KOA með því að samþætta fjölpátta myndgreiningareiginleika og vélanámsaðferðir og þannig bæta greiningarnákvæmni, endurtekningarhæfni og snemmbúna uppgötvun sjúkdómsins. Segulómun (MRI) og tölvusneiðmyndataka (CT) voru framkvæmdar og slitgigt greind innan ramma EU SINPAIN verkefnisins, sem miðar að því að þróa nýstárlegar og sérsniðnar meðferðir við KOA. Myndgreiningargögnin voru unnin í gegnum staðlað ferli fyrir sundurgreiningu, skráningu og eiginleikaútdrátt beina, brjósks og vöðva. Formfræðilegir, þéttleika- og geislafræðilegir eiginleikar voru sameinaðir til að þjálfa og staðfesta stýrð og hálfstýrð námslíkön og bera kennsl á upplýsandi myndgreiningarlífmerki KOA.

Rannsóknin sýnir fram á að sértækir eiginleikar sem myndgreining hefur aflað sér, svo sem breytingar á brjóskeyggingu og áferð, dreifing beinþéttni undir brjóskey og jafnvægi milli vöðva- og vöðvafituvefs, eru næmar vísbendingar um hrörnunarbreytingar í liðnum. Krossvefjagreiningar leiddu í ljós samræmda virkni milli brjóskeyurbrots og aðlögunar stoðkerfisins í kring, sem styður heildræna túlkun á ástandi liðsins. Ennfremur fékkst fram tölfræðilegur áreiðanleiki og stöðugleiki á réttmæti útdreginna meginlegra eiginleika, með raunhæfum breytingum á flokkun gagna.

Með því að sameina háþróaða myndgreiningu og gervigreind, leggja þessar niðurstöður sitt af mörkum til að styrkja áreiðanleika ákveðinna myndrænna atriða fyrir hlutlægt og persónulegt mat á liðbólgu. Reynd aðferðafræði eykur skilning á vefjatengdum víxlverkunum í þróun slitgigtar og styður við nauðsynlegar breytingar í átt að nákvæmri greiningu og markvissum meðferðarúrræðum í slitgigtarmedferð.

Leitarorð: snemmbúin greining slitgigtar í hné, sneiðmyndarannsókn, segulóm-rannsókn, fjölpátta myndgreining, vélanám, mælanlegt mat á slitgigt

Contents

Abstract	iii
Contents	vi
List of Figures	x
List of Tables	xii
List of Abbreviations	xv
Acknowledgments	xvii
List of Publications by the Author	xix
Publications Included in the Thesis	xix
Other Publications by the Author	xx
1 Introduction	1
1.1 Knee Osteoarthritis and Knee Cartilage Evaluation	1
1.1.1 Clinical Overview	1
1.1.2 KOA Assessment	2
1.1.3 Treatment Strategies and Importance of Early Diagnosis	4
1.2 EU project SINPAIN	5
1.2.1 Cartilage Assessment and Digitalization within SINPAIN	7
1.3 AI and Imaging in Healthcare and Precision Medicine	8
1.3.1 Use of Machine Learning in KOA Assessment	9
1.3.2 Use of radiomics in medical imaging and KOA	10
1.4 Research Questions and Objectives	11
2 Quantitative Measures of Knee Osteoarthritis	13
2.1 Image Segmentation	13
2.2 Image Registration	14
2.3 Cohort of data and knee scans	15
2.4 Segmentation and Registration of Bones and Cartilages	17
2.4.1 Bone and Cartilage Segmentation	17
2.4.2 Registration of MRI with CT	19
2.4.3 Refinement of Bone Segmentations	22

2.5	Muscle segmentation	24
2.6	Extracted Features	26
2.6.1	Basic Morphological and Density Features	27
2.6.2	Radiomic Features	28
2.6.3	Summary of Extracted Quantitative Features	28
2.7	Machine Learning Strategies	30
2.7.1	Used Algorithms	30
2.7.1.1	Support Vector Machine	30
2.7.1.2	Logistic Regression	31
2.7.1.3	Decision Tree	31
2.7.1.4	Random Forest	31
2.7.1.5	Gradient Boosting	32
2.7.1.6	K-Nearest Neighbors	32
2.7.2	Feature Normalization	32
2.7.3	Evaluation Metrics	32
2.7.4	Cross-Validation Methods	34
3	Assessing Cartilage Degeneration	37
3.1	Density and Morphological Features	37
3.1.1	Methods	37
3.1.1.1	Statistical Analysis	37
3.1.1.2	Machine Learning Processes	38
3.1.2	Results	38
3.1.2.1	Statistical Analysis	38
3.1.2.2	Machine Learning Performances	39
3.1.2.3	Permutation Importance Results	40
3.1.3	Discussion	44
3.1.3.1	Performance Metrics Among All ML Models	44
3.1.3.2	Key Predictive Features from Statistical and Machine Learning Analyses	44
3.1.3.3	Clinical Interpretation	45
3.2	Radiomic Features	46
3.2.1	Objectives	47
3.2.2	Methods	47
3.2.2.1	Standardization and features selection	47
3.2.2.2	Machine Learning Analysis	48
3.2.3	Results	48
3.2.3.1	Binary classification using the least correlated radiomic features	48
3.2.3.2	Binary classification using the least correlated radiomic features after feature importance	49
3.2.4	Discussion	52
3.2.5	Limitations	54
4	Interplay between cartilage and other tissues	55
4.1	Bones	55

4.1.1	Methods	55
4.1.2	Results	56
4.1.3	Discussion	57
4.2	Muscles	58
4.2.1	Sarcopenia in KOA	58
4.2.2	Cohort for this study	59
4.2.3	Statistical analysis	59
4.2.4	Results and Discussion	60
4.2.4.1	Controls VS Degenerative	60
4.2.4.2	Gender	61
4.2.4.3	Correlation within the KOA group	62
4.2.4.4	Correlation within the KOA-free group	62
4.2.4.5	Correlation within the whole cohort	62
4.3	Multivariate Analysis of Cross-Tissue Relationships	64
4.3.1	Methodology	64
4.3.2	Results	66
4.3.2.1	Cartilage–Bone Interaction	66
4.3.2.2	Cartilage–Muscle Interaction	67
4.3.3	Discussion	68
5	Features validation and optimization	71
5.1	Stability of Density and Morphological Features	71
5.1.1	Imitating Automatic/Manual Error	72
5.1.2	Results: Statistical Analysis	74
5.1.3	Results: Machine Learning Performances	74
5.1.3.1	Pixel Erode 1	75
5.1.3.2	Pixel Dilate 1	75
5.1.3.3	Pixel Close 1 Erode 1	75
5.1.3.4	Pixel Dilate 1 Close 1	76
5.1.4	Discussion	77
5.1.4.1	Comparison of Performance Metrics Between Ero- sion and Dilation Errors	77
5.1.4.2	Comparison of Performance Metrics Between Au- tomated and Manual Errors	77
5.1.4.3	Erosion (Erode 1, Close 1 Erode 1)	77
5.1.4.4	Dilation (Dilate 1, Dilate 1 Close 1)	78
5.1.4.5	Comparisons between Original dataset and Simu- lated Error datasets	78
5.1.4.6	Importance of Mask Segmentation Precision	78
5.1.4.7	Strengths and Limitation	79
5.2	Assessing the Stability of Radiomic Features	79
5.2.1	Robustness analysis	79
5.2.2	Robustness analysis results	81
5.2.3	Discussion	82
5.3	Feature Selection: Density-Morphology Dataset	94
5.3.1	Feature Selection with Correlation and Statistical Analysis	94

5.3.2	Machine learning	95
5.3.3	Results & Discussion	95
5.3.3.1	Reduced set of data	95
5.3.3.2	Machine learning	98
5.4	Ensemble Feature Selection Method in SINPAIN Data	101
5.4.1	Materials and Methods	101
5.4.1.1	SINPAIN Dataset	101
5.4.1.2	Feature Selection Ensemble Method	101
5.4.2	Results	103
5.4.3	Discussion	105
5.5	Summary and Achievements	106
6	Early Stage KOA using Semi-Supervised Learning	109
6.1	Objectives of this Study	109
6.2	Materials and Methods	110
6.2.1	Cohort and Features Utilized	110
6.2.1.1	Label Propagation and Label Spreading	110
6.2.1.2	Specialists' evaluation: the labeled samples	111
6.2.1.3	Labeling Pipeline	111
6.2.1.4	Supervised Learning and Evaluation Metrics	112
6.2.2	Statistical Analysis	113
6.2.2.1	Label Stability	113
6.2.2.2	Group Comparisons	113
6.3	Results	114
6.3.1	Three-class Classification Performances	114
6.3.2	Label Stability	114
6.3.3	Re-labeled Dataset: Statistical Analysis and Distributions	115
6.4	Discussion	116
6.4.1	SSL Pipeline: Labels Stability and Classification Performances	116
6.4.2	Statistical Results and Clinical Interpretation	119
6.4.3	Limitations	120
7	Limitations, Conclusions and Future Developments	123
7.1	Limitations	123
7.2	Conclusions	124
7.3	Future Developments	125

Bibliography**127****List of Figures**

2.1	Cartilage Segmentation on MRI and correspondent 3D parts	18
2.2	Bone Segmentation on CT and correspondent 3D parts	18
2.3	Registration between MRI and CT	20
2.4	Final Result after Registration and Realignment	21
2.5	Three-dimensional reconstructions of knee compartment of a) degenerative and b) control subjects	21
2.6	3D representation of knee cartilages. On the left a 3D reconstruction of cartilages in a control knee (right and left); on the right a 3D reconstruction of degenerative cartilages (right and left knees). The arrows indicate similarity zones between the right and left cartilages of the same patient.	22
2.7	Full Segmentation and Registration Workflow	22
2.8	Cropped femur and tibia (distal femur and proximal tibia) in coronal and sagittal views.	23
2.9	Cortical distal femur, proximal tibia, and patella in coronal and sagittal views.	23
2.10	Trabecular distal femur, proximal tibia, and patella in coronal and sagittal views.	24
2.11	VM: Vastus medialis; VI: Vastus intermedius; VL: Vastus lateralis; BF: Biceps femoris; SART: Sartorius; GRAC: Gracilis; SEMIM: Semimembranosus, SEMIT: Semitendinosus; GMH: Gastrocnemius medial head; GLH: Gastrocnemius lateral head	25
2.12	Segmentation workflow steps	25
3.1	ROC curves with outer cross-validation for classification models trained on density-geometric features. Reported values correspond to mean AUC $\pm$ standard deviation.	40
3.2	Feature importances for the SVM (linear kernel).	41
3.3	Feature importances for the SVM (RBF kernel).	41
3.4	Feature importances for Logistic Regression.	42
3.5	Feature importances for Decision Tree.	42
3.6	Feature importances for Random Forest.	43
3.7	Feature importances for Gradient Boosting.	43
3.8	Framework design	47

3.9	Barplot of the 50 most important robust features for the best model (Linear SVM) in MRI.	50
3.10	Barplot of the 50 most important robust features for the best model (LR) in CT.	51
3.11	Distribution of the top 50 most predictive radiomic features by cartilage area for a) Logistic regression on CT features and b) SVM on MRI features.	52
4.1	Boxplots distributions of the most significant CT-based bone density features across femur, tibia, and patella regions. Green triangles represent mean values.	57
4.2	Boxplot of IMAT density and volume between D and C classes for biceps femoris muscle	61
4.3	Regression plots showing correlation analyses between IMAT/SMT and BMI of BF muscle considering the whole cohort (D+C).	64
4.4	Canonical variate projections for cartilage–bone (left) and cartilage–muscle (right) analyses.	66
4.5	Absolute loadings of cartilage and bone features on the first canonical variate derived from the cartilage–bone CCA	67
4.6	Absolute loadings of cartilage and muscle features on the first canonical variate derived from the cartilage–muscle CCA	68
5.1	Examples of “Original” (center), “Dilate 1” (left), and “Erode 1” cartilage masks and three-dimensional parts for a patient with KOA (Class: “Degenerative” (D)).	73
5.2	Examples of “Original” (center), “Dilate 1” (left), and “Erode 1” cartilage masks and three-dimensional parts for a healthy patient (Class: “Control” (C)).	73
5.3	Morphological operations on extracted masks.	81
5.4	Robustness score boxplots for each radiomics class for different dilation pixels on MRI features	83
5.5	Robustness score boxplots and trendlines for erosion of different pixel sizes on MRI features	84
5.6	Robustness score boxplots for each radiomics class for different dilation pixels on CT features	85
5.7	Robustness score boxplots and trendlines for erosion of different pixel sizes on CT features	86
5.8	Trends of feature robustness in MRI images under morphological perturbations. Each plot shows the distribution and trend of robustness scores across varying pixel-level dilations and erosions.	87
5.9	Trends of feature robustness in CT images under morphological perturbations. The boxplots and trendlines show how robustness scores change with progressive dilation and erosion of the segmented regions.	88
5.10	Barplot of the 50 most important robust features for the SVM linear model in MRI.	89
5.11	Barplot of the 50 most important robust features for the SVM linear model in CT.	90

5.12 Barplot of the 50 most important robust features for the Logistic Regression model in MRI. 91

5.13 Barplot of the 50 most important robust features for the Logistic Regression model in CT. 92

5.14 Correlation Matrix 96

5.15 System’s General Architecture. The pipeline includes preprocessing, feature selection, merging, and classification stages. 102

5.16 Waterfall Selection section of the algorithm, performing exploration and refinement of features. 102

5.17 Feature merging strategy combining repetition, performance, and weighted criteria. 103

5.18 Classification models used for evaluating the feature selection framework. 103

6.1 Workflow of the Labeling Strategy of Class 0 (C0, healthy), Class 1 (C1, early deg), Class 2 (C2, late deg) 111

6.2 Boxplot distributions of the full labeled dataset for each feature, divided by classes 0, 1, and 2 (along x-axis); y-axis represents HU for density features and mm for Vol/Surf ones. 117

List of Tables

2.1 Sex distribution and mean ($\pm$ StDev) of age and BMI by category (TOT, D, C). 17

2.2 Defined volume to be selected during segmentation 26

2.3 Extracted Features divided by Tissue, ROI and Datasets 29

3.1 p-values for density/geometric features, rounded to 3 significant digits; significant values highlighted in ***bold italic***; *: $p < 0.05$, **: $p < 0.005$, ***: $p < 0.001$ 39

3.2 Performance metrics of different models on the density-geometric dataset, best model score for each metric in bold 40

3.3 Hyperparameter set for each machine learning model 48

3.4 Compared Performances among models in MRI radiomics data with 1272 features 49

3.5 Compared Performances among models in CT radiomics data with 958 features 49

3.6 Compared Performances among models in MRI radiomics data with 50 features 50

3.7	Compared Performances among models in CT radiomics data with 50 features	52
4.1	Significant features (p-value < 0.05) among bone density measurements across the femur, tibia, and patella regions.	56
4.2	Comparison of volume and density of skeletal muscle (SMT), and adipose (IMAT) tissues in muscles examined between C and D categories: significant p-values	60
4.3	Muscles significantly correlated with BMI and Age based on density and volume parameters related to SMT and IMAT.	63
4.4	Summary of canonical correlation analyses between cartilage–bone and cartilage–muscle datasets, including canonical correlations, Wilks’ Lambda, Hotelling’s Trace, and significance tests.	66
5.1	Mask Modifications using Morphological Functions in MIMICS. . . .	72
5.2	p-values for Original, Erode1, Dilate1, Close1Erode1, and Dilate1Close1 datasets, rounded to 3 significant digits; significant values highlighted in bold italic , *: p < 0.05, **: p < 0.005, ***: p < 0.001.	74
5.3	Performance metrics of different models on the Erode1 Dataset, best model score for each metric in bold	75
5.4	Performance metrics of different models on the Dilate1 Dataset, best model score for each metric in bold.	76
5.5	Performance metrics of different models on the Close1Erode1 Dataset, best model score for each metric in bold.	76
5.6	Performance metrics of different models on the Dilate1Close1 Dataset, best model score for each metric in bold.	77
5.7	Compared Performances among models in MRI radiomics data with 50 most important robust features	82
5.8	Compared Performances among models in CT radiomics data with 50 most important robust features	93
5.9	Permutation analysis applied after feature selection; p-values approximated by 3 significant digits	97
5.10	ML results for SVM and Logistic Regression. Average accuracy and StDev (in the brackets) among the folds are reported.	98
5.11	Coefficients of SVM per feature rounded to three decimal places . . .	99
5.12	Coefficients of Logistic Regression per feature rounded to three decimal places	99
5.13	Coefficients of SVM per feature for reduced dataset	100
5.14	Coefficients of Logistic Regression per feature for reduced dataset . .	100
5.15	Feature number reduction in SINPAIN datasets	104
5.16	CV (k=5) f1-score, before and after feature selection with the SINPAIN dataset.	104
5.17	Test f1-score after feature selection with the SINPAIN dataset	104
6.1	Hyperparameter grids used for cross-validated grid search.	113

6.2 Comparison of 3 classes algorithm’s metrics with mean $\pm$ standard deviation on ten runs, sorted by F1 Weighted score 114

6.3 Performance metrics for Class 0 across models, reported as mean $\pm$ standard deviation 115

6.4 Performance metrics for Class 1 across models, reported as mean $\pm$ standard deviation 115

6.5 Performance metrics for Class 2 across models, reported as mean $\pm$ standard deviation 115

6.6 Post Hoc Analysis Results for Cartilage Metrics (P-values rounded to 3 significant digits) 118

List of Abbreviations

Abbreviation	Meaning
3D	Three-Dimensional
AI	Artificial Intelligence
ANOVA	Analysis of Variance
Avg	Average
BF	Biceps Femoris
BMI	Body Mass Index
C	Control (healthy samples)
CCA	Canonical Correlation Analysis
CT	Computed Tomography
CV	Cross-Validation
D	Degenerative (KOA samples)
DT	Decision Tree
EU	European Union
FemCart	Femoral Cartilage
FN	False Negatives
FP	False Positives
FPR	False Positive Rates
FS	Feature Selection
GB	Gradient Boosting
GMH	Gastrocnemius Medial Head
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLH	Gastrocnemius Lateral Head
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
GRAC	Gracilis
HU	Hounsfield Unit
IMAT	Intermuscular Adipose Tissue
KL	Kellgren–Lawrence
KNN	K-Nearest Neighbors
KOA	Knee Osteoarthritis
LatTibCart	Lateral Tibial Cartilage
LR	Logistic Regression

MedTibCart	Medial Tibial Cartilage
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NGTDM	Neighbouring Gray Tone Difference Matrix
OA	Osteoarthritis
PatCart	Patellar Cartilage
POP	Popliteus
RBF	Radial Basis Function
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
ROI	Region of Interest
SART	Sartorius
SEMIM	Semimembranosus
SEMIT	Semitendinosus
SMT	Skeletal Muscle Tissue
Surf	Surface Area
SVM	Support Vector Machine
SSL	Semi-Supervised Learning
STL	Standard Triangulation Language
StDev	Standard Deviation
TKA	Total Knee Arthroplasty
TN	True Negatives
TNR	True Negative Rates
TP	True Positives
TPR	True Positive Rates
VM	Vastus Medialis
VI	Vastus Intermedius
VL	Vastus Lateralis
Vol	Volume
Vol/Surf	Volume-to-Surface ratio

Acknowledgments

First of all, I'd like to thank my supervisor, Paolo Gargiulo. When I arrived in Iceland 4 years ago, I was a scared and insecure girl with many dreams but zero courage to pursue them. Paolo, with his empathy, saw through me, considered me, and believed in me. I could not be more grateful for all the opportunities he gave me, which I always accepted with enthusiasm: despite my young age, he never thought I wasn't ready for something, and, therefore, I never felt inadequate. Beyond being an academic guide, Paolo's support and never-ending positivity enlightened my way and gave me confidence. "Those who have a mentor in life are truly fortunate. The path of mentor and disciple leads to personal development and growth."

A huge thanks to my co-supervisor, Halldór Jónsson Jr., for his natural sense of caring, his constant support, his dark humor that made me laugh all the time, and his valuable clinical explanations. His presence in the lab is a significant highlight, creating a sense of moral (and often medical) support.

I thank my family for their constant support over these years we've lived apart: my mum, my dad, and my sister Alessandra. You are the starting point of everything.

A huge thanks to Nicola: your love is so pure. You are probably the first person who didn't make me feel wrong (quite the opposite) when I overshared my thoughts, feelings, and fears. Thank you for making me laugh every time I am feeling down, for your patience in my bad days, and for always being present despite 3500 km of distance. Thank you for always listening to me, supporting me (morally and practically) no matter what, and believing in me every step of the way.

Thanks to all my IBNE fellows, every single person who passed through this lab left something (mostly happy moments, sometimes chaos, but it's all part of the game). Thanks in particular to: Marco, Deborah, Riccardo, Gabriele, Gianmarco, Damiano, Vittorio, Francesca, Simona, Marianna.

A special thanks to Ida: I really don't think you understand how much you meant to me. You made me believe in my capabilities as a "computer engineer". Working next to you, despite the craziness, is a honour and gives me the opportunity to learn something new every day.

Thanks also to Cami (even though she's not part of our lab), who, even after all these years (and all these theses), is never tired of telling me I should stop feeling "never enough".

In this journey, I met amazing people I'd like to mention: Lorena, Luka, Lisa Marie, Luisa, Dan, Sydney, Alva, Stefan, Sebi, Yuri, Marton, Paula, and Clara. At times, you have been like a family to me; you took care of me unconditionally, encouraged me, never judged me, and made me happier than I have ever been. I am, and always will be, infinitely grateful for your love and support over the years, and sometimes, beyond the distance.

Thanks to my beloved Háskólaþórinn: singing with you was an enjoyable break from work and from my overthinking, never-stopping mind.

Thanks to everyone who has been part of this journey and crossed my path.

Last but not least, I'd like to thank the SINPAIN project. This work (and my dream to start a career in research) would have never been possible without the grant received by Horizon Europe (Grant agreement ID: 101057778). Beyond the financial support, being part of this consortium was an incredible opportunity that allowed me to grow both personally and as a researcher in this field. I'd like to thank all the researchers and professors for accepting and integrating me into the group and allowing me to contribute and feel part of an inspiring scientific community.

List of Publications by the Author

Publications Included in the Thesis

Most of the text in this PhD thesis is adapted from peer-reviewed publications authored by me, Federica Kiyomi Ciliberti, at the Institute of Biomedical and Neural Engineering, Reykjavik University, Iceland. These works were produced during my doctoral studies and have been published in peer-reviewed journals and conference proceedings, or are currently under review. As these papers were written in collaboration, portions of the text and concepts were developed jointly with my co-authors. All previously published material is included with their permission.

The publications list follows a chronological order starting from the oldest:

- F. K. Ciliberti, L. Guerrini, A. E. Gunnarsson, et al., “CT- and MRI-based 3D re-construction of knee joint to assess cartilage and bone”, *Diagnostics*, vol. 12, 2022. DOI: 10.3390/diagnostics12020279. [1]
- R. Aubonnet, J. Ramos, M. Recenti, et al., “Toward new assessment of knee cartilage degeneration,” *Cartilage*, vol. 14, no. 3, pp. 351–374, 2023. [2]
- F. Angelone, F. K. Ciliberti, G. P. Tobia, et al., “Innovative diagnostic approaches for predicting knee cartilage degeneration in osteoarthritis patients: A radiomics-based study”, *Information Systems Frontiers*, pp. 1–23, 2024. DOI: 10.1007/s10796-024-10527-5. [3]
- F. K. Ciliberti, I. Maruotto, H. Jónsson jr, M. K. Gíslason, and P. Gargiulo, “Multimetric evaluation of knee cartilage degeneration”, in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, IEEE, 2024, pp. 1022-1027. DOI: 10.1109/MetroXRaine62247.2024.10795901. [4]
- F. K. Ciliberti, G. Pianegonda, H. Jónsson jr, M. K. Gíslason, and P. Gargiulo, “How muscle and intramuscular adipose tissue volume and density progress in knee osteoarthritis-affected subjects: A virtual histology study”, in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, IEEE, 2024, pp. 177–182. DOI: 10.1109/MetroXRaine62247.2024.10796347 [5]

- I. Maruotto, F. K. Ciliberti, P. Gargiulo, and M. Recenti, “Feature selection in healthcare datasets: Towards a generalizable solution”, *Computers in Biology and Medicine*, vol. 196, p. 110 812, Sep. 2025, DOI: 10.1016/j.compbiomed.2025.110812. [6]

Other Publications by the Author

- F. K. Ciliberti, G. Cesarelli, L. Guerrini, et al., “The role of bone mineral density and cartilage volume to predict knee cartilage degeneration”, *European Journal of Translational Myology*, vol. 32, 2022. DOI: 10.4081/e-jtm.2022.10678. [7]
- A. E. Gunnarsson, F. K. Ciliberti, C. Belfiori, et al., “Assessment of Femoral Cartilage Morphological and Topological Features Using Machine Learning”, *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, Rome, Italy, 2022, pp. 277-282, DOI: 10.1109/MetroXRaine54828.2022.9967659. [8]
- F. K. Ciliberti, R. Forni, D. Coato, G. Dolino et al., “Development of Synthetic 3D Printed Knee Joint to Assess Mechanical and Functional Properties of Degenerative Cartilage”, *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, Milano, Italy, 2023, pp. 628-633, doi: 10.1109/MetroXRaine58569.2023.10405695. [9]
- G. Dolino, D. Coato, R. Forni, G. Boretti, F. K. Ciliberti, and P. Gargiulo, “Designing a synthetic 3D-printed knee cartilage: FEA model, micro-structure and mechanical characteristics”, *Applied Sciences*, vol. 14, no. 1, 2024. DOI: 10.3390/app14010331. [10]
- F. Angelone, F. K. Ciliberti, H. Jónsson jr, M. K. Gíslason, and P. Gargiulo, “Knee cartilage degradation in the medial and lateral anatomical compartments: A radiomics study”, in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, IEEE, 2024, pp. 183-188. DOI: 10.1109/MetroXRaine62247.2024.10796456. [11]

Chapter 1

Introduction

The present chapter provides the contextual and theoretical foundation for this thesis. It introduces the clinical and biological aspects of knee osteoarthritis, outlines current approaches to its diagnosis and treatment, and discusses recent advances in medical imaging and artificial intelligence that are reshaping how degenerative joint diseases are studied and managed. The chapter also situates this work within the broader framework of European research initiatives, particularly the SINPAIN project, which aims to develop integrated therapeutic and digital strategies for knee osteoarthritis. Finally, it formulates the research questions and objectives that guide the subsequent chapters, emphasizing the role of multimodal imaging and machine learning in establishing quantitative, reproducible, and personalized assessment tools for knee osteoarthritis.

1.1 Knee Osteoarthritis and Knee Cartilage Evaluation

1.1.1 Clinical Overview

Osteoarthritis (OA) is the most common form of arthritis and a leading cause of pain and disability worldwide, especially among middle-aged and older adults. It is a chronic, progressive, degenerative disease that primarily affects the joints, leading to the breakdown of articular cartilage, changes in the underlying bone (subchondral bone), synovial inflammation, and the formation of bony growths called osteophytes [12], [13]. OA most frequently impacts the knees, hips, hands, and spine, causing symptoms such as joint pain, stiffness, swelling, and reduced function, which can significantly decrease quality of life [14]. Pain in OA is complex, evolving from sporadic weight-bearing discomfort to chronic pain [13], [15].

The knee joint is particularly susceptible to OA, and knee osteoarthritis (KOA) represents the most common form, affecting millions of people worldwide. KOA is a multifaceted disorder involving the articular cartilage, subchondral bone, ligaments, and synovial membranes. It is primarily characterized by cartilage degeneration resulting mostly from nontraumatic daily activity [16], [17]. OA arises from a combination of systemic and local factors. Systemic influences include age, gender (with women being more prone), genetics, and hormonal changes such as those

related to menopause [16]–[20]. Local or biomechanical factors include obesity, joint trauma or surgery, repetitive joint stress, physical activity, muscle weakness or imbalance, baseline cartilage volume, metabolic disorders, and previous OA diagnoses in other joints [13], [15], [16], [18], [21], [22].

KOA symptoms are debilitating and negatively impact quality of life, including pain, tenderness, stiffness, swelling, and decreased mobility [14], [16]. Other clinical indications include quadriceps muscle atrophy, crepitus during movement, limited joint range of motion, and valgus or varus deformities [14], [16]. Chronic pain and reduced mobility also contribute to depression, anxiety, and sleep disturbances associated with physical limitations [23].

KOA is a progressive disease with no known disease-modifying or curative agents [14]. Consequently, symptoms often worsen until total knee arthroplasty (TKA) becomes necessary, typically 9–12 years after symptom onset [24]–[26]. Conservative treatments such as physical therapy, medication, injections, and osteotomies should begin early to delay or prevent the need for surgery [16], [27], [28].

KOA, moreover, represents a substantial burden on healthcare systems, with significant economic implications globally [29]. Patients often face considerable expenses related to ongoing treatment, management, and surgical interventions, which can accumulate over a lifetime. Procedures such as knee replacement and pain management therapies contribute notably to these costs [30], [31]. Patients who qualify for TKA but cannot access it fall into the “treatment gap,” where severe pain, decreased quality of life, and heavy financial burdens persist [25].

The likelihood of developing symptomatic KOA is high, particularly among older adults and individuals with obesity [22]. As populations continue to age and obesity rates increase, the prevalence of osteoarthritis is expected to grow, further amplifying its health and economic impact [15], [32].

Globally, the prevalence of KOA continues to rise. Estimates suggest there will be 642 million individuals diagnosed by 2050 [33]. The prevalence is around 13% in women and 10% in men aged 60 and above, increasing to nearly 40% at age 70 [16], [33]. Despite only 15–50% of patients with radiological KOA reporting symptomatic pain, the disease accounts for 12.1 million years lived with disability (YLD) worldwide, equivalent to 1.4% of global YLD [16], [17], [31].

1.1.2 KOA Assessment

Assessment of cartilage condition and thickness is crucial for both the detection and monitoring of the progression of KOA. Diagnosis is usually based on a combination of clinical evaluation and imaging. Planar X-rays of the knee are routinely used; however, soft tissue is not adequately visualized, nor is this modality sensitive to subtle changes over time [34]. Magnetic resonance imaging (MRI) is the state-of-the-art modality for assessing hyaline cartilage and has seen rapid developments in hardware, pulse sequences, and image analysis [34]–[36]. MRI enables both visual and quantitative evaluation of cartilage volume, thickness, and biochemical composition, while also imaging other tissues involved in KOA such as the subchondral bone, meniscus, and soft tissue. Computed tomography (CT) provides

excellent 3D representations of cortical bone, osteophytes, and calcifications, and has been used to investigate KOA-related joint changes, including trabecular bone remodeling, subchondral cysts, and sclerosis [37].

The severity of KOA is most commonly assessed radiographically by evaluating the degree of joint space narrowing and damage to cartilage and bone. The Kellgren–Lawrence (KL) grading system [38] remains the most widely used classification for radiographic osteoarthritis. It defines five grades (0–4) based on characteristic features observed in X-rays: grade 0 indicates no radiographic features of OA; grade 1 corresponds to doubtful joint space narrowing and possible osteophytic lipping; grade 2 denotes definite osteophyte formation and possible joint space narrowing; grade 3 involves multiple osteophytes, definite joint space narrowing, sclerosis, and possible bony deformity; and grade 4 reflects severe OA with large osteophytes, marked joint space narrowing, pronounced sclerosis, and definite bony deformity. The appearance of a definite osteophyte (KL grade 2) is generally regarded as the threshold for a structural diagnosis of KOA. Despite its simplicity and extensive clinical use, the KL system has notable limitations: it assumes a unidirectional and continuous structural progression of disease, does not account for compartmental variation within the joint, and can be influenced by radiographic positioning or observer subjectivity [34].

To address some of these limitations, the OA Research Society International (OARSI) Atlas system was developed as a semi-quantitative alternative, allowing separate evaluation of osteophytes and joint space narrowing within individual compartments of the knee. The OARSI Atlas provides a more detailed visual reference for grading OA features, improving interobserver reliability compared with KL. However, it primarily focuses on the tibiofemoral joint, often underrepresenting the patellofemoral compartment and thereby underestimating disease burden in cases where anterior knee structures are primarily affected [34].

Another frequently used classification, the Ahlbäck scale [39], was specifically designed to quantify the degree of joint space narrowing as an indicator of cartilage loss. It categorizes disease severity from Grade I (joint space narrowing less than 3 mm) to Grade V (complete obliteration of the joint space and bone attrition exceeding 5 mm). Although the Ahlbäck scale provides a more direct assessment of structural cartilage loss compared with KL, several studies have reported only low-to-moderate inter- and intraobserver reliability, particularly in early-stage KOA [40]. The dependence on subjective interpretation of subtle radiographic features and sensitivity to knee positioning contribute to variability in its application.

Recent work by Wing *et al.* [41] further examined this issue by directly comparing the intra- and inter-observer reliability of KL, Ahlbäck, and the International Knee Documentation Committee (IKDC) classifications grading systems among observers of differing clinical experience. Their results demonstrated substantial to almost perfect inter-observer agreement overall, but also revealed a clear dependence on experience level: senior clinicians showed good to excellent intra-observer reliability, while less experienced observers achieved only moderate reliability. The authors concluded that although these systems perform acceptably in experienced hands, the evaluation of KOA from plain radiographs remains inherently subjective, particularly in borderline or early disease stages. These findings highlight the

limitations of radiographic grading due to its reliance on visual interpretation and observer judgment. They emphasize the need for more objective and quantitative approaches in assessing KOA, such as automated image analysis or MRI-based metrics.

Other comparative studies among these radiographic grading systems (KL, OARSI, and Ahlbäck) have consistently demonstrated only moderate correlation with arthroscopic or intraoperative findings [42], [43]. This discrepancy highlights a general limitation of conventional radiography: while it provides an accessible and cost-effective means of evaluating joint degeneration, it lacks the sensitivity and specificity required to detect early biochemical or microstructural cartilage changes. Consequently, radiographic grading remains primarily a structural, rather than biological, measure of disease burden.

Semi-quantitative MRI-based systems such as the Whole Organ Magnetic Resonance Imaging Score (WORMS) and Knee Osteoarthritis Scoring System (KOSS) integrate multiple MRI features from the entire joint, including cartilage thickness, bone marrow lesions, and subchondral cysts, offering greater sensitivity than traditional radiographic methods [34]. Quantitative MRI measures such as cartilage volume and thickness, treated as continuous variables, require segmentation of the cartilage from high-resolution 3D images. Studies have shown that reduced baseline cartilage thickness in the medial femur predicts longitudinal cartilage thinning [44]. Three-dimensional cartilage surface mapping (3D-CaSM) allows tracking of small (6-month) changes in thickness [45], and recent work has demonstrated the potential of 3D ultrasound, registered with MRI, for quantitative cartilage assessment [46]. Imaging also plays a key role in post-injury evaluation, as knee trauma increases KOA risk. MRI-based systems such as the Anterior Cruciate Ligament Osteoarthritis Score (ACLOAS) [47] integrate both traumatic and degenerative features to monitor injury-induced OA longitudinally.

1.1.3 Treatment Strategies and Importance of Early Diagnosis

Treatment modalities for KOA include physical therapy and exercise, medication, injections, osteotomies, and, in severe cases, partial or total knee replacement operations [27]. Current best practice emphasizes a multimodal strategy combining non-pharmacological and pharmacological interventions, with core recommendations of patient education, structured exercise programs, and weight management for overweight individuals because these consistently improve symptoms and function [27], [48]–[50]. Topical non-steroidal anti-inflammatory drugs (NSAIDs) are the first-line pharmacological option, while oral NSAIDs and intra-articular injections (for example, corticosteroids or hyaluronic acid) are reserved for persistent pain after considering patient comorbidities [48], [50]. Given that KOA is a progressive and currently incurable disease, most patients experience gradual symptom worsening until total knee arthroplasty (TKA) becomes necessary, typically within 9–12 years after symptom onset [14], [24]–[26]. Because KOA evolves slowly, early diagnosis is crucial to allow timely, less invasive interventions such as physiotherapy, lifestyle modification, and medication, which may slow structural degeneration, alleviate symptoms, and delay or reduce the need for TKA [14]. Assessing KOA

severity remains challenging because traditional radiographic grading systems (for example, KL and Ahlbäck) are subjective and insensitive to early cartilage damage, so advances in MRI, CT, and quantitative imaging are essential to develop more sensitive, objective diagnostic criteria [38], [39], [51]. Recent regenerative strategies, such as intra-articular injections of platelet-rich plasma (PRP) and cell-based therapies, particularly mesenchymal stem cells, have shown short- to mid-term promise in supporting cartilage repair and improving pain and function up to one year post-treatment, although long-term efficacy and disease-modifying potential remain under investigation [52]–[54]. In this context, the EU Horizon2020 “RESTORE” project was founded to develop smart, nano-enabled 3D biomaterial matrices for cartilage regeneration that integrate bioprinting, nanocarriers, and stimuli-responsive materials [55]. One of its scientific outputs is the Evenbratt *et al.* review, which underscores the promise of stem cell-based cartilage regeneration, particularly through strategies involving induced pluripotent stem cells and biomolecular gradients—for achieving durable functional repair [56].

Conservative management should begin as early as possible to delay or prevent surgical intervention [16], [28]. When these measures fail and patients qualify for TKA but cannot access it, they fall into the “treatment gap,” a period marked by severe pain, diminished quality of life, and increased financial burden [25].

In summary, early and accurate diagnosis is fundamental for optimal KOA care. It enables targeted, non-invasive treatment before irreversible joint damage occurs and provides a critical window for emerging regenerative therapies (like those championed by RESTORE) that may transform disease progression, improve patient outcomes, and reduce healthcare costs.

1.2 EU project SINPAIN

Building on the progress achieved through the RESTORE project, which advanced nanotechnology-based and stem cell-driven strategies for cartilage regeneration [56], the SINPAIN initiative represents a new step toward integrated, multimodal therapies for KOA. SINPAIN (“A cost-effective combined advanced therapy to treat knee osteoarthritis”) is a Horizon Europe Health project launched in May 2022. With a total budget of approximately €5.3 million, the project will run for 54 months, concluding in November 2026. It brings together twelve partners from nine countries under the coordination of CIDETEC Nanomedicine (Spain) [57].

The overarching goal of SINPAIN is to develop an innovative therapeutic platform for knee osteoarthritis (KOA) that integrates small interfering RNA (siRNA) therapeutics with nanocarrier-based delivery systems and advanced intra-articular hyaluronic acid (IA-HA). This combined strategy seeks to simultaneously target inflammation and pain while incorporating digital tools to enable patient stratification and personalized treatment approaches.

KOA affects more than 500 million people globally, yet existing therapies remain primarily palliative, offering only temporary symptom relief and often accompanied by limited efficacy or undesirable side effects. No curative treatment is currently available. SINPAIN addresses this unmet clinical need through a multifaceted strat-

egy that targets key molecular drivers of pain and inflammation, particularly nerve growth factor (NGF) and interleukin-1 β (IL-1 β), while employing nanocarriers for precise intra-articular siRNA delivery. In addition, the project combines gene-level modulation with an improved IA-HA formulation designed to enhance lubrication and provide regenerative support. Image-derived features and machine learning techniques are integrated to refine diagnosis and tailor interventions to individual patient profiles.

Scientifically, the project combines wet-lab, preclinical, and computational research streams. Initial efforts focus on the design of siRNA molecules and nanocarriers based on lipid and polymer formulations, optimized for silencing efficiency, biocompatibility, and targeting specificity. Parallel work involves the engineering of IA-HA with superior rheological and dynamic properties, which are incorporated into successive generations of nanotherapeutic combinations. Mechanistic studies are supported by single-cell transcriptomics of OA joint tissues (including analyses of sex-specific differences) and by the development of advanced 3D tissue models and bioreactor systems that reproduce the complex biomechanical environment of the joint, including cartilage–pericyte and cartilage–neuron interactions under dynamic loading.

Preclinical validation forms another critical component, where the efficacy and toxicity of the developed therapies are assessed in relevant *in vivo* OA models. Alongside these experimental activities, the project adopts a quality-by-design (QbD) approach to guide scalable manufacturing and ensures alignment with the complex regulatory landscape that governs hybrid advanced therapies. Complementary to the biological and regulatory aspects, SINPAIN also advances digital health tools through radiomic analysis of X-ray, CT, and MRI data. These imaging-derived biomarkers are integrated into machine learning models that enable stratification of patients and support data-driven clinical decision-making.

It is divided in the following Work Packages (WP):

- WP1 - Development of RNA-based Nanoformulations: siRNA nanoformulations (NGF, IL-1 β), HA-(nano)emulsions, *in vitro* safety.
- WP2 - Immunological Interaction and Safety Studies of the Advanced Nanotherapeutics: IA-HA engineering and generation of nanotherapeutics (Gen1–4).
- WP3 - *In vitro* Mode of Action for Osteoarthritis: Single-cell profiling, 3D OA models, silencing efficacy studies.
- WP4 - Bioreactor-based Model for Potency Assay: Bioreactor potency models (chondrocyte–pericyte, chondrocyte–neuron under load).
- WP5 - *In vivo* validation: *In vivo* validation of efficacy, safety, and toxicity.
- WP6 - Regulation and Quality-by-Design: Regulatory mapping and quality-by-design manufacturing.
- WP7 - Cartilage Assessment and Digitalization: Patient-specific Profiling: Imaging-features pipeline and patient profiling.

- WP8–10: Communication, exploitation, management, and ethics.

The SINPAIN consortium consists of twelve partners spanning nine countries, combining expertise from academic, clinical, and industrial sectors. The coordinator, CIDETEC Nanomedicine (Spain), is joined by research institutions including INSERM (France), Saarland University and Eurice (Germany), University of Liverpool (UK), Reykjavik University (Iceland), Università del Piemonte Orientale (Italy), i3S (Portugal), AO Research Institute Davos (Switzerland), and UCLA (USA). Industrial and clinical translation are supported by OZ Biosciences (France) and Asphalion (Spain).

Despite its promise, the project faces several challenges. Achieving durable and safe intra-articular delivery of siRNA remains a demanding task, both scientifically and technically. The hybrid nature of the therapeutic approach introduces regulatory complexities that require careful navigation to achieve compliance across different jurisdictions. Furthermore, the inherent heterogeneity of osteoarthritis progression, shaped by genetic, mechanical, and lifestyle factors, poses an additional obstacle to achieving uniformly effective treatment outcomes.

If successful, SINPAIN will deliver a first-of-its-kind combination therapy that unites molecular, mechanical, and digital strategies within a single therapeutic framework. Beyond its direct clinical implications, the project has the potential to establish a new paradigm for the treatment of degenerative joint diseases, significantly improving the quality of life of KOA patients across Europe and beyond.

1.2.1 Cartilage Assessment and Digitalization within SINPAIN

A distinguishing element of the SINPAIN project is its integration of a digital layer to support personalization of therapy. While the therapeutic components (siRNA nanocarriers, improved viscosupplements) target the biological and mechanical dimensions of OA, the digital pipeline aims to capture and model patient heterogeneity. This is essential because OA progression is multifactorial, involving differences in cartilage morphology, bone remodeling, inflammation, innervation, and patient lifestyle factors [57].

The digital layer is designed to:

- Extract quantitative imaging biomarkers from MRI, CT, and X-ray of the knee joint, using radiomic feature extraction.
- Build patient-specific profiles that combine morphological, textural, and functional data.
- Apply machine learning models to stratify patients according to risk and treatment response likelihood.
- Provide a decision-support system to help align the biological nanotherapeutics with the patient's individual disease trajectory.

The present dissertation focuses specifically on this digital personalization pipeline. The work is concerned with:

1. Subject recruitment: Defining inclusion/exclusion criteria, establishing informed consent procedures, and ensuring ethical compliance for patient participation in the imaging cohorts.
2. Data acquisition: Collection of multi-modal imaging datasets, including MRI, CT, and radiographs, from recruited KOA patients.
3. Data processing: Development of pipelines for pre-processing, normalization, and quality assurance of imaging data to enable robust radiomic feature extraction.

1.3 Artificial Intelligence and Imaging in Healthcare and Precision Medicine

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) studies, and nowadays is one of the most important available solutions for processing data. Indeed, it can be defined as the study of how computer algorithms can learn from data and produce models that demonstrate relationships and influences on an item of interest [58]. A widely used formal definition, provided by Tom Mitchell, states that “A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ” [59]. ML techniques aim to improve themselves as much as they can, and they differ in the applied method. We include, for example, supervised, unsupervised, semi-supervised, and reinforcement learning.

In supervised learning, models are trained using a labeled dataset, meaning that each input is associated with a known output or target value. The algorithm learns to map inputs to outputs and can then predict labels for unseen data. Common applications include classification (e.g., spam detection, image recognition) and regression (e.g., predicting prices or trends).

In contrast, unsupervised learning deals exclusively with unlabeled data. The goal is to uncover hidden patterns, structures, or relationships within the data without prior knowledge of outcomes. Typical approaches include clustering (grouping similar data points) and dimensionality reduction (simplifying data representation while preserving essential information).

Semi-Supervised Learning (SSL) lies between supervised (where labeled data is provided) and unsupervised learning (where no labeled data is available). It addresses the challenge of limited labeled data (often costly, time-consuming, or requiring expert and/or subjective judgment) by training models on a small labeled dataset alongside a larger set of unlabeled data. The assumption is that unlabeled data contains valuable structural patterns which, when combined with labeled examples, improve the model’s generalization and efficiency. SSL is frequently used to infer labels for unlabeled data by uncovering underlying relationships and patterns.

Finally, reinforcement learning focuses on learning optimal actions through interaction with an environment. Instead of learning from fixed datasets, an agent learns by receiving feedback in the form of rewards or penalties, gradually improving

its strategy to maximize long-term success. RL is widely applied in fields such as robotics, game playing, and autonomous systems.

As we are required to process more and more data, ML techniques have evolved to accommodate heterogeneous fields. As an example, ML is used for early detection of malware intrusion and other cyber threats [60], or in agricultural contexts, to identify possible diseases in the crops [61]. ML is also used in marketing contexts, where providing recommendations to users is fundamental [62], but also in business and finance, to perform risk assessment or fraud detection [63].

A huge breakthrough has been achieved with the implementation of ML in the biomedical context. In this field, ML is a critical support instrument for diagnosis, allowing for the detection of diseases at earlier stages than traditional methods, improving diagnostic accuracy, reducing costs, and improving patient outcomes across a range of conditions, including cancer, musculoskeletal, and neurodegenerative diseases [64]–[66].

SSL strategies have also become pivotal in biomedicine, with applications in medical image analysis, text classification, and diagnostics. Techniques include pseudo-labeling, consistency regularization, graph-based methods, contrastive learning, and teacher–student models, often built on assumptions such as smoothness, clustering, and low-density separation [67]–[69]. SSL has achieved performance comparable to fully supervised models while requiring fewer annotations, particularly in segmentation, classification, and cancer diagnostics. However, challenges remain, including violations of SSL assumptions, class imbalance, distribution shifts, and model uncertainty [67], [70].

Deep learning has notable applications in medical image analysis, genomics, and disease diagnosis. For example, deep learning models such as convolutional neural networks (CNNs) are widely used for classifying medical images, detecting tumors, segmenting tissues, and aiding in cancer diagnosis from radiological and histopathological images, often achieving accuracy comparable to or surpassing human experts [71], [72]. Other biomedical applications include processing electronic health records using natural language processing, predicting disease outcomes, and supporting treatment selection, especially in oncology, where deep learning integrates diverse data types for precision medicine [73], [74].

1.3.1 Use of Machine Learning in KOA Assessment

The prominent role of ML in biomedical contexts extends to the classification, diagnosis, and treatment planning of KOA. Numerous classification and regression models, deep learning CNNs, and feature engineering techniques such as principal component analysis (PCA) have been trained on medical images to assess disease severity, achieving accuracies typically between 75% and 92% [58], [75]–[79]. Beyond diagnosis, ML algorithms have also been used to predict KOA progression from X-ray images and general patient health information, achieving AUC and F1 scores around 79% across different studies [80], [81].

Previous research has explored diverse data modalities, feature extraction strategies, and learning algorithms for KOA assessment. For example, Abedin *et al.* [77] combined questionnaire data and X-rays using LASSO-based feature selection

and found that CNNs achieved the lowest prediction error among several models. Ashinsky *et al.* [75] employed MRI data with a wavelet network classifier, reporting 75% accuracy, while Donoghue *et al.* [78] used Laplacian Eigenmaps and linear regression on MRI scans, achieving R^2 values up to 0.75. Similarly, Du *et al.* [79] applied PCA to MRI features and tested several classifiers—including artificial neural networks (ANNs), random forests (RF), and support vector machines (SVMs)—obtaining AUCs between 0.76 and 0.79. Marques *et al.* [76] combined texture features and partial least squares (PLS) with Fisher’s linear discriminant analysis, reaching an AUC of 0.92, while Pedoia *et al.* [82] integrated MRI and biomechanical data through topological data analysis, achieving an AUC of approximately 84%. Collectively, these works highlight the versatility of ML techniques in handling multimodal KOA data and demonstrate how careful feature engineering and model validation can yield robust predictive performance.

More recent studies have focused on SSL frameworks for early KOA detection on X-rays, including self-ensembling teacher–student networks [83], [84], semi-supervised anomaly detection [85], and active learning with consistency regularization [86]. These methods achieve performance comparable to or exceeding fully supervised models while requiring fewer labeled samples, effectively mitigating the challenge of limited annotated medical data [84], [86].

Deep learning architectures such as ResNet, EfficientNet, and InceptionV3 are now widely used for feature extraction and classification from MRI and X-ray images [87]–[89]. Ensemble and hybrid approaches, such as the CDK ensemble proposed by Islam and Rony [88], have further enhanced performance, with some models surpassing 99% accuracy in KOA detection [87], [88].

Most works, however, rely primarily on X-ray or MRI data, with relatively few incorporating CT. Although CT effectively captures bone density and structural changes such as osteophytes, sclerosis, and cysts, it remains less suitable for evaluating soft tissues like cartilage, which are central to KOA pathology. Persistent challenges also include managing class imbalance, reducing label noise, and integrating multimodal data for more comprehensive and generalizable assessment frameworks.

1.3.2 Use of radiomics in medical imaging and KOA

Radiomics involves the extraction of quantitative metrics known as radiomic features from medical images to capture tissue properties. It can reveal quantitative information and properties often imperceptible to the human eye, that can result to be relevant in clinical analysis (combined with various types of data) and can offer valuable insights into the characteristics and progression of diseases. The impact of radiomics can be enhanced by the use of ML: integration of artificial intelligence, machine learning, and advanced image analysis techniques to discern meaningful information from medical images. By integrating radiomic data with clinical, genomic and biological information, researchers can unravel complex disease pathways and refine patient stratification strategies [90]. Radiomics has been used for a large number of medical purposes, such as the detection of disease risk factors [91]–[94], detection of genomic patterns [95], [96], the detection of

soft tissue masses, bone tumors and calcifications [97], [98] for various diagnostic imaging techniques. Radiomic features extracted from CT scans and analyzed using machine learning (ML) techniques have proven to be valid and robust across various oncological diagnostic applications, including liver tumors [99] and head and neck cancer [100]. A study focusing on lung cancer emphasized the necessity of identifying optimal machine-learning methods for radiomic applications to generate stable and clinically relevant radiomic biomarkers applicable in clinical practice [101].

To date, radiomics has not been extensively used in the field of KOA. Jiang *et al.* [102] discuss how radiomics can enhance accurate OA diagnosis by converting medical images into quantitative data. The study highlights the limitations of current imaging techniques like X-ray, CT, and MRI in early OA detection and stresses the importance of radiomics in addressing this issue. Furthermore, it reviews current research on radiomics in OA, detailing studies that classify, detect, and predict OA in various joints (e.g., knee, hip, hand, spine). Radiomics offers advantages over traditional clinical variables in detecting specific forms of OA and shows promise in improving diagnostic accuracy and prognosis prediction. In [103], they extracted radiomic features from T2-weighted MRI images of subjects with post-traumatic osteoarthritis and a control group, focusing on subchondral bone and femoral and tibial cartilage. Radiomic analysis revealed a distinctive signature in the cartilage and subchondral bone of various compartments, showing excellent performance in distinguishing between patients and controls. The articles [104] and [105] explore the use of MRI-based radiomics approaches for diagnosing KOA by analyzing subchondral bone. Both approaches achieve high performance in distinguishing between OA subjects and healthy control class.

1.4 Research Questions and Objectives

Despite continuous progress in imaging and computational methods, the clinical assessment of KOA remains largely qualitative and dependent on observer expertise. This subjectivity hinders early diagnosis, objective monitoring, and personalized treatment planning. The overarching aim of this thesis is to explore whether the integration of multimodal imaging data with ML can provide reproducible and quantitative tools for KOA evaluation. Accordingly, the following research questions are addressed:

1. How can multimodal imaging data (MRI and CT) be combined and processed to obtain reproducible, quantitative measures of knee joint morphology and tissue properties?
2. Which image-derived biomarkers effectively differentiate between healthy and degenerative knee tissues, and how can machine learning support their identification?
3. How robust and reproducible are the extracted features and classification outcomes when subject to segmentation errors, variability, and limited data availability?

4. How can machine learning models utilize imaging biomarkers to distinguish between different stages of the disease, including early-stage knee osteoarthritis?

By addressing these questions, the thesis aims to establish a quantitative and data-driven framework for KOA assessment, bridging the gap between imaging biomarkers and clinical interpretation. Ultimately, this work contributes to developing reproducible, interpretable, and clinically meaningful diagnostic tools that support precision medicine in osteoarthritis management.

The objectives of this research are closely aligned with the research questions outlined above and are progressively addressed throughout the thesis. Each chapter contributes to a specific aspect of the overall goal: developing a quantitative and machine learning-based framework for knee KOA assessment.

Chapter 2 introduces the methodological foundation, describing the multi-modal (MRI-CT) segmentation and registration pipeline developed to obtain reliable three-dimensional representations of knee cartilage and bone. Chapter 3 focuses on the extraction and analysis of quantitative imaging biomarkers, including morphological, density-based, and radiomic cartilage features, and their use in supervised machine learning models to differentiate healthy from degenerative tissues. Chapter 4 explores the interplay between cartilage, bones and muscles in cartilage degeneration and KOA development.

Chapter 5 investigates the robustness of the proposed framework, assessing how segmentation variability, operator-dependent factors, and limited data availability influence the stability and reproducibility of extracted features and model performance.

Chapter 6 explores semi-supervised learning as a strategy to uncover an intermediate or early degenerative stage of KOA, which remains difficult to delineate even for clinical experts.

Finally, Chapter 7 summarizes the findings, discusses the clinical implications, and outlines future directions for integrating quantitative imaging and machine learning into precision diagnostics for osteoarthritis.

Chapter 2

Quantitative Measures of Knee Osteoarthritis

This chapter presents new quantitative methods that utilize image processing and machine learning for assessing KOA. These methods are designed to address the limitations of the current diagnostic gold standard and to provide an objective evaluation of cartilage and KOA diagnosis. The method relies on the use of MRI and CT data combined with machine learning methodologies. The purpose of this methodological framework is to obtain reproducible, quantitative measurements of cartilage morphology and density across subjects. MRI provides high contrast for soft tissues, enabling accurate cartilage segmentation, while CT offers precise bone delineation and quantitative density information. The integration of both modalities allows multimodal feature extraction for downstream analysis (described in Chapters 3, 4, 5, 6).

The section 2.4 of this chapter is adapted from two studies: “CT- and MRI-Based 3D Reconstruction of Knee Joint to Assess Cartilage and Bone” [1] published in *Diagnostics*, MDPI, and “Toward New Assessment of Knee Cartilage Degeneration” [2] published in *CARTILAGE*, Sage. The section 2.5 is adapted from the study entitled “How muscle and intramuscular adipose tissue volume and density progress in knee osteoarthritis-affected subjects: a virtual histology study” [5], presented at the IEEE MetroXRaine Conference, 2024, St. Albans, London, UK.

2.1 Image Segmentation

Image segmentation is a key process in medical image analysis. It involves dividing a medical image into distinct regions to identify and isolate anatomical structures or pathological areas. The main purpose of segmentation is to simplify image representation, making it easier to measure and analyze specific organs or tissues and identify anatomical structures and abnormalities. This process supports medical diagnosis, treatment planning, and disease monitoring [106]. In fact, it enables:

- Accurate detection and measurement of tumors, lesions, and other abnormalities.

- Planning and guidance of medical procedures such as surgery and radiation therapy.
- Quantitative analysis of organ volumes and tissue composition.
- Longitudinal studies of disease progression across multiple scans.

In summary, segmentation helps convert raw medical images into meaningful clinical information. There are three main types of segmentation used in medical imaging:

- **Manual Segmentation:** Experts, such as radiologists, manually outline regions of interest. This method is accurate but time-consuming and may vary between observers.
- **Semi-Automatic Segmentation:** Combines human input with computer algorithms. Users provide initial boundaries or seed points, and the software refines the segmentation. It balances accuracy with efficiency.
- **Automatic Segmentation:** Utilizes artificial intelligence, especially deep learning techniques, to segment structures without human intervention. These methods are fast and consistent but require large annotated datasets for training.

Some frequently used methods include thresholding (which separates regions based on intensity value), region growing (which expands from seed points according to pixel similarity), edge detection (which identifies object boundaries using image gradients), and clustering, which groups pixels based on shared features.

Deep Learning Models, especially Convolutional neural networks (CNNs), such as U-Net and V-Net, are widely used for automatic and highly accurate segmentation.

2.2 Image Registration

Image registration is the process of aligning two or more images of the same scene, often taken at different times, from different viewpoints, or by different sensors. It is widely used in fields such as medical imaging, remote sensing, and computer vision to enable accurate comparison, integration, and analysis of image data [107]. The goal is to establish a spatial correspondence between the images so that corresponding features represent the same physical points in the observed scene. This process enables subsequent tasks such as image fusion, change detection, motion tracking, and 3D reconstruction. In other words, image registration finds a transformation $T(\mathbf{x})$ that aligns a moving image I_m to a fixed image I_f such that:

$$I_f(\mathbf{x}) \approx I_m(T(\mathbf{x})). \quad (2.1)$$

Image registration consists of four fundamental stages: feature detection, feature matching, transformation estimation, and image resampling/interpolation:

- Feature detection involves identifying salient structures or landmarks (e.g., edges, corners, keypoints, or anatomical regions) that are robust under varying imaging conditions.
- Feature matching establishes correspondences between these detected features across images using similarity measures that quantify the degree of alignment.
- Transformation estimation determines the spatial mapping function that minimizes misalignment between the images. It can be rigid (the object's shape is preserved and only translation and rotation are necessary), affine (incorporates scaling and shearing transformations, which account for minor geometric variations) or non-rigid (allows for complex local deformations).
- Resampling/Interpolation stage applies the estimated transformation to warp one image onto another, producing a geometrically registered output.

Image registration techniques can be broadly classified into intensity-based and feature-based approaches. Intensity-based methods directly operate on the image pixel values, optimizing a similarity metric (such as sum of squared differences, normalized cross-correlation, or mutual information) to align images. Feature-based methods instead rely on geometric primitives (such as points, curves, or surfaces) extracted from anatomical landmarks or segmented structures, often providing better interpretability and robustness to intensity variations.

Due to the diversity of images to be registered and due to various types of degradations it is impossible to design a universal method applicable to all registration tasks. Every method should take into account not only the assumed type of geometric deformation between the images but also radiometric deformations and noise corruption, required registration accuracy and application-dependent data characteristics.

2.3 Cohort of data and knee scans

CT scans and MRI have been used for this study. Approval for the study was obtained from the Icelandic Bioethics Commission (approval number: VSN-19-050V1)

CT imaging uses X-rays to create detailed cross-sectional images of the body. During a CT scan, an X-ray source rotates around the patient while detectors measure the amount of radiation that passes through the body. A computer then reconstructs these measurements into a three-dimensional (3D) image. This technique is particularly effective for visualizing bones and hard tissues, as X-rays are highly absorbed by dense materials. It is also useful for lung and chest imaging due to the strong contrast between air and tissue. The main advantages of CT scans include fast image acquisition, excellent spatial resolution, and good visualization of dense structures. However, one downside is that subjects are exposed to ionizing radiation during the process.

MRI uses strong magnetic fields and radio waves instead of radiation. It utilizes the natural magnetic properties of hydrogen protons, which are abundant in the body due to the presence of water and fat. When a person is placed inside a strong magnetic field generated by the MRI scanner, the hydrogen protons align with this field. A radiofrequency pulse is then applied, temporarily knocking these protons out of alignment. When the radiofrequency pulse is turned off, the protons return to their original alignment, releasing energy in the form of radio waves. This emitted signal is detected by receiver coils and used to construct detailed images of the body's internal structures, with different tissues producing distinct signals based on their relaxation times [108]. T1 (longitudinal relaxation time) measures how quickly protons realign with the magnetic field after being disturbed. T1-weighted images provide excellent anatomical detail and are particularly useful for visualizing fat, anatomy, and post-contrast enhancement. T2 (transverse relaxation time) measures how quickly protons lose phase coherence in the transverse plane. T2-weighted images are more sensitive to fluid and are especially useful for detecting edema, inflammation, and pathology, as abnormal tissues often have longer T2 times.

Another common technique, proton density (PD) weighting, reflects the concentration of hydrogen protons in tissues and provides good structural detail with less emphasis on relaxation properties. Advanced MRI sequences such as FLAIR, DWI, and fMRI can further enhance diagnostic information by suppressing certain signals or highlighting physiological processes like diffusion or brain activity.

MRI is therefore particularly useful for visualizing soft tissues, including the brain, muscles, cartilage, heart, and internal organs. Some benefits of MRI include its excellent contrast between different types of soft tissues and the fact that it does not involve ionizing radiation. However, MRI scans typically take longer, are sensitive to patient movement, and are more expensive than CT scans.

A total of 93 subjects participated in this study among boht EU funded projects, RESTORE and SINPAIN. The cohort included individuals diagnosed with KOA and healthy volunteers with no previous diagnosis or symptom of degeneration. All participants signed informed consent forms beforehand. All the subjects underwent knee MRI and CT scans. The scans were taken consecutively, with the knee kept in the same fixed position with the use of a splint. Some subjects had both legs scanned, resulting in a total of 133 scanned knees, with 53 labeled as degenerative (D) and 80 as control (C) (60.2% vs 39.8%). Table 2.1 reports more info about the participants. The sample shows a balanced distribution of males and females across all categories. Participants in the D group are notably older than those in the C group, which aligns with the higher prevalence of KOA in older adults. The D group also shows a slightly higher average BMI, consistent with the association between excess body weight and KOA risk. Overall, the demographic characteristics reflect expected patterns for these populations.

CT was performed using a Toshiba Aquilion ONE 320-slice scanner, capable of covering up to 16 cm in a single gantry rotation. Acquisition parameters included a slice thickness of 0.5 mm with a 0.25 mm reconstruction increment. Tube voltage was set to 120 kV and tube current to 250 mA (effective mAs: 125). The scanned region covered approximately 15 cm in the axial plane, centered on the knee joint, with minor adjustments based on patient size. No intravenous contrast was

Table 2.1: Sex distribution and mean ($\pm$ StDev) of age and BMI by category (TOT, D, C).

Category	% M	% F	Age	BMI
TOT	48.98%	51.02%	47.1 $\pm$ 18.2	27.9 $\pm$ 5.6
D	51.16%	48.84%	62.5 $\pm$ 10.3	30.4 $\pm$ 5.1
C	47.27%	52.73%	35.1 $\pm$ 13.3	25.9 $\pm$ 5.1

administered. Initial CT dose index and preliminary dose-length product values were 12.1 mGy and 193.2 mGy*cm, respectively, with final values automatically adjusted by the scanner according to patient anatomy. MRI was conducted using a 3T Siemens Prisma scanner with a dedicated knee coil. A 3D fat-suppressed intermediate-weighted fast spin echo (FSE) sequence was acquired in the axial plane with isotropic 0.6 mm voxels, enabling multiplanar reconstructions. The field of view was 16 cm with a minimum matrix size of 256×256 , covering approximately 14 cm centered at the knee.

2.4 Segmentation and Registration of Bones and Cartilages

2.4.1 Bone and Cartilage Segmentation

All acquired images were processed using a medical imaging software, Mimics (Materialise's Interactive Medical Image Control System). Mimics is a software for processing medical images and creating 3D models. A stack of images can be loaded into the software. If images in the XY plane (axial plane) are loaded, Mimics is able to calculate and create images in the XZ (coronal) and YZ (sagittal) direction; this allows for a three-dimensional sensing of the 2D data [109].

Knee bones of the femur, tibia, and patella, and their corresponding cartilages (femoral, lateral tibia, medial tibia, and patellar) were considered. These parts were segmented (i.e., located in the images) following the same protocol for bones and cartilages, respectively taken from CT scan and MRI.

MRI is suitable for the study of soft tissues, such as cartilage, because it allows proper visualization of these structures. Bones, on the other hand, are more visible on CT images. For this reason, cartilages were segmented from MRI and bones from CT images, although the protocol was the same for both types of images.

The first step was to create a new mask to isolate the specific element (cartilage or bone) from the rest of the image. This was done by setting a density threshold interval to better visualize the part of interest. If necessary, editing operations such as cropping were performed to refine the masks, set boundaries according to the location of the selected part, and improve their fitting accuracy. Region growing was then used to isolate the part, removing any pixels unconnected to the edited mask. In cases where certain regions of the segmented structures overlapped (for example, where the cartilage boundaries were close to or intersecting with the bone masks), Boolean subtraction operations were applied to remove these overlapping areas. A Boolean operation in this context refers to a mathematical

image-processing function that combines or differentiates two masks based on logical operators (such as union, intersection, or subtraction) to define precise regions of interest. This procedure ensured that each mask represented only its corresponding anatomical structure, maintaining clear separation between adjacent tissues, which was particularly important for accurately distinguishing the different cartilage regions. This precise segmentation process aimed to extract meaningful and consistent information from the images, ensuring accurate representation of both bone and cartilage structures within the knee joint.

This procedure was carried out individually for each bone and each cartilage.

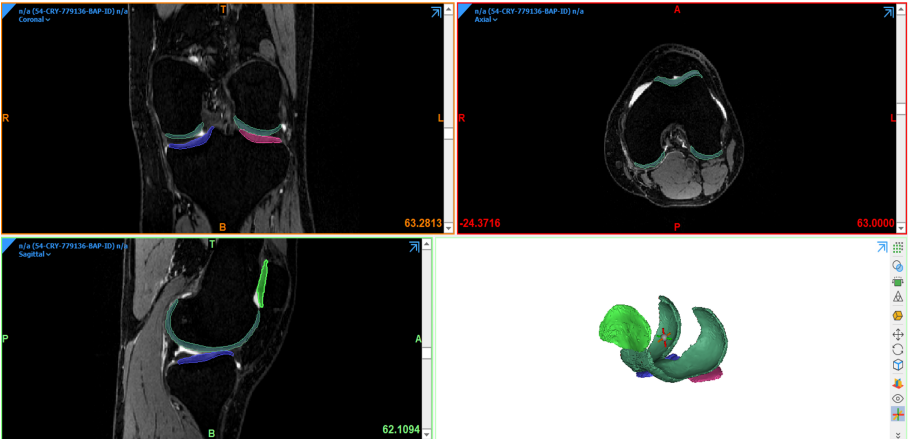


Figure 2.1: Cartilage Segmentation on MRI and correspondent 3D parts

The entire segmentation process was performed precisely to extract meaningful and consistent information from the images. These masks were then converted into very accurate 3D objects. Mimics is able to calculate a 3D model using the segmented structures and the information on pixel size and the distance between

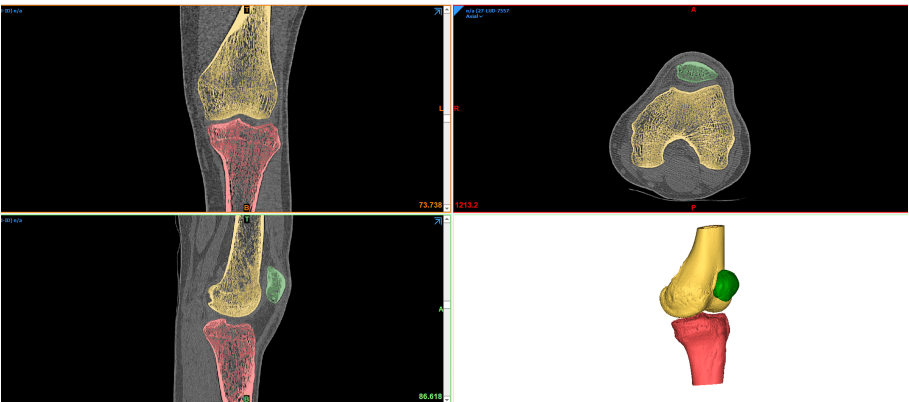


Figure 2.2: Bone Segmentation on CT and correspondent 3D parts

image slices.

Figures 2.1 and 2.2 show the segmentations of the four cartilages on the MRI and three bones on the CT, respectively, and the correspondent 3D parts.

The resulting 3D object is an STL file. STL (StereoLithography interface format, or acronym for "Standard Triangulation Language") is a file format commonly used in rapid prototyping and 3D printing, as it can accurately describe very complex geometries such as medical structures. An STL file represents a solid object whose surface has been discretized into triangles. The file consists of the repeated X, Y, and Z coordinates for each of the three vertices of every triangle, along with a vector to define the normal direction of the triangle. Even when 3D printing is not the goal, STL files are useful for data analysis applications. The geometric detail they preserve enables the extraction of shape-based features and spatial relationships for investigation.

2.4.2 Registration of MRI with CT

Starting from the existing images with segmentations and 3D objects, a new image was created through image registration, aligning MRI data onto CT data. Following Equation 2.1, the CT is the fixed image I_f and the MRI the moving one I_m . The MRI and CT acquisitions were performed with the patient in the exact same position, which facilitated the superimposition of the two modalities onto corresponding anatomical parts.

The Landmark Image Registration tool in Mimics allows the alignment of two image datasets based on corresponding anatomical reference points, known as landmarks. These landmarks are manually selected points that identify the same anatomical locations in both the MRI and CT images, enabling the software to compute a spatial transformation that best aligns the two datasets.

For the fusion method, the Maximum Gray Value (Max GV) approach was used, in which the gray value at each voxel in the resulting image is taken as the maximum between the corresponding gray values of the two input datasets (GV_1 from CT and GV_2 from MRI). This method preserves the most prominent intensity information from both modalities, enhancing the visibility of relevant anatomical structures in the combined image.

To achieve an accurate alignment, at least four anatomical landmarks were manually selected. The most suitable points include the highest and lowest points on the patella bone in the sagittal view, the upper point of the tibial tubercles, and the lateral side of the tibia in the coronal view (Figure 2.3). However, other landmarks could also be used, provided they ensure good anatomical correspondence between the two image sets.

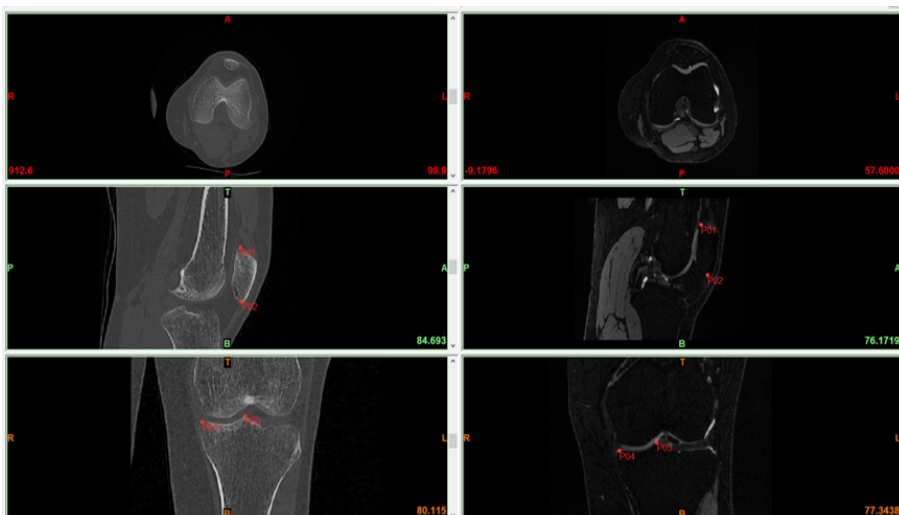


Figure 2.3: Registration between MRI and CT

Once the landmarks were defined, the MRI and CT images were registered, and the corresponding 3D segmented objects (bones from CT and cartilages from MRI) were combined. The result was a newly created CT-based image containing both bone segmentations from CT and cartilage segmentations from MRI, now accurately positioned within a common anatomical framework.

A manual realignment process was then carried out, when needed, to adjust the position of segmented parts, especially the cartilages, so they were properly aligned around the corresponding bones, and to minimize any overlapping. Following this step, new masks were created from the 3D objects to enable further refinements, such as Boolean subtraction operations to eliminate residual overlap between cartilage and bone masks. The final results is visible in Figure 2.4. The final 3D model, generated from these refined masks, provides a realistic and anatomically accurate overview of the knee compartment, with cartilages segmented from MRI accurately mapped and aligned onto the CT anatomy. Figure 2.5 presents a three-dimensional representation of both a D knee compartment and a C one. Figure 2.6 shows a three-dimensional representation of the four different cartilage segments for a degenerative and a control subject for both left and right legs. The full segmentation and registration process is summarized in Figure 2.7.

Quantitative validation of the segmentation and registration accuracy was limited by the absence of manually annotated ground-truth datasets. To ensure anatomical plausibility and consistency across subjects, all segmentations and registrations were visually reviewed and validated by an experienced radiologist. The expert verified the correct alignment of bony and cartilaginous structures across modalities (MRI and CT), anatomical boundaries, and absence of misregistration artifacts. Although this qualitative assessment ensured that the segmented regions and multimodal alignments were anatomically coherent and suitable for subsequent feature extraction and quantitative analyses.

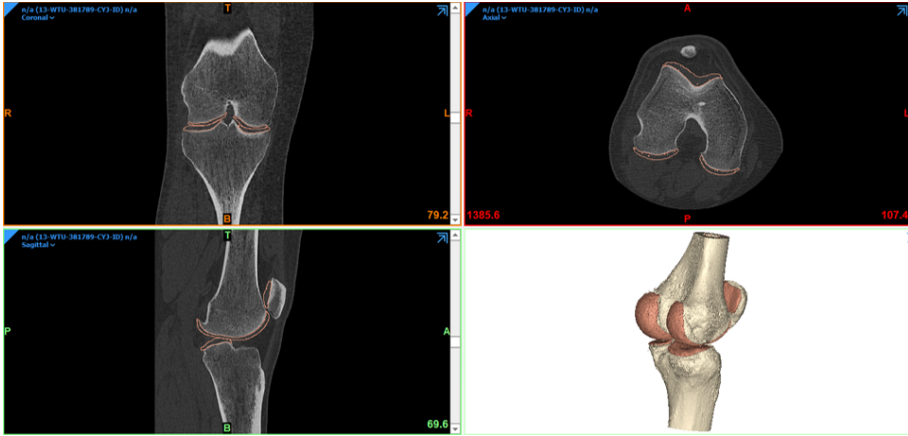
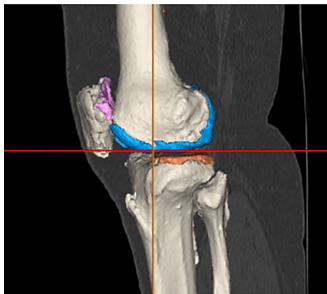
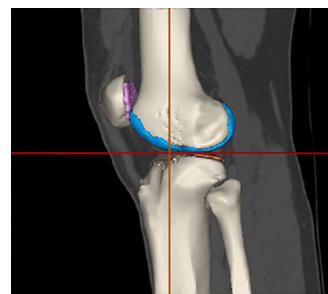


Figure 2.4: Final Result after Registration and Realignment



(a) Degenerative 3D Knee



(b) Control 3D Knee

Figure 2.5: Three-dimensional reconstructions of knee compartment of a) degenerative and b) control subjects

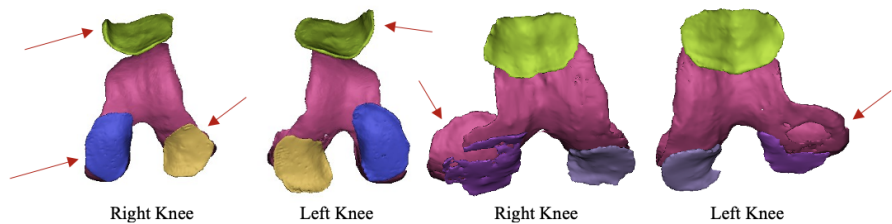


Figure 2.6: 3D representation of knee cartilages. On the left a 3D reconstruction of cartilages in a control knee (right and left); on the right a 3D reconstruction of degenerative cartilages (right and left knees). The arrows indicate similarity zones between the right and left cartilages of the same patient.

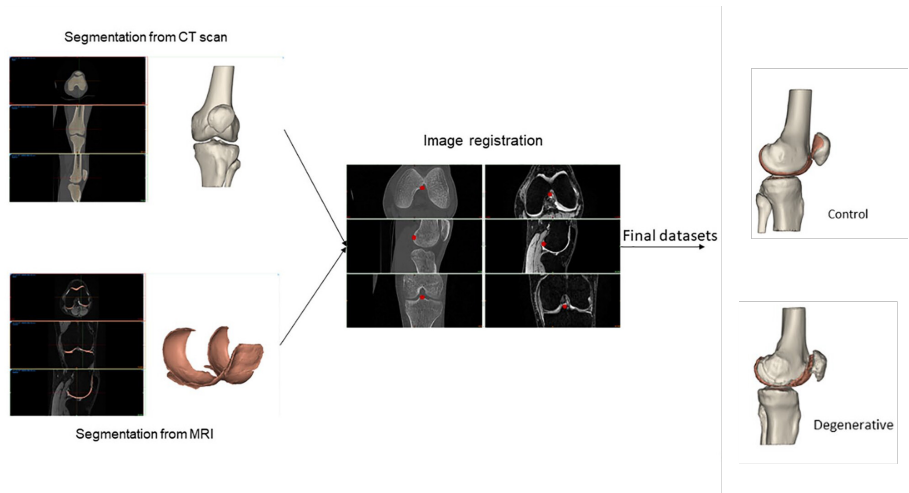


Figure 2.7: Full Segmentation and Registration Workflow

2.4.3 Refinement of Bone Segmentations

Our analysis focuses on the bone regions adjacent to the cartilage. For this reason, we cropped the femur and tibia to isolate ROIs corresponding to the subchondral bone, specifically, the distal femur and the proximal tibia (see Figure 2.8). The patella, on the other hand, was acquired in its entirety, as its small size makes it unaffected by these considerations.

Moreover, we decided to analyze both cortical and trabecular bones to examine the potential differences in how OA degeneration affects different tissue compositions rather than simply considering the bone as a whole. Starting from the previously segmented bones (patella, distal femur, proximal tibia), we applied a density-based classification filter available in Mimics. This filter separates bone tissue according to HU values: voxels with intensities up to 661 HU were classified as trabecular bone, while those above 662 HU were identified as cortical bone. This thresholding approach enables the extraction of two distinct masks corresponding

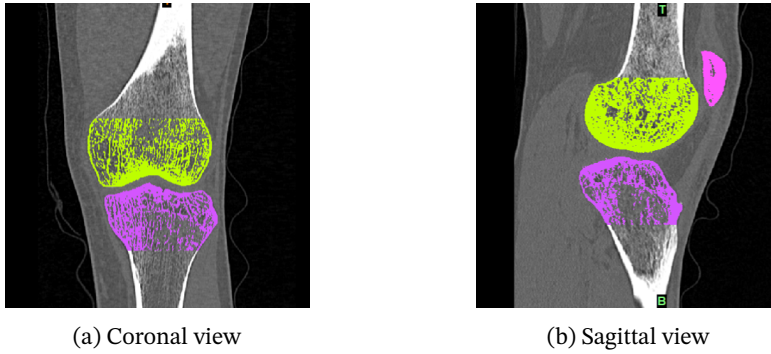


Figure 2.8: Cropped femur and tibia (distal femur and proximal tibia) in coronal and sagittal views.

to the two bone compartments, thereby facilitating the quantitative analysis of their respective morphometric and densitometric characteristics.

The resulting cortical and trabecular bone segmentations are shown in Figures 2.9 and 2.10, respectively.

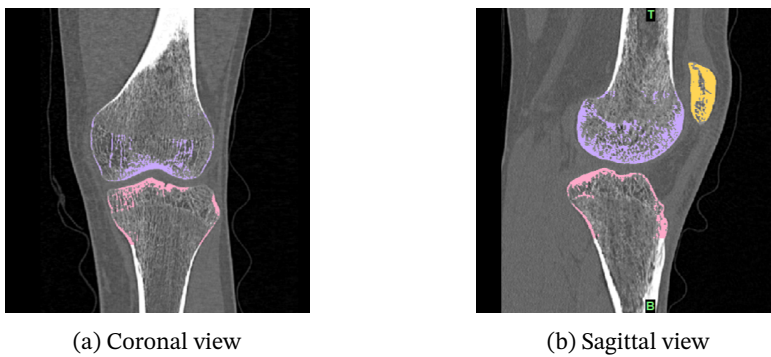


Figure 2.9: Cortical distal femur, proximal tibia, and patella in coronal and sagittal views.

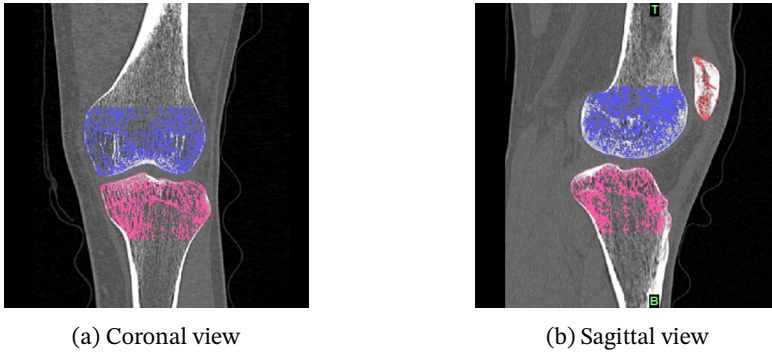


Figure 2.10: Trabecular distal femur, proximal tibia, and patella in coronal and sagittal views.

2.5 Muscle segmentation

Because KOA progression involves not only cartilage degeneration but also muscular atrophy and altered joint loading, we additionally segmented portions of thigh and calf muscles to investigate musculoskeletal interactions in later analyses (see Chapter 4).

Consecutive transaxial CT images of the knee were obtained from subjects with KOA and controls. These images provided data on skeletal muscle tissue (SMT) and IMAT density, as well as the volume of various thigh and calf muscles. The following muscles were considered:

- Vastus Medialis (VM)
- Vastus Intermedius (VI)
- Vastus Lateralis (VL)
- Biceps Femoris (BF)
- Sartorius (SART)
- Gracilis (GRAC)
- Semimembranosus (SEMIM)
- Semitendinosus (SEMIT)
- Popliteus (POP)
- Gastrocnemius Lateral Head (GLH)
- Gastrocnemius Medial Head (GMH)

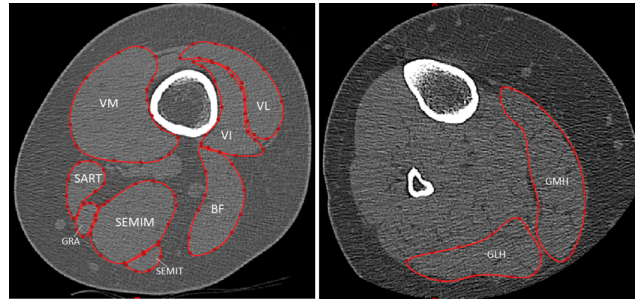


Figure 2.11: VM: Vastus medialis; VI: Vastus intermedius; VL: Vastus lateralis; BF: Biceps femoris; SART: Sartorius; GRA: Gracilis; SEMIM: Semimembranosus, SEMIT: Semitendinosus; GMH: Gastrocnemius medial head; GLH: Gastrocnemius lateral head

All the considered muscle are reported in Figure 2.11.

The used methodology employed a technique known as “virtual histology”, based on segmentation. This technique enables the extraction of histological information comparable to traditional histology but through the analysis of CT images [110], selecting a specific region of interest (ROI) shaped as a cylinder. Many previous studies focused on analyzing a cross-sectional area of the muscle [111]–[114].

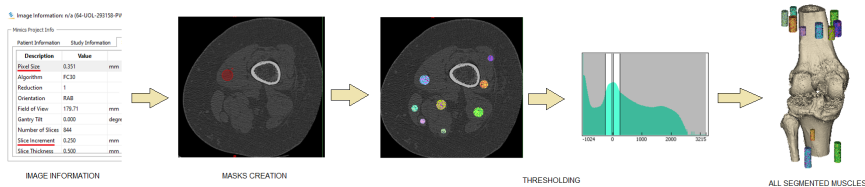


Figure 2.12: Segmentation workflow steps

The employed steps (see Figure 2.12) of the methodology are summarized as follows:

1. Image information: Information about pixel size and slice increment were obtained from each image file. This info will be used to calculate the number of slices to consider when creating the cylinder.
2. Creation of the virtual histology sample: the aim is to create, for every muscle, a small cylinder of fixed volume (volumes for each muscle are reported in Table 2.2 and are different according to muscle size). A circular area is drawn at the location of a muscle, using the Mimics 'Ellipse' tool, with a fixed circumference diameter (width). The cylinders are always taken from the same starting point, relative to each muscle, between different subjects. The volume of the cylinder is calculated according to the following formula

Table 2.2: Defined volume to be selected during segmentation

MUSCLE ROI	VM	VI	VL	BF	SART	SEMIM	SEMIT	GRAC	POP	GMH	GLH
DEFINED VOLUME (mm ³)	2000	1000	500	2000	1000	2000	300	500	100	1500	1500

(Equation 2.2):

$$\text{Vol} = \pi r^2 h \quad (2.2)$$

where r is the radius and h is the height of the cylinder. The square radius and the height are calculated as follows (Equation 2.3), (Equation 2.4):

$$r^2 = \left(\frac{\text{width}}{2} \times \text{pixel size} \right)^2 \quad (2.3)$$

where *width* is the diameter of the selected circumference (base of the cylinder) in pixels and *pixel size* is the dimension in millimeters of every pixel;

$$h = \# \text{slice} \times \text{slice increment} \quad (2.4)$$

where *slice increment* is the dimension in millimeters of the increment of each slice and *#slice* is the number of slices.

Therefore, the number of slices to consider for segmentation can be calculated according to (Equation 2.5):

$$\# \text{slice} = \frac{V \cdot 4}{\pi \cdot (\text{width} \cdot \text{pixel size})^2 \cdot \text{slice increment}} \quad (2.5)$$

The application of this formula allowed to remain consistent and have the same cylinder dimensions on the same considered muscle between different subjects with different image properties.

3. Threshold differentiation between tissue: within the created cylinder, muscle and adipose tissue were separated, according to different density thresholds expressed in Hounsfield units (HU):
 - -250 HU to -20 HU: IMAT
 - +20 HU to +250 HU: SMT
4. Data extraction: from every created thresholded mask, the average and the standard deviation of the density and all the volumes were extracted using Python. The volume of each muscle was normalized by dividing the volume of every mask by the total mm³ taken when segmenting, to be able to compare them. This resulted in an adimensional volume measure.

2.6 Extracted Features

The novelty of this study lies in the integration of 3D objects derived from segmented masks of both CT and MRI images, enabling the reconstruction of a realistic and

anatomically accurate 3D model of the knee compartment. By registering the MRI onto the CT, the segmented cartilages - originally extracted from the MRI - are accurately mapped onto the CT volume. This alignment makes it possible to analyze cartilage radiodensity directly within the CT framework, which is a novel approach. It allows for a quantitative evaluation of cartilage density based on CT data, despite the inherent limitations of CT in visualizing soft tissues. This not only opens new perspectives for assessing cartilage condition using routinely available imaging, but also offers a more cost-effective alternative to repeated or exclusive use of MRI, which is significantly more expensive.

2.6.1 Basic Morphological and Density Features

A first group of features comprised simple morphological and density-based descriptors, which are widely used as baseline quantitative measures in musculoskeletal imaging. These included first-order density statistics (mean and standard deviation of HU pixel values) from the segmented cartilages, as well as volumetric and surface measurements extracted from the 3D reconstructed part. Such descriptors provide an initial characterization of cartilage geometry and radiodensity.

MIMICS enables the direct calculation of radiodensity in HU from a ROI on CT scans. Human tissues absorb or attenuate X-ray beams according to their density. The HU is a radiodensity measurement for CT images based on a linear transformation of the X-ray beam attenuation coefficient. Water corresponds to zero on the HU scale, while bones have values higher than 400 [37]. Therefore, cartilage masks were filtered to a range of 0–300 HU, which provides a suitable range for visualizing soft tissue pixel intensities (cartilage being a type of soft tissue). The cartilage radiodensity was then extracted from the final mask. Volume and surface data were also extracted from 3D objects. Summarizing, for each of the four cartilage the following features were extracted (for a total of 16 features):

1. First-order intensity descriptors:
2.
 - Average (Avg) HU density
 - Standard Deviation (StDev) of HU density
3. Morphological descriptors:
4.
 - Volume (Vol) (mm^3)
 - Surface area (Surf) (mm^2)

From a clinical perspective, cartilage volume and surface area are direct indicators of tissue integrity, loss, and liquid infiltration. Similarly, the Avg HU density reflects underlying tissue composition, such as hydration and mineralization, while its variability (StDev) may capture early structural heterogeneity associated with degeneration. The same density features were extracted from the distal femur, proximal tibia, and patella segmentations, along with density and volume measurements for SMT and IMAT tissues.

2.6.2 Radiomic Features

Beyond simple descriptors, a comprehensive set of radiomic features was extracted from the cartilage segmentations. Radiomics captures higher-order image characteristics by quantifying intensity distributions, texture patterns, and shape descriptors.

Radiomics features were extracted using the PyRadiomics library [115] based on cartilage segmentations obtained from both MRI and CT. The extracted features describe shape, pixel statistics based on gray value histograms, and texture and can be summarized as follows:

- First Order Statistics (19 features)
- Shape-based (3D) (16 features)
- Shape-based (2D) (10 features)
- Gray Level Co-occurrence Matrix (GLCM) (24 features)
- Gray Level Run Length Matrix (GLRLM) (16 features)
- Gray Level Zone Size Matrix (GLSZM) (16 features)
- Neighbouring Gray Tone Difference Matrix (NGTDM) (5 features)
- Gray Level Dependence Matrix (GLDM) (14 features)

In addition to the original image, the described features are also calculated for the wavelet images. The wavelet transform can be used to decompose the original image into different frequencies or scales. This allows textures and other information within specific frequency bands to be examined. This could potentially improve the ability to distinguish between different conditions. Then, a total of 875 features for each of the four cartilages were extracted.

Together, these features extend the analysis towards a richer and more detailed description of cartilage tissue heterogeneity.

Radiomic features, in fact, can capture subtle patterns of tissue heterogeneity that are not detectable by visual inspection or simple morphological measures. In the context of cartilage, radiomics may reflect microstructural alterations, such as changes in collagen organization or water content, that precede gross morphological degeneration. As such, radiomic analysis offers a more comprehensive quantitative characterization of cartilage quality and has the potential to serve as an imaging biomarker for the detection and monitoring of osteoarthritis.

2.6.3 Summary of Extracted Quantitative Features

Table 2.3 summarizes the full set of quantitative features extracted in this work, organized by tissue type, ROI, and dataset. Five independent datasets were generated to capture complementary aspects of knee tissue composition and structure:

- Cartilage - Density-Morphology, including geometric and HU-based parameters derived from CT;
- Cartilage - CT Radiomics, including intensity, texture, and higher-order statistical descriptors derived from CT;
- Cartilage - MRI Radiomics, including analogous radiomic features from MRI;
- Bone - Density, including HU density parameters across femur, tibia, and patella, derived from CT;
- Muscle - Density and Volume, assessing muscle tissue density and volume, including skeletal muscle tissue and intermuscular fat components, derived from CT.

Table 2.3: Extracted Features divided by Tissue, ROI and Datasets

TISSUE	ROI	DATASET	FEATURES
Cartilage	Femoral Cartilage (FemCart)	Density - Morphology Dataset	Avg Density (HU) StDev Density (HU) Vol (mm ³) Surf (mm ²)
	Lateral Tibial Cartilage (LatTibCart) Medial Tibial Cartilage (MedTibCart) Patellar Cartilage (PatCart)	CT Radiomics & MRI Radiomics Datasets	First Order Statistics Shape-based 3D Shape-based 2D GLCM GLRLM GLSZM NGTDM GLDM
Bone	Femur, Tibia, Patella (Total, Cortical, Trabecular)	Bone Dataset	Avg Density (HU) StDev Density (HU)
Muscle	SMT and IMAT of: VM, VI, VL, BF, SART, SEMIM, SEMIT, GRAC, POP, GMH, GLH	Muscle Dataset	Avg Density (HU) StDev Density (HU) Vol (mm ³)

Cartilage datasets will be used in Chapter 3 and Chapter 5, while Bone and Muscle datasets will be analyzed in Chapter 4. Muscle dataset is also used in Chapter 5. Finally, a subset of Cartilage Density-Morphology dataset, merge together Vol and Surf into a volume-to-surface ratio (Vol/Surf) is used in Chapter 5 and Chapter 6.

2.7 Machine Learning Strategies

2.7.1 Used Algorithms

In this work, several supervised machine learning (ML) algorithms were implemented and compared to evaluate their performance in classifying and predicting knee osteoarthritis (KOA) based on quantitative imaging features. The selected models represent diverse methodological families, ranging from linear classifiers to ensemble and instance-based learning approaches. This diversity enables a robust assessment of different decision mechanisms and their generalization ability across heterogeneous datasets.

The evaluated algorithms include:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Gradient Boosting (GB)
- K-Nearest Neighbors (KNN)

These algorithms were chosen for their established effectiveness and interpretability in medical imaging applications, as well as their suitability for small to medium-sized datasets. They span a range of learning paradigms, from linear (SVM, LR) to non-linear and ensemble-based (DT, RF, GB), and instance-based approaches (KNN), providing a comprehensive framework for evaluating KOA-related classification tasks. All analyses were performed in Python 3.12.12 using NumPy, pandas, and scikit-learn libraries.

2.7.1.1 Support Vector Machine

The Support Vector Machine (SVM) is a powerful classification algorithm that seeks to determine the optimal hyperplane that maximizes the margin between two classes [116], [117]. The decision boundary is defined as:

$$w \cdot x + b_0 = 0 \quad (2.6)$$

where w represents the normal vector to the hyperplane and b_0 is the bias term. The coefficients in w indicate the relative importance of each feature for class separation.

Two variants were applied:

- *Linear SVM*: assumes linear separability between classes and provides straightforward interpretability.

- *RBF SVM*: employs the Radial Basis Function (RBF) kernel to project the input data into a higher-dimensional feature space, enabling non-linear separation.

This “kernel trick” enhances model flexibility in complex feature spaces. SVMs are known for their robustness and strong generalization, though they require careful tuning of the regularization parameter C and kernel coefficient γ to balance bias and variance.

2.7.1.2 Logistic Regression

Logistic Regression (LR) is a linear probabilistic model that estimates the likelihood of a sample belonging to a specific class. It applies the logistic (sigmoid) function to map the linear combination of input features into a probability range (0, 1):

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (2.7)$$

A threshold (commonly 0.5) is used to assign the class label. The model coefficients indicate the magnitude and direction of each feature’s contribution to the predicted outcome.

Regularization improves generalization and mitigates overfitting:

- *L1 (LASSO) regularization*: penalizes the absolute values of coefficients, promoting sparsity and implicit feature selection [118].
- *L2 (Ridge) regularization*: penalizes squared coefficient magnitudes, stabilizing correlated features.

The inverse regularization strength C controls model complexity: lower C enforces stronger regularization, while higher C allows greater flexibility.

2.7.1.3 Decision Tree

The Decision Tree (DT) classifier recursively partitions the feature space by selecting optimal thresholds on individual features [119]. Each internal node represents a decision rule, and the leaf nodes correspond to class predictions. DTs are easily interpretable and can capture non-linear relationships, but they are susceptible to overfitting when deep or trained on small, noisy datasets.

2.7.1.4 Random Forest

The Random Forest (RF) algorithm [120] is an ensemble technique that constructs multiple decision trees using the bootstrap aggregating (bagging) principle. Each tree is trained on a random subset of the data and features, and final predictions are obtained by majority voting (classification) or averaging (regression). This ensemble approach enhances robustness, reduces variance, and mitigates the overfitting common in single trees. Additionally, RF provides feature importance scores, offering valuable interpretability in biomedical contexts.

2.7.1.5 Gradient Boosting

Gradient Boosting (GB) is another ensemble method that builds decision trees sequentially, with each new tree trained to minimize the residual errors of the existing model [121]. By iteratively focusing on the hardest-to-predict samples, GB combines weak learners into a strong predictive model. It typically achieves high accuracy but requires careful hyperparameter tuning (e.g., number of estimators, learning rate, and tree depth) to prevent overfitting.

2.7.1.6 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm [122]. For a new input, the model assigns a class based on the majority label among its k nearest neighbors in the feature space, usually using the Euclidean distance metric. The choice of k critically affects performance: small values of k may lead to noise sensitivity, whereas large values may overly smooth class boundaries. KNN is conceptually simple and interpretable but computationally demanding for large datasets, as distance computations are performed at prediction time.

2.7.2 Feature Normalization

Before model training, all quantitative input features were standardized to ensure consistent scaling across datasets and to prevent bias toward features with larger numerical ranges. Normalization is particularly important for algorithms such as SVM and LR, which are sensitive to feature magnitude and variance. In this thesis, z-score normalization (standardization) was applied using the function `StandardScaler()` from the `scikit-learn` library. For each feature x_i , the standardized value x'_i was computed as:

$$x'_i = \frac{x_i - \mu_i}{\sigma_i} \quad (2.8)$$

where μ_i and σ_i denote the mean and standard deviation of feature x_i , respectively. This transformation results in features with zero mean and unit variance, thereby improving convergence and comparability across datasets. This normalization was consistently applied to all machine learning algorithms that required feature scaling (SVM, LR, KNN).

2.7.3 Evaluation Metrics

Let TP , TN , FP , and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. For binary classification, the most commonly used metrics in healthcare are defined as follows:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.9)$$

Measures the overall proportion of correctly classified instances.

- **Sensitivity / Recall / True Positive Rate (TPR):**

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2.10)$$

Indicates the ability of the model to correctly identify positive cases.

- **Specificity / True Negative Rate (TNR):**

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2.11)$$

Reflects the ability of the model to correctly identify negative cases.

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.12)$$

Measures the proportion of predicted positive cases that are actually positive.

- **F1 Score:**

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (2.13)$$

Represents the harmonic mean of precision and sensitivity, balancing false positives and false negatives.

- **Receiver Operating Characteristic - Area Under the Curve (ROC-AUC):**

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), where

$$\text{FPR} = \frac{FP}{FP + TN}. \quad (2.14)$$

The ROC-AUC quantifies the model's ability to discriminate between positive and negative classes across all possible thresholds. A value of 0.5 corresponds to random performance, while 1.0 indicates perfect discrimination.

Although these metrics are most naturally defined in binary classification tasks [123], they can be easily extended to multiclass classification problems, which are also relevant in healthcare applications. For a given class $c \in 0, \dots, C$, the per-class precision and recall are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \quad (2.15)$$

When dealing with unbalanced datasets or multiclass classification, the most common approach is to compute weighted metrics, where each class contribution is weighted by its support N_c (the number of instances belonging to that class). Weighted precision, recall, and F1 score are then defined as:

$$\text{Weighted Precision} = \frac{1}{N} \sum_{c=0}^C (\text{Precision}_c \times N_c), \quad (2.16)$$

$$\text{Weighted Recall} = \frac{1}{N} \sum_{c=0}^C (\text{Recall}_c \times N_c), \quad (2.17)$$

$$\text{Weighted F1} = 2 \times \frac{\text{Weighted Precision} \times \text{Weighted Recall}}{\text{Weighted Precision} + \text{Weighted Recall}}. \quad (2.18)$$

This weighted formulation allows the metrics to account for imbalances in class distribution, making them more reliable in real-world healthcare datasets, where certain conditions may be underrepresented.

2.7.4 Cross-Validation Methods

Cross-validation (CV) is a widely used resampling strategy in machine learning that assesses how well the results of a predictive model generalize to an independent dataset. It provides a more reliable estimate of model performance than a simple train–test split, particularly when the available data are limited. In essence, CV involves partitioning the dataset into multiple subsets (folds), training the model on a subset of the data, and validating it on the remaining portion. This process is repeated multiple times so that each observation is used for both training and validation, thereby reducing the variance associated with any single random partition.

In this work, all experiments employed *stratified k-fold* cross-validation, a variant of CV that ensures each fold maintains approximately the same class proportions as the original dataset. This stratification is especially important for classification tasks with imbalanced classes, as it guarantees that each training and validation subset is representative of the overall class distribution. By preserving these proportions across folds, stratified CV leads to more stable and fair performance estimates and prevents bias toward majority classes.

In the context of limited data samples, identifying the most suitable model and its optimal hyperparameters is of paramount importance. Hyperparameter optimization aims to determine the configuration of model settings that yields the best predictive performance. Unlike model parameters learned during training, hyperparameters control higher-level aspects such as model complexity, regularization strength, or learning rate. A prevalent and systematic approach to this task is *grid search*, which exhaustively explores predefined combinations of hyperparameter values. Each configuration is rigorously evaluated using stratified CV to ensure robust and unbiased assessment, ultimately leading to more reliable and generalizable model performance.

In this thesis, hyperparameter tuning was implemented using `GridSearchCV` from the `scikit-learn` library, which automates this procedure by coupling grid search with internal cross-validation. To further enhance the robustness of performance estimates, a *nested cross-validation* strategy was employed in certain experiments. Nested CV distinctly separates model selection from model evaluation [124], thereby mitigating optimistic bias that may arise from reusing the

same data for both tuning and testing. This is particularly relevant in biomedical applications, where datasets are often small and prone to overfitting.

In the outer layer, stratified CV was applied to maintain class balance during dataset partitioning, ensuring that each fold provides a representative sample of the data. Within each outer fold, an internal grid search—performed using stratified k -fold CV ($k = 5$)—was used to identify the hyperparameter combination yielding the highest cross-validation accuracy. By continuously rotating the validation set across folds while preserving class proportions, this nested structure ensures an objective and unbiased evaluation of model generalization. Using the optimized hyperparameters determined in the inner loop, the final performance of each model was assessed on the held-out data in the outer loop.

Chapter 3

Assessing Cartilage Degeneration

This chapter investigates how quantitative imaging features can characterize cartilage degeneration in KOA. We first analyze conventional morphological and density-based descriptors derived from MRI and CT to assess their diagnostic value. Building on these results, we then explore radiomic features: high-dimensional textural patterns may reveal subtle structural alterations invisible to the naked eye. Together, these analyses provide a foundation for identifying robust imaging biomarkers of KOA. By integrating traditional statistical methods with supervised machine learning, the study investigates which imaging biomarkers are most sensitive to degenerative changes. The proposed analyses aim to move beyond qualitative classification toward a reproducible, data-driven assessment of cartilage health, thereby establishing a basis for objective diagnosis of KOA.

3.1 Density and Morphological Features

This section aims to determine whether simple morphological and density-based features can distinguish degenerative from healthy cartilage. The paragraph has been adapted from the study entitled “Impact of Automated and Manual Segmentation Errors on Knee Osteoarthritis Classification using MRI-registered data on CT scans”, submitted to *Computerized Medical Imaging and Graphics, Elsevier*.

3.1.1 Methods

The Density-Morphology Dataset (Table 2.3), including 16 density and morphological features extracted from cartilages, was utilized in this study.

3.1.1.1 Statistical Analysis

Inferential statistics were initially used to analyze significant differences between classes D and C. For each feature, we evaluated the differences between these two classes using a permutation test [125]. The permutation test, also known as the randomization test, is a non-parametric statistical method that evaluates hypotheses by comparing the observed value of a test statistic to a reference distribution obtained

through random rearrangements, or permutations, of the data. This approach does not rely on specific assumptions about the underlying distribution and is therefore suitable when normality or homogeneity of variance cannot be assumed.

The key idea of the permutation test is to approximate the sampling distribution of a chosen test statistic under the null hypothesis by repeatedly shuffling the data labels. Under the null hypothesis H_0 , the group labels are considered exchangeable, meaning that any observed difference is attributable to random variation. For each permutation, the test statistic is recalculated, generating an empirical distribution that represents what would be expected if the null hypothesis were true. If the observed value of the statistic is sufficiently extreme within this distribution, the null hypothesis can be rejected.

In this analysis, the test statistic was defined as the difference in means between classes D and C. Class labels were randomly shuffled 10,000 times to generate the null distribution, and the p-value was estimated as the proportion of permuted statistics that were as extreme as or more extreme than the observed value. A small p-value indicates that the observed difference is unlikely to have occurred by chance, providing evidence against H_0 .

3.1.1.2 Machine Learning Processes

ML models were implemented for KOA classification.

Six ML models were evaluated on the data: Linear SVM, RBF SVM, LR with L1 LASSO regularization, DT, RF, GB.

All ML analyses were carried out using nested CV. In the inner loop, hyperparameter optimization was performed with GridSearchCV using stratified k-fold cross-validation ($k=5$).

Using the optimized hyperparameters selected in the inner loop, models' performances were evaluated in the outer loop of nested CV (2.7.4).

Performance metrics included accuracy, sensitivity, specificity, precision, F1 score, and the ROC-AUC.

3.1.2 Results

3.1.2.1 Statistical Analysis

Table 3.1 presents the p-values computed for the density-geometric features. p-values < 0.05 were considered significant.

Overall, several features demonstrated statistically significant differences. In particular, the standard deviation of density for all ROIs (FemCart, PatCart, MedTibCart, and LatTibCart) consistently reached high significance ($p < 0.001$). Additionally, MedTibCart showed significant differences in average density ($p = 0.019$), volume ($p < 0.001$), and surface ($p = 0.011$). PatCart volume was also significant ($p < 0.001$), whereas FemCart surface yielded highly significant results ($p < 0.001$). In contrast, average density and volume for most compartments, as well as LatTibCart surface, did not show statistically significant differences.

Table 3.1: p-values for density/geometric features, rounded to 3 significant digits; significant values highlighted in **bold italic**; *: $p < 0.05$, **: $p < 0.005$, ***: $p < 0.001$.

Feature	p-value
FemCart Avg Density	0.919
FemCart StDev Density	0.000***
FemCart Volume	0.245
FemCart Surface	0.000***
PatCart Avg Density	0.987
PatCart StDev Density	0.000***
PatCart Volume	0.000***
PatCart Surface	0.654
MedTibCart Avg Density	0.019*
MedTibCart StDev Density	0.000***
MedTibCart Volume	0.000***
MedTibCart Surface	0.011*
LatTibCart Avg Density	0.103
LatTibCart StDev Density	0.000***
LatTibCart Volume	0.092
LatTibCart Surface	0.310

3.1.2.2 Machine Learning Performances

All ML analyses were performed with 5-fold nested cross-validation, which returned the optimized hyperparameters tuned on the inner folds. Model performances were evaluated, and results are reported as the mean and standard deviation of the scores across the $k=5$ outer folds. For all performance metrics reviewed in this section, the highest score for each performance metric out of the applied ML models is **bolded**.

The evaluation of ML models yielded promising results. Results are reported in Table 3.2. Among all models, the RBF SVM achieved the highest scores in all the metrics, indicating strong overall performance. Using the F1 Weighted score as a benchmark, the second best model was Linear SVM, which reached 0.80 ± 0.11 . The tree-based models (GB, DT and RF) all showed similar performances, each reaching an F1 Weighted score around 0.77. Although this represents a drop of approximately 10 percentage points compared to the best performing, their results remain robust and acceptable. The ROC analysis (Figure 3.1) demonstrates that the RBF SVM achieved the highest performance ($AUC = 0.91 \pm 0.06$), followed closely by Linear SVM ($AUC = 0.89 \pm 0.07$). Logistic Regression also performed well ($AUC = 0.87 \pm 0.08$). Ensemble methods, including Random Forest ($AUC = 0.84 \pm 0.08$) and Gradient Boosting ($AUC = 0.84 \pm 0.11$), showed comparable results, whereas the Decision Tree model achieved lower discriminative ability ($AUC = 0.78 \pm 0.09$).

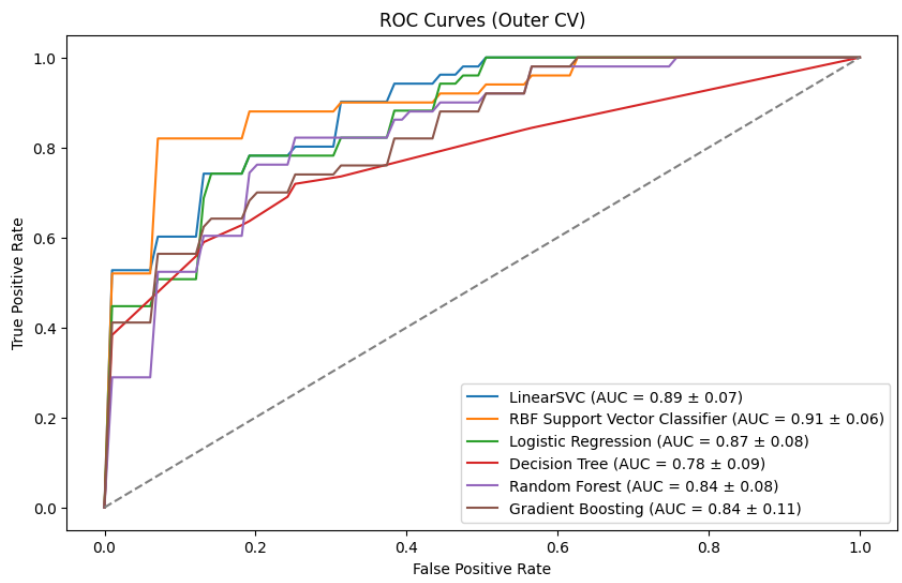


Figure 3.1: ROC curves with outer cross-validation for classification models trained on density-geometric features. Reported values correspond to mean AUC $\pm$ standard deviation.

Table 3.2: Performance metrics of different models on the density-geometric dataset, best model score for each metric in bold

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Weighted	ROC AUC
Linear SVM	0.80 ± 0.10	0.77 ± 0.15	0.73 ± 0.17	0.85 ± 0.12	0.80 ± 0.11	0.89 ± 0.07
RBF SVM	0.86 ± 0.07	0.87 ± 0.12	0.78 ± 0.12	0.91 ± 0.09	0.86 ± 0.07	0.91 ± 0.06
Logistic Regression	0.78 ± 0.10	0.76 ± 0.15	0.67 ± 0.17	0.86 ± 0.11	0.78 ± 0.10	0.87 ± 0.08
Decision Tree	0.78 ± 0.09	0.75 ± 0.16	0.67 ± 0.12	0.85 ± 0.09	0.77 ± 0.09	0.78 ± 0.09
Random Forest	0.78 ± 0.05	0.79 ± 0.13	0.63 ± 0.19	0.88 ± 0.10	0.77 ± 0.06	0.84 ± 0.08
Gradient Boosting	0.78 ± 0.12	0.77 ± 0.19	0.61 ± 0.19	0.89 ± 0.09	0.77 ± 0.12	0.84 ± 0.11

3.1.2.3 Permutation Importance Results

Permutation importance analysis was performed across the six ML models using the outer folds of a nested cross-validation scheme. For each classifier, the decrease in prediction performance following random permutation of each feature was used to estimate its relative contribution to the model’s decision process. The results were averaged across five outer folds to ensure stability and reproducibility. The resulting permutation importance plots for each classifier are shown in Figures 3.2, 3.3, 3.4, 3.5, 3.6, 3.7.

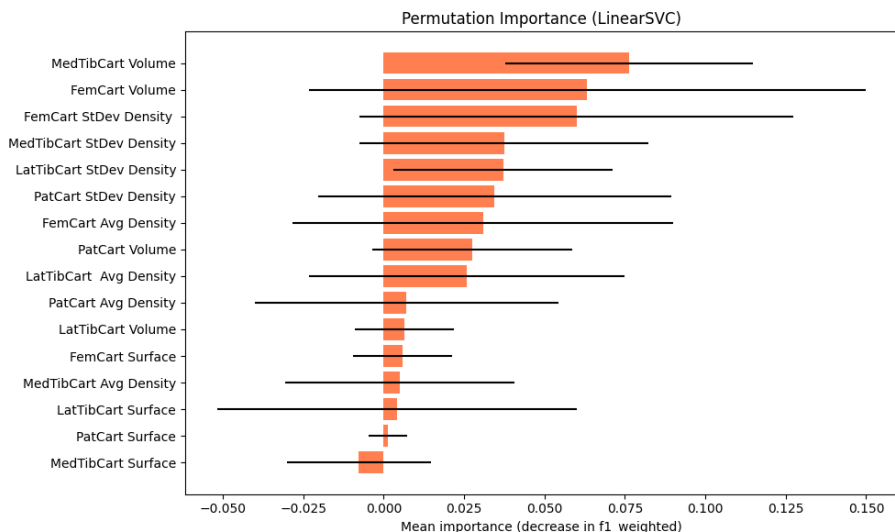


Figure 3.2: Feature importances for the SVM (linear kernel).

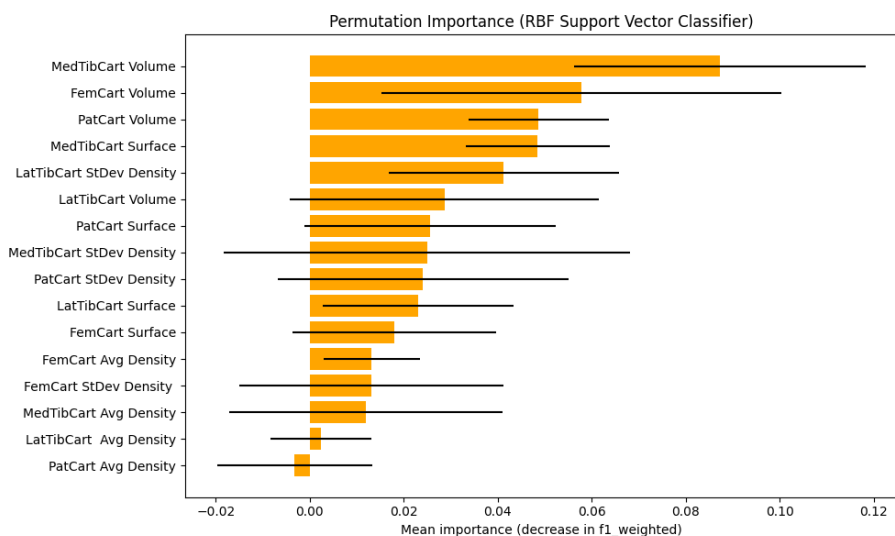


Figure 3.3: Feature importances for the SVM (RBF kernel).

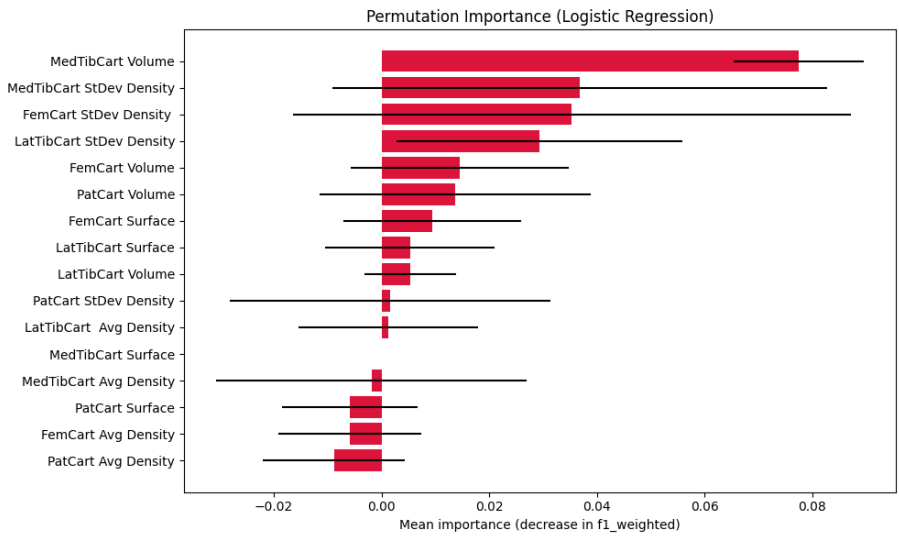


Figure 3.4: Feature importances for Logistic Regression.

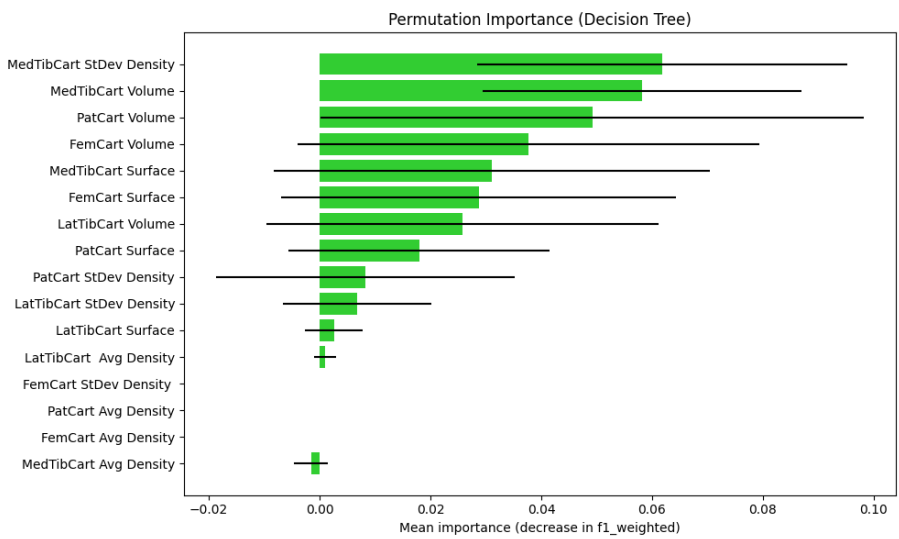


Figure 3.5: Feature importances for Decision Tree.

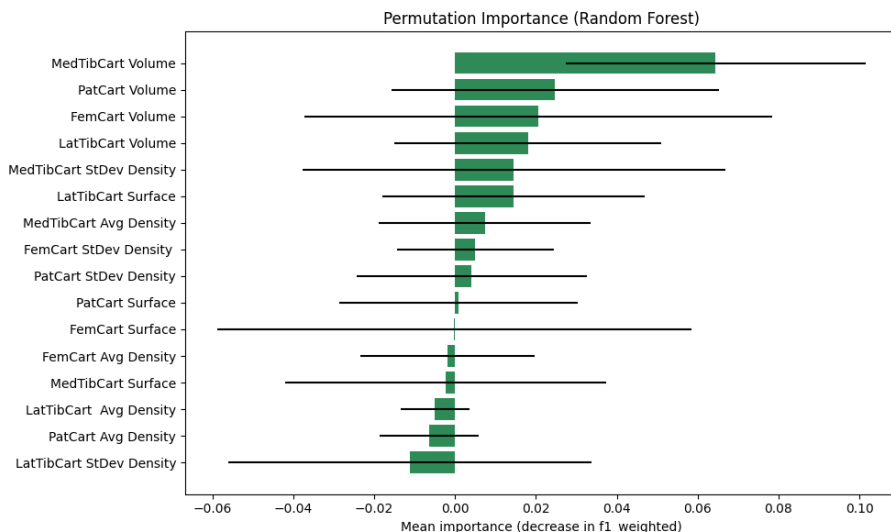


Figure 3.6: Feature importances for Random Forest.

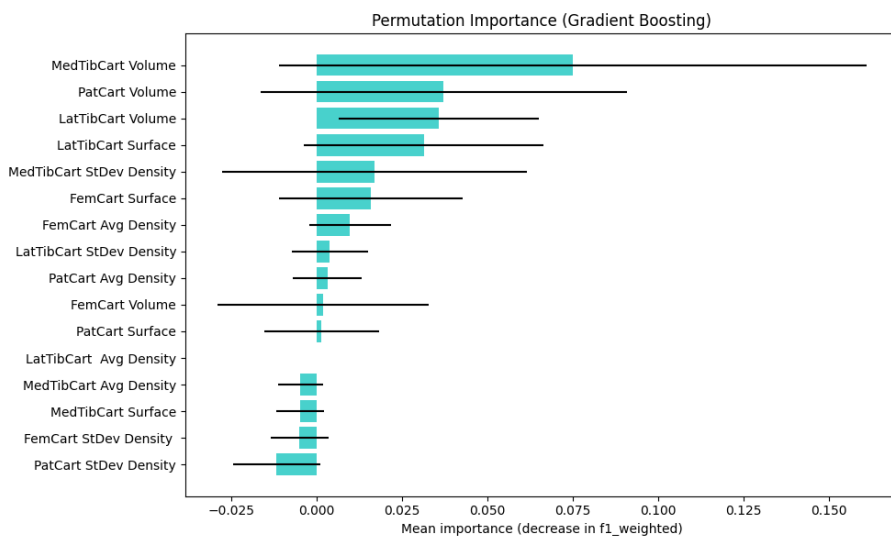


Figure 3.7: Feature importances for Gradient Boosting.

Across all models, cartilage volume-related features emerged as the most influential predictors for classification performance. In particular, MedTibCart Volume consistently ranked among the top features in every model, often accompanied by PatCart Volume, FemCart Volume, and occasionally LatTibCart Volume. These findings indicate that volumetric measures are the dominant morphological markers for differentiating the studied categories. Features capturing density variation,

such as MedTibCart StDev Density and FemCart StDev Density, demonstrated moderate importance, particularly for linear models (LR, Linear SVM) and tree-based methods (DT). This suggests that the heterogeneity of cartilage density provides complementary discriminative information to volume-based measures. Conversely, average density features (Avg Density) exhibited near-zero or slightly negative importance across all classifiers, indicating limited or no contribution to model performance.

While surface-related features (e.g., MedTibCart Surface, FemCart Surface) occasionally appeared among the higher-ranked variables, especially for the Decision Tree and SVM models, their overall influence was secondary compared to volume measures. The broader standard deviations observed for these features suggest a greater variability in their relevance depending on model architecture. Negative importance values were observed for a few features, primarily those with minimal predictive contribution. These small negative means are expected, reflecting random fluctuations around zero when a feature provides no meaningful signal. Importantly, no feature showed a systematically strong negative contribution across folds or models.

3.1.3 Discussion

3.1.3.1 Performance Metrics Among All ML Models

3.1.3.2 Key Predictive Features from Statistical and Machine Learning Analyses

The combined results of the statistical analysis and machine learning models highlight a coherent pattern in the predictive relevance of the cartilage morphometric and densitometric features. Both approaches independently converge on the importance of cartilage volume and density variability (standard deviation of density) as primary discriminative factors between the studied groups.

From the statistical point of view (see Table 3.1), several features exhibited significant differences across categories, with the standard deviation of density in all regions of interest (FemCart, PatCart, MedTibCart, and LatTibCart) consistently achieving strong statistical significance ($p < 0.001$). This suggests that textural heterogeneity of cartilage, rather than mean intensity values, captures critical tissue-level differences. The Medial Tibial Cartilage (MedTibCart) compartment, in particular, showed multiple significant parameters, including average density ($p = 0.019$), volume ($p < 0.001$), and surface ($p = 0.011$), reinforcing its dominant role in cartilage alteration patterns. PatCart Volume and FemCart Surface also showed high significance ($p < 0.001$), further emphasizing structural remodeling in these compartments.

The machine learning permutation importance analysis (Figures 3.2, 3.3, 3.4, 3.5, 3.6, 3.7) corroborates these findings. Across all evaluated classifiers, MedTibCart Volume consistently ranked as the most influential feature, often followed by PatCart and FemCart volumes. Furthermore, MedTibCart StDev Density and other density-variability metrics repeatedly appeared among the top-ranked predictors, especially in linear and tree-based models. In contrast, average density

and most surface measures contributed minimally to classification performance, paralleling their limited statistical significance. The consistency of volume and density variability features across distinct algorithms underscores their robustness and generalizability as markers of tissue degeneration.

The convergence of results across diverse model families supports the robustness of these findings and underscores the central importance of volumetric cartilage metrics in characterizing the studied condition.

Taken together, these complementary analyses suggest that cartilage volume reduction and increased density heterogeneity are the most reliable indicators of pathological change. The strong agreement between univariate statistical tests and model-agnostic feature importance supports the biological plausibility of these findings and strengthens their interpretative value. The prevalence of MedTibCart features across different methods highlights this area as a significant site of change. This finding aligns with the medial tibial compartment being particularly vulnerable to degenerative joint processes. Overall, integrating classical statistical significance testing with model-based interpretability provides a comprehensive understanding of feature relevance, bridging inferential validity with predictive utility.

3.1.3.3 Clinical Interpretation

The joint interpretation of statistical and ML results provides valuable insights into the structural and compositional alterations underlying cartilage degeneration.

The consistent prominence of MedTibCart across both analyses underscores its central role in the disease process. This region is anatomically exposed to higher mechanical loading during gait and weight-bearing [126], [127], which may explain its heightened susceptibility to degenerative changes. Indeed, greater medial knee loading has been linked to accelerated medial tibial cartilage loss and more pronounced structural defects [128], [129]. These findings align with finite element and musculoskeletal modeling evidence showing concentrated stress and pressure within the medial tibial compartment during typical gait [130], [131]. Research consistently shows that KOA is associated with a reduction in cartilage volume, particularly in the medial tibial compartment. Longitudinal studies have demonstrated that individuals with higher joint loads and a higher body mass index experience a greater loss of medial tibial cartilage volume over time, highlighting the vulnerability of this region in the progression of KOA [132], [133]. Advanced imaging and mapping techniques reveal that cartilage thickness is generally reduced in areas with more severe radiographic and MRI-detected pathology, with the medial tibia being one of the most commonly affected subregions [134], [135]. In addition to volume loss, KOA is characterized by increased heterogeneity in cartilage tissue properties. Texture analysis of cartilage T2 mapping shows that osteoarthritic cartilage exhibits greater heterogeneity (higher standard deviation and entropy of T2 values) compared to healthy cartilage, indicating more variable tissue composition and structure in affected regions [136]. This heterogeneity in cartilage damage and morphology is especially pronounced in the medial tibiofemoral joint, where both the prevalence and severity of damage vary widely among individuals with KOA

[135]. These findings underscore the importance of considering both cartilage volume loss and increased density or compositional heterogeneity in medial tibial cartilage when assessing KOA.

The identification of patellar and femoral cartilage volumes as additional influential predictors indicates that disease-related remodeling extends beyond a single compartment, supporting the concept of a multifocal degenerative pattern. Both femoral and patellar cartilage volumes are reduced in individuals with KOA, though the patterns and influencing factors can differ between these compartments. Studies using MRI-based measurements show that as KOA severity increases, the relative cartilage volume percentage in the patella decreases (from 17.37% in healthy controls to 15.71% in severe OA), and a similar trend is observed in the femur, though the reduction is less pronounced (from 57.28% to 62.45%, with some variability due to measurement methods) [137]. Longitudinal research indicates that patellar cartilage volume is lost at an average rate of about 4.5% per year, with higher rates of loss in women, those with higher BMI, and those reporting more pain at baseline [138].

Femoral cartilage volume, particularly in the medial femoral compartment, also decreases with increasing KOA grade, except for early (grade 1) lesions, which may show transient volume increases due to cartilage swelling or edema [139].

Moreover, the limited importance of average density features across all analyses implies that global mean tissue intensity fails to capture the localized structural disruptions that characterize early disease stages. Instead, variability-based metrics appear more sensitive to subtle compositional heterogeneity, potentially offering a more precise imaging biomarker of cartilage health.

Clinically, these findings suggest that combining quantitative volumetric and textural assessments could enhance diagnostic precision and sensitivity to early cartilage damage, providing a foundation for improved risk stratification and monitoring of disease progression. The alignment between statistical significance and predictive importance further strengthens the translational potential of these imaging-derived biomarkers for integration into computational diagnostic pipelines or longitudinal clinical studies.

3.2 Radiomic Features

The features discussed in section 3.1 provided an interpretable baseline for KOA classification. In the following section, the aim is to explore higher-order radiomic descriptors and determine whether they can capture subtle cartilage textural patterns associated with degeneration, thereby potentially improving classification accuracy.

This paragraph has been adapted from the study entitled “Innovative Diagnostic Approaches for Predicting Knee Cartilage Degeneration in Osteoarthritis Patients: A Radiomics-Based Study”, published on *Information System Frontiers*, Springer, as part of the Special Issue: *eXtended Reality and Artificial Intelligence in Medicine and Rehabilitation* [3].

3.2.1 Objectives

The primary objective of this work is to explore a novel approach using radiomic analysis applied to CT and MRI imaging modalities to assess cartilage degradation in KOA. By combining CT and MRI modalities and using radiomics, a technique whose application in cartilage studies is limited, we seek to establish a robust methodology for in-depth evaluation of cartilage. Our goal is to employ radiomics to identify reliable biomarkers for early detection and characterization of KOA-related cartilage changes.

3.2.2 Methods

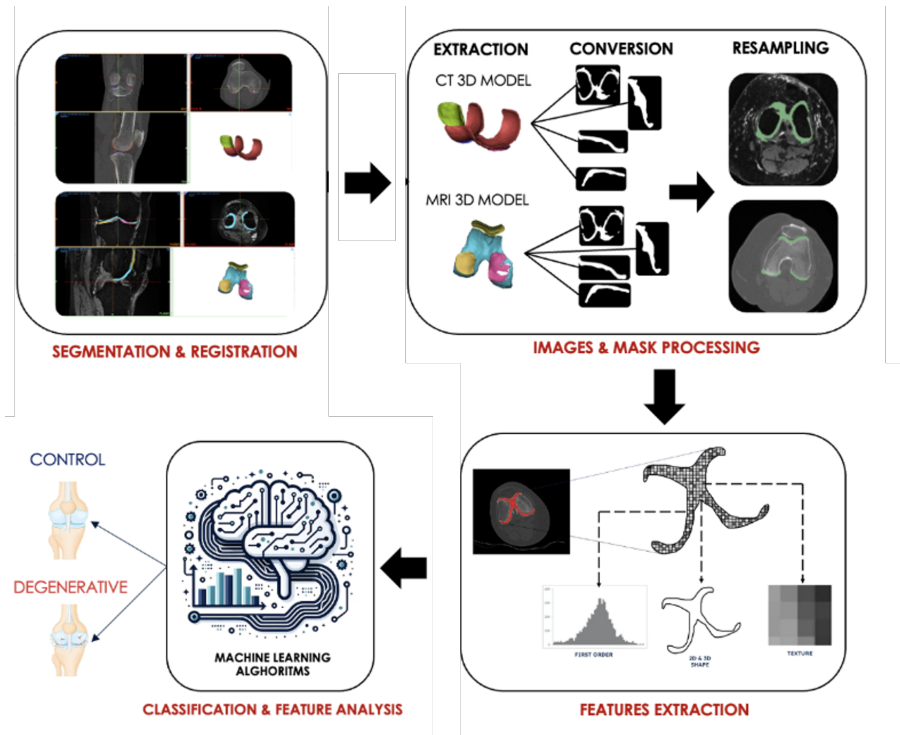


Figure 3.8: Framework design

3.2.2.1 Standardization and features selection

The workflow followed in the present study is shown in Figure 3.8. The datasets used are the CT Radiomic and MRI Radiomic datasets (Table 2.3), focusing on femoral, lateral tibial, medial tibial, and patellar cartilages. A total of 133 knees were employed, 53 belong to class D, while 80 belong to class C.

Once the radiomic features were extracted, feature normalization (see 2.7.2) was carried out. Given the high number of features extracted compared to the number of

samples considered, accurate feature selection is a fundamental step, as it improves computational efficiency, model performance, and avoids overfitting. In this work, the correlation matrix, a statistical approach for quantifying the linear correlation between variables, was first evaluated using the Pearson correlation coefficient [140], excluding all features with a correlation greater than 90%. Afterward, a feature importance analysis was performed by selecting only the most important features for the classification.

3.2.2.2 Machine Learning Analysis

Based on the most significant features calculated, a first ML analysis was conducted for the classification of D and C subjects, considering all features extracted from all cartilages. The following algorithms have been used: RF, Linear SVM, LR, GB.

We used a "nested grid search" coupled with a K-fold CV (see 2.7.4) to tune hyperparameters and evaluate performances. The algorithms tested and the hyperparameters tuned are shown in Table 3.3. To ensure comprehensive evaluation, both the mean and standard deviation were considered in all cases. This approach produces a more generalized and reliable performance metric, effectively mitigating the risks associated with overfitting, particularly on the validation set.

Table 3.3: Hyperparameter set for each machine learning model

MODEL	HYPERPARAMETERS			
	<i>n_estimators</i>	<i>max_depth</i>	<i>min_samples_split</i>	<i>min_samples_leaf</i>
Random Forest	10, 50, 100, 200, 400, 500	None, 5, 10, 20, 30, 50, 100	2, 5, 10, 15	1, 2, 4
SVM	<i>C</i>		<i>Gamma</i>	
Linear	0.01, 0.1, 1, 10, 100, 1000		1, 0.1, 0.01, 0.001, 0.0001	
Logistic Regression	<i>C</i>		<i>Penalty</i>	<i>Solver</i>
	0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000		l1, l2	newton-cg, lbfgs, liblinear, sag, saga
Gradient Boosting	<i>n_estimators</i>	<i>max_depth</i>	<i>learning_rate</i>	<i>subsample</i>
	50, 100, 200	3, 5, 7, 10	0.001, 0.01, 0.1, 0.2	0.7, 0.8, 0.9, 1.0
	<i>alpha</i>	<i>max_iter</i>	<i>Penalty</i>	<i>tol</i> <i>eta0</i>

3.2.3 Results

3.2.3.1 Binary classification using the least correlated radiomic features

The correlation matrix was calculated as the first feature selection method to eliminate all highly correlated features, especially with a linear correlation greater than 90%. It allowed us to reduce the dataset from 3500 initial features, relating to all four cartilages, to 1272 features in MRI imaging and 958 features in CT imaging. Using a 10-fold nested CV, as explained in Section 3.2.2.2, the evaluation metrics

of four different ML algorithms are shown in Table 3.4 for MRI imaging and in Table 3.5 for CT imaging. As can be seen, the best performances in the classification between D and C patients can be obtained with the Linear SVM algorithm for the radiomic features extracted from the MRI, obtaining an average accuracy of 90.25%, and with the logistic regression for the radiomic features extracted from the CT, obtaining an average accuracy of 91.33%.

Table 3.4: Compared Performances among models in MRI radiomics data with 1272 features

Model	Accuracy	Specificity	Precision	Recall	F1-score
Random Forest	84.97 (± 9.30)	84.97 (± 9.30)	88.11 (± 7.45)	86.98 (± 7.60)	86.48 (± 8.08)
Linear SVM	90.25 (± 7.03)	90.25 (± 7.03)	92.34 (± 5.04)	90.50 (± 6.82)	90.36 (± 6.89)
Logistic Regression	87.30 (± 6.03)	87.30 (± 6.03)	89.26 (± 4.82)	86.80 (± 5.56)	86.76 (± 5.55)
Gradient Boosting	82.56 (± 11.78)	82.56 (± 11.78)	86.33 (± 9.00)	84.84 (± 9.84)	83.98 (± 10.81)

Table 3.5: Compared Performances among models in CT radiomics data with 958 features

Model	Accuracy	Specificity	Precision	Recall	F1-score
Random Forest	80.64 (± 7.58)	80.64 (± 7.58)	84.77 (± 5.90)	82.65 (± 6.88)	81.97 (± 7.30)
Linear SVM	85.51 (± 8.49)	85.51 (± 8.49)	87.78 (± 6.47)	87.21 (± 7.19)	86.75 (± 7.93)
Logistic Regression	91.33 (± 7.35)	91.33 (± 7.35)	92.10 (± 6.06)	91.77 (± 6.35)	91.65 (± 6.56)
Gradient Boosting	82.64 (± 7.24)	82.64 (± 7.24)	85.42 (± 5.63)	84.19 (± 6.23)	83.70 (± 6.86)

3.2.3.2 Binary classification using the least correlated radiomic features after feature importance

The set of features considered for model training was further reduced by also carrying out a feature importance analysis following the correlation matrix. In Figure 3.9 it is possible to observe that the first most important features for MRI, using SVM algorithm, mainly refer to the patellar cartilage and are mainly belonging to the shape and GLSZM classes. For the CT images, using the Logistic Regression algorithm, a pattern of features is not so evident, in fact, there is an almost similar distribution between the patellar, femoral, and tibial cartilage, and mainly shape, GLSZM and GLCM features appear, as can be seen in Figure 3.10. In Figure 3.11 the distribution of the 50 most imparting features is shown based on the classes of radiomic features per cartilage area considered, both for CT and MRI images.

The analysis reveals that among the most important features are those related to texture matrices, particularly those derived from the wavelet transform, and those related to shape. This underlines the relevance of specific texture patterns and morphological aspects in the analysed cartilage for the clinical conditions of interest.

The predominance of texture features derived from the wavelet transform may reflect their ability to describe the complexity of cartilage tissue at different scales of analysis, which is essential for identifying subtle but clinically relevant patterns.

Furthermore, the importance of shape-related features highlights how the shape and geometric structure of cartilage may play a critical role in the assessment of joint health. By training the models with the first 50 most important features we obtain the results shown in Table 3.6 and Table 3.7. The performances, as expected, improved, reaching very high results for all the evaluation metrics considered, with an average of around 99% for all in the MRI and 96% in the CT.

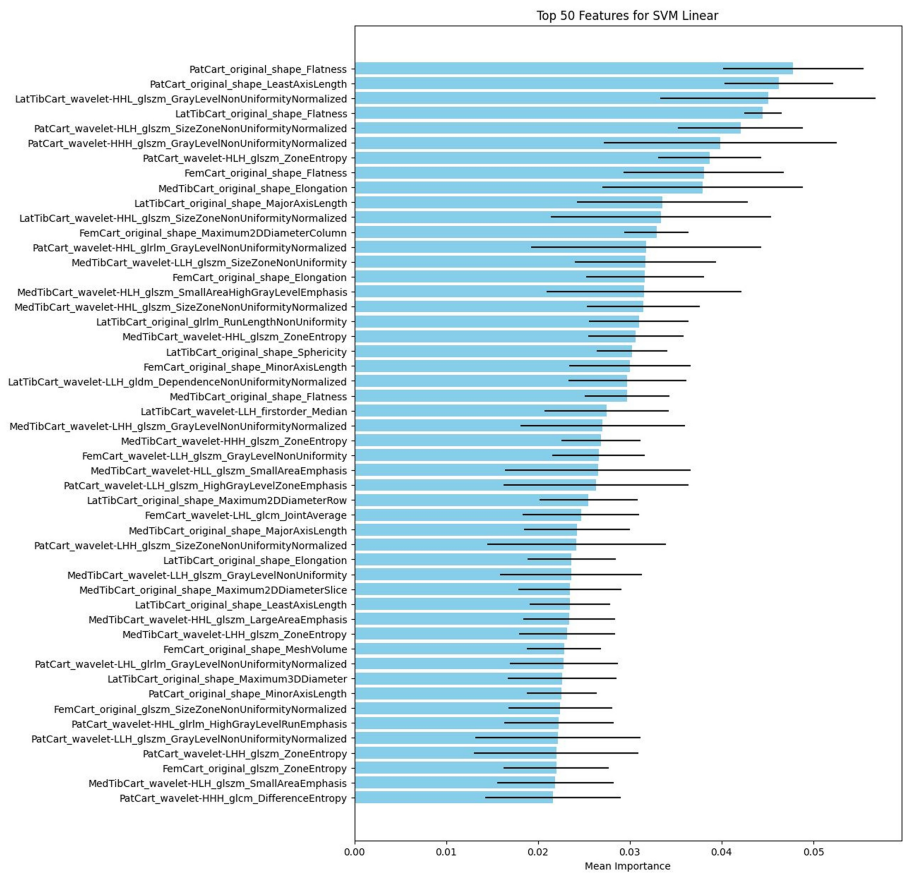


Figure 3.9: Barplot of the 50 most important robust features for the best model (Linear SVM) in MRI.

Table 3.6: Compared Performances among models in MRI radiomics data with 50 features

Model	Accuracy	Specificity	Precision	Recall	F1-score
Random Forest	88.36 (±5.79)	85.37 (±6.85)	89.16 (±5.71)	88.36 (±5.79)	87.90 (±6.10)
Linear SVM	95.61 (±4.28)	94.68 (±5.09)	95.80 (±4.20)	95.61 (±4.28)	95.54 (±4.35)
Logistic Regression	99.26 (±1.48)	99.00 (±2.00)	99.30 (±1.40)	99.26 (±1.48)	99.25 (±1.50)
Gradient Boosting	86.14 (±5.91)	84.07 (±6.25)	86.28 (±5.97)	86.14 (±5.91)	85.89 (±6.03)

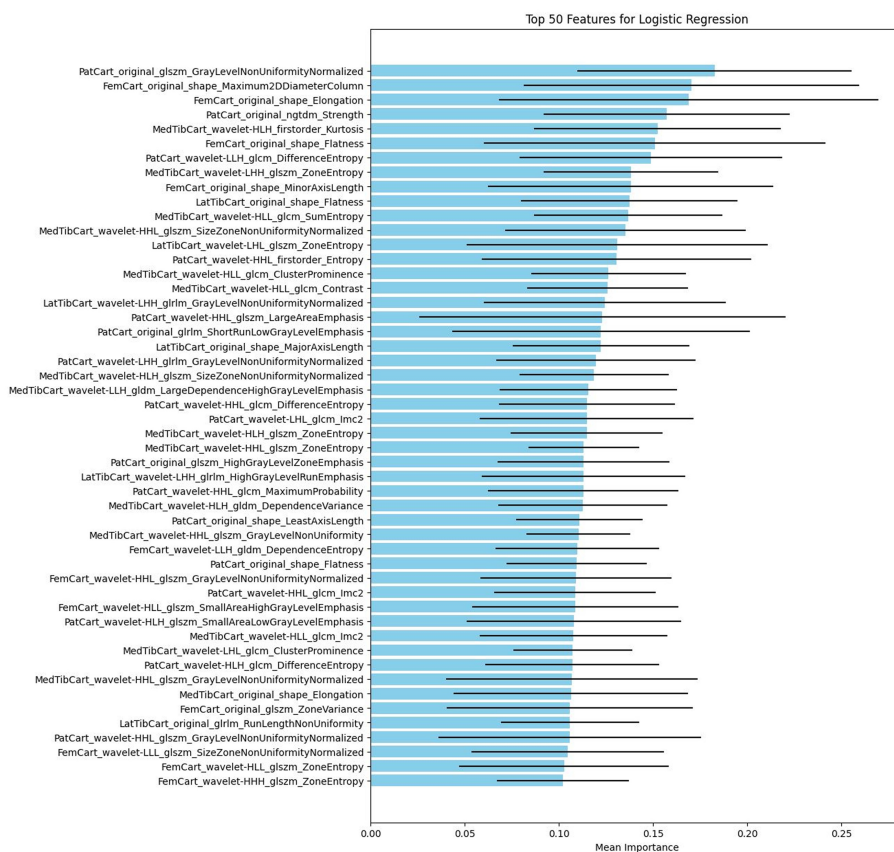


Figure 3.10: Barplot of the 50 most important robust features for the best model (LR) in CT.

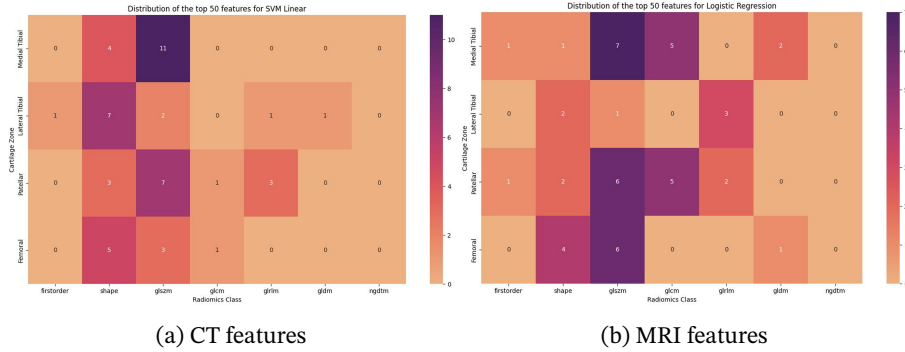


Figure 3.11: Distribution of the top 50 most predictive radiomic features by cartilage area for a) Logistic regression on CT features and b) SVM on MRI features.

Table 3.7: Compared Performances among models in CT radiomics data with 50 features

Model	Accuracy	Specificity	Precision	Recall	F1-score
Random Forest	83.22 (± 7.27)	81.96 (± 7.97)	84.38 (± 6.09)	83.22 (± 7.27)	83.03 (± 7.33)
Linear SVM	96.24 (± 3.38)	95.36 (± 4.27)	96.44 (± 3.22)	96.24 (± 3.38)	96.19 (± 3.42)
Logistic Regression	95.50 (± 3.68)	94.69 (± 4.40)	95.65 (± 3.57)	95.50 (± 3.68)	95.46 (± 3.71)
Gradient Boosting	85.56 (± 3.06)	85.63 (± 4.79)	88.02 (± 3.15)	85.56 (± 3.06)	85.39 (± 3.30)

3.2.4 Discussion

The present work shows a classification of knees with cartilage degradation compared to a control group, using radiomics features and ML algorithms with a multi-modal approach. In the first instance, a classic ML approach was followed, paying attention to an accurate subdivision of the dataset for training and an accurate selection of features.

The primary challenge is avoiding overfitting, which occurs when a model, after being trained on a specific dataset, becomes overly adapted to its peculiarities, thereby losing its predictive power on new and unseen data. In this work, a great effort was made to make the model as reliable and generalizable as possible. There is a risk that the model may inadvertently learn to recognize knee-specific images, particularly if images of both knees from the same patient are included in separate training and testing sets. The correlation between the features was then evaluated and a feature importance analysis was subsequently carried out to select the non-redundant and most predictive features for the ML models.

The predominance of texture features derived from the wavelet transform may reflect their ability to describe the complexity of cartilage tissue at different scales of analysis, which is essential for identifying subtle but clinically relevant patterns. Furthermore, the importance of shape-related features underscores the critical role that the shape and geometric structure of cartilage play in assessing joint health. Features related to wavelet texture matrices capture information regarding the variation of pixel intensity levels across different scales and orientations. This can

be clinically interpreted as a measure of the heterogeneity of cartilage composition and internal structure, potentially reflecting microscopic changes associated with pathological processes, such as KOA. The presence of particular texture patterns may indicate changes in cartilage composition that precede visible morphological changes. Shape features are representative of cartilage geometry and size. These may reflect structural changes in cartilage, such as thickening, deformation, or volume reduction, which are hallmarks of degenerative disease or injury. It is known that KOA cartilages present changes in composition due to increased water content and reduction in collagen and proteoglycan network [141]–[144]. This would explain both the structural changes and the texture differences in our cohort. Cartilage thickness tends to decrease in OA subjects [145]. In clinical terms, the importance of these features supports the idea that radiomics may serve as a powerful non-invasive tool for early detection of joint disease and monitoring disease progression. In particular, greater variability in texture features may be associated with the early stages of cartilage damage, while significant alterations in shape features may indicate a more advanced condition.

Given the proposed multimodal approach, particular considerations can also be made on the differences between the classification on MRI and CT images. For a purely visual evaluation by a clinician, diagnoses based on CT only consider space between joints and bony structures, but it does not allow clear visualization of the cartilaginous structure. From our study, however, it emerges that with radiomic techniques the classification of degenerated knees is not only possible but also very efficient from CT, almost comparable to the results obtained with radiomic features extracted from MRI. This gives space to new perspectives.

Comparison of our results with the existing literature is challenging for several reasons. As previously highlighted, existing studies on the use of radiomics to predict knee degeneration are limited. Among them, most of the studies either focused on different ROIs compared to ours or only considered MRI data. Xie *et al.* [103] proposed a methodology that used combined radiomics models for the cartilage and subchondral bone regions across various knee compartments (lateral femur, lateral tibia, medial femur, and medial tibia) on T2 mapping images. Their approach demonstrated excellent performance in distinguishing post-traumatic OA patients from healthy controls. However, this study focused exclusively on patients predisposed to post-traumatic OA after anterior cruciate ligament reconstruction, whereas our cohort does not specifically include or differentiate subjects based on post-traumatic OA status. In [104] the authors aimed to develop an MRI-based radiomics predictive model for identifying KOA by analyzing the tibial and femoral subchondral bone. Radiomic features were extracted from different regions of the lateral and medial compartments of the femur condyle and tibial plateau. The model reached an area under the curve (AUC) score of 0.961. In [105] the authors assessed tibiofemoral OA extracting MRI-based radiomic features (in particular, shape and texture features) from proximal tibia ROI, reaching an AUC score of 0.80. Cui *et al.* [146] extracted radiomic features from knee cartilage portion manually segmented on MRI. The considered ROI were the medial and lateral compartments of the tibiofemoral joints and the patellofemoral joints. The accuracy of ML model on the training set reached 94% of accuracy.

In contrast to these studies, our innovative approach incorporates segmented knee cartilage on both CT and MRI, achieving high results, particularly on MRI features. The precise segmentation process on MRI and subsequent accurate registration on CT resulted in a unique multimodal dataset. This approach facilitates new comparisons and represents a significant advance in the field. It is the first study available to extract radiomic features of cartilage from CT and the first to compare the two imaging modalities for predicting KOA after a precise segmentation process. This paves the way for future research in the field of radiomics, exploring the combined use of features extracted from CT and MRI.

3.2.5 Limitations

Some limitations should be acknowledged. Firstly, the imbalance between the classes C and D within the dataset may have influenced the performance of the machine learning algorithms. Class imbalance can lead to biased model training, where the classifier tends to favor the majority class (in this case, the healthy class), potentially resulting in reduced sensitivity to detect the minority class (patients with KOA). Strategies such as oversampling the minority class or adjusting class weights during model training could be explored to address this issue in future studies.

Secondly, the quality of the medical imaging data, particularly MRI and CT scans, is crucial for accurate radiomics analysis. Variations in image resolution, noise levels, and artifacts can impact the extraction and reliability of radiomic features. Ensuring standardized imaging protocols and rigorous quality control measures during image acquisition can help mitigate these challenges.

Furthermore, the study's limited sample size and the number of subjects pose potential limitations. With fewer subjects, there is a risk of overfitting the machine learning models to the specific characteristics of the dataset, which may limit the generalizability of the findings to broader patient populations. Increasing the sample size and subjects' diversity could enhance future investigations' robustness and external validity. Such advancements will enhance the efficiency of early diagnosis and improve patient outcomes.

Chapter 4

Interplay between cartilage and other tissues

The knee joint represents a complex biomechanical system in which multiple tissues, including cartilage, bone, and muscle, interact dynamically to maintain structural stability and functional mobility. In degenerative conditions such as KOA, these tissues do not deteriorate independently; rather, their pathological changes are interrelated and mutually reinforcing. In this chapter, we examine the interdependence between cartilage and its surrounding tissues, with a focus on how bone and muscle morphology, density, and composition change in the presence of degeneration. The first section presents an analysis of bone features derived from CT imaging, investigating potential densitometric and volumetric differences between subjects with KOA and healthy subjects. The subsequent section examines the characteristics of muscle and intramuscular adipose tissue, assessing their relationships with BMI and age. Through statistical and machine learning analyses, this chapter aims to provide a comprehensive understanding of the multi-tissue remodeling processes that accompany and possibly drive the progression of cartilage degeneration.

4.1 Bones

Bone health and composition play a key role in the structural and mechanical balance of the knee joint. Understanding how osteoarthritic degeneration affects bone tissue is crucial to understanding the interplay between bone remodeling and cartilage loss in KOA. In this section, we analyzed quantitative bone parameters extracted from CT imaging of the femur, tibia, and patella to evaluate differences between the D and C groups.

4.1.1 Methods

Quantitative bone density metrics were extracted from CT images of the distal femur, proximal tibia, and patella. For each region, mean and standard deviation of

HU were computed for the whole bone (which refers to the combined measurement of trabecular and cortical bone), cortical bone, and trabecular bone, resulting in a total of 18 quantitative descriptors Table 2.3. Prior to statistical testing, normality of distribution was assessed using the Shapiro–Wilk test. Depending on distributional characteristics, either an independent samples t-test (for normally distributed variables) or a Mann–Whitney U test (for non-normally distributed variables) was applied to compare groups. Statistical significance was set at $p < 0.05$. Boxplots were generated to visualize group-wise differences and dispersion for the most significant features.

4.1.2 Results

Several bone-related parameters exhibited significant differences between the degenerative and control groups. Table 4.1) only reports the significant ones and the test used.

Table 4.1: Significant features (p -value < 0.05) among bone density measurements across the femur, tibia, and patella regions.

Feature	p-value	Test Type
Femur_Cut StDev HU	0.003	Mann-Whitney U test
Tibia_Cut Avg HU	0.001	t-test
Patella Avg HU	0.013	t-test
Patella StDev HU	0.016	t-test
Cortical_Femur_Cut Avg HU	0.005	t-test
Cortical_Tibia_Cut StDev HU	0.006	t-test
Cortical_Patella Avg HU	0.006	t-test
Cortical_Patella StDev HU	0.018	Mann-Whitney U test
Trabecular_Tibia_Cut Avg HU	0.007	Mann-Whitney U test
Trabecular_Tibia_Cut StDev HU	0.004	Mann-Whitney U test
Trabecular_Patella Avg HU	0.005	t-test

The analysis of the femur bone showed significant differences only in the variation of density (standard deviation) for the total bone, as well as the mean density in the cortical area. In contrast, the tibia displayed an opposite pattern, showing significance in the mean density values for the total bone and in the density variation specifically within the cortical region. The trabecular tibia demonstrated both significant mean density and standard deviation. The patella, on the other hand, exhibited significant results across almost all ROIs analyzed, with both mean and standard deviation of density being significant. The only exception was the trabecular variation, which was not significant.

Visual inspection of the boxplots (Figure 4.1) shows that the femur has both an increased mean and variation of density in samples presenting KOA. Tibia bone has the same increasing trend for the total Avg density, but inverse for the density StDev in the cortical bone (D group shows lower variation). Trabecular tibia Avg confirms the previous trend, while the StDev follows the femur trend and presents

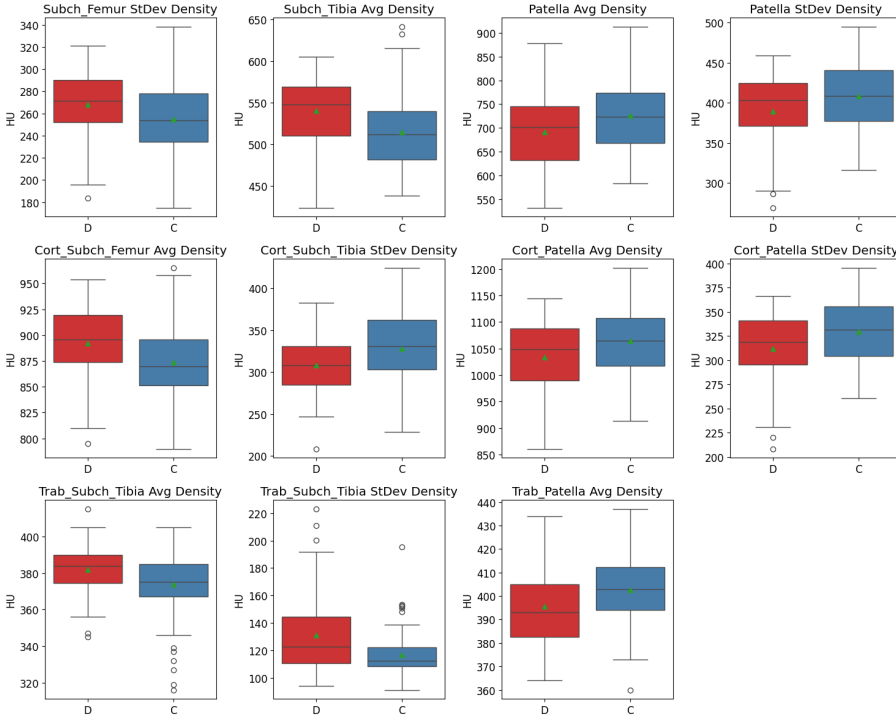


Figure 4.1: Boxplots distributions of the most significant CT-based bone density features across femur, tibia, and patella regions. Green triangles represent mean values.

higher variation in D subjects. Patella shows really consistent results, with D class presenting increased mean and variation in density in all the statistically significant ROIs.

4.1.3 Discussion

The quantitative analysis of CT-derived bone density metrics revealed several densitometric alterations in osteoarthritic samples compared to controls. These alterations reflect the pathological remodeling that occurs in subchondral bone during KOA progression. Across the femur, tibia, and patella, both the mean HU values and their within-region variability (standard deviation) showed significant group-level differences, indicating that KOA affects not only overall mineralization but also the heterogeneity of bone tissue.

In the femur, the D group exhibited higher HU variation and increased cortical density. The elevated average density of the subchondral cortical femur likely results from abnormal bone remodeling and sclerosis driven by mechanical overload and molecular dysregulation. Chronic joint stress induces thickening of the subchondral bone as an adaptive response; however, the imbalance between bone

formation and resorption leads to net bone accumulation [147], [148]. Elevated signaling molecules such as TGF- β 1 and PDGF-BB further stimulate osteoblast activity, angiogenesis, and mineralization, contributing to sclerosis and increased bone mineral density [149]. This mechanism also applies to the tibial plateau, explaining the similar increase in density observed in the D samples of our dataset.

Bone sclerosis and abnormal remodeling in KOA are particularly evident in the trabecular bone of the tibial subchondral region, where both average density and heterogeneity are increased [148]. Regions affected by bone marrow lesions typically exhibit lower mineralization and greater density variability, indicating hypo-mineralized bone undergoing active remodeling [150]. Previous research confirms that density heterogeneity is significantly higher in KOA compared to control cases [151]. The opposite trend, namely the decreased variation in cortical tibial density, may result from intense bone remodeling and the presence of bone marrow lesions, which are characterized by hypo-mineralized subchondral bone [150].

The patella presented the most consistent and pronounced differences across all ROIs, with both mean and standard deviation of HU significantly decreased in the degenerative group. Severe KOA (with attrition) is often associated with lower overall bone mineral density and a higher prevalence of osteopenia and osteoporosis [152]. Moreover, patients with KOA who experience moderate-to-severe pain at rest show significantly lower patellar subchondral bone mineral density [153]. This aligns with our cohort, in which D samples represent individuals diagnosed with KOA after presenting with pain, potentially explaining the observed decrease in both mean density and density variation.

Overall, the patterns observed are complex and multifactorial. Nonetheless, these findings reinforce the value of radiodensitometric biomarkers for characterizing KOA-related bone changes. They demonstrate that CT-derived HU features, even without volumetric or shape information, can effectively capture key aspects of pathological remodeling. Importantly, the results suggest that early degenerative changes manifest more prominently in tissue density and heterogeneity. The consistency of these differences across multiple joint regions supports the integration of HU-based features into multi-tissue predictive models, combining bone, cartilage, and soft-tissue parameters for a comprehensive assessment of joint degeneration.

4.2 Muscles

This paragraph has been adapted from the study entitled “How muscle and intramuscular adipose tissue volume and density progress in knee osteoarthritis-affected subjects: a virtual histology study” [5], presented at the *IEEE MetroXRaine Conference, 2024, St. Albans, London, UK*.

4.2.1 Sarcopenia in KOA

Sarcopenia is one of the conditions that is involved in worsening KOA. It is characterized by the loss of muscle mass, strength, and function that occurs with aging.

It is a natural part of the aging process, but it can also be accelerated by factors such as physical inactivity, poor nutrition, chronic illness, and certain medications [154]–[156]. Recent studies have uncovered a significant correlation between KOA and the presence of intramuscular fat, also known as intramuscular adipose tissue (IMAT) [111], [157], [158]. This relationship sheds new light on the complex mechanisms underlying the progression and severity of the disease. Intramuscular fat refers to the accumulation of adipose tissue within the muscle fibers themselves, rather than just around the muscles. While some levels of fat infiltration into muscles are considered normal, excessive IMAT has been associated with various metabolic disorders and musculoskeletal conditions, including KOA. Studies have shown that individuals with KOA often exhibit higher levels of intramuscular fat in the muscles surrounding the affected joint, particularly the quadriceps and hamstring muscles [159][160]. Furthermore, intramuscular fat is thought to produce inflammatory cytokines and adipokines that contribute to the inflammatory processes occurring within the joint, thus accelerating cartilage degeneration and disease progression. Several studies have highlighted the significant impact of gender and age on intramuscular fat levels [112].

In light of this, we aim to investigate if there is any change in the muscle and intramuscular fat density and volume in the thigh and calf muscles of both patients affected by KOA and healthy individuals, and how these features are correlated with BMI and age. The physical state of the leg muscles is studied through the use of an innovative histology based on segmentation, a process that allows for the same information to be obtained from a common histology through the analysis of CT images, not requiring invasive procedures in patients. As other studies proposed [161] this method should be more investigated and validated also because it paves the path for future research, promoting animal-free experimentation and promising superior, expedited outcomes compared to conventional techniques [114].

4.2.2 Cohort for this study

The cohort included 55 D samples and 77 C samples. The prevalence of obesity in the C group was 16.9%, whereas in the D group was substantially higher at 40%.

We categorized the two cohorts based on the BMI of the patients. Using a BMI threshold of 24.9 kg/m^2 for normal weight and 29.9 kg/m^2 for overweight, we observed that in the control group, 57.1 % were within the normal weight range, 26 % were overweight, and 16.9 % were classified as obese. In contrast, among KOA patients, only 5.6 % fell within the normal weight range, with 55.6 % categorized as overweight and 38.9 % classified as obese.

The dataset used for this study is the Muscle dataset Table 2.3, comprising density and volumetric descriptors from SMT and IMAT of the segmented muscle portions (see 2.5).

4.2.3 Statistical analysis

Statistical analyses were performed. Differences between the test and control groups were assessed using either independent samples t-test (for normally distributed

Table 4.2: Comparison of volume and density of skeletal muscle (SMT), and adipose (IMAT) tissues in muscles examined between C and D categories: significant p-values

Muscle	Tissue	Density	Volume	Muscle	Tissue	Density	Volume
VM	SMT	-	-	SEMIT	SMT	-	-
	IMAT	-	0.017		IMAT	-	0.001
VI	SMT	-	4.62E-04	GRAC	SMT	-	-
	IMAT	0.003	2.01E-05		IMAT	-	0.005
VL	SMT	-	-	POP	SMT	-	-
	IMAT	-	0.002		IMAT	0.026	0.002
BF	SMT	-	0.005	GMH	SMT	-	3.35E-04
	IMAT	0.035	0.001		IMAT	0.001	3.31E-07
SART	SMT	-	-	GLH	SMT	-	-
	IMAT	-	0.006		IMAT	-	0.008
SEMIM	SMT	-	3.67E-05				
	IMAT	0.001	1.78E-06				

data) or Mann-Whitney U test (for non-normal distribution). These analyses were also applied to compare data between genders within both control and case groups. Furthermore, linear regression was conducted to explore potential correlations between the dependent variables of muscle and IMAT values of volume and density, and the explanatory variables BMI and age. This analysis was performed separately for both the KOA-free and KOA groups, as well as for the entire cohort.

4.2.4 Results and Discussion

4.2.4.1 Controls VS Degenerative

As initial analysis, we employed a t-test for independent samples, preceded by a Shapiro-Wilk test to confirm normal distribution. In instances where normality was not confirmed, we utilized a Mann-Whitney U test. These tests were conducted individually for each muscle, assessing both SMT, IMAT density and volume. At a significance level of 95 %, statistically significant differences were observed in SMT volume for the VI, BF, SEMIM, GMH ($p < 0.05$ for all). Similarly, significant differences were found in IMAT volume for all muscles ($p < 0.05$). No statistically significant differences were noted in SMT density across any muscle group. However, the VI, BF, SEMIM, POP, and GMH displayed notable disparities in IMAT density ($p < 0.05$ for all).

These results present significant distinctions in muscle and adipose tissue between C and D groups, highlighting the potential impact of degenerative conditions on muscle composition. IMAT values are indeed really significant both for volumes and densities as shown in Table 4.2. Figure 4.2 shows the trend for BF IMAT density and volume where, indeed, it is possible to notice that the IMAT density decreases with degeneration, while the volume increases. The same trend is noticed for all the significant muscles. On the other hand, the lack of significant differences in SMT

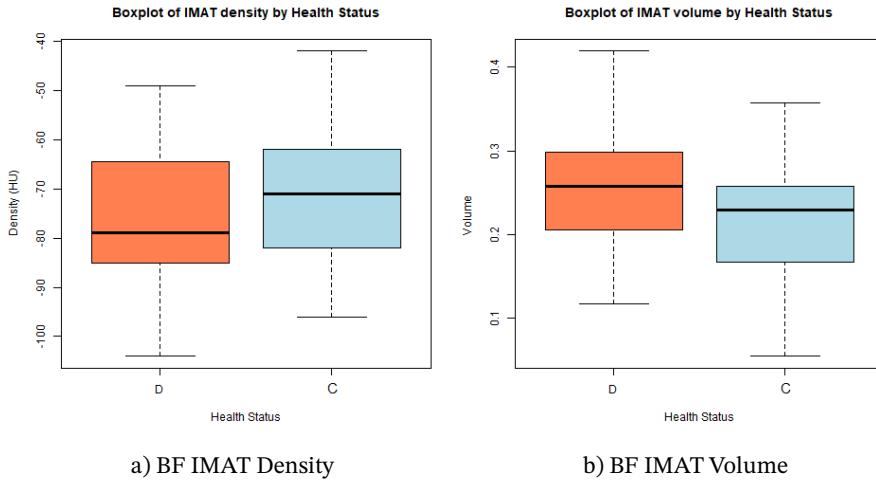


Figure 4.2: Boxplot of IMAT density and volume between D and C classes for biceps femoris muscle

density suggests that the whole structural integrity of the muscle tissue remains relatively unchanged despite significant reductions in volume for some muscles such as VI, BF, SEMIM. Surprisingly these results are in contrast with previous studies [114], which instead outline a marked difference and loss of muscle density. This may be because our study focused on examining small portions of muscle rather than the entire muscle mass. A solution could be to adopt the nonlinear tri-modal regression analysis of radiodensitometric distributions, as cited in the work of Edmunds *et al.* [162]. The authors extracted several geometric features from the radiodensitometric distribution of segmented muscle. Adopting this methodology could provide more in-depth information about the ROI of the muscle, as it does not focus only on the mean HU and StDev values. This may lead to more accurate results.

4.2.4.2 Gender

Following this, we stratified the cohorts according to the gender of the subjects. We then assessed whether there were significant differences in the mean volume and density of the tissues analyzed in our study. As previously described, we employed a t-test for independent samples, preceded by a Shapiro-Wilk test to confirm normal distribution. In examining gender differences within the control cohort, our analysis revealed no significant disparities across all IMAT variables (volume and density) or SMT volume. However, noteworthy variations were observed in the SMT density of the three Vastus, BF, SART, GRAC, SEMIM, SEMIT, and POP muscles ($p < 0.05$).

Different findings emerged when gender differences within the degenerate group were examined. Only VM showed a significant discrepancy in SMT volume (p

< 0.05), while no notable variations in density were observed in either IMAT or SMT. However, the VL and GRAC displayed noteworthy differences in IMAT volume ($p < 0.05$). Gender differences in muscle and fat composition were more pronounced in the control group than in the degenerative group. As some studies already showed [163]–[165], differences between females and males are to be expected, mostly if we consider muscle mass and strength as parameters. To explain why in the D group no differences in terms of SMT and IMAT density and volume were noticed, we might assume that the degenerative condition has a leveling effect, reducing the natural variability observed between genders in a healthy population. Further investigation needs to be conducted to verify this hypothesis.

4.2.4.3 Correlation within the KOA group

For the density and volume values of each tissue, we conducted a linear regression analysis in the KOA group to study their possible correlation with BMI and age. The results showed that muscle density values were significantly and positively correlated with BMI for all muscles ($p < 0.05$). In contrast, only the SART and GMH showed a significant correlation with age ($p < 0.05$) with this correlation displaying an inverse trend. Regarding IMAT density, all muscles displayed a negative significant correlation with BMI ($p < 0.05$), but no correlation was found with age. For SMT volume, the VM, POP, and GMH did not show a significant correlation with BMI whereas the others displayed a slight inverse trend. In addition, only the SART showed a significant and positive correlation with age ($p < 0.05$). IMAT volume positively correlated with BMI for every muscle analyzed ($p < 0.05$) except for the GMH, and no correlation was found with age.

4.2.4.4 Correlation within the KOA-free group

The same analysis was conducted on the cohort of healthy subjects. The significant results indicated strong positive correlations between SMT density with BMI for all the muscles ($p < 0.05$), and a positive one considering IMAT density ($p < 0.05$), while no correlation was found between SMT density and age. IMAT density values negatively correlated with age ($p < 0.05$), except for the VM, VI, VL, SART, and GRAC muscles. Regarding SMT volume, each muscle showed an inverse correlation with BMI ($p < 0.05$), except for the VM, and with age, except for the VI, SART, and GRAC. The IMAT volumes of all muscles positively correlated with BMI and age ($p < 0.05$), except for the VI, which did not reach the threshold of significance concerning age.

4.2.4.5 Correlation within the whole cohort

Finally, the last regression analysis was conducted considering the entire study population, both healthy and degenerative. A strong positive correlation was found between SMT densities and BMI for each muscle ($p < 0.05$), whereas the relationship with age was significant only for the Gastrocnemius muscles, displaying an inverse trend. Adipose densities were also negatively correlated with BMI and age ($p < 0.05$), except for the VM and SART densities, which did not reach significance

Table 4.3: Muscles significantly correlated with BMI and Age based on density and volume parameters related to SMT and IMAT.

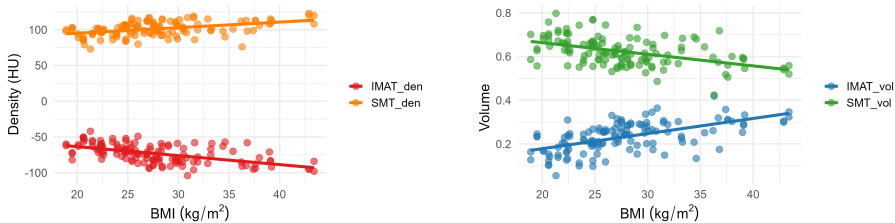
		BMI		Age	
		Muscle	Correlation	Muscle	Correlation
D	SMT density	All	positive (p<0.05)	SART, GMH	negative (p<0.05)
	IMAT density	All	negative (p<0.05)	-	-
	SMT volume	All (except VM, POP, GMH)	negative (p<0.05)	SART	positive (p<0.05)
	IMAT volume	All (except GMH)	positive (p<0.05)	-	-
C	SMT density	All	positive (p<0.05)	-	-
	IMAT density	All	negative (p<0.05)	All (except VM, VI, SART, POP)	negative (p<0.05)
	SMT volume	All (except VM)	negative (p<0.05)	All (except VI, SART, POP)	negative (p<0.05)
	IMAT volume	All	positive (p<0.05)	All (except VI)	positive (p<0.05)
D+C	SMT density	All	positive (p<0.05)	GMH, GLH	negative (p<0.05)
	IMAT density	All	negative (p<0.05)	All (except VM, SART)	negative (p<0.05)
	SMT volume	All	negative (p<0.05)	All (except SART, SEMIT)	negative (p<0.05)
	IMAT volume	All	positive (p<0.05)	All	positive (p<0.05)

concerning age. SMT volumes were negatively correlated with BMI and age ($p < 0.05$), except for the SART and SEMIT, which did not reach the significance threshold concerning age. IMAT volumes showed a significant positive correlation with both BMI and age for each muscle under consideration ($p < 0.05$). Taking Biceps femoris muscle as an example, in Figure 4.3a the linear regression analyses reveal that SMT density increases with higher BMI across the whole cohort. On the other hand, IMAT density shows a significant decrease as BMI increases, indicating denser tissue. As depicted in Figure 4.3b, which also refers to the BF muscle, the volume of SMT significantly decreases with higher BMI within the D+C category, while the quantity of IMAT increases

In the study comprising three cohorts, a consistent correlation emerged between SMT density and increasing BMI across various muscles, indicating denser muscle and adipose tissues with higher BMI. However, tissue volume changes with BMI varied depending on the tissue type, with IMAT volume increasing exponentially while muscle volume decreased. Table 4.3 highlights significant correlations between SMT and IMAT densities with BMI, suggesting BMI as a potential non-invasive marker for assessing musculoskeletal health, particularly in KOA [166]. Correlation with age in group D did not lead to significant results. However, when the analysis was conducted on the entire cohort, significance emerged for almost all parameters examined. This may be explained by the fact that when the two groups were combined, age variability increased. Increased age variability may reveal trends that are not evident in individual groups, which in our case have

a narrower age range. Therefore, although the correlation is indeed present but weak, it may not be detected in groups with a narrow age range, emerging instead when the range is wider.

The overall analysis enhances the importance of IMAT (which results to be highly significant both in the comparison D vs C and in the linear regression analysis with BMI) in the progression of KOA, indicating increased fat infiltration and important changes in the whole tissue density.



(a) Correlation analysis between IMAT and SMT density and BMI of BF muscle (b) Correlation analysis between IMAT and SMT volume and BMI of BF muscle

Figure 4.3: Regression plots showing correlation analyses between IMAT/SMT and BMI of BF muscle considering the whole cohort (D+C).

This study provides valuable insights into the interaction between KOA, muscle composition, and fat infiltration. The innovative histology-based segmentation technique allows detailed and noninvasive analysis of muscle and fat characteristics, highlighting significant differences between healthy and degenerative groups.

4.3 Multivariate Analysis of Cross-Tissue Relationships

4.3.1 Methodology

To investigate the multivariate relationships between cartilage and other joint tissues, a Canonical Correlation Analysis (CCA) was performed. This approach allows for the identification of latent canonical variates that maximize the shared variance between two multivariate datasets. In this context, CCA was applied to explore both the interplay between cartilage and bone, and the interplay between cartilage and muscle.

Referring to Table 2.3, three feature sets were considered: Cartilage Density–Morphology Dataset, Bone Dataset, Muscle Dataset. Prior to analysis, all features were normalized (z-scored, see 2.7.2) Two separate CCA models were implemented:

- Cartilage–Bone CCA, assessing the multivariate associations between cartilage density and morphology metrics and bone density features (on total, trabecular and cortical subchondral ROI).

- Cartilage–Muscle CCA, assessing associations between cartilag edensity and morphology metrics and muscle morphology and composition, including SMT and IMAT volumes and IMAT densities.

For each model, canonical correlations and associated canonical variates were computed using the `scikit-learn` implementation of CCA in Python. Feature loadings and cross-loadings were extracted to identify the features most strongly contributing to each canonical dimension.

To assess the statistical significance of the canonical relationships, two complementary multivariate tests were applied. Specifically, Wilks' Lambda (Λ) and Hotelling's Trace (T^2) were derived from the canonical correlations to quantify the overall strength of association between the two variable sets. Wilks' Lambda represents the proportion of total variance not explained by the canonical functions. It is defined as the product of the unexplained variance terms associated with each canonical correlation:

$$\Lambda = \prod_{i=1}^m (1 - r_i^2) \quad (4.1)$$

where m denotes the number of canonical functions. Values of Λ range from 0 to 1, with smaller values indicating a stronger multivariate association (i.e., less unexplained variance). When Λ approaches 1, the relationship between the two sets of variables is weak or absent. The approximate F -statistic and corresponding p -value are subsequently derived from Λ to evaluate statistical significance. F -statistic represents the ratio between explained and unexplained multivariate variance. Larger F values indicate stronger associations between the two variable sets. The corresponding p -value expresses the probability of observing such an F under the null hypothesis of no relationship. In this study, very small p -values ($p < 0.001$) confirmed that the canonical relationships between tissue domains were highly significant.

Hotelling's Trace quantifies the total shared variance across all canonical functions. In the simplified form adopted in this study, it is computed as the sum of squared canonical correlations:

$$T^2 = \sum_{i=1}^m r_i^2 \quad (4.2)$$

This formulation provides an intuitive measure bounded between 0 and m , where higher values correspond to stronger overall multivariate coupling. This bounded version ($T^2 = \sum r_i^2$) was preferred to the classical one in this study for interpretability and comparability with Wilks' Lambda.

Together, Wilks' Lambda and Hotelling's Trace provide complementary perspectives on the canonical relationships: Λ emphasizes the residual variance not captured by the model, while T^2 summarizes the total amount of variance shared across all canonical dimensions.

4.3.2 Results

Table 4.4: Summary of canonical correlation analyses between cartilage–bone and cartilage–muscle datasets, including canonical correlations, Wilks’ Lambda, Hotelling’s Trace, and significance tests.

Tissue Pair	r_1	r_2	r_3	Wilks’ Λ	F-value	p-value	Hotelling’s T^2
Cartilage–Bone	0.797	0.690	0.558	0.132	5.47	1.2×10^{-10}	1.42
Cartilage–Muscle	0.951	0.924	0.885	0.003	53.50	1.1×10^{-16}	2.54

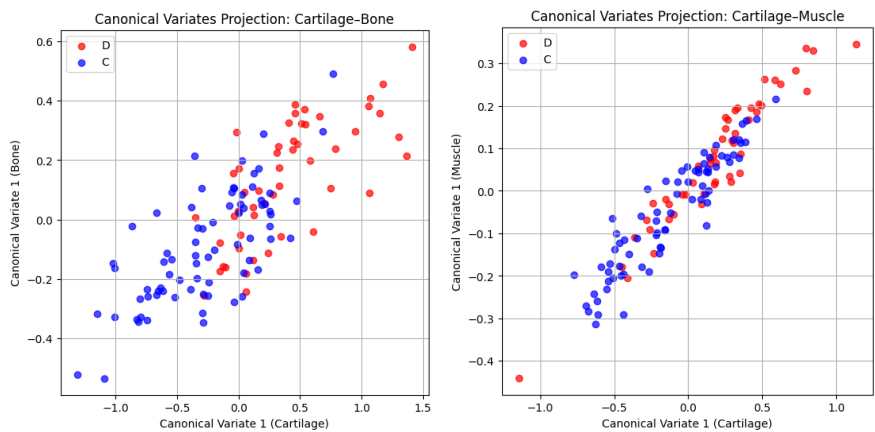


Figure 4.4: Canonical variate projections for cartilage–bone (left) and cartilage–muscle (right) analyses.

4.3.2.1 Cartilage–Bone Interaction

A CCA was first performed to investigate the multivariate relationships between cartilage and subchondral bone features. The analysis revealed a significant canonical association between the two tissue compartments (Table 4.4). The first canonical correlation ($r_1 = 0.797$), indicated a strong linear dependency between the derived canonical variates of cartilage and bone, with subsequent modes ($r_2 = 0.690$, and $r_3 = 0.558$) also contributing meaningfully to the shared variance structure. The overall multivariate test was significant (Wilks’ $\Lambda = 0.132$, $F = 5.47$, $p = 1.2 \times 10^{-10}$; Hotelling’s Trace = 1.42), confirming a robust interdependence between cartilage and bone parameters.

The canonical variate projection (Figure 4.4) illustrates the separation of D and C groups in the space defined by the first pair of canonical variates. Although partial overlap remained, subjects with disease tended to cluster toward higher scores along both variates, suggesting coordinated deviations in cartilage and bone characteristics under pathological conditions.

Feature loading patterns (Figure 4.5) identified the most influential variables driving the canonical relationship. On the cartilage side, patellar cartilage volume

and medial tibial cartilage volume exhibited the largest absolute loadings, followed by the standard deviation of density within the medial and lateral tibial regions.

For bone features, the dominant contributors were patellar average density, trabecular and cortical patellar density, and subchondral femoral average density. Notably, both average and variability (standard deviation) measures were heavily loaded, reflecting the importance of local bone microarchitecture in modulating the osteochondral response. The relatively larger bone loadings (up to $|\text{loading}| = 2.76$) suggest that subchondral bone features, and in particular patella bone, exert a stronger influence on the shared canonical structure than the cartilage variables, underscoring the biomechanical coupling across the osteochondral junction.

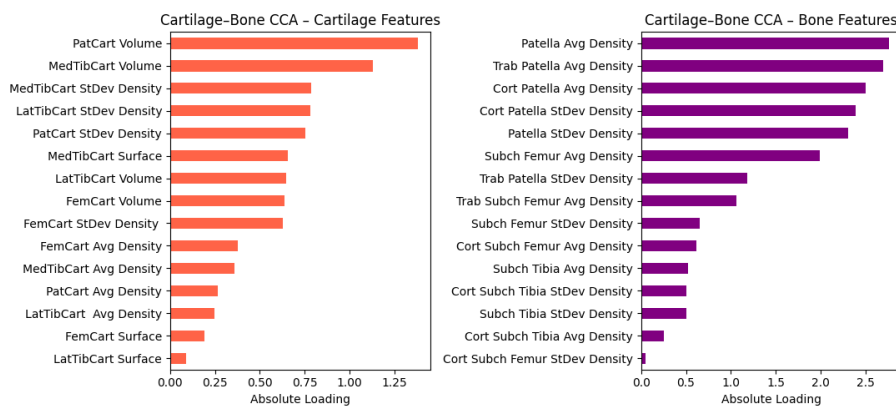


Figure 4.5: Absolute loadings of cartilage and bone features on the first canonical variate derived from the cartilage–bone CCA

4.3.2.2 Cartilage–Muscle Interaction

The second CCA examined the multivariate coupling between cartilage and muscle datasets. As summarized in Table 4.4, the strength of association was markedly higher than for the cartilage–bone relationship. The first canonical correlation was extremely strong ($r_1 = 0.951$), with additional modes also exhibiting high inter-tissue correlations ($r_2 = 0.924$, $r_3 = 0.885$). The overall multivariate model was highly significant (Wilks' $\Lambda = 0.003$, $F = 53.50$, $p = 1.1 \times 10^{-16}$; Hotelling's Trace = 2.54), indicating a very tight linear correspondence between cartilage and muscle features.

The canonical variate projection (Figure 4.4) revealed a near-perfect linear relationship between the first canonical variates of cartilage and muscle. The D and C groups were again distinguishable, with diseased participants displaying higher canonical scores, suggesting a joint modulation of muscle composition and cartilage structure under pathological loading or metabolic alterations.

Feature loadings (Figure 4.6) highlighted medial tibial cartilage standard deviation of density and femoral cartilage surface area as the key cartilage contributors, consistent with local compositional heterogeneity and joint surface remodeling.

On the muscle side, the most prominent loadings corresponded to VI and VM IMAT density measures, particularly average and standard deviation of density. These features exhibited the highest absolute loadings ($|\text{loading}|$ up to 4.15), identifying IMAT heterogeneity as a dominant driver of the cartilage–muscle canonical association.

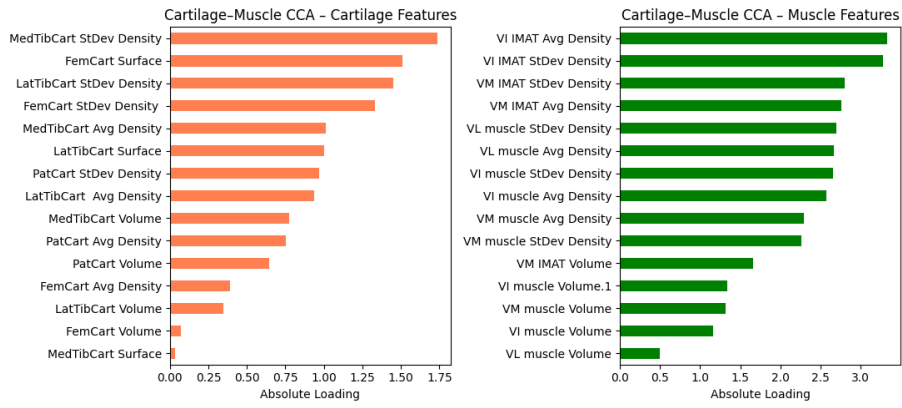


Figure 4.6: Absolute loadings of cartilage and muscle features on the first canonical variate derived from the cartilage–muscle CCA

4.3.3 Discussion

The significant association between cartilage and bone variables (Wilks’ $\Lambda = 0.132$, $p < 10^{-10}$) supports the hypothesis that cartilage degeneration and subchondral bone adaptation are tightly coupled processes [1,3].

The first canonical mode was dominated by cartilage volume and density heterogeneity, paired with high loadings in bone density metrics, particularly at the patella and subchondral femur. This aligns with the concept of osteochondral unit remodeling, where cartilage thinning and proteoglycan loss coincide with increased bone density and trabecular thickening [167], [168]. Altered load transmission through degenerated cartilage may drive bone sclerosis, while increased subchondral stiffness can further accelerate cartilage breakdown, indicating a dynamic, bidirectional interaction [147], [168]. The bilateral but non-linear relationship observed here likely reflects both adaptive and maladaptive remodeling, consistent with the progressive mechanical uncoupling of the osteochondral interface reported in advanced KOA [167], [169].

The second analysis demonstrated a remarkably strong multivariate association between cartilage and muscle features (Wilks’ $\Lambda = 0.003$, $p < 10^{-16}$), indicating a high degree of mechanical and compositional interdependence between these tissues. The first canonical mode ($r_1 = 0.951$) showed that changes in cartilage microstructure, particularly density heterogeneity and surface morphology, were closely aligned with variations in quadriceps muscle composition, especially IMAT density within the vastus intermedius and vastus medialis. Poor muscle quality,

defined by increased IMAT infiltration, has been consistently associated with cartilage degeneration and worse clinical outcomes in KOA [170], [171].

Increased IMAT within the vastus medialis correlates with elevated MRI T2 relaxation times, indicating poorer cartilage matrix quality independent of muscle volume [170]. Longitudinal and experimental evidence further support the association between quadriceps atrophy and strength loss with progressive cartilage and meniscal degeneration, suggesting that preserving muscle strength may help mitigate structural deterioration [172]–[174].

These findings highlight a mechanical interdependence between muscle composition and cartilage integrity. IMAT infiltration impairs muscle contractility and shock absorption, reducing joint stability and increasing localized stress on the articular cartilage surface.

Taken together, these results suggest that cartilage, bone, and muscle change jointly and undergo coordinated structural adaptations during KOA progression. Subtle alterations in subchondral bone density variability appear to reflect early osteochondral remodeling, complementing cartilage-based imaging biomarkers. At the same time, poor muscle composition and cartilage deterioration are mechanically linked phenomena, likely mediated by changes in joint loading patterns and reduced capacity for dynamic load modulation.

Chapter 5

Features validation and optimization

The assessment of KOA through ML-based approaches is influenced not only by the intrinsic quality of the imaging and clinical data, but also by a range of human and methodological factors. In particular, two key aspects emerge as potential sources of variability in performance: (i) the role of human operators in data preparation, especially in tasks such as cartilage segmentation, where manual or semi-automatic delineations may introduce errors; and (ii) the availability and selection of input features, which becomes particularly relevant when only reduced or incomplete datasets are available.

In this chapter, we explore the interplay between these two dimensions: human factors and data variability, within the context of the proposed methodological framework. On one side, we examine how segmentation inaccuracies, reflecting operator-dependent variability, can affect the robustness of the system. On the other hand, we investigate scenarios of data scarcity, evaluating the impact of feature selection and dimensionality reduction strategies on model performance. By systematically introducing segmentation errors and testing reduced subsets of features, we aim to assess whether our approach remains stable and reliable under these challenging conditions.

This dual analysis provides a more realistic picture of the constraints and uncertainties encountered in real-world applications. Ultimately, it highlights the importance of designing machine learning strategies for KOA that are resilient to both human variability and limitations in data availability.

5.1 Assessing the Stability of Density and Morphological Features

This section has been adapted from the study entitled “Impact of Automated and Manual Segmentation Errors on Knee Osteoarthritis Classification using MRI-registered data on CT scans”, submitted to *Computerized Medical Imaging and Graphics, Elsevier*.

Table 5.1: Mask Modifications using Morphological Functions in MIMICS.

Operation	Purpose
Erode 1	Erosion (Automated)
Dilate 1	Dilation (Automated)
Close 1 then Erode 1	Erosion (Manual)
Dilate 1 then Close 1	Dilation (Manual)

5.1.1 Imitating Automatic/Manual Error

The aim of the study was to determine whether or not the extensive manual labor and expertise required for precise cartilage segmentations would be necessary to use the original image analysis procedure of CT/MRI to identify KOA. Therefore, two study objectives were identified: to imitate automated error (as if, in an automated process, an image analysis ML model were to extract cartilage mask and part segments with boundaries expanded or dilated by one or many pixels), and to imitate manual error (similar to automated error, but with holes filled in, as to mimic what might be created by an individual with little experience in segmentation or an untrained eye for cartilage boundaries). Both of these errors might also occur due to equipment error, such as low quality or blurry images.

To manipulate images within the MIMICS software, all masks were uploaded onto MIMICS and subject to “erosion” (erode), “dilation” (dilate), “opening” (open), and “closing” (close) operations (or combinations of the operations) using the Morphological Functions tool with 8-bit dilation or erosion. A spherical structural element was used to perform the morphological operations. The erode function eliminates the outer boundary of a mask of the number of pixels selected, whereas the dilate function extends the boundary of the mask. The open function performs an erode followed by a dilate to break small connections in the mask. The close function performs a dilate followed by an open, filling cavities within a mask. 8-bit connectivity performs erode, dilate, open, or close functions on pixels in the surrounding plane (2D) [175] [175]. The erode and dilate functions used individually can mimic automated error by expanding or reducing mask boundaries without filling in holes created by these operations. Using erode and dilate operations in tandem with close operation will fill in holes to better simulate manual error.

The final modifications used to imitate automatic and manual errors are described in Table 5.1 and visualized in Figures 5.1 (for D samples) and 5.2 (C samples). The same 16 features used for the original mask in section 3.1 (Table 2.3) were also obtained from these modified segmentations.

One patient in class “D” had all pixels in the medial tibial cartilage mask eliminated with the use of “Erode 1” and “Close 1 Erode 1” modifications. This instance was removed from “Erode 1” and “Close 1 Erode 1” modification datasets. The final dataset input into ML models contained 79 instances in class “C” and 50 or 51 in class “D”.

To evaluate the robustness of the extracted features, the same statistic tests and ML algorithms as in 3.1.1 were employed.

Figure 5.1: Examples of “Original” (center), “Dilate 1” (left), and “Erode 1” cartilage masks and three-dimensional parts for a patient with KOA (Class: “Degenerative” (D)).

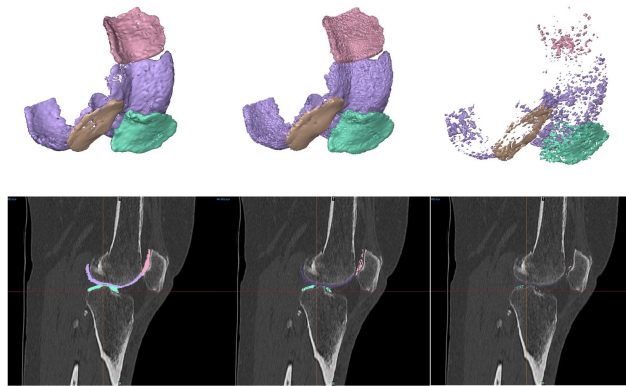
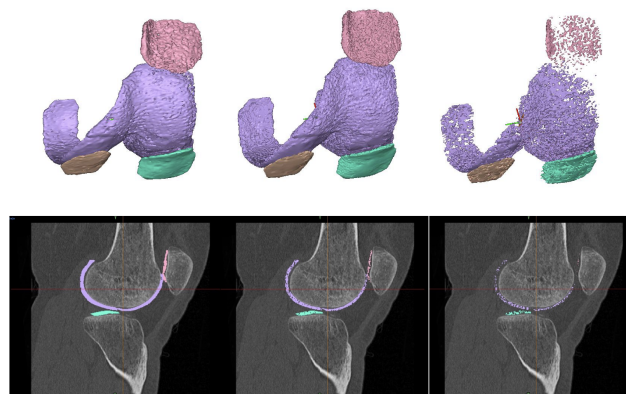


Figure 5.2: Examples of “Original” (center), “Dilate 1” (left), and “Erode 1” cartilage masks and three-dimensional parts for a healthy patient (Class: “Control” (C)).



5.1.2 Results: Statistical Analysis

Table 5.2 presents the p-values computed for modified datasets; p-values from the “Original” Density-Morphology dataset (from Table 2.3 and used in section 3.1) were also reported for comparison. Notably, the standard deviations of the densities maintained significant p-values across all datasets, despite the pixel modifications. The medial tibial cartilage volume and surface also remained significant across the datasets. However, the results were different for other features: the surface area of the femoral cartilage, the volume of the patellar cartilage, and the average HU density of the medial tibial cartilage were significant in the original dataset, but only some of the modified datasets retained these significant results. Certain features showed significance only in specific modified datasets; for instance, ‘PatCart Surface’ was significant only in the Erode1 dataset, and ‘LatTibCart Surface’ was significant only for the Dilate1 dataset. Meanwhile, ‘FemCart Volume’ was significant in the Erode1, Dilate1, and Dilate1Close1 datasets but not in the original dataset.

Table 5.2: p-values for Original, Erode1, Dilate1, Close1Erode1, and Dilate1Close1 datasets, rounded to 3 significant digits; significant values highlighted in ***bold italic***; *: $p < 0.05$, **: $p < 0.005$, ***: $p < 0.001$.

Feature	p-values				
	Original	Erode1	Dilate1	Close1Erode1	Dilate1Close1
FemCart Avg Density	0.919	0.825	<i>0.034*</i>	<i>0.000***</i>	<i>0.031*</i>
FemCart StDev Density	<i>0.000***</i>	<i>0.000***</i>	<i>0.001**</i>	<i>0.000***</i>	<i>0.001**</i>
FemCart Volume	0.245	<i>0.005**</i>	<i>0.001**</i>	0.331	<i>0.000***</i>
FemCart Surface	<i>0.000***</i>	0.299	<i>0.000***</i>	0.178	<i>0.000***</i>
PatCart Avg Density	0.987	0.938	0.400	<i>0.006**</i>	0.361
PatCart StDev Density	<i>0.000***</i>	<i>0.000***</i>	<i>0.001**</i>	<i>0.000***</i>	<i>0.001**</i>
PatCart Volume	<i>0.000***</i>	<i>0.000***</i>	0.077	<i>0.000***</i>	0.100
PatCart Surface	0.654	<i>0.000***</i>	0.424	<i>0.045*</i>	0.684
MedTibCart Avg Density	<i>0.019*</i>	<i>0.003**</i>	0.191	0.796	0.200
MedTibCart StDev Density	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>
MedTibCart Volume	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>
MedTibCart Surface	<i>0.011*</i>	<i>0.000***</i>	<i>0.037*</i>	<i>0.000***</i>	<i>0.034*</i>
LatTibCart Avg Density	0.103	0.148	0.552	0.785	0.533
LatTibCart StDev Density	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>	<i>0.000***</i>
LatTibCart Volume	0.092	<i>0.000***</i>	0.492	0.065	0.530
LatTibCart Surface	0.310	0.142	0.116	0.566	0.148

5.1.3 Results: Machine Learning Performances

All ML analyses were performed on the four datasets simulating automatic and manual errors. The same cross validation (5-fold nested CV with hyperparameters tuning) and the same model performances as in subsubsection 3.1.1.2 (‘Original dataset’) were employed.

5.1.3.1 Pixel Erode 1

Erode1, the first ‘automated’ error introduced, was applied as a one-pixel erosion to the cartilage masks of 130 samples. Out of these, 129 samples (79 in class C and 50 in class D) retained all their cartilage masks after the modification. However, one patient in class D lost all remaining pixels in their medial tibial cartilage mask as a result of the modification.

The ML evaluated on the dataset, including the features extracted from the eroded masks, did not reach the same scores as the original dataset, but still remained pretty high in scores. Specifically, the highest weighted F1 score was 0.83 ± 0.04 for the Linear SVM model. Although this score is lower than the best model’s score on the original dataset, it was still higher than the score obtained by the same model on the original dataset. LR and RBF SVM also recorded similarly high scores. RF and GB models achieved scores similar to those on the original dataset, with a weighted F1 score of 0.78. DT model performed the worst across all metrics (Table 5.3).

Table 5.3: Performance metrics of different models on the Erode1 Dataset, best model score for each metric in bold

Erode1						
Model	Accuracy	Precision	Sensitivity	Specificity	F1 Weighted	ROC AUC
Linear SVM	0.83 ± 0.04	0.82 ± 0.11	0.74 ± 0.05	0.89 ± 0.08	0.83 ± 0.04	0.86 ± 0.04
RBF SVM	0.81 ± 0.07	0.79 ± 0.11	0.72 ± 0.07	0.87 ± 0.08	0.81 ± 0.06	0.86 ± 0.07
Logistic Regression	0.82 ± 0.04	0.81 ± 0.12	0.74 ± 0.05	0.87 ± 0.09	0.82 ± 0.04	0.88 ± 0.04
Decision Tree	0.71 ± 0.06	0.63 ± 0.08	0.64 ± 0.14	0.75 ± 0.10	0.70 ± 0.06	0.72 ± 0.06
Random Forest	0.78 ± 0.05	0.76 ± 0.12	0.70 ± 0.11	0.84 ± 0.12	0.78 ± 0.05	0.89 ± 0.06
Gradient Boosting	0.78 ± 0.10	0.72 ± 0.14	0.76 ± 0.05	0.79 ± 0.16	0.78 ± 0.09	0.86 ± 0.06

5.1.3.2 Pixel Dilate 1

Dilate1 is the other automated error introduced. In this case, no samples were lost; therefore, the new dataset still comprises 130 subjects. Table 5.4 shows that the best model remains RBF SVM on most of the metrics, maintaining a high score of 0.81 ± 0.06 on F1. Linear SVM reached performances similar to the ones in the Original dataset (0.80), while LR was slightly higher (0.79). All the other models scored lower results with respect to the original dataset (even if RF showed the highest specificity of the dataset, but only slightly higher than the original one and at the expense of a pretty low sensitivity). Also in this case, DT had the lower scores, not reaching 0.6 in most of the metrics.

5.1.3.3 Pixel Close 1 Erode 1

Close1Erode1 is a manual simulated error, composed by a closing operation followed by an erosion on all the segmented cartilages. Even in this case, due to the applied erosion, one D sample had to be removed because it lost all pixels. Results on this new dataset are shown in Table 5.5. Also in this case, RBF SVM reaches the highest score on all the metrics, with an F1 of 0.84, only 2 percentage points lower

Table 5.4: Performance metrics of different models on the Dilate1 Dataset, best model score for each metric in bold.

Dilate1						
Model	Accuracy	Precision	Sensitivity	Specificity	F1 Weighted	ROC AUC
Linear SVM	0.80 $\pm$ 0.07	0.75 $\pm$ 0.12	0.75 $\pm$ 0.13	0.84 $\pm$ 0.08	0.80 $\pm$ 0.07	0.90 $\pm$ 0.06
RBF SVM	0.82 $\pm$ 0.06	0.80 $\pm$ 0.13	0.72 $\pm$ 0.08	0.87 $\pm$ 0.09	0.81 $\pm$ 0.06	0.91 $\pm$ 0.07
Logistic Regression	0.79 $\pm$ 0.08	0.75 $\pm$ 0.12	0.75 $\pm$ 0.15	0.82 $\pm$ 0.10	0.79 $\pm$ 0.08	0.90 $\pm$ 0.06
Decision Tree	0.58 $\pm$ 0.07	0.46 $\pm$ 0.12	0.49 $\pm$ 0.16	0.65 $\pm$ 0.06	0.58 $\pm$ 0.07	0.57 $\pm$ 0.09
Random Forest	0.76 $\pm$ 0.04	0.77 $\pm$ 0.07	0.57 $\pm$ 0.11	0.89 $\pm$ 0.05	0.75 $\pm$ 0.04	0.83 $\pm$ 0.08
Gradient Boosting	0.73 $\pm$ 0.07	0.70 $\pm$ 0.16	0.57 $\pm$ 0.11	0.84 $\pm$ 0.08	0.72 $\pm$ 0.07	0.82 $\pm$ 0.07

than the original dataset, and with less variability across folds (StDev = 0.02 versus 0.07 in the original dataset). In this case, RF is the second best model, reaching a F1 of 0.81 ± 0.06 , comparable to the highest outcomes of the original dataset, but significantly better than the same algorithm applied to that dataset. Worst model is again DT, with 0.70 ± 0.07 of F1.

Table 5.5: Performance metrics of different models on the Close1Erode1 Dataset, best model score for each metric in bold.

Close1Erode1						
Model	Accuracy	Precision	Sensitivity	Specificity	F1 Weighted	ROC AUC
Linear SVM	0.77 $\pm$ 0.06	0.74 $\pm$ 0.13	0.68 $\pm$ 0.07	0.83 $\pm$ 0.09	0.77 $\pm$ 0.06	0.85 $\pm$ 0.06
RBF SVM	0.84 $\pm$ 0.02	0.83 $\pm$ 0.10	0.78 $\pm$ 0.10	0.89 $\pm$ 0.07	0.84 $\pm$ 0.02	0.90 $\pm$ 0.04
Logistic Regression	0.79 $\pm$ 0.06	0.73 $\pm$ 0.10	0.76 $\pm$ 0.05	0.81 $\pm$ 0.08	0.79 $\pm$ 0.06	0.86 $\pm$ 0.04
Decision Tree	0.71 $\pm$ 0.07	0.61 $\pm$ 0.06	0.62 $\pm$ 0.19	0.76 $\pm$ 0.02	0.70 $\pm$ 0.07	0.73 $\pm$ 0.09
Random Forest	0.81 $\pm$ 0.06	0.76 $\pm$ 0.10	0.76 $\pm$ 0.05	0.84 $\pm$ 0.08	0.81 $\pm$ 0.06	0.88 $\pm$ 0.06
Gradient Boosting	0.79 $\pm$ 0.11	0.73 $\pm$ 0.13	0.72 $\pm$ 0.17	0.84 $\pm$ 0.08	0.79 $\pm$ 0.11	0.86 $\pm$ 0.08

5.1.3.4 Pixel Dilate 1 Close 1

The other manual simulated error is given by a dilation of 1 pixel followed by a closing operation of 1 pixel. For this simulated error, Table 5.6 displays different trends than the other datasets. Although the RBF SVM maintains high performance, the best weighted F1 score was achieved by Linear SVM, with a value of 0.81 ± 0.06 . Additionally, Linear SVM produced the highest ROC AUC of 0.92 ± 0.05 . This score is not only the highest for this dataset, but also the highest overall, including results from the Original dataset. LR reached the highest accuracy and sensitivity for this dataset, while RF reached the highest precision and specificity, but again at the expense of a pretty low sensitivity. DT is confirmed as the model with the lowest scores.

Table 5.6: Performance metrics of different models on the Dilate1Close1 Dataset, best model score for each metric in bold.

Dilate1Close1						
Model	Accuracy	Precision	Sensitivity	Specificity	F1 Weighted	ROC AUC
Linear SVM	0.81 $\pm$ 0.06	0.78 $\pm$ 0.09	0.73 $\pm$ 0.11	0.86 $\pm$ 0.07	0.81 $\pm$ 0.06	0.92 $\pm$ 0.05
RBF SVM	0.80 $\pm$ 0.07	0.78 $\pm$ 0.14	0.70 $\pm$ 0.11	0.86 $\pm$ 0.10	0.80 $\pm$ 0.07	0.90 $\pm$ 0.07
Logistic Regression	0.82 $\pm$ 0.08	0.78 $\pm$ 0.10	0.75 $\pm$ 0.19	0.86 $\pm$ 0.07	0.81 $\pm$ 0.09	0.92 $\pm$ 0.05
Decision Tree	0.68 $\pm$ 0.06	0.60 $\pm$ 0.10	0.49 $\pm$ 0.19	0.80 $\pm$ 0.06	0.66 $\pm$ 0.07	0.65 $\pm$ 0.09
Random Forest	0.79 $\pm$ 0.06	0.84 $\pm$ 0.11	0.59 $\pm$ 0.13	0.93 $\pm$ 0.06	0.78 $\pm$ 0.07	0.86 $\pm$ 0.05
Gradient Boosting	0.75 $\pm$ 0.11	0.69 $\pm$ 0.13	0.67 $\pm$ 0.19	0.81 $\pm$ 0.09	0.75 $\pm$ 0.11	0.79 $\pm$ 0.10

5.1.4 Discussion

5.1.4.1 Comparison of Performance Metrics Between Erosion and Dilation Errors

Dilation errors typically had a stronger negative effect on performance metrics than erosion errors, especially F1 scores. This discrepancy emphasizes the impact of including the boundary outside of the cartilage boundary in image-based ML classification. The cartilage boundary is where KOA-driven degeneration occurs, which should help model performance, but capturing too wide an outer boundary might include non-cartilage tissues which negatively impact model performance.

Errors which simulated erosion miss the diverseness of degenerative cartilage boundaries, which is less present in the boundaries of healthy cartilage, but draw greater focus to the central pixels of the mask which might provide insight into the composition of the inner cartilage as a result of KOA pathophysiology.

5.1.4.2 Comparison of Performance Metrics Between Automated and Manual Errors

5.1.4.3 Erosion (Erode 1, Close 1 Erode 1)

The datasets which mimicked automated errors returned higher performance metric scores for F1, specificity, and accuracy, although differences between manual and automatic erosion errors always differed by less than 0.025. The non-tree-based models (LR, RBF SVM, Linear SVM) returned the highest F1 scores for automated erosion error (Erode 1), whereas RBF SVM and RF returned the highest F1 scores when working with the Close 1 Erode 1 dataset. For both datasets, models returned F1 scores within a small range, with the exception of DT which returned a markedly lower score: 0.78-0.83 for Erode 1, DT F1 score = 0.70; 0.77-0.84 for Close 1 Erode 1, DT F1 score = 0.66).

The similarities between F1 scores for automatic and manual erosion errors reflect that the inner cartilage structure presents a degree of similarity which is not widely influenced by the filling of holes within mask segmentations. Further exploration of the effects of larger erosion errors might illustrate the extent of this similarity as more central pixels' volumes, surface areas and HU values are given more weight in ML-based classification.

5.1.4.4 Dilation (Dilate 1, Dilate 1 Close 1)

Manual dilation errors returned higher average scores than automated dilation errors for all performance metrics, although once again the difference between scores was marginal (0.04 or less). Also paralleling the comparison between manual and automatic erosion errors, F1 scores for manual and automatic dilation errors were within a small range with the exception of the DT models. For automated dilation error (Dilate 1), F1 scores ranged from 0.75-0.81, with DT F1 = 0.66. For manual dilation error (Dilate 1 Close 1), F1 scores ranged from 0.72-0.81, with DT F1 = 0.58.

The similarities between manual and automated dilation error performance metrics suggests that the outer boundary of segmented cartilage masks do not exhibit strong influence from the filling of holes. Again, further exploration of the effects of larger dilation errors can illustrate the extent of this influence, especially when considering non-cartilage tissues which exist in the outer boundary of the cartilage that might play a role in classifying KOA with image-based ML models.

5.1.4.5 Comparisons between Original dataset and Simulated Error datasets

The original dataset returned the highest F1 score or within 0.03 of the highest F1 score across all models and datasets. For every metric except sensitivity, the original dataset returned the highest average score across all models. In the analysis of the original dataset, every model returned most metric scores of > 0.60 , with the highest score for each metric of every model > 0.75 . No model returned any performance metric scores < 0.5 , or “random chance”. For most ML classifications, scores of 0.7 or more for any performance metric is considered adequate, although the stakes for correct classifications (especially of positive cases) is high in medical settings and higher scores are ideal, especially for sensitivity and F1 score.

Overall, there was little variation across scores between the original dataset and imitated error sets. The only exception to this was with the DT models, in particular when the simulated automatic dilation error (Dilate 1) dataset was analyzed. This suggests that, while the original dataset returned comparatively higher average performance metric scores, most ML models tested in this work would be appropriate for clinical use even with imprecise cartilage segmentation inputs. Expansion to a larger dataset is necessary to fully determine the ability of these models for KOA classification before clinical applications are considered.

5.1.4.6 Importance of Mask Segmentation Precision

The performance of the ML models using imprecise cartilage segmentation data suggests that the precision of the mask segmentations from the original dataset might not be necessary for the image-based KOA classification described in this work. Although most scores procured in this work are promising, the type of error imitation and feature set created variance in these scores which should be noted, as even minute variance in performance metrics of different models can bring significant consequences for healthcare providers and patients. With the exception

of DT, the models achieved scores greater than 0.7 for almost all performance metrics, a score considered acceptable in most ML applications. However, stakes and consequences of incorrect classifications for KOA diagnosis are too high to accept these scores without replication of these methods on larger datasets.

The results of this work are important for clinical considerations. Precise image analysis segmentations for KOA classification are expensive in finances, time, and personnel, even with access to experts of navigating MIMICS software. Therefore, these results suggest that cartilage mask segmentation in MIMICS software could be an accessible and efficient implementation for KOA classification.

5.1.4.7 Strengths and Limitation

Both a strengthening and limiting aspect of the work is the use of MIMICS scripting to imitate automated or manual error. Only “erode”, “dilate”, “open” and “close” operators provided in MIMICS were used to modify cartilage masks. However, this limitation allowed for an identically sized set (relative to the original subject set) of imitated “errors” to be created. In future work, instead of estimating what imprecise masks might look like, actual segmentations created by users of the software might be used to identify the types of errors that occur (such as imprecise segmentation or low image quality). This data could be given to the ML models to determine if cartilage degeneration can still be identified with these ML models.

Some of the models returned nearly identical scores for performance metrics. It is likely that the similar scores were achieved due to the small test set size for all models. Running the ML models on larger datasets would likely correct this occurrence, allowing for deeper analysis and identification of true model performance rankings.

5.2 Assessing the Stability of Radiomic Features

This section has been adapted from the study entitled “Innovative Diagnostic Approaches for Predicting Knee Cartilage Degeneration in Osteoarthritis Patients: A Radiomics-Based Study”, published on *Information System Frontiers*, Springer, as part of the Special Issue: *eXtended Reality and Artificial Intelligence in Medicine and Rehabilitation* [3].

The aim here is to evaluate the performance and robustness of radiomic features under different imaging conditions, highlighting the potential of this approach to improve diagnostic accuracy.

The same features extracted in section 3.2 (Table 2.3) were used.

5.2.1 Robustness analysis

Following the statistical approach proposed in [176], a robustness analysis of the radiomics features analyzed in section 3.2 was conducted to improve the reliability of the analysis. In particular, the variability and robustness of radiomics features as a function of cartilage segmentation, both in MRI and CT images, were investigated by varying the shape of the segmentation through morphological operators and

simulating erosion or dilation of the area. Figure 5.3 shows a visual representation of the effects of morphological operations, specifically dilation and erosion, on segmented cartilages: each panel demonstrates the effect of different levels of dilation and erosion on what could be interpreted as variation in segmenting biological structures. Starting from the top, the first red block shows the effect of dilation, expanding the boundaries of the ROI from 5 to 1 pixels, showing progressively lower levels of dilation. The middle row labeled 'Original' represents the baseline. Following, in the blue block, there are the effects of erosion from 1 to 5 pixels. Such manipulations are crucial for exploring the robustness of radiomics features since changing the shape and size of ROIs reflects intra- and inter-operator variability in segmentation, which is common in practical clinical applications.

Following these alterations, a full extraction of radiomic features is performed on both the original and modified ROIs. Then, a Shapiro-Wilk test ($\alpha = 0.05$) was performed to evaluate the normality of the distribution of radiomic features, and a t-test or Mann-Whitney statistical test was performed (depending on the normality result) to compare the distributions of radiomic features between original ROI and dilated ROI and between original ROI and eroded ROI. Based on the p-value obtained, a robustness score in the range between 0 and 1 was obtained, as detailed in [176]. Robustness score values closer to 1 indicate greater robustness of the feature classes examined.

Boxplots and trend lines were represented to see how the robustness score varies as the amplitude of the erosion or dilation varies.

An innovative method was then developed to integrate robustness into the calculation of feature importance for ML models. This method prioritizes features not only based on model significance but also on their stability and data variations, such as those introduced by noise or changes to ROI segmentation.

Typically, feature importance in ML is determined by each feature's contribution to the model's predictive accuracy, but this traditional approach often overlooks the stability or robustness of features to fluctuations in the input data, especially when the sample of data is not very large. In this new methodology, robustness is quantified using the negative logarithm of the average p-value of the robustness score, which inversely correlates a feature's weight with its robustness: a lower p-value, suggesting a lower likelihood that the feature's robustness is random, results in a higher weight, and the negative logarithm transformation scales the p-values to a more understandable range, where smaller p-values (indicating stronger evidence against the null hypothesis and affirming the robustness of the feature) produce larger weights. By multiplying the original feature importance (from 3.2) by this robustness weight, importance scores are recalibrated to favor features that are both predictive and robust. Features with lower robustness (higher p-value) suffer a penalty, decreasing their modified importance score. The reweighted system may reveal that some features previously perceived as crucial are less important when robustness is taken into account, while others gain importance due to their enhanced robustness, suggesting a more reliable set of predictors.

Statistical tests were carried out in Python using the SciPy library [177].

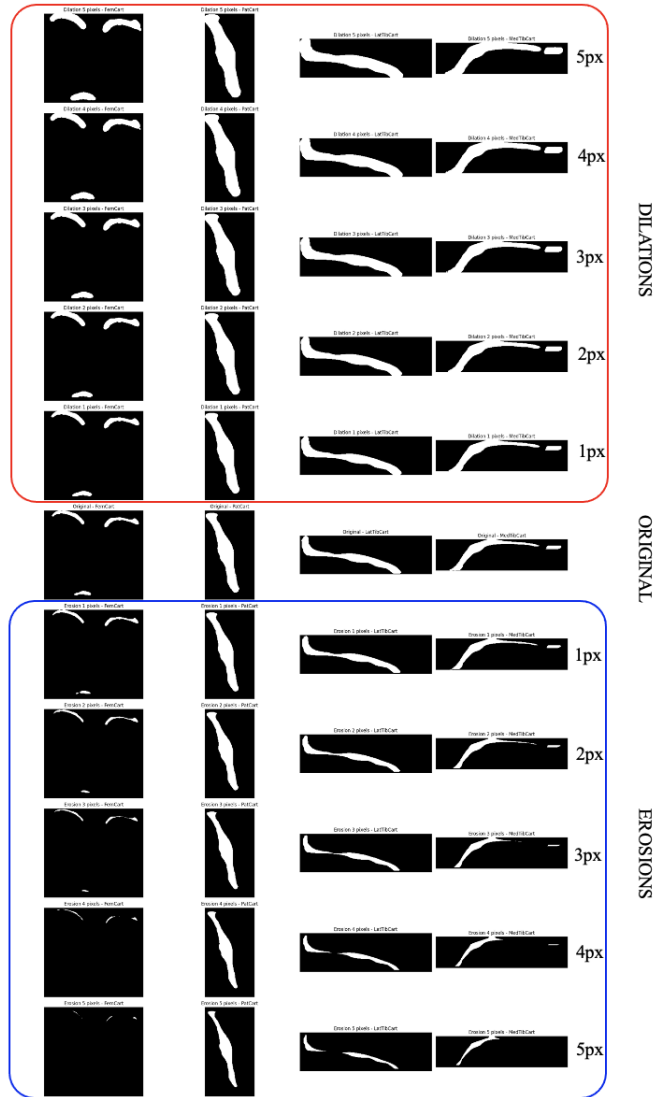


Figure 5.3: Morphological operations on extracted masks.

5.2.2 Robustness analysis results

After introducing a controlled noise by dilating and eroding the edges of the segmentation and extracting the radiomic features for each dilated, eroded and original ROI, a robustness score was produced based on the p-value resulting from the Mann-Whitney test. In particular, boxplots were produced for variations of pixels

from 1 to 5, which facilitate robustness comparisons between classes of features. In particular, in Figure 5.4 and Figure 5.6 it is possible to observe the boxplots of robustness score for the segmentations with edge dilation from 1 to 5 pixels, respectively for MRI and CT. In Figure 5.5 and Figure 5.7, however, the boxplots of robustness score for the segmentations with edge erosion from 1 to 5 pixels are shown. For both erosion and dilation, considering that a robustness score closer to 1 indicates greater robustness of the feature classes, it is possible to observe that the robustness decreases as the degree of increase of the morphological operation increases. This is an expected result because greater alterations of the ROI determine a higher variability in the extracted radiomic features. However, it is interesting to note how erosion is more affected by this phenomenon. Observing Figures 5.5 and 5.7, for an alteration of 5 pixels, the robustness score drops much more than the analog in Figures 5.4 and 5.6. Trend lines, shown in Figure 5.8 and Figure 5.9, have also been graphed to show the trend of the robustness score as the degree of variation in pixels in erosion/dilation varies. In these plots, a more robust feature class would show a flatter trend line, indicating that its robustness score is less affected by changes in ROI.

As can be seen from the boxplots and trend plots, the first-order features and the GLCM features remain more robust for all dilation amplitudes, with greater robustness for small variations. In the case of erosion, however, we can observe greater robustness in the case of first-order, NGDTM and GLCM features.

By combining the robustness analysis with the feature importance analysis, we obtain new weighted feature importance scores. The 50 most important robust features for each best model are shown in the barplots in Figure 5.10, Figure 5.11, Figure 5.12, and Figure 5.13.

By retraining the ML algorithms with these new more important and robust features, we obtain the metrics summarized in the Tables 5.7 and 5.8. As expected, the tables show lower results than previous performances, in which the importance of features was evaluated solely on the basis of meticulously executed segmentation data. However, moving away from an ideal situation and considering the great variability of results during the segmentation process, having more generalizable and more robust features becomes crucial, despite lower but still optimal results.

Table 5.7: Compared Performances among models in MRI radiomics data with 50 most important robust features

Model	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	78.96 (± 6.8)	75.58 (± 5.9)	79.63 (± 6.4)	78.96 (± 5.3)	78.53 (± 6.1)
SVM Linear	85.55 (± 5.3)	84.42 (± 4.1)	83.74 (± 6.1)	85.55 (± 4.0)	85.49 (± 4.2)
Logistic Regression	88.68 (± 1.48)	88.10 (± 2.00)	88.74 (± 5.3)	88.68 (± 4.5)	87.66 (± 3.3)
Gradient Boosting	72.61 (± 4.9)	70.86 (± 5.1)	72.72 (± 6.1)	72.61 (± 5.8)	72.39 (± 6.2)

5.2.3 Discussion

The objective of the robustness analysis is to enhance the reliability of radiomic features by assessing their stability against segmentation variations in MRI and CT images. Indeed, segmentation is affected by observer variability: individual

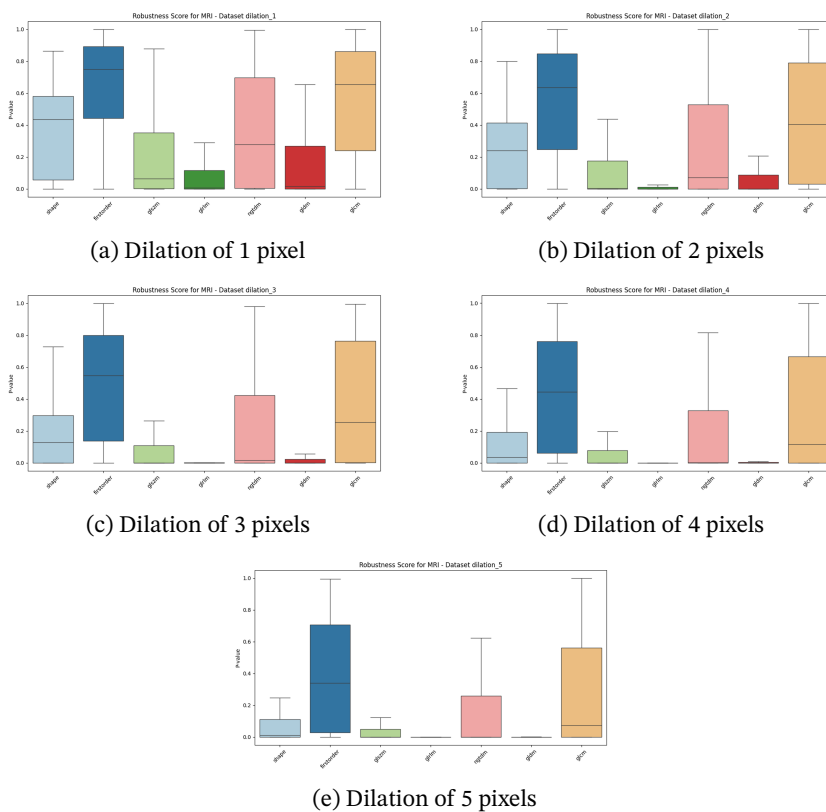


Figure 5.4: Robustness score boxplots for each radiomics class for different dilation pixels on MRI features

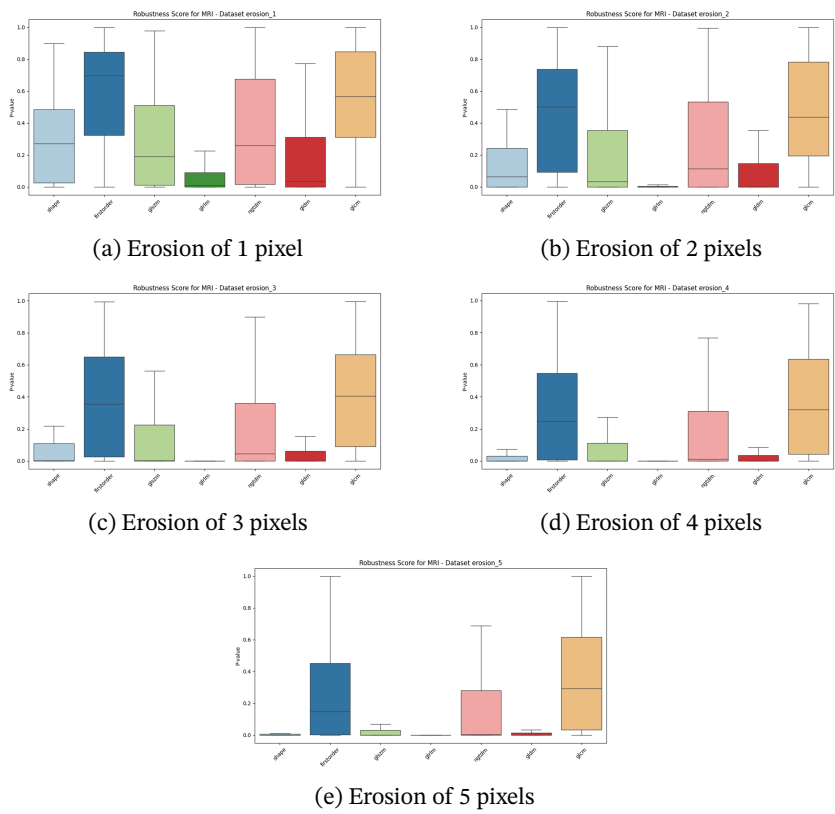


Figure 5.5: Robustness score boxplots and trendlines for erosion of different pixel sizes on MRI features

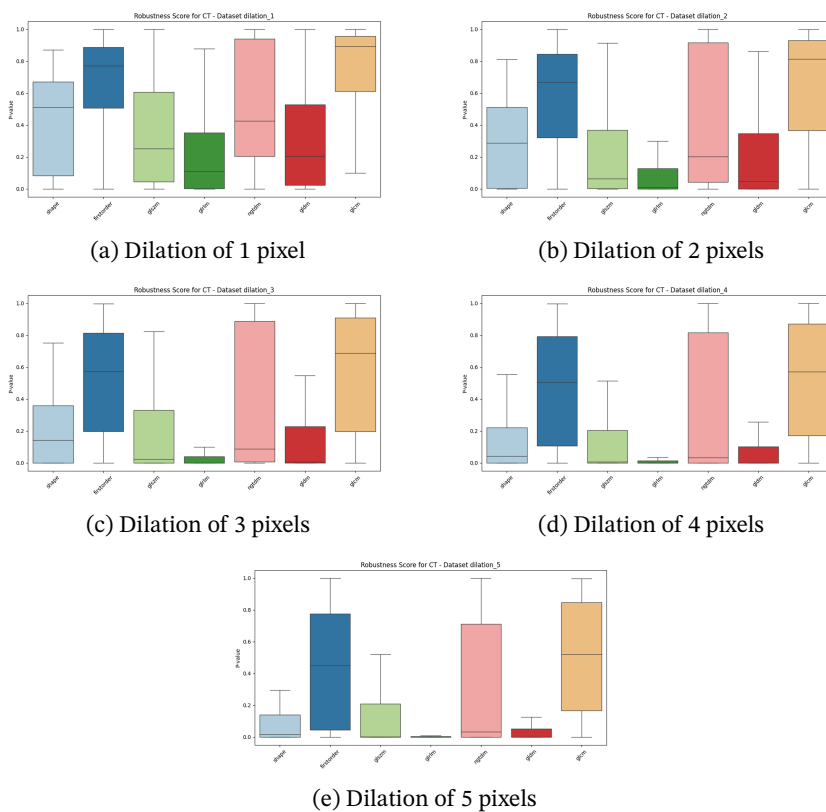
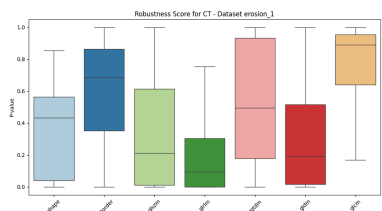
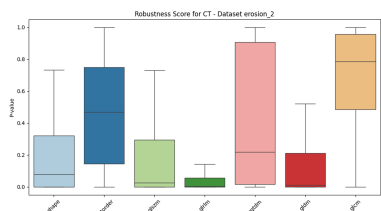


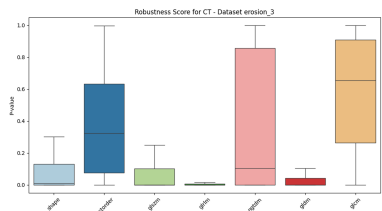
Figure 5.6: Robustness score boxplots for each radiomics class for different dilation pixels on CT features



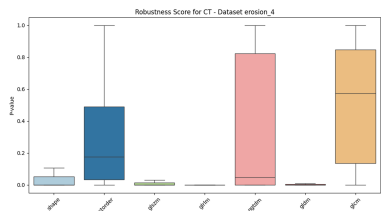
(a) Erosion of 1 pixel



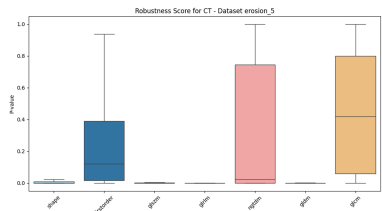
(b) Erosion of 2 pixels



(c) Erosion of 3 pixels

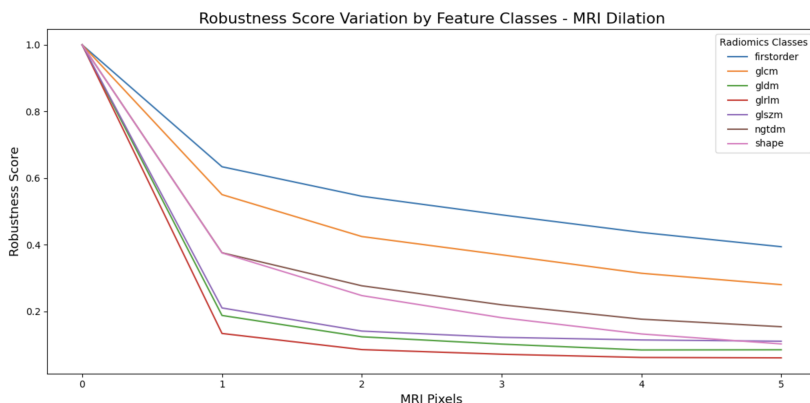


(d) Erosion of 4 pixels

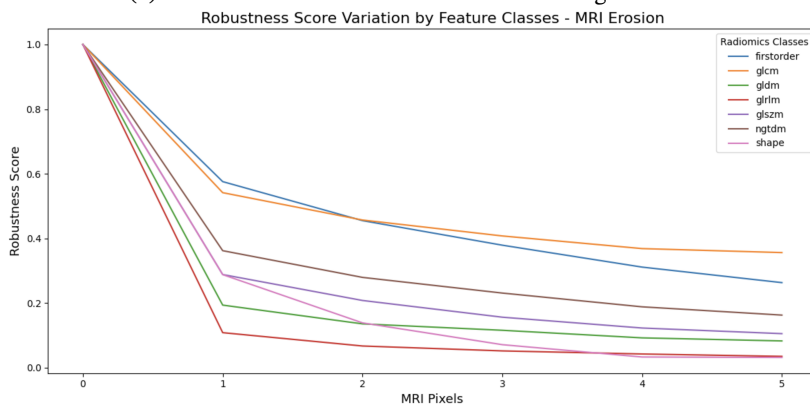


(e) Erosion of 5 pixels

Figure 5.7: Robustness score boxplots and trendlines for erosion of different pixel sizes on CT features

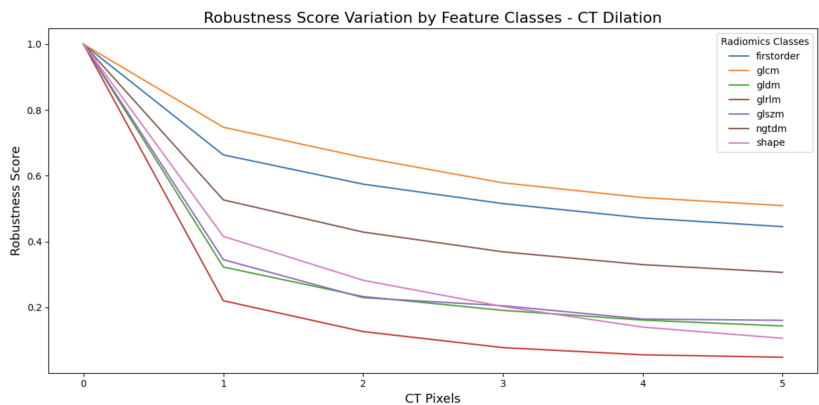


(a) Trend of robustness scores under increasing dilation.

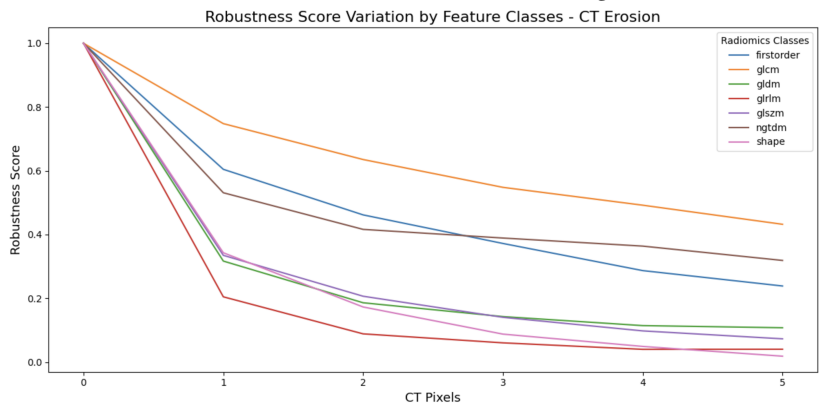


(b) Trend of robustness scores under increasing erosion.

Figure 5.8: Trends of feature robustness in MRI images under morphological perturbations. Each plot shows the distribution and trend of robustness scores across varying pixel-level dilations and erosions.



(a) Trend of robustness scores under increasing dilation.



(b) Trend of robustness scores under increasing erosion.

Figure 5.9: Trends of feature robustness in CT images under morphological perturbations. The boxplots and trendlines show how robustness scores change with progressive dilation and erosion of the segmented regions.

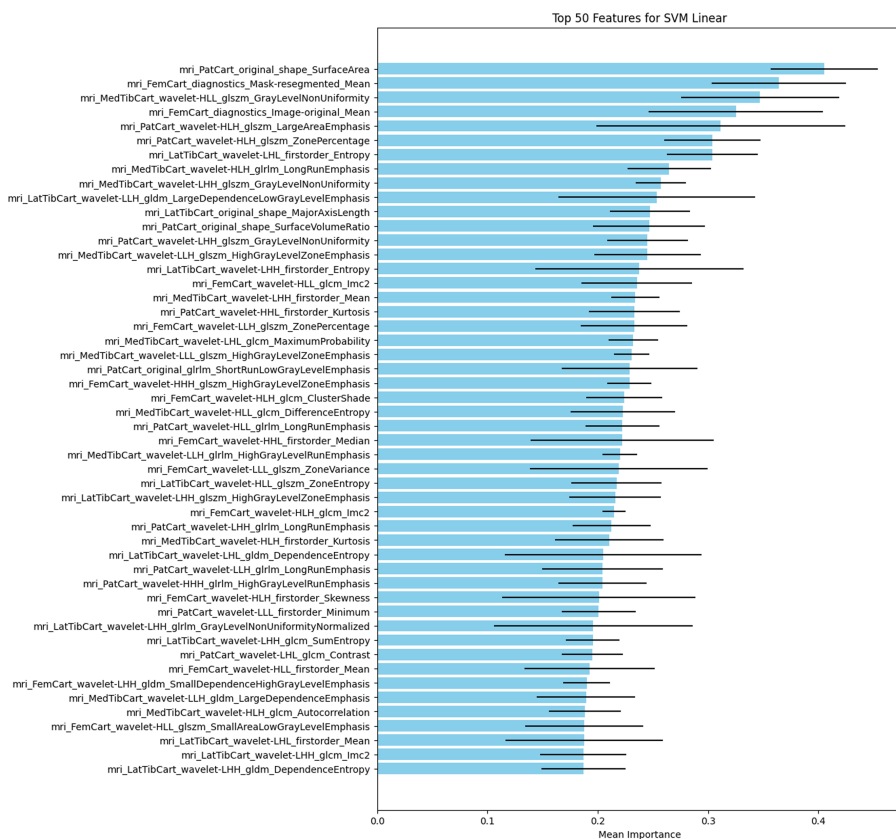


Figure 5.10: Barplot of the 50 most important robust features for the SVM linear model in MRI.

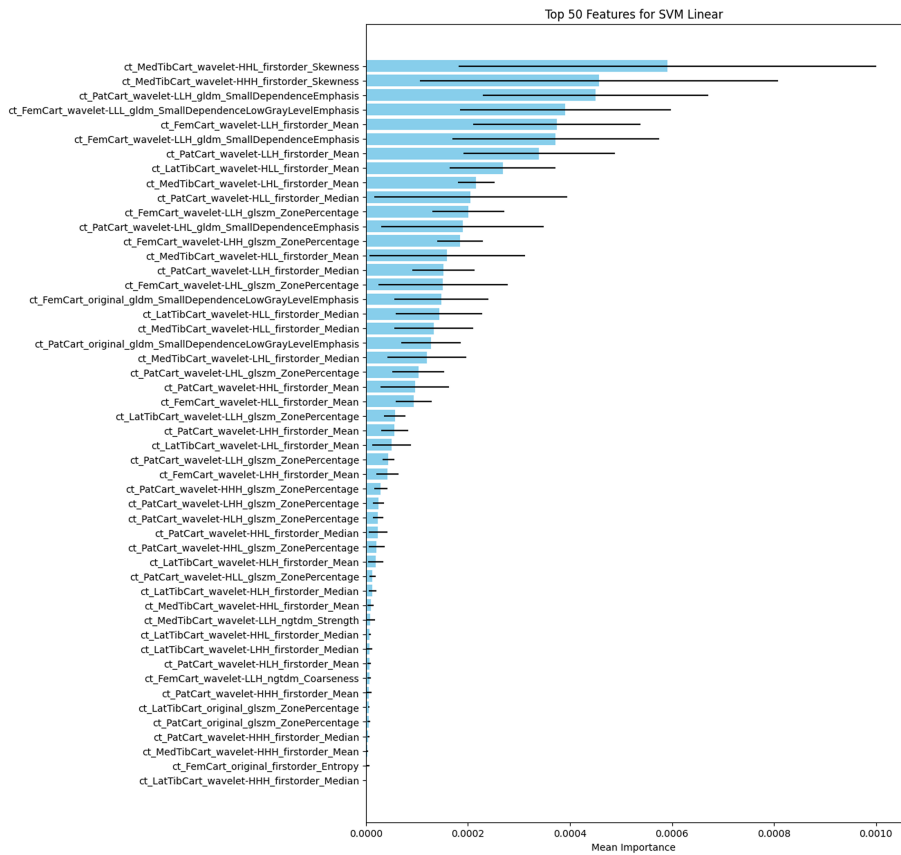


Figure 5.11: Barplot of the 50 most important robust features for the SVM linear model in CT.

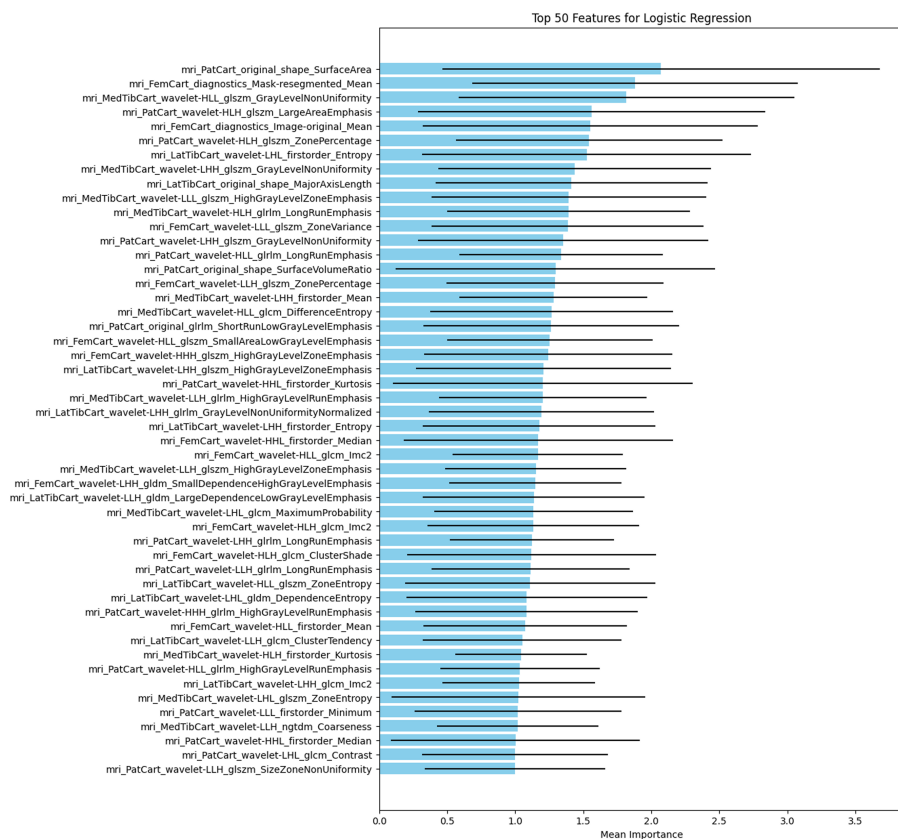


Figure 5.12: Barplot of the 50 most important robust features for the Logistic Regression model in MRI.

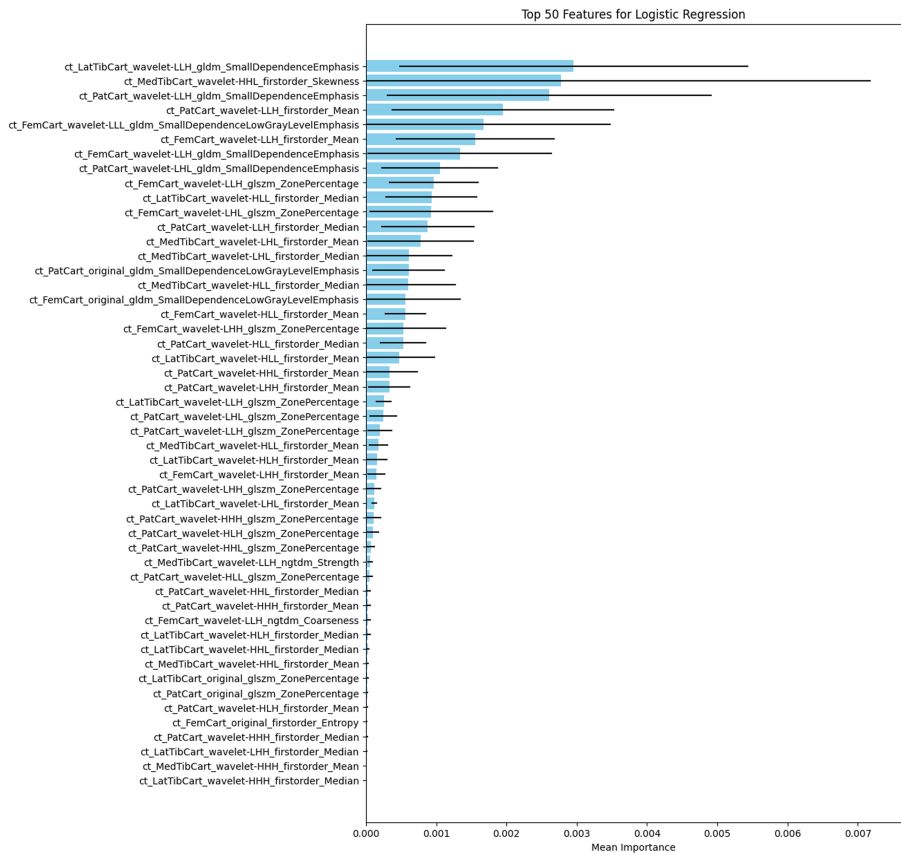


Figure 5.13: Barplot of the 50 most important robust features for the Logistic Regression model in CT.

Table 5.8: Compared Performances among models in CT radiomics data with 50 most important robust features

Model	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	74.95 (± 6.5)	73.76 (± 6.8)	76.05 (± 7.2)	74.95 (± 6.2)	74.80 (± 7.4)
SVM Linear	88.02 (± 3.38)	88.28 (± 4.27)	88.21 (± 3.9)	88.02 (± 4.3)	87.98 (± 4.2)
Logistic Regression	87.34 (± 5.6)	87.59 (± 5.3)	87.48 (± 4.6)	87.33 (± 4.3)	86.30 (± 4.1)
Gradient Boosting	74.14 (± 5.7)	74.21 (± 6.9)	76.42 (± 5.3)	75.14 (± 6.2)	78.99 (± 7.1)

observers (either within the same observer over time - intra-observer, or between different observers - inter-observer) may interpret and delineate structures in medical images in different ways. Assessing robustness against such variability is critical to ensuring the reliability of radiometric features in clinical applications. This was achieved by evaluating how changes in segmentation, induced by morphological operations such as dilation and erosion, affect the stability of radiomic features. Robustness scores were calculated using statistical tests and integrated into the feature importance calculation for machine learning models. The findings reveal that first-order and GLCM features are the most robust, whereas shape features are highly sensitive to segmentation changes. Ideally, robustness is indicated by a p-value greater than 0.5 across all morphological changes (from 1 to 5 pixels). Figure 5.4 and Figure 5.5 illustrate this trend for GLCM and first-order features under both erosion and dilation on MRI. Figure 5.6 and Figure 5.7 show the same positive trend for first-order, NGTDM and GLDM features on CT. Other features exhibit a different trend, with p-values decreasing to 0 as the dilation/erosion pixels increase.

More specifically, all feature classes show a downward trend, indicating a decrease in robustness scores as the degree of morphological operation increases. This is expected since more extensive morphological alterations can lead to greater changes in the features extracted from the altered ROIs. Initial robustness scores to 1-pixel change provide information on the intrinsic stability of each feature class to slight morphological changes.

Based on the trend plots and the previous discussed boxplots, the first-order and GLCM feature classes appear to be the most robust in the analysis of dilation and erosion. This would indicate that these classes maintain their feature values more consistently despite morphological changes applied to the images. First-order features, which describe the distribution of individual pixel values without considering spatial relationships (such as energy, entropy, or average pixel intensity), may be inherently less sensitive to minor changes in ROI delineations because they capture the global statistics of the ROI rather than local models. This may explain why they appear to remain more stable across different levels of morphological operations. GLCM features, on the other hand, are based on the frequency of pixel values occurring at a certain distance and orientation within the image, which describes the texture, this suggests that the structural information they encapsulate is preserved even when the ROI is modified. This could mean that the texture patterns are sufficiently distinct and not easily disrupted by ROI dilation. In particular, a strong difference can be observed between the robustness scores of

dilation and erosion; in fact, features remain more robust for dilation. In general, this is easily explained by the fact that, compared to dilation, in erosion, information is removed by reducing the region of interest. In particular, with very high erosion (3 to 5 pixels), as shown in Figure 5.3, the features change significantly, affecting the results of the analysis. These features are likely to provide reliable and consistent information for the diagnosis of conditions such as osteoarthritis in CT and MRI scans, despite variations in image segmentation. From the boxplots in Figures 5.4, 5.5, 5.6, and 5.7, it appears that first-order features are generally the most robust against segmentation variations. This robustness suggests that pixel intensity values within the ROI are reliable despite small changes in segmentation, while texture features are more sensitive to changes in segmentation. Interestingly, some texture features, such as GLCM and NGDTM, hold up better than others, such as GLRLM and GLSZM, when these small changes occur. In contrast, shape features are much more sensitive to segmentation, meaning that inaccurate segmentation could significantly affect the results. As expansion and erosion increase from 1 to 5 pixels, the variance of robustness of almost all features increases. This indicates higher sensitivity and potentially lower reliability under conditions of more significant variation in ROIs. In conclusion, the robustness analysis applied in our study provides further validation of the reliability and consistency of our results. The generalization of the results gives room for new perspectives in diagnosis.

5.3 Feature Selection and Limited Data: Density-Morphology Dataset

After assessing the robustness of the features, we will next examine how dimensionality reduction and feature selection affect model performance when the training data is limited.

In this analysis, we will train and compare linear SVM and LR classifiers using both the original Density-Morphology Dataset (Table 2.3) and a reduced feature set. The goal is to determine whether a more compact and clinically meaningful subset of features can achieve diagnostic accuracy comparable to that of the full feature set. This consideration is particularly relevant for medical imaging studies, where the available data is often limited.

This section has been adapted from the study entitled “Multimetric Evaluation of Knee Cartilage Degeneration” [4], presented at the *IEEE MetroXRAINE Conference, 2024, St. Albans, London, UK*.

5.3.1 Feature Selection with Correlation and Statistical Analysis

Feature Selection (FS) aims to obtain a smaller clinically relevant feature set. To do so, we calculate p-values and perform correlation analysis on the features, and they yield a reduced feature set.

Correlation analysis determines the strength and direction of the relationship between two variables. The correlation coefficient ranges from -1 to 1; a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative

correlation, and a value of 0 indicates no correlation. The correlation coefficients among the analyzed features are used to exclude from the feature set the feature whose information capacity has already been expressed by other features.

The correlation measures the relation of two variables, and, indeed, our goal in this step is to remove related features that do not provide more information when evaluating the belonging class in addition to the ones already considered.

After reducing the features, permutation analysis (see 3.1.1.1) was applied again on the new subset.

5.3.2 Machine learning

ML analysis allows features to be ranked according to their importance in the classification. The selected algorithms are SVM and LR.

The classifiers are sensitive to scaling. Thus, the data are standardized through z-scaling to preserve relative spatial distribution, a key element for SVM.

The metric used to evaluate the classification quality for both algorithms is the F1 weighted score. The evaluation is performed in a stratified cross-validation using 5 folds.

For feature importance, the algorithms' coefficients are extracted, and, more specifically, they represent the weights assigned to each input variable for linear SVM and the impact of each feature on the probability estimation for LR. Larger absolute values indicate a stronger influence on the prediction made by the model. By sorting these coefficients in descending order of their absolute values, it is possible to identify which features have the most significant impact on the predicted outcomes.

5.3.3 Results & Discussion

5.3.3.1 Reduced set of data

Building on previous considerations, we have decided to reduce the feature set by removing features with minimal impact on the KOA's future diagnosis and by creating new features from existing values.

From the correlation analysis (Figure 5.14), both tibialis cartilage volumes and surfaces showed a correlation coefficient of 0.87. Although the other cartilages do not exhibit high correlation coefficients, they have been combined for consistency. A new feature, volume-to-surface ratio (Vol/Surf), has been introduced to replace the use of both volume and surface area separately. This is not a proper feature selection; however, reducing dimensionality (even by merging features) can improve further analysis. Additionally, this measure expresses the compactness and is proportional to the exposed surface. It can be interpreted as the degeneration status of the deteriorated cartilage. Indeed, as described by Belluzzi *et al.* [178], the changes start from the superficial zone and vertical clefts are created, thus changing the volume-surface ratios. Based on the study by Otterness *et al.* [179], there is a correlation between the areas and volumes of knee cartilage in healthy subjects, with a specific constant value describing this relationship for different types of

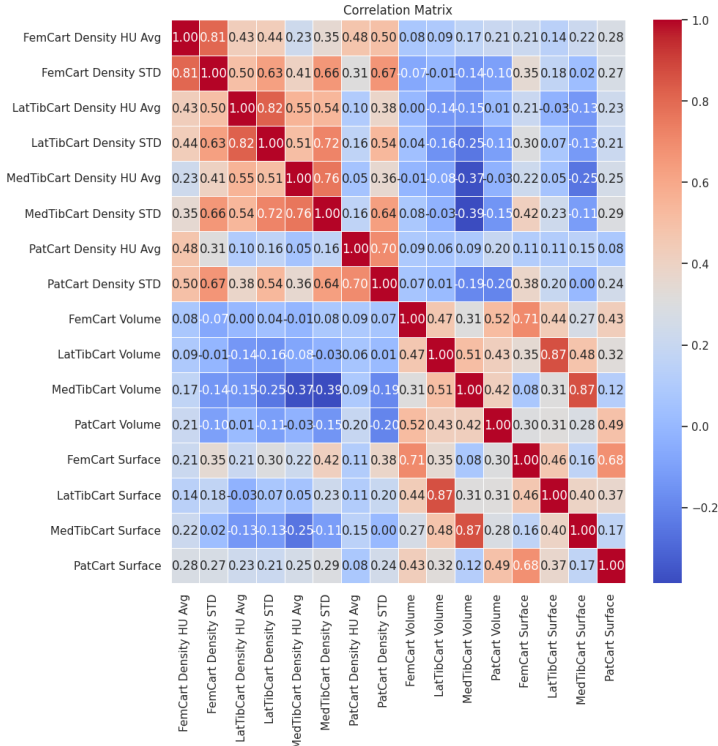


Figure 5.14: Correlation Matrix

cartilage. However, in our study, which includes both healthy subjects and patients with degenerative conditions, the values for these two groups differ. In patients with degenerative conditions, the relationship between cartilage area and volume may also stabilize to a different constant value after degeneration. This variation can be used to classify the patients, and our results show that this classification is significant, as demonstrated in Table 5.10.

Another interesting correlation result is in cartilage densities: the mean and standard deviation are significantly correlated at more than 0.7 for each cartilage. This leads to the possibility of removing the density mean or density StDev from the feature set. Since the correlation analysis doesn't provide more information about which feature to choose, we checked their p-values from the statistical data test previously showed in Chapter 3 (Table 3.1). We have decided to exclude the means from the feature set because the StDevs p-values are lower than the mean ones. The decision is substantiated by the understanding, as previously described, that the StDev of cartilage density reflects the degree of density variation within the cartilage mask, which can offer more informative insights for degeneration classification.

Therefore, the reduced set of features resulted to be composed as follows:

- FemCart Density StDev
- LatTibCart Density StDev
- MedTibCart Density StDev
- PatCart Density StDev
- FemCart Vol/Surf
- LatTibCart Vol/Surf
- MedTibCart Vol/Surf
- PatCart Vol/Surf

A permutation test was applied to the new subset of data, and it is shown in Table 5.9. It is possible to notice that all the features are highly significant, with almost all the features have a p-value < 0.001 .

Table 5.9: Permutation analysis applied after feature selection; p-values approximated by 3 significant digits

Feature	p-value
FemCart StDev Density	0.000
PatCart StDev Density	0.000
MedTibCart StDev Density	0.000
LatTibCart Avg Density	0.000
FemCart Vol/Surf	0.006
LatTibCart Vol/Surf	0.000
MedTibCart Vol/Surf	0.000
PatCart Vol/Surf	0.000

When compared with the broader analysis presented in Table 3.1, the permutation results demonstrate that the dimensionality reduction did not affect statistical sensitivity. In fact, by focusing on density variability (StDev) and compactness (Vol/Surf), the reduced set emphasizes the most biologically meaningful descriptors of cartilage degeneration. A particularly noteworthy outcome is that some features that were previously non-significant (such as those derived from the lateral tibial cartilage or the femoral cartilage) became significant after being transformed into the Vol/Surf ratio. This suggests that combining correlated geometric variables can uncover underlying relationships and enhance their discriminative ability. Therefore, the dimensionality reduction process served not only as a simplification step but also as a means of feature enhancement.

These results reinforce the hypothesis that heterogeneity in cartilage density is a key indicator of early degenerative changes. Similarly, the Vol/Surf ratios provide an interpretable geometric measure of cartilage thinning and surface irregularity, which collectively capture both compositional and morphological degeneration

aspects. Overall, the reduced feature set preserves the diagnostic power of the original dataset while enhancing parsimony and interpretability, thereby supporting its use in the following machine learning analyses.

5.3.3.2 Machine learning

Both ML algorithms were trained on both the full dataset and the reduced one, and their performances were compared on a 5-fold cross-validation.

Table 5.10: ML results for SVM and Logistic Regression. Average accuracy and StDev (in the brackets) among the folds are reported.

	All features	Reduced features
Linear SVM	0.81 (0.09)	0.81 (0.03)
Logistic Regression	0.78 (0.09)	0.78 (0.11)

Table 5.10 reveals minimal performance change between using the full set of features and the reduced set. This indicates that our feature selection approach aligns well with the objectives of our study, as reducing features does not significantly impact model performance. This capability is particularly valuable for diagnostic applications, enabling us to extract fewer features while maintaining high accuracy.

The SVM evaluates features’ coefficients during its optimization problem to find the best hyperplane dividing the classes. Indeed, each coefficient corresponds to a feature and represents its relevance in determining the decision boundaries. They are real numbers and factors that multiply the corresponding feature value of the sample. Depending on their absolute value and scale, they can give more *importance* to a feature than another one. Thanks to scaling, which reports the values in comparable scales, the coefficients can be compared with each other, ranking the features according to the ML model (Table 5.11). The same considerations can be applied to LR, even though its performance is slightly less, with its feature ranking in Table 5.12.

Table 5.11: Coefficients of SVM per feature rounded to three decimal places

Feature	coefficients
LatTibCart Volume	-1.488
MedTibCart Volume	-1.389
LatTibCart Surface	1.074
LatTibCart Density StDev	0.738
FemCart Volume	0.656
FemCart Surface	0.574
MedTibCart Surface	0.530
LatTibCart Density HU Avg	-0.504
PatCart Volume	-0.450
PatCart Density HU Avg	-0.432
PatCart Density StDev	0.345
FemCart Density HU Avg	-0.275
FemCart Density StDev	0.263
MedTibCart Density HU Avg	-0.167
PatCart Surface	-0.143
MedTibCart Density StDev	0.0523

Table 5.12: Coefficients of Logistic Regression per feature rounded to three decimal places

Feature	coefficients
MedTibCart Volume	-1.445
LatTibCart Volume	-1.183
LatTibCart Surface	0.973
PatCart Volume	-0.851
FemCart Surface	0.837
FemCart Volume	0.835
LatTibCart Density StDev	0.640
LatTibCart Density HU Avg	-0.572
FemCart Density StDev	0.558
MedTibCart Density StDev	0.447
PatCart Density StDev	0.434
FemCart Density HU Avg	-0.420
MedTibCart Density HU Avg	-0.412
PatCart Density HU Avg	-0.362
MedTibCart Surface	0.314
PatCart Surface	-0.208

Table 5.13: Coefficients of SVM per feature for reduced dataset

Feature	coefficients
MedTibCart Vol/Surf	-1.095
FemCart Vol/Surf	0.879
PatCart Vol/Surf	-0.688
LatTibCart Vol/Surf	-0.683
FemCart Density StDev	0.183
LatTibCart Density StDev	0.171
MedTibCart Density StDev	0.128
PatCart Density StDev	-0.010

Table 5.14: Coefficients of Logistic Regression per feature for reduced dataset

Feature	coefficients
MedTibCart Vol/Surf	-1.336
FemCart Vol/Surf	1.073
PatCart Vol/Surf	-0.823
LatTibCart Vol/Surf	-0.782
FemCart Density StDev	0.482
PatCart Density StDev	0.198
LatTibCart Density StDev	0.126
MedTibCart Density StDev	0.080

In the reduced dataset, as shown in Tables 5.13 and 5.14, the ratio volume/surface was the most predictive one for all four considered cartilages. In particular, medial tibial cartilage was the most important. It is also possible to notice that the volume of medial tibial cartilage is the first or the second most predictive in both the algorithms when considering the full set of data (Table 5.11, Table 5.12). Looking at Table 3.1, it is possible also to notice that all the features related to medial tibial cartilage are significant, and in particular, StDev and volume are the ones with the higher p-values in all the datasets. This enhances the importance of looking at the medial compartment for diagnosis purposes.

As enhanced by Everhart *et al.* [180], medial compartment defects progress faster than lateral defects in older adults with minimal to moderate KOA. Stoddart *et al.* [181] also suggested that medial tibiofemoral disease is more common and almost always significant in the development of the disease.

Looking at Table 5.13 and Table 5.14, FemCart Vol/Surf is the second most important feature for the classification in both tested algorithms. The algorithms evaluate dimensionless coefficients to balance the classification, reflecting the importance relative to the scale of each feature in the specific context of the dataset. It is critical to understand that the importance derived from ML models is not an absolute indicator of the importance of a feature, but rather reflects the optimization process of the model. Therefore, the observed importance of FemCart Vol/Surf in

the models highlights its role in this analysis and emphasizes the need to interpret ML results within the context of the dataset and algorithm principles.

The comparative analysis demonstrated that reducing the feature set did not compromise model performance, confirming that the selected subset retained the most relevant diagnostic information. Both SVM and LR achieved similar accuracies on the reduced dataset, highlighting the robustness of the chosen features. The analysis highlighted the importance of the standard deviation of cartilage density and the medial tibial cartilage as critical indicators of degeneration. The volume-to-surface ratio of cartilage provided valuable insights into morphological changes and disease progression.

5.4 Ensemble Feature Selection Method in SINPAIN Data

In the following section, we introduce a new feature selection strategy that employs an innovative ensemble method combined with machine learning techniques. The analysis focuses on the density-morphology dataset, on the CT and MRI radiomic datasets and on muscle features.

This section has been adapted from the study entitled “Feature Selection in Healthcare Datasets: Towards a Generalizable Solution” [6], published in *Computers in Biology and Medicine*.

5.4.1 Materials and Methods

A novel ensemble FS approach was applied to the SINPAIN datasets to evaluate model robustness and generalizability under conditions of feature redundancy and limited data availability.

5.4.1.1 SINPAIN Dataset

The SINPAIN data used in this study include 102 legs (61 C and 41 D) and four distinct feature categories:

- Cartilages Density Morphology Dataset (here defined as ‘Cartilages’)
- Muscle Dataset (here defined as ‘Muscles’)
- Cartilage MRI Radiomics (here defined as ‘Radiomics MRI’)
- Cartilage CT Radiomics (here defined as ‘Radiomics CT’)

Features description can be found in Table 2.3.

5.4.1.2 Feature Selection Ensemble Method

The proposed ensemble FS pipeline, whose full architecture is shown in Figure 5.15, consists of two main stages: a feature selection step and a feature merging step,

producing the final reduced feature set used for classification. The detailed implementation and algorithmic design are described in [6], and are summarized below.

The dataset was split into training (80%) and test (20%) subsets using a stratified sampling strategy to maintain class balance. All features were standardized using Z-score normalization (subsection 2.7.2) to enhance model convergence and comparability across features.

The first stage of the FS pipeline, inspired by a “waterfall” model (Figure 5.16), performs a preliminary exploration based on Gradient Boosting (GB) feature importance to identify the most informative variables. The second stage applies a refinement procedure, reducing redundancy and retaining features that maximize classification performance across multiple random seeds. This ensemble strategy mitigates bias due to random initialization and allows assessment of feature stability over repeated runs.

The resulting feature subsets from all runs are aggregated and merged following three ranking criteria—repetition, performance stability, and weighted importance—as illustrated in Figure 5.17. The intersection of these ranked subsets forms the final reduced feature set provided as input to the classification models.

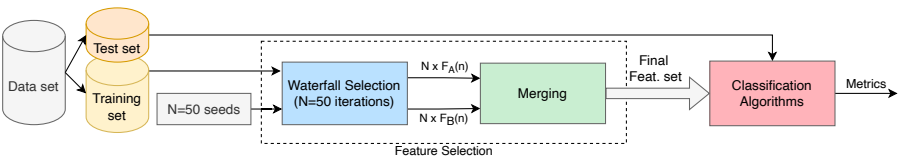


Figure 5.15: System’s General Architecture. The pipeline includes preprocessing, feature selection, merging, and classification stages.

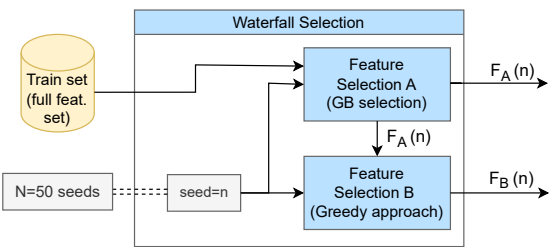


Figure 5.16: Waterfall Selection section of the algorithm, performing exploration and refinement of features.

Two interpretable classification models—Random Forest (RF) and Support Vector Machine (SVM)—were trained on both the full and reduced datasets to evaluate the FS method’s effectiveness (Figure 5.18). Both models were validated in 5-fold cross-validation and tested on unseen data. Evaluation metrics included

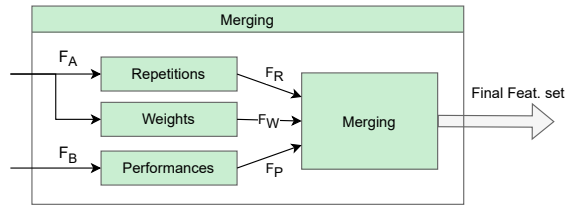


Figure 5.17: Feature merging strategy combining repetition, performance, and weighted criteria.

accuracy, precision, recall, and F1-score, with the latter prioritized for its balance between false positives and false negatives, particularly relevant in clinical settings.

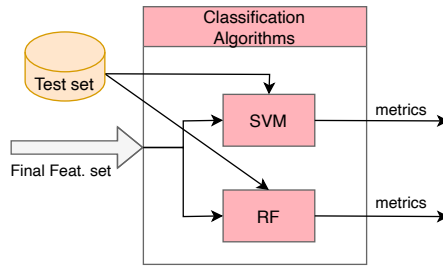


Figure 5.18: Classification models used for evaluating the feature selection framework.

5.4.2 Results

SVM and RF models are trained in CV on the original datasets and the reduced datasets obtained from the proposed FS process. In the SINPAIN dataset, the features are 1082: the most present feature group is the radiomics, mainly from MRI images, completely overshadowing the other categories composed of only 66 and 16 features.

The performances reported in Table 5.16 are not directly derived from the FS approach. "Before feature selection" refers to the model's performance when trained on the original dataset, which includes all the features. "After feature selection" refers to the model's performance when trained on the reduced dataset, which contains only the features chosen by the FS process. The performances of the two models are F1 scores on a 5-fold CV, and we consider them as mean and StDev. These results can be analyzed alongside the test set performances on unseen data.

The pipeline has been fed by a total of 102 samples, 41 belonging to the positive class and 61 to the negative one. The positive class refers to people diagnosed with KOA. One positive sample is missing in the radiomics CT dataset, leading to the training of 60 samples from the healthy class and 41 degenerate in the complete

Table 5.15: Feature number reduction in SINPAIN datasets

Algorithm execution	Starting features	Selected features
SINPAIN cartilages	16	8 (50%)
SINPAIN muscles	66	32 (48%)
SINPAIN radiomics MRI	581	237 (41%)
SINPAIN radiomics CT	420	150 (36%)
SINPAIN complete	1082	341 (32%)

dataset. However, the ratio between classes is kept, and this difference can be neglected.

The performances are evaluated in CV (see Table 5.16) and on the test set (see Table 5.17).

Table 5.16: CV (k=5) f1-score, before and after feature selection with the SINPAIN dataset.

Feature set	RF		SVM	
	Full set	Final set	Full set	Final set
Cartilages	0.67 ± 0.13	0.73 ± 0.08	0.66 ± 0.12	0.80 ± 0.07
Muscles	0.62 ± 0.15	0.65 ± 0.12	0.58 ± 0.10	0.67 ± 0.11
Radiomics MRI	0.75 ± 0.19	0.83 ± 0.11	0.85 ± 0.08	0.92 ± 0.06
Radiomics CT	0.75 ± 0.16	0.81 ± 0.10	0.75 ± 0.17	0.86 ± 0.07
Complete	0.79 ± 0.14	0.82 ± 0.11	0.86 ± 0.12	0.94 ± 0.07

Considering the complete dataset, the final composition of the dataset is 52.2% from radiological MRI features, 38.4% CT features, and 7.9% from muscle features, 1.5% cartilage features. Corresponding to 178 MRI features, 131 CT, 27 muscle, and 5 cartilage ones.

Table 5.17: Test f1-score after feature selection with the SINPAIN dataset

Feature set	RF	SVM
Cartilages	0.83	0.81
Muscles	0.79	0.70
Radiomics MRI	0.95	0.95
Radiomics CT	0.88	0.90
Complete	0.82	0.94

5.4.3 Discussion

The objective of the work is to reduce the feature number for clinical evaluations.

FS is considered effective when it identifies the most informative features that provide the most relevant data for the task. In ML terms, this means selecting the smallest possible feature subset that still allows the model to achieve accurate classification. In other words, the goal is to retain only the essential features that contribute significantly to the model's performance while discarding any that do not add value or could even introduce noise. AI-driven approaches can also focus on important characteristics that physicians would otherwise neglect because of the high amount of information that can be misleading for diagnosis. In the SINPAIN datasets, the amount of information that can be extracted from an image sometimes complicates the creation of a metric system for the clinical assessment of cartilage. After FS, AI algorithms' performance may worsen. Still, it is neglectable with respect to the reduction amount, which is at least 43% in our tests, as shown in Table 5.15. However, this possibility never arises in this work because, as Table 5.16 shows, the performances never worsen after FS; they increase or remain unchanged. Even with the least effective network, the KOA classification using muscles is close to random when using the SVM model. However, after FS, the performance improves from 58% to 67%, while maintaining the same StDev. The strong results in CV may indicate overfitting, even though the goal of reducing features is to mitigate its risk. So, we can look at the test set performance, to obtain a more objective measure on unseen data.

It is interesting to notice that the test performances exceed those of the CV. This could be due to the smaller size of the dataset, which may lead to more variable results across folds (high standard deviation) and less reliable generalization.

Due to their intrinsic differences in model structure and learning mechanisms, RF and SVM can exhibit varying performance on the same data, with RF benefiting from its ensemble approach and ability to handle non-linearities. In contrast, SVM's performance is more sensitive to data distribution and the choice of kernel. These differences highlight that FS remains effective regardless of the chosen algorithm, as RF and SVM perform well on all SINPAIN datasets. Considering Table 5.17, the RF KOA classification with muscle features is more effective than SVM. A possible explanation is the fact that the data are not easily separable by the SVM algorithm because of their spatial distribution. This effect is also present when using the complete subset of features, as it includes muscle information.

Analyzing the performances and reduction, the dataset containing all the features is unbalanced with a ratio of 1082 features to 102 samples, over 10:1; after the FS, the ratio has decreased to 341 to 102, over 3:1. Also in this case, the performances have increased after the FS with an SVM peak performance on the test set of 94% in Table 5.16. The StDev decreased as well, having more accurate ML models. Again, in terms of biosignals percentage, the final set composition complies with the original one. Although, the final set F of the complete dataset, is composed of fewer features from each feature category when compared with the homogeneous FS. For instance, the MRI features extracted when focusing on them are 237, whilst the number of MRI features extracted using the complete one are

only 178. This may be because of emerging properties when in the presence of a composition of different features, leading to the need for less data to represent the same information.

The dataset allows us to test the algorithm's capability to extract the best feature set using a limited number of features. The starting number of cartilage features (density and morphology) is only 16, and the algorithm selects half of it. Not only does it reduce the number of features, but the selected ones significantly improve the performance of RF and SVM in CV, both in the considered metric and StDev. It is unlikely that this behavior is due to overfitting, as the test set's performance confirms.

In section 3.2, FS was performed by extracting the top 50 features from different models. Considering KOA classification on MRI, the reported F1 score in RF in 10-fold CV is $87.90\% \pm 6.10$ and with SVM $95.54\% \pm 4.35$. This higher score may be due to the tuning of hyperparameters, which falls outside the scope of this work.

In section 3.1 we extracted 8 cartilage features from the original 16 and obtained $0.81\% \pm 0.03$ performance metric with SVM and linear kernel. Our approach also reduces the number of features to 8, reaching $0.80\% \pm 0.07$ performance metric. The comparable results show that the selected features contain enough information to be able to classify KOA comparably to a custom feature extraction method.

Overall, the ensemble feature selection approach applied to the SINPAIN dataset successfully reduced dimensionality while maintaining or improving model performance across all feature categories. The results demonstrate that the proposed FS strategy can effectively identify the most informative and stable features even in heterogeneous and limited datasets, supporting the robustness and generalizability of ML models for KOA assessment. The results of our feature selection and classification framework have practical value in clinical decision support, as they help reduce the complexity of data analysis while preserving diagnostic performance. This can assist clinicians and researchers in focusing on the most relevant features, leading to more efficient resource allocation, reduced diagnostic time, and potentially more interpretable ML models in real-world healthcare settings.

5.5 Summary and Achievements

The aim of this chapter was to validate the reliability and generalizability of the extracted imaging features and to optimize their selection for subsequent clinical and machine learning analyses. Each section addressed a specific objective: assessing the impact of segmentation variability, testing feature robustness under morphological perturbations, and evaluating model behavior with reduced set of data. Together, these experiments established a rigorous methodological foundation for ensuring that the proposed imaging biomarkers are reproducible, interpretable, and transferable across datasets and acquisition conditions.

The results confirmed that the developed features are stable to segmentation and preprocessing variations, maintain predictive consistency across analytical pipelines, and can be effectively reduced to compact, informative subsets without compromising classification performance. The optimization and validation strate-

gies presented here strengthen the credibility of the digital assessment framework introduced in previous chapters and support its application in subsequent analyses focused on early KOA identification and patient stratification.

Chapter 6

Assessing Early Stage KOA using Semi-Supervised Learning

While supervised learning has shown promise in classifying healthy and degenerated knees, the early stages of KOA remain particularly challenging to identify due to ambiguous imaging patterns and limited expert-labeled data. This chapter introduces a semi-supervised learning (SSL) that using both labeled and unlabeled multimodal imaging samples, propagates label information across data structures, enabling the discovery of subtle and intermediate disease stages.

The present chapter has been adapted from the paper “A Semi-Supervised Learning Approach for Early Detection of Knee Osteoarthritis Using Multimodal Image Data”, submitted on *IEEE Journal of Biomedical and Health Informatics*.

6.1 Objectives of this Study

Early detection of cartilage degeneration in KOA remains a critical yet unsolved clinical challenge. Precise labeling of degeneration stages from imaging data is limited by the lack of reliable ground truth and the subjective nature of clinical symptoms (which vary widely across individuals) and radiological assessments. Consequently, diagnostic datasets often contain imperfect labels that misrepresent the nuanced disease progression, leading to under- or over-estimation of clinically relevant degeneration.

Our study has two interconnected objectives:

1. To develop a reliable, quantitative surrogate for cartilage ground truth labels through a SSL pipeline, informed by domain knowledge and validated against clinically consistent test cases.
2. To assess imaging-based features derived from combined MRI and CT segmentation to characterize the degeneration stages using robust ML techniques.

Rather than relying on expensive and inconsistent manual annotations, we propose a SSL pipeline that categorizes cartilage status into three levels: healthy, early

degeneration, and late degeneration. Our work addresses the challenges of noisy and imperfect cartilage labels in KOA by employing a graph-based semi-supervised propagation strategy on multimodal data. This approach emphasizes scalability, reproducibility, and interpretability, which are essential for clinical applications. By mapping cartilage features extracted from MRI segmentation onto corresponding CT scans, we facilitate a quantitative analysis of cartilage density and morphology, effectively simulating expert assessments in a systematic and automated way. Ultimately, this SSL-driven stratification provides a clinically meaningful characterization of KOA severity, supporting early and tailored interventions to prevent irreversible cartilage loss. The integration of domain expertise with algorithmic inference ensures that our method can be adapted to other conditions affected by diagnostic uncertainty and label noise.

6.2 Materials and Methods

6.2.1 Cohort and Features Utilized

The initial diagnosis of D subjects was based primarily on the presence or absence of clinical symptoms. Subjects reporting joint pain were evaluated by a physician, who could initiate treatment protocols ranging from conservative therapy to surgical intervention in more advanced cases. However, it is important to acknowledge the subjective nature of pain perception: some individuals may not experience noticeable discomfort until degeneration is well advanced, while others may report significant pain even at early stages of cartilage deterioration. Consequently, although all knees labeled as D exhibit signs of degeneration, the severity and extent of degeneration likely vary across subjects within this group. Similarly, individuals in the control group, who were asymptomatic, may still exhibit early, subclinical degenerative changes that they are unaware of.

Following the results discussed in section 5.3, for this study, the Density-Morphology Dataset (Table 2.3) has been used, but combining together Vol and Surf parameters into a single feature.

Figure 6.1 gives an overview of the entire workflow for the labeling strategy described in the following sections.

6.2.1.1 Label Propagation and Label Spreading

Label Propagation (LP) and Label Spreading (LS) are graph-based SSL techniques particularly useful for small datasets with limited labeled data, as in our case. They rely on the assumption that the structure of the unlabeled data reflects the class structure, using a small set of labeled nodes (seeds) to propagate their labels to nearby unlabeled nodes via graph-encoded similarities. Both methods construct a similarity graph and use a propagation matrix to iteratively diffuse label information across the graph, with labeled nodes guiding the process. The two algorithms differ in how they modify the similarity matrix and apply clamping: LP uses hard clamping (keeping original labels fixed), whereas LS allows a degree of flexibility, by minimizing a loss function with regularization, making it more robust to noise

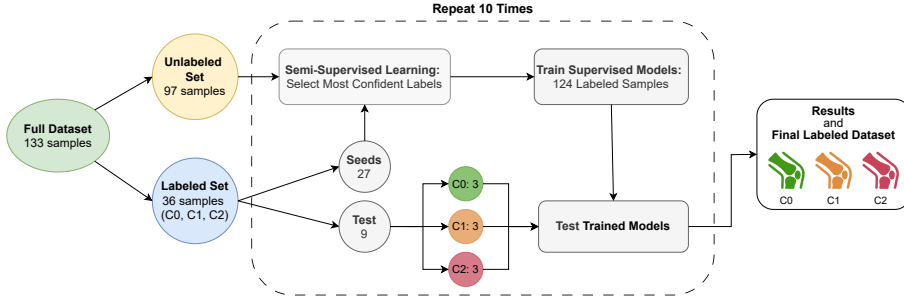


Figure 6.1: Workflow of the Labeling Strategy of Class 0 (C0, healthy), Class 1 (C1, early deg), Class 2 (C2, late deg)

[68]. `sklearn.semi_supervised` scikit learn python library was used to implement the two algorithms.

6.2.1.2 Specialists' evaluation: the labeled samples

To capture a broader spectrum of the disease (rather than the simple healthy vs diseased distinction), a team of specialists, composed of radiologists, orthopedic surgeons, and engineers, independently reviewed MRI and CT scans of the knee joints without any prior knowledge of the patient's medical history or symptoms. The evaluation was conducted according to the following criteria:

- Score 0 (Healthy knee joint): no abnormalities on either CT or MRI
- Score 1 (Early degeneration): signs of osteophytes and/or subchondral sclerosis on CT; signs of chondral irregularity and/or edema on MRI
- Score 2 (Advanced degeneration): large osteophytes, severe subchondral sclerosis and cysts on CT; almost or complete disappearance of joint cartilage and fluid in joint on MRI

In the end, only a subgroup of 36 scans was univocally labeled by all the evaluators:

- Class 0: 9 instances
- Class 1: 11 instances
- Class 2: 16 instances

6.2.1.3 Labeling Pipeline

Out of the 36 labeled samples, 9 (3 per class) are randomly excluded from the dataset to simulate unlabeled data, which will be used as a test set to validate the labeling protocol. The remaining 27 labeled instances are distributed as follows:

- Class 0: 6 instances

- Class 1: 8 instances
- Class 2: 13 instances

Both LP and LS were applied independently to the full set of 124 subjects, using the 27 labeled ones as seeds. The algorithms used the RBF kernel (6.1), with $\gamma = 10$. This value was chosen after testing γ from 0.5 to 20 and evaluating label distributions by average confidence and entropy. As γ increased, the label distributions became sharper: confidence increased while entropy decreased, demonstrating that the models became more certain as nearby labeled points had a stronger and more localized influence. Around $\gamma = 10$, both models reached stable labeling, with minimal changes at higher values. Thus, $\gamma = 10$ was selected as a balance, providing strong, stable, and confident labeling while retaining a neighborhood influence broad enough to capture patterns beyond very local structures.

$$RBF = K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (6.1)$$

K is the kernel function and it measures the similarity between points x and y . It returns a value between 0 and 1. When x and y are very close, $\|x - y\|^2$ is small, so K is close to 1 (high similarity). When x and y are far apart, $\|x - y\|^2$ is large, so K is close to 0 (low similarity). Higher γ means the kernel focuses more on points closer to each other (more localized influence), while lower γ means the influence spreads more broadly.

The label distributions produced by LS and LP models were combined by summing their respective class probabilities. The final predicted label (pseudo-label) for each sample was then determined by selecting the class with the highest combined probability (6.2).

$$\hat{y}_i = \arg \max_{c \in \{0,1,2\}} (p_{i,c}^{(LS)} + p_{i,c}^{(LP)}) \quad (6.2)$$

where $p_{i,c}^{(LS)}$ and $p_{i,c}^{(LP)}$ represent the probabilities assigned to class c for sample i by the LS and LP models, respectively.

6.2.1.4 Supervised Learning and Evaluation Metrics

Supervised learning was applied to the new labeled dataset (healthy vs early degenerate vs late degenerate). Models were trained on a combined set of 124 samples (27 expert-labeled + 97 pseudo-labeled) and evaluated on the 9 held-out expert-labeled samples. To ensure robustness and mitigate variability from relying on a small set of expert labels (36 in total), the entire pipeline (from test split to label propagation and supervised training) is repeated 10 times with different random selections of 9 test subjects. This approach is a form of Monte Carlo cross-validation [182] (also known as repeated random subsampling), where multiple randomized train/test splits are used to better estimate generalization performance. By aggregating results across these 10 repetitions, the evaluation captures both central tendency and variability, yielding a more reliable assessment of the model's predictive performance. We assessed the following supervised learning algorithms: SVM, LR, RF, GB, KNN. All

models were tuned using grid search with 5-fold cross-validation on the training data. When required (e.g., for SVM and KNN), features were scaled using Standard-Scaler to ensure consistent performance. The best-performing hyperparameters were selected based on average cross-validation scores. The hyperparameter grids used for tuning are shown in Table 6.1. Performance metrics included overall accuracy, weighted precision, recall, and F1 score, as well as per-class scores to evaluate class-wise performance.

Table 6.1: Hyperparameter grids used for cross-validated grid search.

Classifier	Hyperparameter Grid
SVM	$C \in \{0.1, 1, 10, 100\}$, kernel $\in \{\text{linear}, \text{rbf}\}$
Logistic Regression	$C \in \{0.01, 0.1, 1, 10, 100\}$
Random Forest	$n_estimators \in \{10, 50, 100\}$, $max_depth \in \{3, 5, 10, \text{None}\}$, $min_samples_split \in \{2, 5, 10\}$
Gradient Boosting	$n_estimators \in \{10, 50, 100\}$, $learning_rate \in \{0.01, 0.1, 1\}$, $max_depth \in \{3, 5, 10\}$
KNN	$k \in \{3, 5, 7, 9\}$, $weights \in \{\text{uniform}, \text{distance}\}$

6.2.2 Statistical Analysis

6.2.2.1 Label Stability

To assess label stability across the ten different independent runs, we computed some metrics:

1. Per-sample agreement: For each unlabeled sample (the set of 97 data points), it checks how often it got the same label across the 10 runs.
2. Fleiss' kappa: widely utilized in the medical field, it is a statistical measure used to evaluate the reliability of agreement among multiple raters (more than two) that classify items into categorical outcomes. It quantifies the extent of agreement among several raters beyond what would be expected by chance [183].
3. Entropy: it measures the uncertainty of each sample across the different runs; the lower the entropy (the lower the uncertainty), the more stable the label.

6.2.2.2 Group Comparisons

As a first evaluation of the newly created SSL labels, inferential statistical tests were employed to evaluate the differences among the three categories: healthy, early degeneration, and late degeneration. Although the dataset was labeled using machine learning methods, traditional inferential statistics were essential for validating group-level differences. One-way ANOVA [184] and the Kruskal-Wallis test [185] were used to assess group differences in normally and non-normally distributed variables, respectively. P-values obtained from these tests provided a quantitative measure of statistical significance on the extracted features. To further explore specific intergroup differences, post hoc analyses using Tukey's HSD

[186] and Dunn's [187] test were conducted, controlling for multiple comparisons. This hybrid approach of machine learning and statistical testing strengthens the reliability of the findings and supports the clinical relevance of the categorized outcomes.

6.3 Results

6.3.1 Three-class Classification Performances

Table 6.2: Comparison of 3 classes algorithm's metrics with mean $\pm$ standard deviation on ten runs, sorted by F1 Weighted score

Model	Accuracy	Precision Weighted	Recall Weighted	F1 Weighted
SVM	0.84 ± 0.11	0.88 ± 0.08	0.84 ± 0.11	0.84 ± 0.11
LR	0.82 ± 0.12	0.87 ± 0.10	0.82 ± 0.12	0.81 ± 0.13
KNN	0.80 ± 0.10	0.83 ± 0.11	0.80 ± 0.10	0.79 ± 0.11
RF	0.79 ± 0.11	0.80 ± 0.15	0.79 ± 0.11	0.77 ± 0.14
GB	0.76 ± 0.17	0.76 ± 0.20	0.76 ± 0.17	0.74 ± 0.18

In each independent run, 124 samples were used for training the classification algorithms, and performances were computed on the test set composed of the 9 held-out samples. The results were combined from the ten runs. Table 6.2 reports the mean and standard deviation of the ten runs for each metric and model on the three-class problem. Table 6.3, Table 6.4, and Table 6.5 show the corresponding results for classes 0 (healthy), 1 (early degeneration), and 2 (severe degeneration).

The highest overall performance was obtained by SVM, with an F1-Weighted of 0.84 ± 0.11 , followed by LR at 0.81 ± 0.13 . The lowest-performing model was GB, with 0.74 ± 0.18 . Accuracy values followed a similar trend, ranging from 0.76 (GB) to 0.84 (SVM). Weighted precision reached its highest values with SVM (0.88 ± 0.08) and LR (0.87 ± 0.10), while recall-weighted was also led by SVM (0.84 ± 0.11) and LR (0.82 ± 0.12).

At the class level, class 0 consistently reached the highest scores across models, with KNN achieving an F1 of 0.97 and SVM 0.95. Class 1 results were lower overall but remained stable, with LR performing best ($F1 = 0.83$) followed by SVM (0.78). For class 2, SVM obtained the highest F1 (0.79), while LR showed a strong precision value (0.94) accompanied by lower recall (0.67).

Taken together, these results show that all algorithms reached acceptable performance, with class 0 producing the most consistent metrics and higher variability observed in classes 1 and 2.

6.3.2 Label Stability

Agreement per-sample across the ten runs was calculated; the average and standard deviation among the runs are 0.911 ± 0.136 , indicating that, on average, 91% of the runs agree on each sample's label, with a slight variability. Fleiss' Kappa is 0.781, indicating strong agreement between runs beyond chance. Finally, the mean

Table 6.3: Performance metrics for Class 0 across models, reported as mean $\pm$ standard deviation

Class 0			
Model	Precision	Recall	F1-score
SVM	0.95 ± 0.11	0.97 ± 0.11	0.95 ± 0.08
LR	0.88 ± 0.15	0.87 ± 0.17	0.86 ± 0.13
KNN	0.95 ± 0.11	1.00 ± 0.00	0.97 ± 0.06
RF	0.88 ± 0.16	0.93 ± 0.14	0.89 ± 0.12
GB	0.85 ± 0.21	0.93 ± 0.14	0.89 ± 0.18

Table 6.4: Performance metrics for Class 1 across models, reported as mean $\pm$ standard deviation

Class 1			
Model	Precision	Recall	F1-score
SVM	0.89 ± 0.17	0.73 ± 0.21	0.78 ± 0.15
LR	0.78 ± 0.17	0.93 ± 0.14	0.83 ± 0.12
KNN	0.82 ± 0.20	0.70 ± 0.25	0.71 ± 0.16
RF	0.76 ± 0.14	0.80 ± 0.23	0.75 ± 0.12
GB	0.75 ± 0.18	0.73 ± 0.21	0.73 ± 0.16

Table 6.5: Performance metrics for Class 2 across models, reported as mean $\pm$ standard deviation

Class 2			
Model	Precision	Recall	F1-score
SVM	0.81 ± 0.18	0.83 ± 0.24	0.79 ± 0.16
LR	0.94 ± 0.12	0.67 ± 0.27	0.74 ± 0.19
KNN	0.73 ± 0.20	0.70 ± 0.29	0.68 ± 0.20
RF	0.77 ± 0.32	0.63 ± 0.29	0.67 ± 0.26
GB	0.68 ± 0.31	0.60 ± 0.34	0.61 ± 0.30

entropy across runs is 0.283 with a standard deviation of 0.385; this indicates that most samples have low uncertainty among the different runs. These outcomes demonstrate that the SSL pipeline provided reliable and consistent labels across runs, with high agreement and low entropy indicating stability across different seed initializations. Only a small subset of samples showed higher variability, suggesting that the risk of label noise in the final dataset remained limited.

6.3.3 Re-labeled Dataset: Statistical Analysis and Distributions

The unlabeled samples were eventually assigned labels based on the majority agreement from multiple runs. The complete dataset (which includes the originally

labeled 36 samples along with 97 newly machine-labeled samples) consists of 42 samples in class 0, 56 samples in class 1, and 35 samples in class 2. This final dataset was analyzed to determine statistically significant differences among the three classes on each considered features. ANOVA/Kruskal–Wallis tests (parametric/non-parametric) showed significant p-values for each evaluated feature; therefore, a post hoc analysis was conducted. For Tukey’s test, p-values adjusted with Bonferroni correction were considered. Results are reported in Table 6.6 and show that the considered features discriminated between healthy, early degeneration, and late degeneration, with the strongest separation observed for the density standard deviation of both tibial cartilages and the volume-over-surface features. Figure 6.2 illustrates these differences, displaying the distributions of all features across the three classes and highlighting clear progressive trends. For volume-to-surface ratio features, a decreasing trajectory is observed from healthy to severe degeneration. Tibial cartilages in particular displayed highly significant reductions ($p < 0.001$).

Density-related features exhibited more complex trends. Average densities varied inconsistently across different groups. Some sites, such as the femoral and patellar areas, showed early decreases followed by later increases, while tibial cartilage demonstrated a consistent upward trend. Significant differences were mainly observed between classes 1 and 2 in the femoral and patellar sites, and between healthy and degenerated knees in the tibial regions. In contrast, the standard deviation of density provided clearer distinctions, revealing significantly higher and increasing heterogeneity in the diseased groups, with significant differences detected in nearly all pairwise comparisons.

Overall, both geometric (volume-to-surface) and heterogeneity (density standard deviation) features showed the most robust ability to separate the three groups, while mean density values displayed greater variability across sites.

6.4 Discussion

The image-extracted features and the defined SSL pipeline used in this work proved their ability in labeling the dataset in three different classes according to the severity and progression of the disease, by capturing consistent patterns within the data.

6.4.1 SSL Pipeline: Labels Stability and Classification Performances

The results demonstrate that the SSL strategy yields stable and consistent outcomes across the ten different runs. The high per-sample agreement (0.911 ± 0.136) suggests that the majority of samples are labeled consistently, regardless of initialization of seeds and test split, with only minor variability. Fleiss’ Kappa of 0.781 confirms substantial agreement beyond chance, demonstrating labeling robustness. Moreover, the low mean entropy (0.283 ± 0.385) suggests that most samples exhibit low uncertainty across runs, implying that the model is confident in its assignments for the majority of the data. Together, these metrics indicate that the labeling procedure is reliable, although the presence of some variability

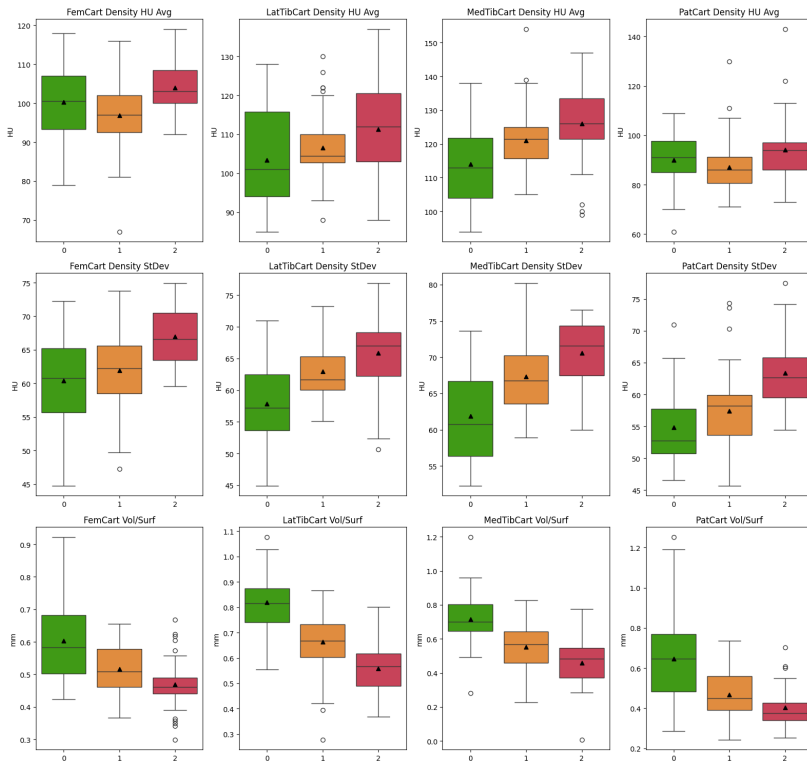


Figure 6.2: Boxplot distributions of the full labeled dataset for each feature, divided by classes 0, 1, and 2 (along x-axis); y-axis represents HU for density features and mm for Vol/Surf ones.

Table 6.6: Post Hoc Analysis Results for Cartilage Metrics (P-values rounded to 3 significant digits)

Feature	Test	Groups	P-value	Reject
FemCart Density HU Avg	Dunn	0 vs 1	0.191	False
		0 vs 2	0.105	False
		1 vs 2	0.000	True
LatTibCart Density HU Avg	Dunn	0 vs 1	0.416	False
		0 vs 2	0.008	True
		1 vs 2	0.215	False
MedTibCart Density HU Avg	Dunn	0 vs 1	0.022	True
		0 vs 2	0.000	True
		1 vs 2	0.055	False
PatCart Density HU Avg	Dunn	0 vs 1	0.067	False
		0 vs 2	1.000	False
		1 vs 2	0.007	True
FemCart Density StDev	Tukey HSD	0 vs 1	0.343	False
		0 vs 2	0.000	True
		1 vs 2	0.000	True
LatTibCart Density StDev	Tukey HSD	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.043	True
MedTibCart Density StDev	Tukey HSD	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.011	True
PatCart Density StDev	Dunn	0 vs 1	0.123	False
		0 vs 2	0.000	True
		1 vs 2	0.000	True
FemCart Vol/Surf	Tukey HSD	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.060	False
LatTibCart Vol/Surf	Dunn	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.001	True
MedTibCart Vol/Surf	Dunn	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.012	True
PatCart Vol/Surf	Dunn	0 vs 1	0.000	True
		0 vs 2	0.000	True
		1 vs 2	0.031	True

and higher entropy for a subset of samples highlights potential areas where the labeling strategy may struggle with ambiguous or harder-to-classify cases. It is important to consider the intrinsic subjectivity that characterizes the disease and, more generally, the biomedical field. A certain degree of disagreement or uncertainty is to be expected, as even field experts may differ in their interpretations of complex or borderline cases. The three-class classification results highlight clear differences in performance across models and classes. At the global level, SVM consistently outperforms the other approaches, achieving the highest weighted F1-score (0.84 ± 0.11), followed by LR (0.81 ± 0.13). Even the lowest-performing model, GB, maintains a reasonable F1-score (0.74 ± 0.18), suggesting that all tested

algorithms are able to capture relevant information from the data despite the small sample size and inherent variability across runs.

When analyzing class-specific performances, distinct patterns emerge. Class 0 (healthy) is identified with high reliability across all models, with near-perfect recall and precision in some cases (e.g., KNN with $F1 = 0.97$, SVM with $F1 = 0.95$), suggesting that healthy samples occupy a clearly separable region in the feature space. In contrast, Classes 1 (early degeneration) and 2 (severe degeneration) exhibit comparable F1-scores (approximately 0.78–0.79), indicating higher similarity and partial overlap in their feature distributions. The class distribution (42, 56, and 35 samples, respectively) is moderately balanced, suggesting that this performance gap is not primarily due to data imbalance but rather reflects the intrinsic difficulty in discriminating between progressive disease stages. Overall, the model demonstrates robust generalization, with strong reliability for healthy samples and a small but consistent margin for improvement in the separation between early and severe degeneration.

These findings show how different models emphasize distinct trade-offs between precision and recall, particularly for the more ambiguous diseased classes. The broader variability across runs, especially in Classes 1 and 2, reflects the difficulty of discriminating progressive disease stages and the challenges of small-sample biomedical evaluation. Nonetheless, the repeated experiments demonstrate that certain models, such as SVM, remain robust across different splits, while others, such as LR, show complementary strengths that could be valuable in ensemble approaches. Overall, healthy samples are reliably identified, while distinguishing between early and severe degeneration is more challenging, although still feasible. AI in pain assessment must be interpretable, multimodal, and clinically viable [188], principles our SSL pipeline addresses with multimodal imaging and reproducibility. Prior work also notes that while healthy controls are easily separated, early- and late-stage KOA are harder to distinguish due to overlapping features and stages' gradual transitions [189].

6.4.2 Statistical Results and Clinical Interpretation

The final dataset of 133 samples validated the employed labeling strategy. As shown in Table 6.6, many features reject the null hypothesis, indicating that the SSL pipeline captures both structural (density) and geometrical (volume/surface) changes across healthy, early, and severe degeneration. This supports that the data-driven stratification reflects biological and morphological differences rather than being arbitrary. Figure 6.2 further shows that labeled samples follow ascending or descending trends for nearly every feature, consistent with expected degeneration progression. This strengthens ML interpretability, suggesting models separate classes statistically while aligning with known pathophysiological mechanisms.

Most Vol/Surf features effectively differentiate healthy, early, and late degeneration, with the femoral cartilage showing some difficulty distinguishing the two degeneration stages, though remaining significant when comparing healthy and degenerative cases. The volume-to-surface ratio indicates that higher values reflect greater tissue volume relative to surface area. Healthy cartilage is thick, smooth,

and evenly distributed, yielding a high Vol/Surf ratio. Osteoarthritic cartilage, conversely, is thinner, fibrillated, and irregular, reducing volume while increasing effective surface area, and thus lowering the Vol/Surf ratio. Multiple studies [190], [191] confirm that cartilage volume is significantly reduced in KOA across knee compartments and that tibial surface area increases, particularly in the presence of osteophytes (bony outgrowths). Early KOA is associated with concurrent cartilage volume loss and bone remodeling [191], [192], and longitudinal studies confirm progressive volume loss [132], [133]. Overall, the downward Vol/Surf trends support KOA pathophysiology and suggest this ratio as a sensitive marker for degeneration.

Density results reveal an interesting trend. Average densities of femoral and patellar regions are less predictive, particularly among healthy and advanced cases, though significant differences exist between classes 1 and 2. This suggests density initially decreases, then increases. While this partly contradicts the literature, several mechanisms may explain it. Early KOA degrades the softer, water-rich regions of cartilage, which are mechanically weaker and less dense [141]. As these regions are lost, remaining cartilage appears denser on imaging due to exclusion of degraded areas, not because healthy tissue densifies. Late-stage KOA also shows increased mineralization and crystal accumulation in the calcified cartilage layer [193], [194], raising HU values without improving mechanical function [141], [194]. Healthy cartilage retains a hydrated, proteoglycan-rich matrix, explaining lower density and variability. Technical factors like joint space narrowing may also affect segmentation. Site-specific differences clarify trends: tibial cartilage shows a more pronounced density increase, possibly due to meniscal protection of femoral cartilage and higher mechanical load on the tibia. Early decreases in femoral and patellar cartilage density may reflect lower mechanical stress, with tibial and trochlear moduli declining earlier [195]. Standard deviation (density heterogeneity) provides more consistent separation across groups. Almost all sites, including tibial and femoral cartilage, show significant differences in StDev between healthy, early, and late KOA, indicating greater intra-cartilage heterogeneity in KOA. This aligns with MRI and histopathological studies reporting higher variability in water content, matrix organization, and density in degenerating cartilage [136]. It is also important to consider that degeneration is not occurring all at once throughout the entire knee compartment, but rather progresses gradually and may affect different regions to varying degrees over time. Notably, heterogeneity may be a more sensitive marker than mean density, particularly for early detection.

Overall, these results highlight that while mean density trends are site-dependent and sometimes counterintuitive, variability measures consistently reflect degeneration. Combined with vol/surf features, which robustly separate all classes, this supports the multifactorial nature of KOA and validates the clinical relevance of the new three-class labels derived from multimodal imaging.

6.4.3 Limitations

This study has some limitations that should be acknowledged. First, the dataset was relatively small (133 knees, with only 36 expert-labeled), which may limit the generalizability of the results and increase the variability in model performance.

Additionally, class imbalance among the degeneration stages may have affected the stability and accuracy of the model, particularly for early and severe cases.

The reliance on pseudo-labels also introduces potential bias and uncertainty, especially for borderline samples. Furthermore, the lack of longitudinal data prevented an assessment of disease progression.

Finally, although the framework showed strong performance, clinical translation will require validation in larger and more diverse cohorts, as well as in real-world diagnostic settings, to ensure its robustness and applicability in clinical practice.

Chapter 7

Limitations, Conclusions and Future Developments

7.1 Limitations

Despite the promising results discussed in this thesis, several limitations must be acknowledged.

The dataset used in this study was relatively limited in size and heterogeneity, which may restrict the generalization of the models to broader and more diverse populations. Moreover, it lacked longitudinal follow-up, which limited the ability to assess temporal changes and validate whether the proposed imaging biomarkers can reliably predict disease progression. Additionally, demographic and physiological confounders such as BMI, age, and gender were not explicitly corrected for, even though these factors are known to strongly influence the overall knee condition, including joint morphology, bone density, and cartilage composition.

The cohort also presented a small imbalance between healthy and degenerative cases, which may have influenced the performance of some classification models despite the use of cross-validation strategies. Additionally, most of the data were acquired under controlled imaging conditions and standardized protocols, which may not fully reflect the variability typically encountered in real clinical environments.

One significant limitation is that the segmentation and registration processes are done manually, which can be time-consuming. While manual segmentation has its advantages, especially considering the non-homogeneous shapes and varying morphology of osteoarthritic cartilage among patients, it is essential to develop an automated process that matches the accuracy of manual segmentation, even though this presents challenges. Optimizing and speeding up these initial steps is crucial for providing patients with quicker and more accessible results from knee scans.

It is important to underline that, although the registration and feature extraction pipelines were carefully optimized, the generalizability of the results across different scanners, acquisition parameters, and reconstruction algorithms has yet to

be established. Variations in voxel resolution, field strength, and image contrast can impact the reproducibility of the extracted radiomic and morphological features. Additionally, while this thesis addressed the robustness of features in relation to segmentation variability, the reproducibility of feature extraction using various software implementations and preprocessing pipelines remains a significant challenge in quantitative imaging research.

Additionally, although the study employs cross-validation and robustness analyses, it lacks external validation on independent cohorts, which is crucial for confirming the reproducibility and real-world applicability of the proposed imaging and machine learning frameworks.

One final limitation relates to the interpretability of machine learning models. Although feature importance analyses have been conducted, the relationship between certain high-dimensional features and their biological significance is often complex and not always intuitive. This complexity can hinder the direct translation of model outputs into actionable clinical insights. Additionally, there is an ongoing challenge concerning the trade-off between interpretability and predictive performance; more complex models tend to achieve higher accuracy but often at the expense of reduced transparency.

7.2 Conclusions

This dissertation aimed to develop and validate quantitative imaging and machine learning approaches for the assessment of knee osteoarthritis, focusing on reproducibility, explainability, and early diagnosis. Within the framework of the EU Horizon SINPAIN project, the work established a digital pipeline for extracting imaging biomarkers from CT and MRI data to enable patient-specific profiling and treatment stratification.

The thesis addressed the five research questions defined in 1.4. The first research question, addressed in Chapter 2, focused on how multimodal imaging data from MRI and CT could be combined and processed to obtain reproducible quantitative measures of knee joint morphology and tissue properties. A complete imaging pipeline was established, including segmentation, registration, and feature extraction procedures that enabled 3D reconstructions of cartilage, bone, and muscle. This pipeline allowed consistent computation of morphological, density, and radiomic descriptors across imaging modalities, providing a foundation for subsequent analyses.

The second research question, examined in Chapters 3 and 4, investigated which quantitative imaging biomarkers best characterize tissue degeneration and how machine learning enhances their identification. Analyses revealed that cartilage volume, especially in the tibial and patellar regions, along with density variability and texture-based radiomic features, were the most discriminative markers of degeneration. Machine learning models, particularly SVM and LR, effectively ranked these features and achieved strong diagnostic accuracy. Cross-tissue analyses further demonstrated significant correlations between cartilage heterogeneity, bone

densitometric changes, and muscle fatty infiltration, underscoring the multifactorial and interconnected nature of knee osteoarthritis.

The third research question, explored in Chapter 5, evaluated the robustness and reproducibility of the proposed framework under segmentation variability, operator-dependent factors, and limited data conditions. Controlled perturbations of segmentation masks demonstrated that while most morphological and density-based features remained stable, radiomic features were more sensitive to boundary variations. Despite these perturbations, classification performance remained consistent, highlighting the reliability of the quantitative approach even with small datasets and imperfect segmentations.

The fourth research question, addressed in Chapter 6, examined whether machine learning models could reliably distinguish between different stages of knee osteoarthritis using quantitative imaging biomarkers. Semi-supervised learning successfully revealed hidden structure in the data, grouping samples with similar degenerative patterns and producing stable, clinically meaningful labels. Supervised models trained on these labels reached strong performance, with the SVM classifier achieving a weighted F1-score of 0.84 ± 0.11 . Statistical analysis further supported clear separability among disease stages, particularly for cartilage density variability and volume-to-surface ratio, which showed progressive changes with degeneration.

Overall, the thesis demonstrates that integrating quantitative multimodal imaging with machine learning enables a more objective, reproducible, and interpretable evaluation of knee osteoarthritis. The proposed framework enhances diagnostic accuracy and paves the way for a more personalized approach to joint health. By capturing the specific structural and morphological features of each knee, it helps establish an individual knee profile that can support customized monitoring, risk evaluation, and treatment options. Consequently, the future of osteoarthritis care may shift from standard assessments to personalized solutions, where quantitative imaging guides prevention and therapy based on each person's unique joint characteristics.

7.3 Future Developments

Future work will aim to address the limitations previously discussed, and to further enhance the clinical applicability of the proposed framework. Increasing the size and diversity of the dataset would help validate the findings across broader populations and imaging settings, improving the robustness of the models. Moreover, including longitudinal data will be particularly important to observe how knee osteoarthritis evolves over time and to test whether the proposed imaging biomarkers can truly predict disease progression.

Automating parts of the pipeline, especially the segmentation step, could make the process faster and reduce its dependence on manual input, making it more suitable for clinical use.

Future analyses should also account for demographic and physiological factors (such as BMI, age, gender) as well as clinical history and lifestyle. These factors,

combined with biochemical and imaging data, could enhance the clinical relevance of the proposed approach and pave the way for a quantitative risk index for early diagnosis and personalized management of knee osteoarthritis.

Future research should further investigate the interactions between cartilage, bone, and muscle within the knee joint. Building on the findings from Chapter 4, upcoming studies will aim to enhance the quantitative analysis across these tissues. This will involve integrating multimodal imaging biomarkers and advanced ML models to better understand the systemic and biomechanical relationships that contribute to joint degeneration. By taking a multi-tissue approach, we hope to improve the characterization of early-stage KOA, identify potential cross-tissue biomarkers, and achieve a more comprehensive understanding of the entire joint pathology.

Future extensions of this work could also explore multimodal musculoskeletal modeling, integrating not only cartilage, bone, and muscle but also tendons and soft-tissue biomechanics into a unified computational framework. Such models would enable a more holistic understanding of joint function and degeneration. Moreover, incorporating longitudinal monitoring through repeated imaging and patient follow-up would make it possible to track disease trajectories, evaluate intervention effects, and refine predictive models for personalized prognosis.

Moreover, future developments should focus on integrating unsupervised and deep learning methodologies. Unsupervised learning can help uncover hidden patterns and latent disease phenotypes without depending on predefined clinical labels, providing new insights into early or atypical manifestations of KOA. Additionally, deep learning architectures can be used for end-to-end feature learning directly from multimodal imaging data. This approach has the potential to outperform traditional handcrafted features in terms of both accuracy and generalizability.

Bibliography

- [1] F. K. Ciliberti, L. Guerrini, A. E. Gunnarsson, *et al.*, “Ct-and mri-based 3d reconstruction of knee joint to assess cartilage and bone,” *Diagnostics*, vol. 12, no. 2, p. 279, 2022.
- [2] R. Aubonnet, J. Ramos, M. Recenti, *et al.*, “Toward new assessment of knee cartilage degeneration,” *Cartilage*, vol. 14, no. 3, pp. 351–374, 2023.
- [3] F. Angelone, F. K. Ciliberti, G. P. Tobia, *et al.*, “Innovative diagnostic approaches for predicting knee cartilage degeneration in osteoarthritis patients: A radiomics-based study,” *Information Systems Frontiers*, pp. 1–23, 2024.
- [4] F. K. Ciliberti, I. Maruotto, H. j. Jónsson, M. K. Gíslason, and P. Gargiulo, “Multimetric evaluation of knee cartilage degeneration,” in *Proceedings of IEEE MetroXRaine 2024*, St Albans, London, UK, Oct. 2024.
- [5] F. K. Ciliberti, G. Pianegonda, H. Jonsson, M. K. Gislason, and P. Gargiulo, “How muscle and intramuscular adipose tissue volume and density progress in knee osteoarthritis-affected subjects: A virtual histology study,” in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, IEEE, 2024, pp. 177–182.
- [6] I. Maruotto, F. K. Ciliberti, P. Gargiulo, and M. Recenti, “Feature selection in healthcare datasets: Towards a generalizable solution,” *Computers in Biology and Medicine*, vol. 196, p. 110 812, Sep. 2025, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2025.110812. [Online]. Available: <http://dx.doi.org/10.1016/j.compbiomed.2025.110812>.
- [7] F. Ciliberti, G. Cesarelli, L. Guerrini, *et al.*, “The role of bone mineral density and cartilage volume to predict knee cartilage degeneration,” *European Journal of Translational Myology*, vol. 32, 2022. DOI: 10.4081/ejtm.2022.10678.
- [8] A. E. Gunnarsson, F. K. Ciliberti, C. Belfiori, *et al.*, “Assessment of femoral cartilage morphological and topological features using machine learning,” in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)*, 2022, pp. 277–282. DOI: 10.1109/MetroXRaine54828.2022.9967659.

- [9] F. K. Ciliberti, R. Forni, D. Coato, *et al.*, “Development of synthetic 3d printed knee joint to assess mechanical and functional properties of degenerative cartilage,” in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, 2023, pp. 628–633. DOI: 10.1109/MetroXRAINE58569.2023.10405695.
- [10] G. Dolino, D. Coato, R. Forni, G. Boretti, F. K. Ciliberti, and P. Gargiulo, “Designing a synthetic 3d-printed knee cartilage: Fea model, micro-structure and mechanical characteristics,” *Applied Sciences*, vol. 14, no. 1, 2024, ISSN: 2076-3417. DOI: 10.3390/app14010331. [Online]. Available: <https://www.mdpi.com/2076-3417/14/1/331>.
- [11] F. Angelone, F. K. Ciliberti, H. j. Jónsson, M. K. Gíslason, and P. Gargiulo, “Knee cartilage degradation in the medial and lateral anatomical compartments: A radiomics study,” in *Proceedings of IEEE MetroXRAINE 2024*, St Albans, London, UK, Oct. 2024.
- [12] L. Yue and J. Berman, “What is osteoarthritis?” *JAMA*, vol. 327 13, p. 1300, 2022. DOI: 10.1001/jama.2022.1980.
- [13] Y. Zhang and J. M. Jordan, “Epidemiology of osteoarthritis,” *Clinics in Geriatric Medicine*, vol. 26, no. 3, pp. 355–369, 2010, Osteoarthritis, ISSN: 0749-0690. DOI: <https://doi.org/10.1016/j.cger.2010.03.001>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0749069010000261>.
- [14] R. Geng, J. Li, C. Yu, *et al.*, “Knee osteoarthritis: Current status and research progress in treatment,” *Experimental and therapeutic medicine*, vol. 26, no. 4, pp. 1–11, 2023.
- [15] T. Neogi, “The epidemiology and impact of pain in osteoarthritis,” *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1145–1153, 2013, Pain in Osteoarthritis, ISSN: 1063-4584. DOI: <https://doi.org/10.1016/j.joca.2013.03.018>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458413007607>.
- [16] H. Hsu and R. Siwec, *Knee Osteoarthritis*. StatPearls Publishing, 2024.
- [17] J. Martel-Pelletier *et al.*, “Osteoarthritis,” *Nature Reviews Disease Primers*, vol. 2, no. 1, 2016.
- [18] D. Prieto-Alhambra *et al.*, “Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: Influences of age, gender and osteoarthritis affecting other joints,” *Annals of the Rheumatic Diseases*, vol. 73, no. 9, pp. 1659–1664, 2014. DOI: 10.1136/annrheumdis-2013-203355.
- [19] L. Callahan *et al.*, “Limited educational attainment and radiographic and symptomatic knee osteoarthritis: A cross-sectional analysis using data from the johnston county (north carolina) osteoarthritis project,” *Arthritis Research & Therapy*, vol. 12, no. 2, R46, 2010. DOI: 10.1186/ar2950.

- [20] C. Palazzo, C. Nguyen, M.-M. Lefevre-Colau, F. Rannou, and S. Poiraudau, "Risk factors and burden of osteoarthritis," *Annals of physical and rehabilitation medicine*, vol. 59, no. 3, pp. 134–138, 2016.
- [21] E. M. Roos and N. K. Arden, "Strategies for the prevention of knee osteoarthritis," *Nature Reviews Rheumatology*, vol. 12, no. 2, pp. 92–101, 2016.
- [22] L. Murphy, T. A. Schwartz, C. G. Helmick, *et al.*, "Lifetime risk of symptomatic knee osteoarthritis," *Arthritis Care & Research: Official Journal of the American College of Rheumatology*, vol. 59, no. 9, pp. 1207–1213, 2008.
- [23] M.-E. Pickering *et al.*, "Sleep disturbances and osteoarthritis," *Pain Practice*, vol. 16, no. 2, pp. 237–244, 2016. DOI: 10.1111/papr.12283.
- [24] D. C. Crawford *et al.*, "Conservative management of symptomatic knee osteoarthritis: A flawed strategy?" *Orthopedic Reviews*, vol. 5, no. 1, Feb. 22, 2013.
- [25] N. J. London *et al.*, "Clinical and economic consequences of the treatment gap in knee osteoarthritis management," *Medical Hypotheses*, vol. 76, no. 6, Jun. 1, 2011.
- [26] J. T. Evans *et al.*, "How long does a knee replacement last? a systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up," *The Lancet*, vol. 393, no. 10172, Feb. 16, 2019.
- [27] E. Ringdahl and S. Pandit, "Treatment of knee osteoarthritis," *American family physician*, vol. 83, no. 11, pp. 1287–1292, 2011.
- [28] J.-P. Pelletier *et al.*, "Risk factors for the long-term incidence and progression of knee osteoarthritis in older adults: Role of nonsurgical injury," *Therapeutic Advances in Chronic Disease*, vol. 14, 2023.
- [29] M. Cross, E. Smith, D. Hoy, *et al.*, "The global burden of hip and knee osteoarthritis: Estimates from the global burden of disease 2010 study," *Annals of the rheumatic diseases*, vol. 73, no. 7, pp. 1323–1330, 2014.
- [30] E. Losina *et al.*, "Lifetime medical costs of knee osteoarthritis management in the united states: Impact of extending indications for total knee arthroplasty," *Arthritis Care & Research*, vol. 67, no. 2, pp. 203–215, 2015. DOI: 10.1002/acr.22412.
- [31] V. Leifer, J. Katz, and E. Losina, "The burden of oa-health services and economics," *Osteoarthritis and Cartilage*, vol. 30, no. 1, pp. 10–16, 2022. DOI: 10.1016/j.joca.2021.09.003.
- [32] L. Marcucci and C. Reggiani, "Increase of resting muscle stiffness, a less considered component of age-related skeletal muscle impairment," *European journal of translational myology*, vol. 30, no. 2, p. 8982, 2020.
- [33] J. Steinmetz *et al.*, "Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: A systematic analysis for the global burden of disease study 2021," *The Lancet Rheumatology*, vol. 5, no. 9, e557–e567, 2023. DOI: 10.1016/S2665-9913(23)00108-9.

- [34] F. W. Roemer, S. Demehri, P. Omoumi, *et al.*, "State of the art: Imaging of osteoarthritis—revisited 2020," *Radiology*, vol. 296, no. 1, pp. 5–21, 2020.
- [35] M. D. Crema, F. W. Roemer, M. D. Marra, *et al.*, "Articular cartilage in the knee: Current mr imaging techniques and applications in clinical practice and research," *Radiographics*, vol. 31, no. 1, pp. 37–61, 2011.
- [36] R. J. Schmitz, H.-M. Wang, D. R. Polprasert, R. A. Kraft, and B. G. Pietrosimone, "Evaluation of knee cartilage thickness: A comparison between ultrasound and magnetic resonance imaging methods," *The Knee*, vol. 24, no. 2, pp. 217–223, 2017.
- [37] J. Johnston, B. Masri, and D. Wilson, "Computed tomography topographic mapping of subchondral density (ct-tomasd) in osteoarthritic and normal knees: Methodological development and preliminary findings," *Osteoarthritis and Cartilage*, vol. 17, no. 10, pp. 1319–1326, 2009, ISSN: 1063-4584. DOI: <https://doi.org/10.1016/j.joca.2009.04.013>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458409001113>.
- [38] J. H. Kellgren, J. Lawrence, *et al.*, "Radiological assessment of osteo-arthritis," *Ann Rheum Dis*, vol. 16, no. 4, pp. 494–502, 1957.
- [39] S. Ahlback, "Osteoarthritis of the knee. a radiographic investigation," *Acta Radiol Diagn (Suppl)*, vol. 277, pp. 7–72, 1968.
- [40] M. Galli, V. De Santis, and L. Tafuro, "Reliability of the ahlbäck classification of knee osteoarthritis," *Osteoarthritis and Cartilage*, vol. 11, no. 8, pp. 580–584, 2003, ISSN: 1063-4584. DOI: [https://doi.org/10.1016/S1063-4584\(03\)00095-5](https://doi.org/10.1016/S1063-4584(03)00095-5). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458403000955>.
- [41] N. Wing, N. van Zyl, M. Wing, R. Corrigan, A. Loch, and C. Wall, "Reliability of three radiographic classification systems for knee osteoarthritis among observers of different experience levels," *Skeletal Radiology*, vol. 50, pp. 399–405, 2020. DOI: 10.1007/s00256-020-03551-4.
- [42] R. W. Wright, M. Group, *et al.*, "Osteoarthritis classification scales: Interobserver reliability and arthroscopic correlation," *JBJS*, vol. 96, no. 14, pp. 1145–1151, 2014.
- [43] O. J. Keenan, G. Holland, J. F. Maempel, J. F. Keating, and C. E. Scott, "Correlations between radiological classification systems and confirmed cartilage loss in severe knee osteoarthritis," *The Bone & Joint Journal*, vol. 102, no. 3, pp. 301–309, 2020.
- [44] F. Eckstein, M. H. Le Graverand, H. Charles, *et al.*, "Clinical, radiographic, molecular and mri-based predictors of cartilage loss in knee osteoarthritis," *Annals of the rheumatic diseases*, vol. 70, no. 7, pp. 1223–1230, 2011.
- [45] J. W. MacKay, J. D. Kaggie, G. M. Treece, *et al.*, "Three-dimensional surface-based analysis of cartilage mri data in knee osteoarthritis: Validation and initial clinical application," *Journal of Magnetic Resonance Imaging*, vol. 52, no. 4, pp. 1139–1151, 2020.

- [46] S. Papernick, R. Dima, D. Gillies, C. Appleton, and A. Fenster, "Reliability and concurrent validity of three-dimensional ultrasound for quantifying knee cartilage volume," *Osteoarthritis and cartilage open*, vol. 2, no. 4, p. 100 127, 2020.
- [47] F. W. Roemer, R. Frobell, L. S. Lohmander, J. Niu, and A. Guermazi, "Anterior cruciate ligament osteoarthritis score (acloas): Longitudinal mri-based whole joint assessment of anterior cruciate ligament injury," *Osteoarthritis and cartilage*, vol. 22, no. 5, pp. 668–682, 2014.
- [48] R. R. Bannuru, M. Osani, E. Vaysbrot, *et al.*, "Oarsi guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis," *Osteoarthritis and cartilage*, 2019. DOI: 10 . 1016/j . joca . 2019 . 06 . 011.
- [49] L. O. Dantas, T. de Fátima Salvini, and T. McAlindon, "Knee osteoarthritis: Key treatments and implications for physical therapy.," *Brazilian journal of physical therapy*, 2020. DOI: 10 . 1016/j . bjpt . 2020 . 08 . 004.
- [50] N. Arden, T. Perry, R. R. Bannuru, *et al.*, "Non-surgical management of knee osteoarthritis: Comparison of esceo and oarsi 2019 guidelines," *Nature Reviews Rheumatology*, vol. 17, pp. 59–66, 2020. DOI: 10 . 1038 / s41584 - 020 - 00523 - 9.
- [51] A. Litwic, M. H. Edwards, E. M. Dennison, and C. Cooper, "Epidemiology and burden of osteoarthritis," *British Medical Bulletin*, vol. 105, no. 1, pp. 185–199, Jan. 2013, ISSN: 0007-1420. DOI: 10 . 1093/bmb/lds038. eprint: <https://academic.oup.com/bmb/article-pdf/105/1/185/757775/lds038.pdf>. [Online]. Available: <https://doi.org/10.1093/bmb/lds038>.
- [52] U. Anil, D. Markus, E. Hurley, *et al.*, "The efficacy of intra-articular injections in the treatment of knee osteoarthritis: A network meta-analysis of randomized controlled trials.," *The Knee*, vol. 32, pp. 173–182, 2021. DOI: 10 . 1016/j . knee . 2021 . 08 . 008.
- [53] Y. Song, J. Zhang, H. Xu, *et al.*, "Mesenchymal stem cells in knee osteoarthritis treatment: A systematic review and meta-analysis," *Journal of Orthopaedic Translation*, vol. 24, pp. 121–130, 2020. DOI: 10 . 1016/j . jot . 2020 . 03 . 015.
- [54] G. Copp, K. Robb, and S. Viswanathan, "Culture-expanded mesenchymal stromal cell therapy: Does it work in knee osteoarthritis? a pathway to clinical success," *Cellular and Molecular Immunology*, vol. 20, pp. 626–650, 2023. DOI: 10 . 1038/s41423-023-01020-1.
- [55] RESTORE Consortium, *Restore project*, Accessed: 2025-04-14, 2024. [Online]. Available: <https://restoreproject.eu/>.
- [56] H. Evenbratt, L. Andreasson, V. Bicknell, M. Brittberg, R. Mobini, and S. Simonsson, "Insights into the present and future of cartilage regeneration and joint repair," *Cell Regeneration*, vol. 11, no. 1, p. 3, 2022.
- [57] *Sinpain project website*, <https://www.sinpain-project.eu>, Accessed: 2025-10-01, 2022.

- [58] C. Kokkotis *et al.*, “Machine learning in knee osteoarthritis: A review,” *Osteoarthritis and Cartilage Open*, vol. 2, no. 3, Sep. 1, 2020.
- [59] T. M. Mitchell, *Machine learning, International Edition* (McGraw-Hill Series in Computer Science). McGraw-Hill, 1997, ISBN: 978-0-07-042807-2. [Online]. Available: <https://www.worldcat.org/oclc/61321007>.
- [60] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. AlQahtani, P. A. Watters, and A. Ng, “Cybersecurity data science: An overview from machine learning perspective,” *J. Big Data*, vol. 7, no. 1, p. 41, 2020. DOI: 10.1186/S40537-020-00318-5. [Online]. Available: <https://doi.org/10.1186/s40537-020-00318-5>.
- [61] I. Attri, L. K. Awasthi, and T. P. Sharma, “Machine learning in agriculture: A review of crop management applications,” *Multim. Tools Appl.*, vol. 83, no. 5, pp. 12 875–12 915, 2024. DOI: 10.1007/S11042-023-16105-2. [Online]. Available: <https://doi.org/10.1007/s11042-023-16105-2>.
- [62] V. de J. Duarte, S. Zuniga-Jara, and S. Contreras, “Machine learning and marketing: A systematic literature review,” *IEEE Access*, vol. 10, pp. 93 273–93 288, 2022. DOI: 10.1109/ACCESS.2022.3202896. [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3202896>.
- [63] M. F. Dixon, I. Halperin, P. Bilokon, *et al.*, *Machine learning in finance*. Springer, 2020, vol. 1170.
- [64] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: A practical introduction,” *BMC medical research methodology*, vol. 19, no. 1, p. 64, 2019.
- [65] J. Ker, L. Wang, J. Rao, and T. C. C. Lim, “Deep learning applications in medical image analysis,” *IEEE Access*, vol. 6, pp. 9375–9389, 2018. DOI: 10.1109/ACCESS.2017.2788044. [Online]. Available: <https://doi.org/10.1109/ACCESS.2017.2788044>.
- [66] E. M. Nwanosike, B. R. Conway, H. A. Merchant, and S. S. Hasan, “Potential applications and performance of machine learning techniques and algorithms in clinical practice: A systematic review,” *International journal of medical informatics*, vol. 159, p. 104 679, 2022.
- [67] V. Cheplygina, M. de Bruijne, and J. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Medical Image Analysis*, vol. 54, pp. 280–296, 2018. DOI: 10.1016/j.media.2019.03.009.
- [68] Z. Song, X. Yang, Z. Xu, and I. King, “Graph-based semi-supervised learning: A comprehensive review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 8174–8194, 2021. DOI: 10.1109/TNNLS.2022.3155478.
- [69] J. E. van Engelen and H. Hoos, “A survey on semi-supervised learning,” *Machine Learning*, vol. 109, pp. 373–440, 2019. DOI: 10.1007/s10994-019-05855-6.

- [70] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: A brief introduction," *Frontiers of Computer Science*, vol. 13, pp. 669–676, 2019. DOI: 10.1007/s11704-019-8452-2.
- [71] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis.," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017. DOI: 10.1146/annurev-bioeng-071516-044442.
- [72] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, 2023. DOI: 10.3390/cancers15143608.
- [73] A. Esteva, A. Robicquet, B. Ramsundar, *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019. DOI: 10.1038/s41591-018-0316-z.
- [74] K. A. Tran, O. Kondrashova, A. Bradley, E. Williams, J. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine*, vol. 13, 2021. DOI: 10.1186/s13073-021-00968-x.
- [75] B. G. Ashinsky, M. Bouhrara, C. E. Coletta, *et al.*, "Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative," *Journal of Orthopaedic Research*, vol. 35, no. 10, pp. 2243–2250, 2017. DOI: 10.1002/jor.23519.
- [76] J. Marques, H. K. Genant, M. Lillholm, and E. B. Dam, "Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from mri," *Magnetic Resonance in Medicine*, vol. 70, no. 2, pp. 568–575, 2013. DOI: 10.1002/mrm.24477.
- [77] J. Abedin, J. Antony, K. McGuinness, *et al.*, "Predicting knee osteoarthritis severity: Comparative modeling based on patient's data and plain x-ray images," *Scientific Reports*, vol. 9, no. 1, p. 5761, 2019. DOI: 10.1038/s41598-019-42215-9.
- [78] C. R. Donoghue, A. Rao, A. M. J. Bull, and D. Rueckert, "Manifold learning for automatically predicting articular cartilage morphology in the knee with data from the osteoarthritis initiative (oai)," in *Medical Imaging 2011: Image Processing*, vol. 7962, SPIE, 2011, 79620E. DOI: 10.1117/12.878266.
- [79] Y. Du, R. Almajalid, J. Shan, and M. Zhang, "A novel method to predict knee osteoarthritis progression on mri using machine learning methods," *IEEE Transactions on Nanobioscience*, vol. 17, no. 3, pp. 228–236, 2018. DOI: 10.1109/TNB.2018.2840082.
- [80] Y. Du *et al.*, "A novel method to predict knee osteoarthritis progression on mri using machine learning methods | ieee journals & magazine | ieee xplore," *IEEE Transactions on NanoBioscience*, vol. 17, no. 3, 2018.
- [81] A. Tiulpin *et al.*, "Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data," *Scientific Reports*, vol. 2019, no. 9-1, p. 9, Dec. 27, 2019.

- [82] V. Pedoia, J. Haefeli, K. Morioka, *et al.*, “Mri and biomechanics multidimensional data analysis reveals $R_2-R_{1\rho}$ as an early predictor of cartilage lesion progression in knee osteoarthritis,” *Journal of Magnetic Resonance Imaging*, vol. 47, no. 1, pp. 78–90, 2018. DOI: 10.1002/jmri.25750.
- [83] J. Huo, L. Si, X. Ouyang, *et al.*, “A self-ensembling framework for semi-supervised knee cartilage defects assessment with dual-consistency,” pp. 200–209, 2020. DOI: 10.1007/978-3-030-59354-4_19.
- [84] J. Huo, X. Ouyang, L. Si, *et al.*, “Automatic grading assessments for knee mri cartilage defects via self-ensembling semi-supervised learning with dual-consistency,” *Medical image analysis*, vol. 80, p. 102508, 2022. DOI: 10.1016/j.media.2022.102508.
- [85] N. Belton, A. Lawlor, and K. M. Curran, “An ai system for continuous knee osteoarthritis severity grading: An anomaly detection inspired approach with few labels,” *Artificial intelligence in medicine*, vol. 167, p. 103138, 2024. DOI: 10.1016/j.artmed.2025.103138.
- [86] A. M. Raisuddin, H. H. Nguyen, and A. Tiulpin, “Deep semi-supervised active learning for knee osteoarthritis severity grading,” *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2022. DOI: 10.1109/ISBI52829.2022.9761668.
- [87] H. A. Alshamrani, M. Rashid, S. S. Alshamrani, and A. H. Alshehri, “Osteonet: An automated system for predicting knee osteoarthritis from x-ray images using transfer-learning-based neural networks approach,” *Healthcare*, vol. 11, 2023. DOI: 10.3390/healthcare11091206.
- [88] M. S. Islam and M. Rony, “Cdk: A novel high-performance transfer feature technique for early detection of osteoarthritis,” *Journal of Pathology Informatics*, vol. 15, 2024. DOI: 10.1016/j.jpi.2024.100382.
- [89] M. Sohail, M. M. Azad, and H. Kim, “Knee osteoarthritis severity detection using deep inception transfer learning,” *Computers in biology and medicine*, vol. 186, p. 109641, 2024. DOI: 10.1016/j.compbimed.2024.109641.
- [90] M. Mayerhoefer, A. Materka, G. Langs, *et al.*, “Introduction to radiomics,” *The Journal of Nuclear Medicine*, vol. 61, pp. 488–495, 2020. DOI: 10.2967/jnumed.118.222893.
- [91] F. Angelone, C. Ricciardi, G. Gatta, *et al.*, “Breast density analysis on mammograms: Application of machine learning with textural features,” in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)*, IEEE, 2022, pp. 295–300.
- [92] F. Angelone, A. M. Ponsiglione, M. P. Belfiore, *et al.*, “Evaluation of breast density variability between right and left breasts,” in *Atti del Convegno Nazionale di Bioingegneria 2023*, Jun. 2023.
- [93] M. Sansone, R. Fusco, F. Grassi, *et al.*, “Machine learning approaches with textural features to calculate breast density on mammography,” *Current Oncology*, vol. 30, no. 1, pp. 839–853, 2023.

- [94] I. Cetin, Z. Raisi-Estabragh, S. E. Petersen, *et al.*, “Radiomics signatures of cardiovascular risk factors in cardiac mri: Results from the uk biobank,” *Frontiers in Cardiovascular Medicine*, vol. 7, p. 591 368, 2020.
- [95] Z. Liu, T. Duan, Y. Zhang, *et al.*, “Radiogenomics: A key component of precision cancer medicine,” *British Journal of Cancer*, vol. 129, no. 5, pp. 741–753, 2023.
- [96] S. Saxena, B. Jena, N. Gupta, *et al.*, “Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine,” *Cancers*, vol. 14, no. 12, p. 2860, 2022.
- [97] P. Eslami, C. Parmar, B. Foldyna, *et al.*, “Radiomics of coronary artery calcium in the framingham heart study,” *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, e190119, 2020.
- [98] C. Lei, W. Wei, Z. Liu, *et al.*, “Mammography-based radiomic analysis for predicting benign bi-rads category 4 calcifications,” *European journal of radiology*, vol. 121, p. 108 711, 2019.
- [99] L. E. Sánchez, L. Rundo, A. Gill, M. Hoare, E. M. Serrão, and E. Sala, “Robustness of radiomic features in ct images with different slice thickness, comparing liver tumour and muscle,” *Scientific Reports*, vol. 11, 2021. DOI: 10.1038/s41598-021-87598-w.
- [100] C. Parmar, P. Grossmann, D. Rietveld, M. Rietbergen, P. Lambin, and H. Aerts, “Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer,” *Frontiers in Oncology*, vol. 5, 2015. DOI: 10.3389/fonc.2015.00272.
- [101] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. Aerts, “Machine learning methods for quantitative radiomic biomarkers,” *Scientific Reports*, vol. 5, 2015. DOI: 10.1038/srep13087.
- [102] T. Jiang, S.-H. Lau, J. Zhang, *et al.*, “Radiomics signature of osteoarthritis: Current status and perspective,” *Journal of Orthopaedic Translation*, vol. 45, pp. 100–106, 2024, ISSN: 2214-031X. DOI: <https://doi.org/10.1016/j.jot.2023.10.003>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2214031X23000773>.
- [103] Y. Xie, Y. Dan, H. Tao, *et al.*, “Radiomics feature analysis of cartilage and subchondral bone in differentiating knees predisposed to posttraumatic osteoarthritis after anterior cruciate ligament reconstruction from healthy knees,” *BioMed Research International*, vol. 2021, 2021. DOI: 10.1155/2021/4351499.
- [104] Z. Xue, L. Wang, Q. Sun, *et al.*, “Radiomics analysis using mr imaging of subchondral bone for identification of knee osteoarthritis,” *Journal of Orthopaedic Surgery and Research*, vol. 17, no. 1, p. 414, 2022.
- [105] J. Hirvasniemi, S. Klein, S. Bierma-Zeinstra, M. W. Vernooij, D. Schiphof, and E. H. Oei, “A machine learning approach to distinguish between knees without and with osteoarthritis using mri-based radiomic features from tibial bone,” *European radiology*, vol. 31, pp. 8513–8521, 2021.

- [106] J. Ma, Y. He, F. Li, L.-J. Han, C. You, and B. Wang, “Segment anything in medical images,” *Nature Communications*, vol. 15, 2023. DOI: 10.1038/s41467-024-44824-z.
- [107] B. Zitová and J. Flusser, “Image registration methods: A survey,” *Image Vis. Comput.*, vol. 21, pp. 977–1000, 2003. DOI: 10.1016/S0262-8856(03)00137-9.
- [108] A. Berger, “How does it work?: Magnetic resonance imaging,” *BMJ*, vol. 324, pp. 35–35, 2002. DOI: 10.1136/BMJ.324.7328.35.
- [109] *Materialise Mimics Core | 3D Medical Image Segmentation Software — materialise.com*, <https://www.materialise.com/en/healthcare/mimics/mimics-core>, [Accessed 24-06-2025].
- [110] C. Ricciardi, J. Jónsson H., D. Jacob, *et al.*, “Improving prosthetic selection and predicting bmd from biometric measurements in patients receiving total hip arthroplasty,” *Diagnostics*, vol. 10, no. 10, p. 815, 2020. DOI: 10.3390/diagnostics10100815.
- [111] A. Yoshiko, M. Hioki, N. Kanehira, *et al.*, “Three-dimensional comparison of intramuscular fat content between young and old adults,” *BMC Medical Imaging*, vol. 17, no. 12, p. 12, 2017. DOI: 10.1186/s12880-017-0185-9.
- [112] J.-Y. Hogrel, Y. Barnouin, N. Azzabou, *et al.*, “Nmr imaging estimates of muscle volume and intramuscular fat infiltration in the thigh: Variations with muscle, gender, and age,” *Age (Dordrecht, Netherlands)*, vol. 37, no. 3, p. 9798, 2015. DOI: 10.1007/s11357-015-9798-5.
- [113] B. Mohajer, M. Dolatshahi, K. Moradi, *et al.*, “Role of thigh muscle changes in knee osteoarthritis outcomes: Osteoarthritis initiative data,” *Radiology*, vol. 305, no. 1, pp. 169–178, 2022.
- [114] K. Edmunds, M. Gíslason, I. Arnadottir, A. Marcante, F. Piccione, and P. Gargiulo, “Quantitative computed tomography and image analysis for advanced muscle assessment,” *Eur J Transl Myol*, vol. 26, no. 2, p. 6015, 2016. DOI: 10.4081/ejtm.2016.6015.
- [115] J. J. 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.
- [116] S. Suthaharan and S. Suthaharan, “Support vector machine,” *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.
- [117] scikit-learn developers, *Support vector machines — scikit-learn documentation*, 2024. [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html#classification>.
- [118] scikit-learn developers, *L1 vs l2 regularization — scikit-learn documentation*, 2024. [Online]. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_l1_l2_sparsity.html.

- [119] L. Rokach and O. Maimon, "Decision trees," *The Annals of Statistics*, vol. 29, no. 5, 2001.
- [120] L. Breiman, "Random forests," *Machine Learning*, vol. 45, 2001.
- [121] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, 2001.
- [122] D. T. Larose and C. D. Larose, "K-nearest neighbor algorithm," in *Discovering Knowledge in Data: An Introduction to Data Mining*. 2014, pp. 149–164. DOI: 10.1002/9781118874059.ch7.
- [123] S. A. Hicks, I. Strümke, V. Thambawita, *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, Apr. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-09954-8. [Online]. Available: <http://dx.doi.org/10.1038/s41598-022-09954-8>.
- [124] D. Krstajic, L. Buturovic, D. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics*, vol. 6, 2014. DOI: 10.1186/1758-2946-6-10.
- [125] C. A. Holt and S. Sullivan, "Permutation tests for experimental data," *Experimental Economics*, pp. 1–38, 2023. DOI: 10.1007/s10683-023-09799-6.
- [126] K. Bennell, K. Bowles, Y. Wang, F. Cicuttini, M. Davies-Tuck, and R. Hinman, "Higher dynamic medial knee load predicts greater cartilage loss over 12 months in medial knee osteoarthritis," *Annals of the Rheumatic Diseases*, vol. 70, pp. 1770–1774, 2011. DOI: 10.1136/ard.2010.147082.
- [127] M. Creaby, M. Creaby, Y. Wang, *et al.*, "Dynamic knee loading is related to cartilage defects and tibial plateau bone area in medial knee osteoarthritis," *Osteoarthritis and cartilage*, vol. 18 11, pp. 1380–5, 2010. DOI: 10.1016/j.joca.2010.08.013.
- [128] A. Esrafilian, K. Halonen, C. Dzialo, *et al.*, "Effects of gait modifications on tissue-level knee mechanics in individuals with medial tibiofemoral osteoarthritis: A proof-of-concept study towards personalized interventions," *Journal of Orthopaedic Research*, vol. 42, pp. 326–338, 2023. DOI: 10.1002/jor.25686.
- [129] K. Halonen, C. Dzialo, M. Mannisi, M. Venäläinen, M. D. Zee, and M. Andersen, "Workflow assessing the effect of gait alterations on stresses in the medial tibial cartilage - combined musculoskeletal modelling and finite element analysis," *Scientific Reports*, vol. 7, 2017. DOI: 10.1038/s41598-017-17228-x.
- [130] H. Iijima, H. Iijima, T. Aoyama, *et al.*, "Subchondral plate porosity colocalizes with the point of mechanical load during ambulation in a rat knee model of post-traumatic osteoarthritis," *Osteoarthritis and cartilage*, vol. 24 2, pp. 354–63, 2016. DOI: 10.1016/j.joca.2015.09.001.

- [131] B. S. Crook, A. Collins, N. K. Lad, C. Spritzer, J. R. Wittstein, and L. DeFrate, "Effect of walking on in vivo tibiofemoral cartilage strain in acl-deficient versus intact knees.," *Journal of biomechanics*, vol. 116, p. 110210, 2020. DOI: 10.1016/j.jbiomech.2020.110210.
- [132] N. Brisson, E. Wiebenga, P. Stratford, *et al.*, "Baseline knee adduction moment interacts with body mass index to predict loss of medial tibial cartilage volume over 2.5 years in knee osteoarthritis," *Journal of Orthopaedic Research*, vol. 35, 2017. DOI: 10.1002/jor.23564.
- [133] G. Zhai, J. Pelletier, M. Liu, *et al.*, "Activation of the phosphatidylcholine to lysophosphatidylcholine pathway is associated with osteoarthritis knee cartilage volume loss over time," *Scientific Reports*, vol. 9, 2019. DOI: 10.1038/s41598-019-46185-w.
- [134] M. P. Jansen, T. Turmezei, K. Dattani, *et al.*, "Cartilage thickness distribution and its dependence on demographic, radiographic, and mri structural pathology in knee osteoarthritis-data from the imi-approach cohort.," *Skeletal radiology*, 2025. DOI: 10.1007/s00256-025-04907-4.
- [135] F. Roemer, D. Felson, J. Stefanik, *et al.*, "Heterogeneity of cartilage damage in kellgren and lawrence grade 2 and 3 knees: The most study.," *Osteoarthritis and cartilage*, 2022. DOI: 10.1016/j.joca.2022.02.614.
- [136] G. Blumenkrantz, R. Stahl, R. Stahl, *et al.*, "The feasibility of characterizing the spatial distribution of cartilage t(2) using texture analysis.," *Osteoarthritis and cartilage*, vol. 16 5, pp. 584–90, 2008. DOI: 10.1016/j.joca.2007.10.019.
- [137] W. Hou, J.-p. Zhao, R. He, *et al.*, "Quantitative measurement of cartilage volume with automatic cartilage segmentation in knee osteoarthritis," *Clinical Rheumatology*, vol. 40, pp. 1997–2006, 2020. DOI: 10.1007/s10067-020-05388-7.
- [138] F. Cicuttini, A. Wluka, Y. Wang, and S. Stuckey, "The determinants of change in patella cartilage volume in osteoarthritic knees.," *The Journal of rheumatology*, vol. 29 12, pp. 2615–9, 2002.
- [139] O. Baysal, T. Baysal, A. Alkan, Z. Altay, and S. Yoloğlu, "Comparison of mri graded cartilage and mri based volume measurement in knee osteoarthritis.," *Swiss medical weekly*, vol. 134 19-20, pp. 283–8, 2004. DOI: 10.4414/SMW.2004.10546.
- [140] I. Cohen, Y. Huang, J. Chen, *et al.*, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.
- [141] M. Cooke, B. Lawless, S. W. Jones, and L. Grover, "Matrix degradation in osteoarthritis primes the superficial region of cartilage for mechanical damage.," *Acta biomaterialia*, vol. 78, pp. 320–328, 2018. DOI: 10.1016/j.actbio.2018.07.037.

- [142] M. Danalache, R. Kleinert, J. Schneider, *et al.*, “Changes in stiffness and biochemical composition of the pericellular matrix as a function of spatial chondrocyte organisation in osteoarthritic cartilage,” *Osteoarthritis and Cartilage*, vol. 27, no. 5, pp. 823–832, 2019, ISSN: 1063-4584. DOI: <https://doi.org/10.1016/j.joca.2019.01.008>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458419300342>.
- [143] S. Saarakkala, P. Julkunen, P. Kiviranta, J. Mäkitalo, J. Jurvelin, and R. Korhonen, “Depth-wise progression of osteoarthritis in human articular cartilage: Investigation of composition, structure and biomechanics,” *Osteoarthritis and Cartilage*, vol. 18, no. 1, pp. 73–81, 2010, ISSN: 1063-4584. DOI: <https://doi.org/10.1016/j.joca.2009.08.003>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458409002106>.
- [144] P.-J. Wu, M. I. Masouleh, D. Dini, *et al.*, “Detection of proteoglycan loss from articular cartilage using brillouin microscopy, with applications to osteoarthritis,” *Biomed. Opt. Express*, vol. 10, no. 5, pp. 2457–2466, May 2019. DOI: 10.1364/BOE.10.002457. [Online]. Available: <https://opg.optica.org/boe/abstract.cfm?URI=boe-10-5-2457>.
- [145] S. Kubakaddi, K. Ravikumar, and D. Harini, “Measurement of cartilage thickness for early detection of knee osteoarthritis (koa),” in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)*, IEEE, 2013, pp. 208–211.
- [146] T. Cui, R. Liu, Y. Jing, J. Fu, and J. Chen, “Development of machine learning models aiming at knee osteoarthritis diagnosing: An mri radiomics analysis,” *Journal of Orthopaedic Surgery and Research*, vol. 18, no. 1, p. 375, 2023.
- [147] Y. Hu, X. Chen, S. Wang, Y. Jing, and J. Su, “Subchondral bone microenvironment in osteoarthritis and pain,” *Bone Research*, vol. 9, 2021. DOI: 10.1038/s41413-021-00147-z.
- [148] S. Mastbergen, A. Ooms, T. Turmezei, *et al.*, “Subchondral bone changes after joint distraction treatment for end stage knee osteoarthritis,” *Osteoarthritis and cartilage*, 2022. DOI: 10.1016/j.joca.2021.12.014.
- [149] G. Dai, H. Xiao, J. Liao, *et al.*, “Osteocyte $\text{tgf}\beta 1$ - $\text{smad}2/3$ is positively associated with bone turnover parameters in subchondral bone of advanced osteoarthritis,” *International Journal of Molecular Medicine*, vol. 46, pp. 167–178, 2020. DOI: 10.3892/ijmm.2020.4576.
- [150] Y.-R. Lee, D. Findlay, D. Muratovic, and J. Kuliwaba, “Greater heterogeneity of the bone mineralisation density distribution and low bone matrix mineralisation characterise tibial subchondral bone marrow lesions in knee osteoarthritis patients,” *Bone*, p. 115979, 2021. DOI: 10.1016/j.bone.2021.115979.

- [151] J. Hirvasniemi, J. Thevenot, V. Immonen, *et al.*, “Quantification of differences in bone texture from plain radiographs in knees with and without osteoarthritis,” *Osteoarthritis and cartilage*, vol. 22 10, pp. 1724–31, 2014. DOI: 10.1016/j.joca.2014.06.021.
- [152] K. N. Linde, K. Puhakka, B. Langdahl, *et al.*, “Bone mineral density is lower in patients with severe knee osteoarthritis and attrition,” *Calcified Tissue International*, vol. 101, pp. 593–601, 2017. DOI: 10.1007/s00223-017-0315-y.
- [153] W. Burnett, S. Kontulainen, C. McLennan, *et al.*, “Patella bone density is lower in knee osteoarthritis patients experiencing moderate-to-severe pain at rest,” *Journal of Musculoskeletal and Neuronal Interactions*, vol. 16, pp. 33–39, 2016.
- [154] J. Yang, P. Liu, S. Wang, *et al.*, “Causal relationship between sarcopenia and osteoarthritis: A bi-directional two-sample mendelian randomized study,” *European Journal of Medical Research*, vol. 28, p. 327, 2023. DOI: 10.1186/s40001-023-01322-0.
- [155] J. Liao, J. Chen, W. Xu, *et al.*, “Prevalence and associations of sarcopenia, obesity and sarcopenic obesity in end-stage knee osteoarthritis patients,” *Journal of Health, Population and Nutrition*, vol. 42, p. 108, 2023. DOI: 10.1186/s41043-023-00438-7.
- [156] L. Zhang, C. Zhang, J. Zhang, A. Liu, P. Wang, and J. Xu, “A bidirectional mendelian randomization study of sarcopenia-related traits and knee osteoarthritis,” *Clinical Interventions in Aging*, vol. 18, pp. 1577–1586, Sep. 2023. DOI: 10.2147/CIA.S424633.
- [157] M. A. Belzunce, J. Henckel, A. D. Laura, and A. Hart, “Intramuscular fat in gluteus maximus for different levels of physical activity,” *Scientific Reports*, vol. 11, no. 1, p. 21 401, Nov. 2021. DOI: 10.1038/s41598-021-00790-w.
- [158] M. G. Pedroso, A. C. de Almeida, J. B. Aily, *et al.*, “Fatty infiltration in the thigh muscles in knee osteoarthritis: A systematic review and meta-analysis,” *Rheumatology International*, vol. 39, pp. 627–635, 2019. DOI: 10.1007/s00296-019-04271-2.
- [159] D. Kumar, D. C. Karampinos, T. D. MacLeod, *et al.*, “Quadriceps intramuscular fat fraction rather than muscle size is associated with knee osteoarthritis,” *Osteoarthritis and Cartilage*, vol. 22, no. 2, pp. 226–234, 2014, ISSN: 1063-4584. DOI: 10.1016/j.joca.2013.12.005.
- [160] Z. Khatun *et al.*, “The role of muscle and tendon in predicting cartilage degeneration and tendinopathy,” in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)*, Rome, Italy, 2022, pp. 289–294. DOI: 10.1109/MetroXRINE54828.2022.9967653.

- [161] I. Vivodtzev, L. Moncharmont, R. Tamisier, *et al.*, “Quadriceps muscle fat infiltration is associated with cardiometabolic risk in copd,” *Clinical Physiology and Functional Imaging*, vol. 38, pp. 788–797, 2018. DOI: 10.1111/cpf.12481.
- [162] K. J. Edmunds, I. Arnadottir, M. K. Gislason, U. Carraro, and P. Gargiulo, “Nonlinear trimodal regression analysis of radiodensitometric distributions to quantify sarcopenic and sequelae muscle degeneration,” *Research Article*, 2016.
- [163] W. Chen, C. Li, Y. Wang, *et al.*, “Comparison of the asymmetries in muscle mass, biomechanical property and muscle activation asymmetry of quadriceps femoris between patients with unilateral and bilateral knee osteoarthritis,” *Frontiers in Physiology*, vol. 14, 2023, ISSN: 1664-042X. DOI: 10.3389/fphys.2023.1126116.
- [164] A. G. Culvenor, D. T. Felson, J. Niu, *et al.*, “Thigh muscle specific-strength and the risk of incident knee osteoarthritis: The influence of sex and greater body mass index,” *Arthritis Care & Research (Hoboken)*, vol. 69, no. 8, pp. 1266–1270, Aug. 2017. DOI: 10.1002/acr.23182.
- [165] I. Janssen, S. B. Heymsfield, Z. Wang, and R. Ross, “Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr,” *Journal of Applied Physiology*, vol. 89, no. 1, pp. 81–88, 2000. DOI: 10.1152/jappl.2000.89.1.81.
- [166] W. Visser, M. den Heijer, M. Reijnierse, R. de Mutsert, F. Rosendaal, and M. Kloppenburg, “Fri0323 the association of fat mass and skeletal muscle mass with knee oa: The neo study,” *Annals of the Rheumatic Diseases*, vol. 71, no. Suppl 3, pp. 423–423, 2013.
- [167] A. E. Peters, R. Akhtar, E. Comerford, and K. Bates, “The effect of ageing and osteoarthritis on the mechanical properties of cartilage and bone in the human knee joint,” *Scientific Reports*, vol. 8, 2018. DOI: 10.1038/s41598-018-24258-6.
- [168] S. Goldring, “Alterations in periarticular bone and cross talk between subchondral bone and articular cartilage in osteoarthritis,” *Therapeutic Advances in Musculoskeletal Disease*, vol. 4, pp. 249–258, 2012. DOI: 10.1177/1759720x12437353.
- [169] Y. Hu, Y. E. Yu, H. J. Cooper, *et al.*, “Mechanical and structural properties of articular cartilage and subchondral bone in human osteoarthritic knees,” *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, 2024. DOI: 10.1093/jbmr/zjae094.
- [170] S. Okada, M. Taniguchi, M. Yagi, *et al.*, “Degeneration of the cartilage quality is correlated with a higher intramuscular fat infiltration of the vastus medialis in older adults with pre-to-mild knee osteoarthritis,” *European journal of radiology*, vol. 183, p. 111930, 2025. DOI: 10.1016/j.ejrad.2025.111930.

- [171] M. Taniguchi, Y. Fukumoto, M. Yagi, *et al.*, "A higher intramuscular fat in vastus medialis is associated with functional disabilities and symptoms in early stage of knee osteoarthritis: A case-control study," *Arthritis Research & Therapy*, vol. 25, 2023. DOI: 10.1186/s13075-023-03048-0.
- [172] J. Xu, G.-r. She, T. Gui, *et al.*, "Knee muscle atrophy is a risk factor for development of knee osteoarthritis in a rat model," *Journal of Orthopaedic Translation*, vol. 22, pp. 67–72, 2019. DOI: 10.1016/j.jot.2019.10.003.
- [173] H. Zhang, T. Shuai, J. Wang, K. Li, and Y. Nie, "Relationship between 2 years of muscle strength decrease and volume loss of menisci and cartilage according to knee pain in mild and moderate knee osteoarthritis," *American Journal of Physical Medicine & Rehabilitation*, vol. 104, pp. 1–7, 2024. DOI: 10.1097/phm.0000000000002507.
- [174] Z. Gong, J. Li, Z.-J. He, *et al.*, "Quadriceps strength is negatively associated with knee joint structural abnormalities—data from osteoarthritis initiative," *BMC Musculoskeletal Disorders*, vol. 23, 2022. DOI: 10.1186/s12891-022-05635-9.
- [175] Materialise NV, *Mimics student edition course book*, Mimics Handbook, Student Edition - Version 21.0, Materialise, 2020. [Online]. Available: <http://uc.materialise.com/mimics>.
- [176] A. M. Ponsiglione, F. Angelone, F. Amato, and M. Sansone, "A statistical approach to assess the robustness of radiomics features in the discrimination of mammographic lesions," *Journal of Personalized Medicine*, vol. 13, no. 7, p. 1104, 2023.
- [177] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.
- [178] E. Belluzzi, S. Todros, A. Pozzuoli, P. Ruggieri, E. L. Carniel, and A. Berardo, "Human cartilage biomechanics: Experimental and theoretical approaches towards the identification of mechanical properties in healthy and osteoarthritic conditions," *Processes*, vol. 11, no. 4, 2023, ISSN: 2227-9717. DOI: 10.3390/pr11041014. [Online]. Available: <https://www.mdpi.com/2227-9717/11/4/1014>.
- [179] I. Otterness, M.-P. Le Graverand, and F. Eckstein, "Allometric relationships between knee cartilage volume, thickness, surface area and body dimensions," *Osteoarthritis and Cartilage*, vol. 16, no. 1, pp. 34–40, 2008, ISSN: 1063-4584. DOI: <https://doi.org/10.1016/j.joca.2007.05.010>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063458407001896>.
- [180] J. Everhart, M. M. Abouljoud, S. Poland, and D. Flanigan, "Medial compartment defects progress at a more rapid rate than lateral cartilage defects in older adults with minimal to moderate knee osteoarthritis (oa): Data from the oa initiative," *Knee Surgery, Sports Traumatology, Arthroscopy*, pp. 1–9, 2018. DOI: 10.1007/s00167-018-5202-1.

- [181] J. C. Stoddart, O. Dandridge, A. Garner, J. Cobb, and R. V. van Arkel, "The compartmental distribution of knee osteoarthritis - a systematic review and meta-analysis.," *Osteoarthritis and cartilage*, 2020. DOI: 10.1016/j.joca.2020.10.011.
- [182] Q.-S. Xu and Y.-Z. Liang, "Monte carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001, ISSN: 0169-7439. DOI: [https://doi.org/10.1016/S0169-7439\(00\)00122-2](https://doi.org/10.1016/S0169-7439(00)00122-2). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169743900001222>.
- [183] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?" *BMC Medical Research Methodology*, vol. 16, 2016. DOI: 10.1186/s12874-016-0200-9.
- [184] A. Chatzi and O. Doody, "The one-way anova test explained," *Nurse researcher*, vol. 32, no. 4, 2024.
- [185] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010.
- [186] A. Nanda, B. B. Mohapatra, A. P. Mahapatra, A. P. K. Mahapatra, and A. Mahapatra, "Multiple comparison test by tukey's honestly significant difference (hsd): Do the confident level control type i error," *International Journal of Statistics and Applied Mathematics*, 2021. DOI: 10.22271/MATHS.2021.V6.I1A.636.
- [187] A. Dinno, "Nonparametric pairwise multiple comparisons in independent groups using dunn's test," *The Stata Journal*, vol. 15, pp. 292–300, 2015. DOI: 10.1177/1536867X1501500117.
- [188] M. Cascella, A. M. Ponsiglione, V. Santoriello, *et al.*, "Expert consensus on feasibility and application of automatic pain assessment in routine clinical use," *Journal of Anesthesia, Analgesia and Critical Care*, vol. 5, no. 1, p. 29, 2025, doi:10.1186/s44158-025-00249-8.
- [189] A. Mahmoudian, L. Lohmander, H. Jafari, and F. Luyten, "Towards classification criteria for early-stage knee osteoarthritis: A population-based study to enrich for progressors.," *Seminars in arthritis and rheumatism*, vol. 51 1, pp. 285–291, 2020. DOI: 10.1016/j.semarthrit.2020.11.002.
- [190] D. Ünalmiş, N. Acer, S. Yilmaz, A. Tokpınar, S. Doğan, and H. Demir, "The calculation of the femoral condyle cartilage volume and surface area in patients with osteoarthritis," *erciyes medical journal*, 2020. DOI: 10.14744/etd.2020.75768.
- [191] C. Ding, P. Garner, F. Cicuttini, F. Scott, H. Cooley, and G. Jones, "Knee cartilage defects: Association with early radiographic osteoarthritis, decreased cartilage volume, increased joint surface area and type ii collagen breakdown.," *Osteoarthritis and cartilage*, vol. 13 3, pp. 198–205, 2005. DOI: 10.1016/J.JOCA.2004.11.007.

- [192] H. Bonakdari, J. Pelletier, F. Abram, and J. Martel-Pelletier, "A machine learning model to predict knee osteoarthritis cartilage volume changes over time using baseline bone curvature," *Biomedicines*, vol. 10, 2022. DOI: 10.3390/biomedicines10061247.
- [193] M. Finnilä, S. D. Gupta, M. Turunen, *et al.*, "Mineral crystal thickness in calcified cartilage and subchondral bone in healthy and osteoarthritic human knees," *Journal of Bone and Mineral Research*, vol. 37, pp. 1700–1710, 2022. DOI: 10.1002/jbmr.4642.
- [194] J. Legrand, C. Marzin, T. Neogi, L. Norberciak, J. Budzik, and T. Pascart, "Associations of changes in knee hyaline cartilage composition measured with dual-energy computed tomography in gout, aging and osteoarthritis," *Cartilage*, vol. 15, pp. 283–292, 2023. DOI: 10.1177/19476035231172152.
- [195] A. Linus, P. Tanska, E. Nippolainen, *et al.*, "Site-specific elastic and viscoelastic biomechanical properties of healthy and osteoarthritic human knee joint articular cartilage.," *Journal of biomechanics*, vol. 169, p. 112 135, 2024. DOI: 10.1016/j.jbiomech.2024.112135.