Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Development of Radiomics Workflow for Head and Neck Cancer Patients Prognostic Models

Master's Final Degree Project

Ugnė Balčiūnaitė

Project author

assoc. prof. dr. Benas Gabrielis Urbonavičius

Supervisor

Kaunas, 2025

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Development of Radiomics Workflow for Head and Neck Cancer Patients Prognostic Models

Master's Final Degree Project

Medical Physics (6213GX001)

Ugnė Balčiūnaitė

Project author

assoc. prof. dr. Benas Gabrielis Urbonavičius

Supervisor

assoc. prof. dr. Teresa Moskaliovienė

Reviewer

Kaunas, 2025

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Author's name and surname

Development of Radiomics Workflow for Head and Neck Cancer Patients Prognostic Models

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Ugnė Balčiūnaitė / *Confirmed electronically*

Balčiūnaitė Ugnė. Development of Radiomics Workflow for Head and Neck Cancer Patients Prognostic Models. Master's Final Degree doc. dr. Benas Gabrielis Urbonavičius; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Health sciences, (Medical technologies).

Keywords: delta radiomics, machine learning, prognostic model.

Kaunas, 2025. 78 pages.

Summary

Medical physicists play an important role in integrating imaging, treatment planning, and therapy monitoring into cancer care. With the increased availability of high-resolution medical imaging and advanced computational tools, radiomics has emerged as a potent non-invasive method for quantifying tumor features. Delta radiomics, which examines changes in imaging features during or after therapy, sheds light on treatment-induced biological impacts. However, the practical application of delta radiomics in ordinary clinical workflows is still limited, primarily due to the complexity of data processing, diversity in radiomic feature extraction procedures, and the lack of defined, clinically validated implementation paths.

The aim of this master's thesis was to create and test a reproducible delta radiomics-based process that would assist medical physicists in prognostic modelling for patients with head and neck cancer. Rather than focusing solely on prediction accuracy, the goal was to develop a simple and adaptable approach for clinical usage. In this work, pre- and post-treatment medical imaging data were used to identify delta radiomic features to verify the performance of the developed workflow. Two independent machine learning models were created: one to determine which medication causes the greatest radiomic alterations, and another to discover variables linked with patient survival. These modelling tasks not only revealed the workflow's power to recognize clinically meaningful patterns but also emphasized its potential to discover non-invasive imaging biomarkers for treatment monitoring and survival prediction.

The developed workflow includes key steps, such as balancing the dataset using the synthetic oversampling method SMOTE, correlation analysis to reduce over-sampling and feature selection using recursive feature removal with Random Forest and XGBoost and evaluating the performance of these methods. The CatBoost method is then used to build classification models based on the given features. Finally, Mann-Whitney U, Kruskal-Wallis and Dunn post hoc tests are used to assess the statistical significance and discriminatory power of the features included in the final models.

This thesis advances medical physics by proposing a robust, useable delta radiomics methodology for developing imaging-based prognostic models. It emphasizes the critical role of medical physicists in enabling data-driven, individualized treatment evaluation and response assessment. By increasing the incorporation of quantitative imaging biomarkers into clinical radiation workflows, this study adds to the ongoing enhancement of customized oncology care and the employment of AI techniques in routine clinical practice.

Balčiūnaitė Ugnė. Radiomikos darbo eigos kūrimas galvos ir kaklo vėžiu sergančių pacientų prognozavimo modeliams. Magistro baigiamasis projektas vadovas doc. dr. Benas Gabrielis Urbonavičius; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Sveikatos mokslai, (Medicinos technologijos).

Reikšminiai žodžiai: delta radiomika, mašininis mokymasis, prognostinis modelis.

Kaunas, 2025. 78 p.

Santrauka

Medicinos fizikai atlieka itin svarbų vaidmenį integruojant vaizdinimo technologijas, gydymo planavimą ir terapijos stebėseną į visapusišką onkologinio gydymo procesą. Didėjant aukštos raiškos medicininių vaizdų ir pažangių skaitmeninių analizės įrankių prieinamumui, radiomika tampa vis svarbesniu neinvaziniu metodu, leidžiančiu kiekybiškai įvertinti navikų savybes. Delta radiomika, tirianti vaizdavimo požymių pokyčius gydymo metu arba po jo, suteikia reikšmingų įžvalgų apie biologinį atsaką į terapiją. Nepaisant to, praktinis delta radiomikos taikymas klinikinėje veikloje tebėra ribotas dėl sudėtingų duomenų apdorojimo etapų, nevienodų požymių išskyrimo metodologijų bei standartizuotų ir kliniškai patvirtintų įgyvendinimo gairių stokos

Šio magistro darbo tikslas - sukurti ir išbandyti atkuriamą delta radiomikos procesą, kuris padėtų medicinos fizikams modeliuoti galvos ir kaklo vėžiu sergančių analizės atvejus. Siekiant užtikrinti ne tik prognostino modelio tikslumą, bet ir praktinę naudą, sukurta paprasta, klinikinėje aplinkoje lengvai pritaikoma radiomikos modelio kūrimo eiga. Šiame tyrime radiomikos požymiams nustatyti buvo naudojami medicininiai vaizdai atlikti prieš gydymą ir po jo, siekiant patikrinti sukurto darbo eigos efektyvumą. Jos išbandymui sukurti du nepriklausomi mašininio mokymosi modeliai: vienas skirtas nustatyti, kuris vaistas sukelia didžiausius radiominius pokyčius, o kitas - atrasti kintamuosius, susijusius su paciento išgyvenamumu. Šios modeliavimo užduotys ne tik atskleidė darbo eigos galią atpažinti kliniškai reikšmingus modelius, bet ir pabrėžė jos potencialą atrasti neinvazinius vaizdinimo biomarkerius, skirtus gydymo stebėsenai ir išgyvenamumo prognozavimui.

Šiame darbe siūlomas metodas apima sistemingą analizės eigą, kurią sudaro duomenų rinkinio subalansavimas taikant sintetinio duomenų generavimo technikas (pvz., SMOTE), duomenų filtravimas naudojant koreliacinę analizę, skirtas sumažinti perteklinių požymių įtaką bei požymių atranka, atliekama rekursyvaus eliminavimo metodu pasitelkiant „Random Forest“ ir „XGBoost“ algoritmus. Atrinkti požymiai buvo panaudoti kuriant klasifikavimo modelius, pagrįstus „CatBoost“ algoritmu. Siekiant įvertinti galutinai atrinktų požymių statistinį reikšmingumą ir jų diskriminacinę galią, buvo taikomi Mann-Whitney U, Kruskal-Wallis ir Dunno post hoc testai.

Šiuo tyrimu siekta prisidėti prie medicinos fizikos srities pažangos, pateikiant patikimą ir praktikoje pritaikomą delta radiomikos darbo eigą, skirtą vaizdavimo duomenimis grįstų prognostinių modelių kūrimui. Šiame darbe akcentuojama medicinos fizikų reikšmė diegiant duomenimis grįstą ir individualizuotą gydymo atsako vertinimą klinikinėje aplinkoje. Tobulinant kiekybinių vaizdinių biologinių žymenų integravimą į radioterapijos klinikinę praktiką. Be to, šis tyrimas prisideda prie personalizuotos onkologinės priežiūros tobulinimo bei dirbtinio intelekto metodų diegimo kasdienėje medicinos praktikoje.

Table of contents

List of figures **7**
List of tables **8**
List of abbreviations **9**
Introduction **10**
1. Literature review **12**
1.1. History of Radiomics 12
1.2. Steps of Radiomics process 18
1.2.1. Image acquisition 18
1.2.2. Image preprocessing 19
1.2.3. Region of interest segmentation. 19
1.2.4. Feature extraction. 20
1.2.5. Feature analysis 21
1.3. Fundamental limitations and potential improvements in the Radiomics process 22
1.3.1. Lack of Standardization 22
1.3.2. Small and Heterogeneous Datasets 23
1.3.3. Reproducibility and Validation 24
1.3.4. Clinical integration 25
1.3.5. Overfitting in Machine Learning Models 26
1.4. Radiomics in Medical Physics field 27
2. Methods **32**
2.1. Data set 32
2.2. Radiomics workflow 33
2.2.1. Verifying data's balance 33
2.2.2. Synthetic data generation 33
2.2.3. Selection of significant variables 35
2.2.4. Development of prognostic model 40
3. Results **44**
3.1. Analysis of Delta Features Changes Across Treatment Types 44
3.2. Delta Radiomic Features as Predictors of Survival 51
Conclusions **59**
Recommendations **60**
References **62**
Appendices **79**

List of figures

Fig. 1. An example of how the multiparametric MRI framework works when multiple pathologies are analyzed[21]. 13
Fig. 2. representation of five distinct types of mpRAD framework features using both first and second order statistical analysis [21]. 14
Fig. 3. Radiomic feature maps from a patient with a cancerous lesion were produced using both single and multiparametric methods [21]. 15
Fig. 4. The left image shows a Spearman ranking of delta-radiomics features, whereas the right image shows a heatmap of the same features[24]. 16
Fig. 5. Comparison of the accuracy of three models: dynamic, static, and ResNet18 + LSTM [25]. 17
Fig. 6. Illustration of Radiomics process workflow [40]. 18
Fig. 7. DCA of prediction models built from training and validation datasets (a, b) [95] 28
Fig. 8. Random Forest algorithm scheme [134]. 36
Fig. 9. Schematical representation of eXtreme Gradient Boost [136]. 37
Fig. 10. Illustration of ROC curves for each class using One-vs-Rest ROC strategy [155] 39
Fig. 11. Confusion matrix examples for binary and multiclass classification problems [159]. 40
Fig. 12. Illustration of Categorical Boost approach [162] 41
Fig. 13. Redundancy reduction using pairwise correlation analysis (The higher scale correlation matrices are shown in Appendix 1.) 45
Fig. 14. Macro-average ROC curves comparing multi-class model performance using RFE-RF and XGBoost feature selection methods. 46
Fig. 15. Significant delta radiomic features identified by the CatBoost model for treatment type classification. 47
Fig. 16. Performance of the CatBoost approach 48
Fig. 17. Variation in significant delta radiomics features across different treatment types 49
Fig. 18. The boxplot of delta radiomic features across chemotherapy treatment types with Kruskal–Walli's test results. 50
Fig. 19. Redundancy reduction by using correlation analysis (The higher scale correlation matrices are shown in Annex 1.) 52
Fig. 20. Comparison of the performance of two ML approaches using ROC Curves 53
Fig. 21. The most significant features by CatBoost 54
Fig. 22. The ROC curve for CatBoost model in binary classification 55
Fig. 23. Variation in significant delta radiomics features across survival outcomes 56
Fig. 24. Mann–Whitney U test for delta features differentiating deceased and surviving patients . 57

List of tables

Table 1. Traditional Radiomics features [60] 21
Table 2. Clinical data 32
Table 3. Categories of main delta features 32
Table 4. Comparison of different data generating techniques [118]. 34
Table 5. Interpretation of model's performance based on AUC value [149, 150]. 38
Table 6. Evaluation metrics [158–160]. 40
Table 7. The frequencies of target variable 44
Table 8. Performance metrics of feature selection algorithms 45
Table 9. Performance metrics of CatBoost algorithm. 48
Table 10. Statistical comparison of delta features across different treatment types. 51
Table 11. The frequencies of the target variable 52
Table 12. A comparison of performance metrics of feature selection algorithms. 52
Table 13. Performance metrics to evaluate the CatBoost approach. 55
Table 14. Statistical significance of feature differences between survived and deceased patients . 57

List of abbreviations

Abbreviations:

CNN – Convolutional Neural Network

mpRAD – Multiparametric Radiomics

mpMRI – Multiparametric Magnetic Resonance Imaging

DRF – Delta Radiomics Features

ROI – Region of Interest

ADC – Apparent Diffusion Coefficient

SVM – Support Vector Machine

ROC – Receiver Operating Characteristic

FNN – Feedforward Neural Network

LDA – Linear Discriminating Analysis

LASSO – Least Absolute Shrinkage and Selection Operator

RNN – Recurrent Neural Network

LSTM – Long-Short-Term Memory

AUC – Area Under the Curve

FBP – Filtered Back Projection

IFT – Inverse Fourier Transform

IR – Iterative Reconstruction

DLCT – Dual-Layer Computed Tomography

VOI – Volume of Interest

ADASYN – Adaptive Synthetic Sampling

HNC – Head and Neck Cancer

DCA – Decision Curve Analysis

PCA – Principal Component Analysis

Introduction

Cancer is an all-encompassing term for a wide-ranging and multidimensional disease characterized by the uncontrolled and persistent development of aberrant cells. This destructive process impairs the structural and functional integrity of the surrounding tissues and can spread systemically via hematogenous and lymphatic channels, eventually resulting in the production of new malignant tumors known as metastases [1, 2]. Currently, the diagnosis and treatment of carriage disease is regarded as a serious health problem, since it accounts for one in six fatalities worldwide, and rates of mortality continue to rise fast despite substantial advancements in early detection and treatment [3, 4]. The cornerstone of effective oncology is an accurate diagnosis, which enables individual patient treatment planning, comprehensive risk assessment and evidence-based clinical decision-making [5, 6]. Traditional prognostic models, on the other hand, have limitations in terms of characterizing tumor biology because they primarily depend on clinical and histological characteristics that might not accurately represent the complex and heterogeneous tumor structure. These obstacles can diminish the ability to correctly forecast illness development and response to treatment, resulting in suboptimal treatment changes and poor patient outcomes [7–9].

With rapid advances in medical imaging and computational analysis, radiomics are now recognized as an effective method for improving cancer prognosis. Radiomics is the extraction of high-dimensional quantitative features from different imaging modalities, transforming medical images into data capable of identifying clinically meaningful biomarkers and improving diagnostic and prognostic modeling [10, 11]. These features quantify a variety of tissue heterogeneity, texture, shape, and intensity using numerical values derived from pixel (2D) or voxel (3D) intensity levels (bits) [12]. However, one significant problem in radiomics is the susceptibility of these quantitative aspects to differences in imaging methods and acquisition conditions. Given that these radiomic features are fundamentally numerical representations of image intensity distributions, variations in imaging procedures, such as algorithms used for reconstruction, image acquisition settings, or scanner types, can cause significant discrepancies in feature values, even for identical anatomical regions. This heterogeneity undermines the repeatability and dependability of radiomics-based models, emphasizing the necessity for rigorous standardization of imaging processes, strong feature selection techniques, and extensive validation to assure clinical application [13–15].

To improve the superiority and predictive accuracy of static radiomic analysis while also addressing existing weaknesses, a new branch of delta radiomics was established relatively recently and has already proven to be a viable technology. Unlike traditional radiomics, which examines features at a single time point, delta radiomics focuses on quantitative changes in radiomic features across time between two or more imaging time points, such as before, during, or after therapy. These changes may indicate minor biological reactions to treatment, such as tumor reduction, necrosis, or changes in tissue heterogeneity, which may not be visible but can be detected quantitatively. As a result, delta characteristics provide extra information on tumor dynamics, making them useful biomarkers for predicting therapy response and patient outcomes [16]. The final master's project will leverage delta radiomics elements to create a thorough methodology for developing a prediction model for head and neck cancer cases. To demonstrate how the workflow employing delta radiomics features performs, two models were developed: one to evaluate which treatment type causes the most significant changes, and another to find which features are connected with patient survival outcomes.

Aim: to build a workflow of developing a prognostic model using delta radiomics features in head and neck cancer patients.

Tasks:

- to perform a comprehensive analysis of the evolution of radiomics and critically evaluate each stage of the radiomics workflow, including its methodological limitations.
- to design and implement a delta radiomics workflow adapted to the imaging and clinical characteristics of head and neck cancer patients.
- to validate a machine learning–based prognostic model utilizing the designed delta radiomics features workflow.

1. Literature review

1.1. History of Radiomics

Radiomics has its beginnings in the early 1970s, when scientists first experimented with texture analysis for image classification. To be more specific, radiomics began in 1973, when some researchers proposed using texture features to identify images. Later on, in 1995, researchers began utilizing convolutional neural networks (CNNs) to identify lung nodules, indicating that computer algorithms could be trained to recognize medical images [17, 18]. In the late 2000s, researchers attempted to establish a correlation between tumor images and genetic types. At the time, most of the research was conducted on very small datasets and lacked external validation, which meant that the identified radiomic patterns were based on just tiny datasets from individual organizations and could not be validated by data from other organizations. Radiomics in oncology has progressed significantly as medical imaging technologies have advanced [19]. In 2012, Lambin coined the term radiomics to characterize the extraction of physiologically relevant quantitative information from medical imaging. He proposed that radiomics features (RF), which are invisible components of the tissue infrastructure of imaged objects, could be a useful tool for studying cancer with computed tomography (CT), magnetic resonance imaging (MRI), and other techniques. According to Lambino, such imaging investigations are easily replicable and allow for in vivo viewing and quantitative measurement of RF across the scanned array. As a result, they could help to advance a personalized precision medicine approach to cancer diagnosis in each patient, as well as serial assessment and prediction of therapy response. Nonetheless, radiomics has not gained widespread acceptance as a reliable component of cancer assessment. The complexity of radiomics and its validation caused numerous concerns at the time, including the reproducibility and generalizability of texture analysis and other critical components of the radiomics signature [18].

Robert A. Gillies, Patricia E. Kinahan, and Hedvig Hricak published a study in 2014 entitled “Decoding Tumor Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach”. This publication was the first scientific paper describing the combined use of radiomics and machine learning. This research is critical in the field of radiomics, which entails extracting many quantitative variables from medical images to aid tumor characterization, prognosis, and treatment prediction. According to the scientists, radiomics is found on the notion that medical images are more than just images; they contain quantitative data that indicate the underlying biology of cancer. Researchers can study patterns in photographs that are invisible to the naked eye by extracting characteristics using computer approaches. This paper sought to show how quantitative information collected from imaging can be used to decipher tumor phenotypes noninvasively. The researchers hoped that radiomics had the potential to bridge the gap between imaging and genetics, which would lead to personalized treatment recommendations based on the extracted radiomics features. During the study, the researchers aimed to extract a wide range of parameters from imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). During the study, the researchers found that malignant tumors are often heterogeneous, i.e. they consist of compartments with different biological characteristics. Radiomics can assist quantify this heterogeneity, which can provide valuable information regarding the tumor's aggressiveness and likelihood of responding to various treatments. The research also examined how quantitative imaging features can serve as imaging biomarkers. These indicators have the potential to replace invasive tissue samples in the assessment of tumor features. In a 2014 paper, researchers stated that radiomic features can be used for a variety of clinical applications, including disease prognosis, predictive

modelling of cancer yield to treatment, and the use of radiomic features to categorize patients into different risk categories based on their tumor's radiomic profile [20].

In 2018, a Multiparametric radiomics framework (mpRAD) was introduced by authors Parekh V. S. and Jacobs M.A. (2018). It was designed to extract radiomic features from high-dimensional imaging datasets. Basically, it means that this framework allows us to extract features from multiple types of imaging information, for instance CT, MRI or PET, while standard radiomics frequently uses single imaging modalities, which may not adequately capture the complex tissue features found in diverse disorders. Th mpRAD overcomes this restriction by combining different imaging characteristics, allowing for a more thorough examination of tissue heterogeneity. Several examples of how mpRadiomics looks by combining several imaging techniques can be seen in Fig. 1 [21].

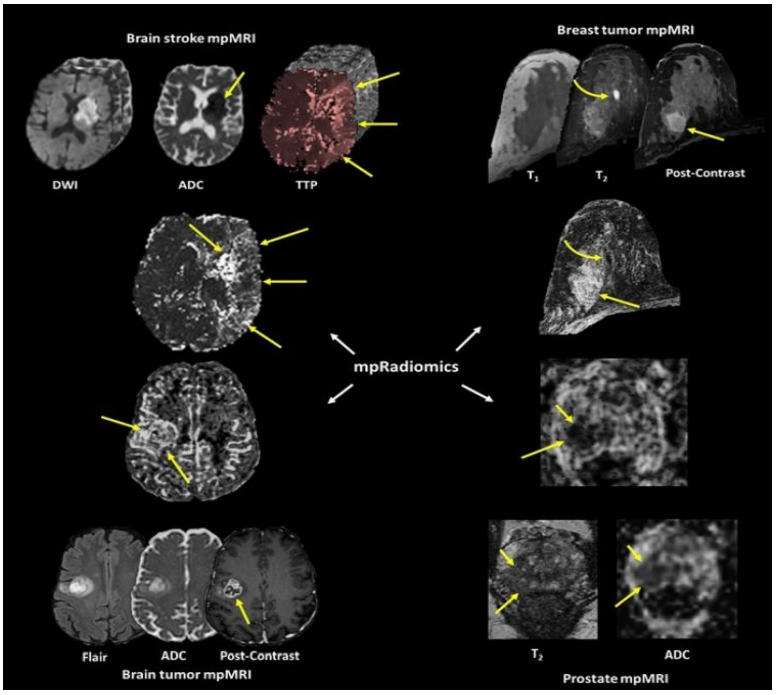


Fig. 1. An example of how the multiparametric MRI framework works when multiple pathologies are analyzed [21].

In this study scientists applied mpRAD to several distinct clinical scenarios, from which one was related to breast cancer. In this clinical breast cancer dataset was with one hundred and thirty-eight patients with breast lesions, which were scanned using multi parametric MRI (mpMRI). The authors were seeking to check if the frameworks can process multiparametric MRI data for improved classification and quantification of tissue characteristics while addressing computational efficiency and statistical validation. Fig 2 depicts the features of five different types of mpRAD systems, using both first- and second-order statistical analysis. The left side shows the formation of normal breast tissue signatures for damaged and healthy tissue. On the right side, the mpRAD features are outlined as the first-order statistics of the radiomic tissue signatures, the probability matrix of the tissue signatures, and the features of the covariance matrix of the tissue signatures, which assess the complex interaction between different tissue signatures. The complex inter-parametric interactions are estimated using the first-order statistics of the complex interaction network of tissue signatures as well as the characteristics of the tissue attribute connection matrix. Two sorts of yellow arrows can

also be seen. In this situation, straight arrows show lesion tissue, and curved yellow arrows indicate glandular tissue.

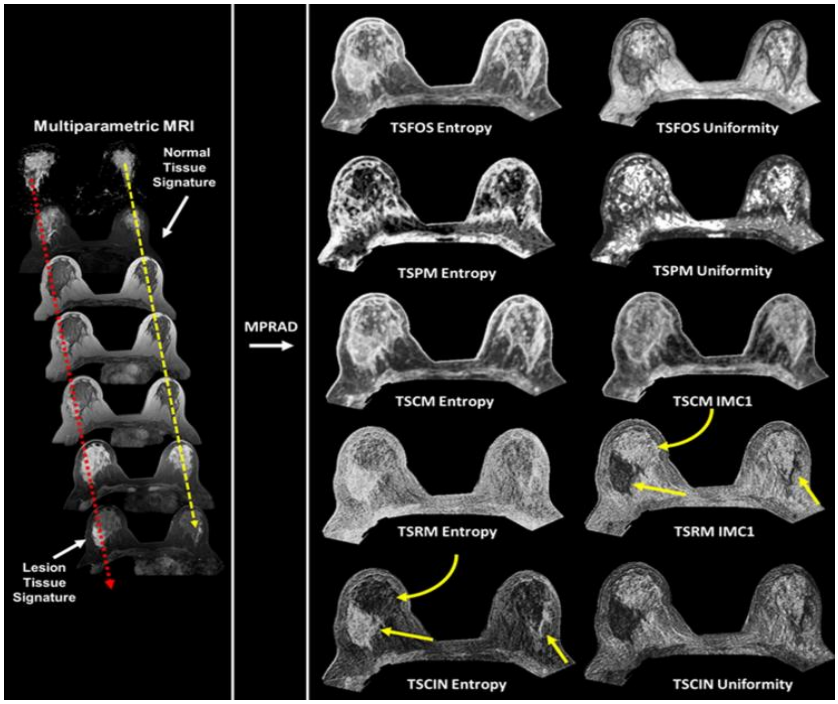


Fig. 2. representation of five distinct types of mpRAD framework features using both first and second order statistical analysis [21].

Using multiparametric radiomics analysis, isoSVM (a mix of Isomap and SVM), and statistical analysis, the authors demonstrated substantial variations between multiparametric MRI, single parametric radiomics, and multiparametric radiomics. Fig 3 shows one of the examples obtained. It is also worth noting that the patients who participated in the study underwent a biopsy to confirm the results. Of the 138 individuals, 97 had biopsy-proven cancer, while 41 had benign tumors. According to the authors, the apparent diffusion coefficient ADC map and Pharmacokinetic contrast enhancement parameters (PK-DCE) scores differed considerably between benign and malignant tumors. Fig. 3 depicts single and mpRAD feature maps from a typical patient with a malignant tumor in the upper extremity of the right breast and a benign-appearing cyst situated medial to the lesion (curved yellow arrow). As stated by the researchers, the cyst appears equally visible on the T2 and ADC maps, which is consistent with known MRI tissue features associated with cysts. The cyst is likewise dark at T1 and has negative enhancement of contrast on the dynamic contrast enhanced DCE images, indicating a lack of blood vessels. Furthermore, the lesion tissue looks heterogeneous on MRI imaging, with lower ADC values and higher PK-DCE features. Individual radiological images contain some textural elements, however when viewed alongside mpRAD radiological images, the textural portrayal of normal and diseased tissue is significantly different. According to the researchers' findings, the cyst has a lower entropy in mpRAD than in single radiological images. The cyst's reduced entropy reflects the observation that a homogeneous structure has less clutter, resulting in lower entropy. This is clear when inspecting a heterogeneous lesion with greater entropy levels. As a result, mpRAD can discriminate between the qualities of healthy and injured tissues, allowing for a more in-depth investigation of their structure and a comprehension of what modifications are occurring in the human body [21].

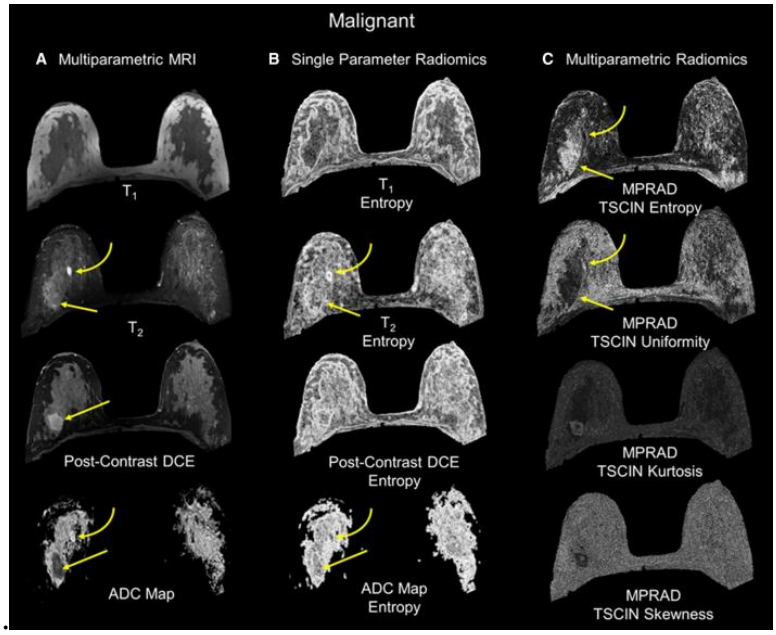


Fig. 3. Radiomic feature maps from a patient with a cancerous lesion were produced using both single and multiparametric methods [21].

2019, Delta-radiomics, which is known as a revolutionary technique to medical imaging and oncology was created. The initial research that introduced this notion cantered on using machine learning approaches to examine changes in radiomic characteristics over time, particularly in response to treatments like chemoradiotherapy. Several studies confirmed delta-radiomics' capacity to predict treatment responses by examining the net changes in radiomic characteristics collected from longitudinal imaging data of pancreatic cancer patients [22]. Since its inception, delta-radiomics has grown in popularity and has been utilized in several trials to discriminate between pre-cancerous and invasive diseases, to assess immunotherapy response, and to enhance the accuracy of predicting cancer treatment outcomes [23]. Nasief H., Zheng C., and colleagues (2019) propose that variations in radiomic properties over time on longitudinal images, known as delta radiomics, can be utilized as a biomarker to predict response to treatment. The researchers verified the process's efficiency by retrospectively evaluating daily non-contrast CT scans from 90 pancreatic cancer patients who had standard CT and chemoradiotherapy. A total of 2 520 CT scans (28 daily fractions per patient) and their pathological outcomes were examined. The study also involved ROI segmentation, which yielded over 1 300 radiomic characteristics. The researchers used Spearman correlations to obtain highly correlated delta-radiomic features (DRF). Results of correlation are depicted in Fig. 4. The researchers utilized linear regression models to establish correlations between the selected DRFs and pathological response, and the authors employed a t-test and linear mixed-effects models to identify which DRFs showed a significant difference from the first portion.

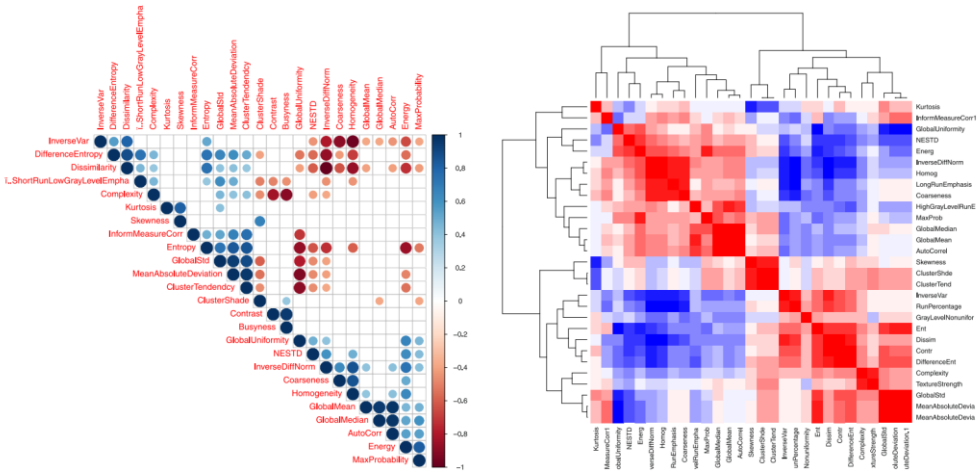


Fig. 4. The left image shows a Spearman ranking of delta-radiomics features, whereas the right image shows a heatmap of the same features [24].

To create the liability prediction model, the researchers used Bayesian neural networks. The model was trained on 50 patients using leave-one-out cross-validation. Linear regression models were employed to find associations between specific DRFs and pathological reactions. The T-test and linear mixed-effects methods were used to determine which DRFs differed significantly from the initial proportion. A Bayesian neural network was used to build the liability prediction model. The model was trained with 50 patients using leave-one-out cross-validation. The area under the curve known as the ROC curve was used to evaluate performance. External verification was conducted using data from the final 40 cases. The data revealed that 13 DRFs underwent the tests and showed significant alterations after 2-4 weeks of treatment. The researchers found that the normalized entropy, standard deviation difference, kurtosis, and roughness indices performed best in distinguishing between positive and negative responders (CV-AUC = 0.94). The researchers believe that with more research and larger data sets, delta radiomics might emerge into a biomarker for early treatment response prediction [24].

Dynamic radiomics was discovered in 2021 as an expansion of classic static radiomics to incorporate a temporal dimension in the paper "A new methodology to extract quantitative time-related features from tomographic images" by Qu H., Shi R., and colleagues [25]. This concept appears to be quite similar to delta radiomics, however there is one significant difference between the two techniques. Delta radiomics, as previously discussed, examines changes in features between two discrete time points (e.g., pre-treatment and post-treatment), whereas dynamic radiomics extract and model's continuous temporal changes from multiple time points to provide a more detailed assessment of disease progression. The main reason for developing dynamic radiomics was to solve the limitations of static radiomics in evaluating sickness progression. Cancer is a dynamic disease that evolves over time, therefore static images are insufficient to depict the changes. In this paper, the researchers present a novel dynamic radiomics feature extraction approach that uses time-dependent tomographic images of a single patient, emphasizing changes in image features over time and characterizing them as distinct dynamic characteristics for diagnostic or prognostic purposes. To compare the accuracy of dynamic radiomics to static radiomics, investigator-selected dynamic features were tested in three different clinical scenarios. Breast cancer axillary lymph node metastasis prognosis, colorectal cancer liver metastasis gene mutation status prognosis, and neoadjuvant chemotherapy efficacy prognosis - to evaluate the performance of the selected dynamic features to traditional 2D and 3D static features.

The researchers applied a variety of machine learning techniques, including Feedforward Neural Network (FNN), Random Forest (RF), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). The study also employed the Least Absolute Shrinkage and Selection Operator (LASSO), which was utilized to select features, minimize data dimension, and evaluate the effectiveness of ROC curve models. The Friedman test was employed to compare the performance of different models (with the same type of dynamic features). In addition, a two-layer time-dependent Recurrent Neural Network (RNN) model known as LSTM was created and compared to the study's dynamic model. The LSTM model was applied to both the original images and the extracted image features, which were processed using ResNet18 (pre-trained on the ImageNet dataset). The study compared these different machine learning algorithms using various feature sets, such as static and dynamic characteristics, to determine which combinations gave the highest accuracy for each of the clinical prediction tasks. These findings revealed that models constructed using dynamic features outperformed static models, especially if discrete features were included in analysis. As shown in Fig. 5, the dynamic model outperforms the other two models by a wide margin, especially for breast and bowel cancer cases, with accuracies of 0.9 and 0.8, but slightly below the neoadjuvant chemotherapy case, where the accuracy is just 0.65. ResNEt18+LSTM results are mixed. It can be observed that the neoadjuvant case has the highest accuracy, above 0.7, but it is possible to observe that the results are more variable. The static model performs the worst in this study, since the accuracy of all cases is below 0.7.

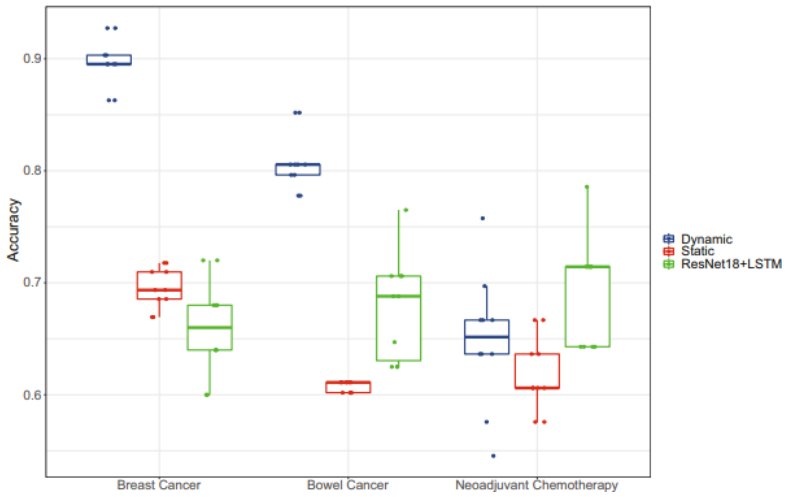


Fig. 5. Comparison of the accuracy of three models: dynamic, static, and ResNet18 + LSTM [25].

The combined use of radiomics and artificial intelligence has the potential to significantly improve personalized and precision medicine, enhancing the latter's accuracy and decision-making process. The combination of these complimentary approaches will open new possibilities regarding the treatment of a variety of diseases, including early identification to personalized therapy planning [26]. One of the primary benefits of merging radiomics along with artificial intelligence is the development of more accurate and reliable prediction models, which may lead to more frequent and dependable use of prediction models soon for treatment of patients [27, 28]. Early and more precise diagnosis, better risk stratification, and more informed treatment decisions can all lead to better patient outcomes [29]. Integrating radiomics data with additional biomarkers such as genetic, clinical, and demographic data allows AI-based predictive models to provide a deeper view of each patient's unique characteristics [30]. This can help to build a fully personalized method for healthcare, with diagnostic, prognostic, and therapeutic procedures adapted to each patient's unique traits [31]. However, in this

situation, AI models offer both significant advantages and limits that must be addressed in the future. One of the primary issues is that training effective models necessitates huge, high-quality databases, which are frequently challenging to come by in the industry of medical imaging. There is also a risk of overfitting, which occurs when models perform effectively with training data but badly on new, untested data [32, 33]. Furthermore, the complexity and absence of disclosure of AI algorithms, particularly deep learning models, make it challenging for clinicians to evaluate and trust the results. Data privacy and security considerations impede the incorporation of AI into the therapeutic environment since sensitive patient information must be secured. Finally, regulatory constraints and the necessity for defined protocols may impede the use of based on artificial intelligence radiomics in normal practice[34]. However, by addressing the root causes and devising solutions, as well as incorporating artificial intelligence and radiomics as an essential tool for diagnosis into daily medical practice, clinical choice-making and medical imaging could be elevated to new heights [35]. Medical professionals and researchers can use quantitative imaging and powerful computational approaches to open new paths for early disease identification, more accurate risk classification, and personalized therapy planning. This will eventually lead to a more precise and personalized healthcare system. As these fields progress, there is huge potential to enhance patient outcomes and transform healthcare delivery [34, 36–38].

1.2. Steps of Radiomics process

Radiomics is the process of translating medical images (such as CT scans, MRIs, or X-rays) into numbers with the intention to detect hidden patterns that clinicians may miss with the human eye. Detailed features are retrieved from the images to define the texture, shape, and intensity of the tissue or mass, which can aid in illness prediction, treatment response assessment, and patient outcomes[39]. Radiomics analysis is a multi-step process (Fig. 6).

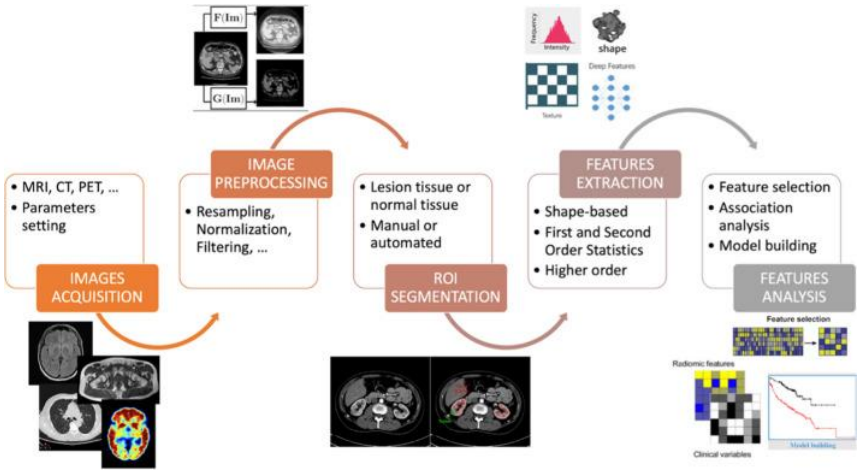


Fig. 6. Illustration of Radiomics process workflow [40].

1.2.1. Image acquisition

Medical images are the outcome of a two-dimensional or three-dimensional imaging procedure involving numerous modalities. Radiomics is dependent on both data source and modality (CT, MRI, PET, etc.) [41]. Furthermore, acquisition of the exact same modality can differ based on the protocols as well as equipment used. For example, MRI is very useful in the analysis of specific diseases including soft tissue problems because it offers information about the structural characteristics, functional metabolism, and dynamic alterations in tissues. At the same time, CT is the preferred

imaging modality for examining structural alterations in organs and evaluating the impact of chemo on patients. PET is well-known for its exceptional ability to assess the beginning stage, define the entire tumor volume, and assess the response to treatment after therapy. Several factors influence the acquisition of images and radiomic feature retrieval, including scanner hardware, reconstruction settings, and acquisition parameters. Furthermore, diversity in images acquisition occurs even within an identical modality, based on the tools and techniques used [39, 42].

1.2.2. Image preprocessing

Pre-processing of medical images is a critical step in the radiomics process, where medical images are prepared for analysis, their quality is improved, and further processing tasks are ensured. The medical image itself is the result of several processes, all of which contribute in different ways to the overall result. Understanding that what radiologist's study is not a real object, but only an image of a real subject, can be crucial for the advancement of radiomics research. Medical images are usually reconstructed using mathematical algorithms from raw data captured by physical detectors. This raw data is the physical image of the subject, which is also monitored by the interrogation system, filtered according to the detector characteristics and processed by all the devices in the digital acquisition and transmission chain [40].

A variety of computational approaches are used to rebuild images. Filtered Back Projection (FBP) is a technique extensively used in CT and PET that combines numerous projections and applies a statistical filter to improve detail. While it is quick, it is sensitive to noise and distortion. Fourier Transform-Based Reconstruction, which is used in MRI, employs the Inverse Fourier Transform (IFT) to convert frequency-domain information into spatial-domain images. Iterative Reconstruction (IR), used in CT, PET, and MRI, changes the image by continually comparing rebuilding data to raw data, reducing noise and radiation dosage while enhancing computation [43, 44]. Deep Learning-Based Reconstruction uses AI and deep learning models, such as convolutional neural networks, to rebuild images from under-sampled or noisy datasets. Deep Learning-Based Reconstruction combines AI with deep learning models, particularly convolutional neural networks, to reconstruct images from undersampled or noisy information. This technique is utilized for low-dose CT, MRI, and super-resolution medical imaging [45, 46].

1.2.3. Region of interest segmentation.

The computer vision approach known as image segmentation labels each pixel in an image to ensure that pixels with the same label have certain traits in common [47]. In other words, the primary goal of this method is to categorize every pixel in the image to group pixels that are comparable together. In contrast to analyzing the entire image all at once, such granular technique enhances the image analysis skills by concentrating on the pertinent portions of the image [48]. In the medical field in particular, segmentation enables medical practitioners to draw attention to diseased regions or anatomical features in an image, facilitating more precise image evaluation and interpretation. Segmentation can be used to determine the extent of a tumor, track its development, and evaluate how it responds to treatment [49]. Segmentation can assist surgeons in planning procedures and enhancing the precision and effectiveness of surgical care by evaluating intricate anatomical data. Image segmentation is employed as well in radiotherapy to precisely define the target regions and the surrounding healthy tissue to calculate the radiation dose [50, 51]. Several different methods can be used to segment medical images. The simplest technique is manual segmentation, which is done by

skilled professionals by precisely identifying the zones of interest. Excellent anatomical expertise and previous experience are required for this procedure. Segmentation that is semi-automatic is thought to be marginally better. With this approach, automated processes are first used, and then manual changes are made. It combines the benefits of both human expertise and automation. Advanced algorithms are used in the automatic segmentation approach to automatically define and identify regions of interest eliminating the need for human interaction. Its tremendous efficiency is one of its defining characteristics [52].

To accomplish accurate and consistent segmentation in medical imaging, different computational algorithms have been created developed over time, each one having its own strengths and limitations. Traditional segmentation methods in medical image analysis include thresholding, edge detection, region-based, clustering, and graph-based techniques. Thresholding converts grayscale images into binary by setting intensity thresholds, with global and local variations adapting to different lighting conditions. Edge-based segmentation detects object boundaries through first- and second-order derivative operators. Region-based methods, such as Seeded Region Growing and Split-and-Merge, segment images based on pixel similarity and spatial proximity. Clustering-based segmentation groups similar pixels using hierarchical or partitional clustering, including k-means and fuzzy c-means. Lastly, graph-based approaches, like Minimum Spanning Tree and Graph Cuts, represent images as graphs to optimize segmentation by minimizing edge weights. While effective, these techniques often struggle with complex medical images, requiring more advanced approaches for improved accuracy [53].

1.2.4. Feature extraction.

In radiomics, feature extraction refers to the methodical process of calculating a variety of quantitative characteristics from biomedical images following their segmentation into distinct areas of interest. These quantitative properties are derived using the pixel (2D) or voxel (3D) intensity values found in medical imaging, which represent the grayscale level or intensity of tissue at certain spatial locations. The numerical values obtained during this method serve as the foundation for assessing tissue heterogeneity and other critical properties that may reveal information about its biological behavior and medicinal implications [54, 55].

Since various characteristics associated with the tissue or lesion might provide insight into its biological behavior and its medical ramifications, the goal is to document them [56]. Given the meaning of radiomic characteristics, most of them do not exist in radiologists' lexicon. In this setting, it needs to be noted that radiomics represents a hypothesis-free technique. This means that no a priori assumptions are made about the clinical importance of the characteristics, which are determined automatically by image processing algorithms designed by specialists. The approach's goal is to identify previously undetected visual patterns employing these diagnostic, non-semantic elements and then classify them based on their most discriminative ones, which is also known as the construction of a radiomic signature. This process is defined by two types of radiomic characteristics: traditional and deep features [57]. Traditional characteristics are predetermined or developed by image processing professionals and can be classified into five categories: size, shape, first-order statistics, second-order statistics, and high-order statistics. Table 1 shows some of the traditional radiomic properties.

As can be observed, there are five distinct sets of attributes. The size and form characteristics indicate the geometric parameters of the area of interest. First-order statistics, derived from pixel or voxel intensity values, capture intensity distribution metrics such as energy, entropy, skewness, and kurtosis, which describe tissue grayscale properties. Second-order features reveal textural patterns and structural structure by examining interactions between neighboring pixels or voxels, which are commonly represented by matrices such as the gray level co-occurrence matrix, run length matrix, and size zone matrix. Higher-order features use sophisticated transformations to determine links between numerous pixels or voxels, resulting in more detailed structural data [58, 59].

Table 1. Traditional Radiomics features [60].

Feature category	Size	Shape	First-order statistics	Second-order statistics	High-order statistics
Radiomic features	• Area • Volume • Maximum 3D diameter • Surface area • Major/minor axis length	• Elongation • Flatness • Sphericity • Spherical disproportion	• Energy • Entropy • 10^{th} percentile • 90^{th} percentile • Skewness • Kurtosis	Gray level: • Co-occurrence matrix • Run length matrix • Size zone matrix	• Autoregressive model • Haar wavelet

The second type is deep features, which have gained significant popularity with the rapid advancement of deep learning algorithms. These algorithms automatically generate and select features within their own layers to accomplish a specific task, eliminating the need for additional human intervention. Scientific studies suggest that deep features outperform traditional features [58, 60, 61].

1.2.5. Feature analysis

Feature analysis is the systematic process of locating, extracting, and assessing significant features from data to improve machine learning and image processing performance. This comprehensive method focuses on identifying the most informative features for applications such as classification, segmentation, and pattern recognition [62]. By removing duplicate or unnecessary variables, feature analysis improves model performance while minimizing computational complexity and algorithm training time. LASSO, Random Forest, SVM, and Logistic Regression models tend to be used for similar tasks [58, 63, 64]. Performance indicators such as accuracy, precision, specificity, sensitivity and AUROC are used to evaluate model robustness [65].

When data sets are limited, the model can be overfitted. In this case, additional data generation techniques can help, such as adding random functions to the data or generating synthetic data using SMOTE [66]. This can help to train and generalize the model. It is also important to mention that many redundant features can also lead to overfitting, where the model captures noise rather than significant patterns, resulting in poor prediction results. In this instance, approaches such as PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), or feature selection methods (e.g., mutual information, recursive feature deletion) might aid in dimensionality reduction, resulting in a balanced dataset [67].

In the radiomics workflow, a thorough understanding of machine learning algorithms is required to choose the best strategy for a specific dataset, as the performance of different algorithms varies based on the nature and structure of the data.

1.3. Fundamental limitations and potential improvements in the Radiomics process

The future of radiomics in medical physics has a huge potential to change the way diseases are diagnosed, treated and monitored, especially in oncology, but also in a wide range of other medical fields. Advances in data science, machine learning and imaging technologies will further propel radiomics to become a key component of personalized medicine. While radiomics is a fast-developing duo of science and artificial intelligence, it also faces several challenges that could shape the evolution of science. The content of this section will go over the major issues and limitations of this science, as well as how they may influence the future of radiomics in the quickly evolving world of artificial intelligence and technology.

1.3.1. Lack of Standardization

A current limitation of radiomics is the lack of uniformity in medical image processing procedures. Researchers Cobo M., Fernández-Miranda P. M., Bastarrika G., Iglesias L. L. conducted a review of current image processing and combination techniques, together with the limitations of radiomics, and argue that standardization of medical imaging standards and procedures is necessary for the proper operation of radiology and CT systems. According to the researchers, standardization is hampered mostly by two factors: mathematical definitions of radiomic characteristics and medical images. For example, there is currently no standard mathematical description of radiomic properties. Medical images, on the other hand, are based on physical markers but capture the same phenomenon (illness) in different ways depending on the technique utilized and the patient. Although the latter issue cannot be eliminated, it can be lessened by standardizing the machines, which is accomplished by establishing the same imaging parameters according to standard protocols. The scientists are also sure that implementing standard methods will help to reduce radiation dose [68].

According to Li X. T. and Huang R, the rapid development and use of radiomics models has resulted in a lack of a uniform methodology for analysis and reporting. In their publication, the researchers review the reproducibility and repeatability of radiomics models for cancer monitoring and provide details on image processing and feature extraction in the several studies that have been performed so far. Although most researchers apply cross-validation or independent validation datasets to verify the repeatability of trained radiomic models, their generalizability may be hampered if used to datasets that differ significantly from the ones used for training and validation. Developing and implementing standard data collection methods can help minimize the variances limiting generalizability [69]. According to the authors, the biggest issues stem from discrepancies in imaging equipment, making it impossible to standardize. To eliminate these disparities, standard methods have been devised, such as the Brain Tumor Imaging Protocol (BTIP) for magnetic resonance imaging [70]. IBSI-compliant software solutions such as PyRadiomics and RadCaT help to achieve uniform differentiation of radiomic features. Nevertheless, significant functional differences exist across the various software systems, emphasizing the need for additional standardization. According to the scientists, open access datasets, like the Cancer Image Archive, help greatly to standardization by increasing model training and validation. However, achieving uniformity across all stages, from acquisition to processing, remains a challenge.

1.3.2. Small and Heterogeneous Datasets

Machine-learning models for prediction based on specific patient characteristics have the potential to significantly enhance patient outcomes prior to or during therapy. The latter models assist in improving tailored treatment decisions, which can lead to enhanced patient results [71, 72]. Over a decade of research has demonstrated that radiomics can offer prognostic data regarding patients' illnesses [73, 74]. Despite encouraging outcomes, radiomics presents a hurdle in gathering and accumulating sufficient (large-scale) imaging data. This is due to technological changes, a lack of protocol standardization, and variances between institute organizations and protocols. Imaging data obtained at a single organization often ends up in a more uniform dataset, such as images captured with identical settings using the same scanner. However, due to the dataset's homogeneity, such a model may fail to successfully include diverse data from outside sources [75]. Timmeren J., Carvalho S., and colleagues performed a study to demonstrate the possible issues that can arise during a retrospective multicenter investigation analyzing cohort and clinical features when the available dataset is small and varied. In this study, researchers used PET along with CT images from non-small cell lung cancer (NSCLC) patients who had 18F-fluorodeoxyglucose (FDG) PET/CT scans prior to and during chemotherapy to identify and assess the radiomic potential to detect and predict early response. As the researchers point out, PET/CT images obtained during therapy are not commonly accessible in clinical practice, hence the dataset contains an insignificant number of medical images. However, to perform the study, the scientists gathered the information from three distinct institutes, allowing them to obtain small subgroups, resulting in a diverse dataset. During the investigation, variations among cohorts were evaluated based on clinical variables and the cohort difference model. A LASSO regression model was also utilized to analyze the possible predictive influence on the general survival of radiomic characteristics from CT or FDG-PET, as well as relative or absolute variances among scans at two time periods. In addition, the effectiveness of five alternative classifiers was tested across all image sets. The results revealed that patient characteristics differed significantly between groups. Additionally, the cohort's difference model revealed statistically significant variations between the cohorts. Either LASSO or any of the investigated classifiers generated a clinically acceptable model for prediction that could be verified against the given datasets. In other words, the study's models cannot evaluate the probable link between radiomic characteristics in FDG-PET/CT medical imaging and overall survival. The researchers believe that this outcome was impacted by a lack of data as well as the various patient features. This study cannot establish the accuracy of prediction models, yet it does demonstrate the need for cohort estimate when data is obtained from several institutions [76].

When dealing with medical image datasets, frequent obstacles typically appear by class imbalance or lack of data. The reasons for these issues remain in the relatively low incidence of illnesses, difficulties in acquiring images that meet, or issues associated with the difficult and laborious tasks of image labelling and segmentation [77]. Imbalanced or smaller data sets may exert a severe impact on machine learning model performance, resulting in overfitting, biased outcomes, and erroneous results. Therefore, it is critical to address these difficulties while developing ML models. Scientists Iacono F. L., Maragna R., Pontone G., and Corino V.D.A. noted that when working with a limited number of samples, data augmentation procedures are typically utilized on the data used for training to reduce the danger of overfitting [78, 79]. The researchers' major goal was to offer a new way of balancing and data augmentation that used disruptions like dilatation, contour randomization, or erosion in the regions of interest of cardiac CT scans. Radiomic features were extracted from the

altered ROIs, resulting in new data. This method has been tested to solve the clinical challenge of separating cardiac amyloidosis (CA) from aortic stenosis (AS) and hypertrophic cardiomyopathy (HCM). The study sample comprised of 74 patients: 21 CA, 32 AS, and 21 HCM. During the investigation, the researchers extracted 107 radiomic characteristics from each original and distorted ROI. The CA-AS dataset had been balanced using a perturbation-based technique that included random oversampling, adaptable synthetic (ADASYN), and synthetic minority oversampling (SMOTE). The same approaches were used for CA and HCM dataset augmentation. Features were tested for reliability, redundancy, and significance using five feature selection approaches (least absolute shrinkage, p-value and selection operator (LASSO), semi-supervised LASSO, principal component analysis (PCA), and semi-supervised PCA). A SVM conducted the classification tasks, and its performance was tested using tenfold cross validation. The results showed that the perturbation-based method outperformed the other methods in terms of F1 score (CA-AS - 80%; CA-HCM - 86%) and balanced accuracy (by evaluating area under the curve AUC) (CA-AS - 0.91; CA-HCM - 0.92). Thus, the researchers' findings show that ROI perturbation is a viable strategy for addressing data rebalancing and data augmentation challenges.

1.3.3. Reproducibility and Validation

Reproducibility refers to getting the same results while performing radiomics analysis under similar settings, which is critical for ensuring the reliability, validity, and clinical relevance of radiomics-based models. If radiomics features and models are not reproducible, they cannot be used in real-world medical applications. Poor repeatability can be due to a variety of variables, including imaging variability, artifacts from patient mobility, image processing approaches, alternative segmentation methods, and the software used to extract radiomic characteristics [80, 81]. In a review article, authors Park J. E., Park S. Y., and colleagues discuss the primary issues associated with repeatability and feature selection. One of the most significant issues is the reproducibility of segmentation, which is regarded as the most crucial and difficult stage in the radiomics process. As the authors suggest, because radiological characteristics are taken from regions of interest in medica images, variations in segmentation (manual, semi-automated, or fully automatic) might result in significant variances in feature extraction. According to the scientists, irregular segmentation might cause changes in feature values, which can compromise the validity of radiomics-based models. The study found that semi-automatic segmentation had greater reproducibility of the extraction of features (ICC, 0.85 ± 0.15) than manual segmentation (ICC, 0.77 ± 0.17). Nevertheless, despite semi-automatic segmentation, radiomic feature repeatability is not optimum, necessitating more automated segmentation improvements. Researchers also propose the fact that reproducibility of segmentation might differ based on the type of cancer. The researchers also emphasize the importance of computational considerations such as feature extraction, bias management, intensity range, and bin discretization in improving the repeatability of radiomics features across multiple imaging modalities. The researchers found that grey-level discretization had significant effects on the reliability of the features in perfusion CT, while the effect was comparable but less noticeable in PET. MRI was extremely sensitive to preprocessing adjustments, with no feature achieving a consistency correlation coefficient (CCC) larger than 0.85 across 33 investigated voxel sizes, grey-level discretization, and quantization approaches. MRI was extremely sensitive to preprocessing adjustments, with no feature achieving a CCC larger than 0.85 across 33 investigated voxel sizes, grey-level discretization, and quantization approaches. Furthermore, the study revealed that discretizing PET scan values into 64 bins, an instance of overbinning, did not enhance predictive information, indicating that excessive binning

does not necessarily improve predictive performance. As a result, an appropriate balance of computational effectiveness and feature reliability must be maintained, as raising the bin size beyond a particular threshold does not necessarily provide extra diagnostic value [82].

Researchers Thomas H.M.T., Wang H.Y., and others investigated methods to determine whether radiomic qualities are stable and reproducible. According to the researchers, the reproducibility of different test results has been a source of contention for some time [83]. In their study, the researchers attempted to investigate the reproducibility of CT texture characteristics used in radiomics through the comparison of two feature extraction applications, using the MATLAB toolbox and PyRadiomics, applied to independent patient CT scan datasets. The researchers employed two datasets for the study: the first had medical images of 31 individuals diagnosed with lung cancer, and the other contained 137 patients diagnosed with head and neck cancer. The medical images used in this investigation were manually segmented. The 43 standard characteristics of radiomics accessible through MATLAB and PyRadiomics were derived utilizing two intensity level quantization approaches, both with and without an intensity restriction. Following that, both examples were ranked for every characteristic over all quantization parameter combinations, and the Spearman's rank factor was determined. Reproducibility has been defined as the degree to which a highly correlated variable in one dataset is also highly correlated in another, and vice versa. The researchers discovered that out of the 43 radiomic features presented, 29 were highly reproducible across both MATLAB and PyRadiomics applications. 18/43 of the radiomic features reported were shared by both datasets, implying that they might be independent of the precise location of the malignant tissue. The researchers believe that significant radiomics features should be chosen based on reproducibility, and that the set of features used in their study, as well as the proven methods for analyzing dataset reproducibility, are beneficial.

1.3.4. Clinical integration

Numerous measurement methods are used in healthcare nowadays to help discover illness characteristics that physicians are unable to find through manual inspection. Imaging techniques span most of the range of biological activities that may be observed, from the entire body to individual molecules. By directly quantifying the tumor's imaging phenotype at a spatial scale that does not exceed the spatial resolution of the imaging modality used, radiomics aims to offer surrogate, associated knowledge of various aspects of the disease, such as tumor grade, histological and inherited subtype, and predicted outcome. Although most research fails to consider the biological foundation of the implicit correlations that enable radiomic predictions, these features demonstrate the changes that take place at various scales in radiomic data. In 2021, researchers M.R. Tomaszewski and R.J. Gillies published a report outlining a significant recent effort to biologically validate radiomics results. Gene expression data, expression of proteins from immunohistochemistry staining, micro histological texture, and physiological tumor habitat are the four primary categories of biological correlations and techniques that are utilized to inform the biological foundation of radiomics. Based to scientists, in a conventional radiomic pipeline, an indicator of outcome produced and validated in an independent training set can then be investigated for its link with a specific biological measurement such as gene expression. This method has the potential to improve the model by informing the likely outcome prediction procedure retrospectively. Conversely, biological correlation can be employed freely for model construction, leading in a radiomic signature which is inversely related to outcome due to the biological connection, such as tumor oxygen deprivation's negative prognostic value, as demonstrated in the studies. Although both techniques provide valuable

confirmation as well as understanding of tumor biology, the following approach's more hypothesis-driven approach may be more objective for outcome analysis. Contrary to the scientists, many radiomic studies have failed to verify their findings beyond using an independent test sample. This tendency adds to low reproducibility and limited impact. Scientists propose that, as the value of biological information in radiomic signatures becomes more widely recognized, a uniform validation process be developed and applied across the radiomics community. Moving forward, all published studies should seek to include such analysis, either during model creation or later verification, to propose an assumption for the biological procedure behind the observed correlation. This will enable the explanation of the biological factors in front of the findings to emerge into an area of standard that will be enforced through the peer-review process [84].

The future of clinical integration and the use of radiomics in medical practice is predicted to change substantially as technology advances, standardization efforts, and a greater focus on individualized medicine. Majumder S., Katz S., Kontos D., and Roshkovan L. in 2024 have presented their perspectives on radiomics and key steps toward integrated healthcare. According to the researchers, radiomics' goal is to easily integrate into clinical workflows and supplement radiological interpretation with quantitative measures. Another essential role of radiomics is to support radiologists by offering high-quality imaging interpretations rather than replacing them. According to the researchers, radiomics holds significant promise not only for screening, diagnosis, and prognosis, but also for predicting and assessing therapy responses. Machine learning or deep learning-assisted radiomics can provide treating clinicians with a better understanding of disease heterogeneity, progress, and therapeutic response on an individual level, allowing them to develop focused treatments—a step toward precision medicine. This is especially interesting in regions with diverse diseases like cancer. As a result, it is critical that academics, developers, and physicians work together to implement this technology into clinical practice. [85]

1.3.5. Overfitting in Machine Learning Models

Overfitting can be a significant challenge in machine learning, especially overfitting is a significant difficulty for machine learning, particularly in radiomics. This problem occurs when the model over-learns the training data, catching noise and outliers instead of frequent patterns, resulting in poor performance with unknown data. Overfitting in radiomics hampers the transition from research to clinical use. High dimensionality is a significant source of over-application. This could lead to a "small-n-large-p" conundrum, in which the number of features (p) greatly exceeds the number of samples (n). This condition increases the risk of overfitting, as the models may become too intricate for the available data [86]. The model's complexity may also be a significant contributor to overfitting. Complex models, such as deep neural networks, are powerful but susceptible to overfitting if not properly tuned. It is worth noting that feature oversampling adds to model overfitting [87]. It is already clear that researchers are willing to try new approaches to limit abuse of the model. The most common way is to try to incorporate strong cross-validation techniques into the model to better measure its performance. Cross-validation is regarded to be particularly helpful since it effectively separates the training and validation datasets, reducing bias in performance predictions [88]. There are also several publications in the literature that suggest that employing a larger dataset, which provides a wider range of training samples, can help to prevent model overfitting. Studies have shown that having at least 10 000 examples can considerably lower the danger of overfitting. However, gathering enormous databases of medical images is incredibly difficult [89]. Furthermore, by

efficiently choosing features and thereby minimizing the size of the research, the model's performance can be improved. Researchers can limit the likelihood of overfitting by focusing on the most relevant characteristics. Principal component analysis (PCA) and LASSO regression are typical techniques for feature refinement[90].

Radiomics application in the clinical setting future depends on machine learning and its ability to reduce overfitting. Standardized imaging methods and feature extraction approaches may finally solve the problem of model overfitting. This may improve the repeatability and reliability of all examinations, hence boosting the model's generalizability [88, 90]. Further research into advanced machine learning methodologies, such as unsupervised learning and transfer learning, is expected to lead to new methods for generating robust models that are relevant to a wide range of patient populations and imaging modalities [87].

1.4. Radiomics in Medical Physics field

Radiomics is a rapidly growing topic in medical imaging that focuses on extracting a wide range of quantitative information from medical images using powerful algorithms. This technology seeks to show the patterns and features of tumors that may not be visible to the human eye, thereby improving clinical decision-making and tailored medicine [91]. The standard radiomics procedure typically includes acquiring or retrieving images, determining the area or volume, extracting quantitative characteristics of the image from the determined, such as shape, size, or texture parameters that are performing statistical analyses on the image features, and assessing any possible relationship among the radiomic features and a significant clinical endpoint [92]. Scientists Wagner, M. W. Namdar K., Biswas A., Mohani P., and Ertl-Wagner B. used radiomics for several purposes, including diagnosis and prognosis. As a result, the output may be a class label, such as tumor subtype, a risk score, such as response to therapy, a time to events in survival evaluation pipelines, or even features designed for use in a hybrid pipeline. According to the researchers, radiometric analysis pipelines can incorporate a range of models. Support vector machines (SVMs), random forests (RF), and forward neural networks are the most used categorization algorithms. According to the research authors, tree-based models are frequently used to categorize tabular data, such as radiomics. A random forest is made up of decision trees, that consist of simple nested if-else statements. These trees segregate the data based on the threshold and numerical value of a feature. However, decision trees are vulnerable to overfitting, which can be mitigated by mixing numerous decision trees to form random forests. In contrast, XGBoost is a boosting-based strategy that progressively builds decision trees to lower prior trees mistakes [93].

Radiomics is a promising area of research that extracts quantitative data from medical images. In other words, radiomics measures tissue and lesion properties, such as homogeneity, which can help to address a clinical condition alone or in combination with other medical elements. The most crucial aspect of this procedure is that it enables the non-invasive measurement of tissue alterations [94]. Researchers Jin D., Ni, X. and their colleagues utilized the radiomics approach to assess the effectiveness of dual-layer computed tomography (DLCT)-based radiomics in predicting epidermal growth factor receptor (EGFR) mutation status in patients with non-small cell lung cancer (NSCLC) [95]. The researchers used DLCT scans and medical records from 115 individuals suffering from NSLPV who were retrospectively collected and randomly allocated to the training cohort. The algorithm's training dataset included 81 patients, whereas the validation dataset included 34. The researchers used LASSO reduction of dimensionality method to create a radiomics model based on

DLCT's radiological properties. The medical and CT data were then combined to create a clinical model. A nomogram was then created by integrating radiomics scores with clinical variables. The researchers employed ROC and DCA curves to evaluate the models' performance and clinical relevance. In the decision-making stage, only six radiomic characteristics and two clinical characteristics were fitted to the model and tested to determine the outcome. The radiomic model, clinical model, and nomogram had AUCs of 0.909, 0.797, and 0.922 in the training dataset, and 0.874, 0.691, and 0.881 in the validation dataset, respectively. The findings suggest that the area under the curve (AUC) for a nomogram and radiomics models were much higher compared to those for the medical model, but the researchers did not detect a statistically significant difference. Lastly, based on the DCA, the scientists determined that the nomogram provided the highest therapeutic benefit at the majority of threshold levels (Fig. 7).

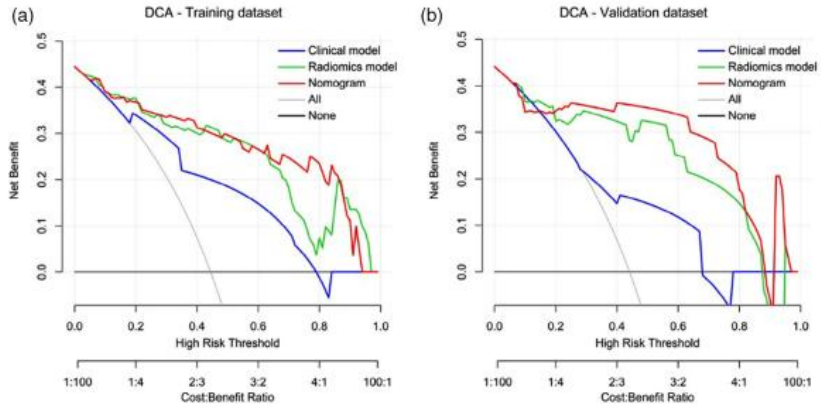


Fig. 7. DCA of prediction models built from training and validation datasets (a, b) [95].

Based on these findings, the researchers propose that a nomogram that combines clinical characteristics and pre-treatment radiomic features obtained through DLCT could aid in the non-invasive assessment of the EGFR mutation status of NSLPV participants.

Renal cell carcinoma, also known as RCC, is the most frequent kind of cancer of the kidneys, causing around 2% of total cancer fatalities globally. A local tumor has a 5-year survival rate of 93%, but this lowers to 75% if it spreads to lymph nodes or other regions of the body, and to 18% once it spreads to the whole. Individuals with pure cell RCC (ccRCC) have a poorer prognosis compared to those with other histological subtypes [96]. Researchers Nazari M., Shiri I., and Zaidi H. created a radiomics-based ML model that uses extracted radiomic features and clinical information to predict the 5-year mortality rate of patients with renal cell carcinoma with clear cells (ccRCC), which spreads to lymph nodes and other body sites. Based on the results, the authors conclude that people with clear cell RCC experience poorer outcomes than individuals with other histological subtypes. The researchers chose 70 ccRCC patients for CT scans throughout the investigation based on images quality and the availability of diagnostic data. Each image was segmented using 3D slicer software, and the VOI was manually segmented. Prior to obtaining feature information, the CT images had been processed to extract several images features, including wavelets, Gaussian Laplace, and intensity values regrouping into 32, 64, and 128 bin levels. Each VOI yielded 2544 3D radiological parameters for each subject. The approach of minimal redundancy and maximum relevance was implemented for feature selection. Combining radiomic properties and clinical information, the XGBoost model was the best of the eight unique models. This is supported by the efficacy metrics AUROC, accuracy,

sensitivity and specificity, with 95% confidence intervals of 0.95-0.98, 0.93-0.98, 0.93-0.96 and 1.0, respectively. Based on this model, scientists developed a powerful radiomics-based classifier capable of accurately predicting the overall survival of RCC individuals as well as the long-term outlook of ccRCC patients. This feature, according to the academics, may aid in finding people at greater risk who require additional treatment and follow-up regimes.

In recent years, both mortality and incidence rates from breast cancer (BC) have increased annually. In the United States, an expected to 313 510 new cases of breast cancer will be identified in 2024, including 310 720 women and 2 790 males [97]. Infiltrating breast cancer (IBC) is one of the most frequent types, accounting for 80% of all diagnoses. It is a tumor that develops as malignant epithelial cells that reside in the breast and is the most common pathogenic type of BC [98, 99]. When identified in the late stages, there is a substantial risk of metastasis. The recurrence rate following treatments is elevated, and the outlook for recovery is poor. As a result, early detection and treatment of IBC are critical for improving survival rates for patients and changing the disease's trajectory. Breast cancer is a highly varied illness, and traditional histological classification does not adequately capture its underlying biology and molecular properties [100]. Here and now, BC is diagnosed mostly through histological investigations, such as biopsy using a needle or surgical tumor removal. Immunohistochemistry, also known as IHC and Fluorescent in Situ Hybridization, known as FISH, are still regarded as the highest-quality benchmarks to identify BC biomarkers and for recognizing abnormal molecular subtypes. However, pathological approaches have some disadvantages, including invasiveness, surgical problems, low repeatability, extended wait periods, and high subjectivity. Most significantly, due to BC's internal heterogeneity, a single- or multiple-point sampling is limited to representing a portion of the tumor and is insufficient to assess the genetic phenotype as well as other molecular biological states across the tumor. Therefore, technologies are needed that allow non-invasive, extensive and advanced monitoring of the biomolecular properties of the tumor to distinguish molecular subtypes of IBC before surgery. In response to the need for such a tool, researchers Liu H., Xia H. and their colleagues aimed to develop a two-dimensional ultrasound radiological model for non-invasive differentiation of the four different molecular subtypes of IBC, thus providing a new and rapid approach for the early diagnosis of patients with IBC and a basis for clinicians to develop individualized pre-operative diagnostic and therapeutic strategies [101]. This study comprised a retrospective review of 210 patients diagnosed with IBC at the same hospital between May 2019 and February 2024, using either surgical or needle biopsy pathology. Patients were permitted to participate in the study if they were diagnosed with a pathological diagnosis of IBC with IHC and/or FISH results along with all clinical and laboratory findings, had only one lesion, had an ultrasound examination within 2 weeks before surgery, and had not received neoadjuvant chemotherapy or any other treatment. The ultrasound images in the investigation were standardized with high-frequency probes. ITK-SNAP was used for image segmentation, and the specialist program Ultrasomics-Platform was used to extract tumor features. To prevent uninformative features, scientists utilized a variety of methods, including LASSO and random forests, to guarantee that the extracted feature dataset was useful and not redundant. The radiomics models were built and optimized using ten distinct machine learning classification algorithms, including k-nearest neighbors, logistic regression, naive Bayes, decision tree, support vector machines, bag-of-feet, random forests, ultra-random trees, AdaBoost, and gradient boosting trees. These classifiers were trained with the best selected features, and the models' accuracy was tested using a five-fold cross-validation procedure. To evaluate their effectiveness, the models were validated with the validation set and assessed based on ROC as well as AUC values. The models constructed revealed that 39, 25,

and 19 optimum features were chosen from the 5 936 features recovered to discriminate both luminal and non-luminal, luminal A and luminal B, HER2 overexpression, and triple-negative (TN) subtypes, respectively. The models' performance reveals a distinction between luminal and non-luminal regions, with a model training AUC of 0.901 and a model validation AUC of 0.752. In terms of separating Luminal A and Luminal B, the training model's AUC value is 0.931, while the validation AUC is 0.773. In terms of separating Luminal A and Luminal B, the training model's AUC value is 0.931, while the validation AUC is 0.773. Higher values can be seen in both HER2 overexpression and TN feature separation compared to other compared pairs. The training and validation models are capable of discriminating HER2 and TN characteristics with high accuracy, as evidenced by AUC values of 0.962 and 0.842, respectively. Thus, based on the researchers' findings, it can be inferred that the models exhibited good discriminatory performance, particularly in distinguishing HER2 overexpression from TNBC, although differentiating between luminal A and B was more challenging due to overlapping characteristics.

Cancer is distinguished by quickly dividing aberrant cells and significant heterogeneity, which influences therapy response and clinical outcomes. Metastatic spread is a leading cause of cancer-related fatalities, leading to lesions of varying genotypes and morphologies, making therapy extremely difficult [102]. Soft tissue sarcomas (STS) are cancers that affect the body's connective and supporting tissues, for example muscles, tendons or ligaments. They are highly rare and difficult to treat. Individuals with advanced or metastatic illness have a very low survival rate [103]. In the therapy of this malignancy, radiographic evaluation for therapeutic efficacy becomes critical to replace failed treatments with alternative, perhaps more active therapies. Radiomic characteristics are especially useful for this assessment because the spatial grouping of increased and not increased voxels that histologically distinguish viable and necrotic tumor components. Individual malignant tumors with distinct metastases may have different features and respond differently to anticancer treatment, although having the same histological classification. Therefore, researchers Geady C., Abbas-Aghababazadeh F., and colleagues conducted a study to investigate the utility of radiomic biomarkers to predict lesion-specific responses to therapy in individuals with multi-metastatic leiomyosarcoma [104]. The study involved 80 patients who had at least one lesion. The amount of contour lesions per patient varied from 1 to 11, however 54 of the 80 individuals had two or more, bringing the total number of contour lesions included in the study to 202. Using segmented CT scans, 6 592 radiomic features were acquired. However, 1 452 characteristics were removed after examining segmentation reliability. Furthermore, after evaluating the radiomic features with a volume as well as coefficient of correlation threshold of 0.20, the finalized dataset contained 32 features. The study employed *"RepeatedStratifiedKFold"* to evaluate the relative prediction performance of LM volume categories. The lesion-specific radiomic models had up to 4.5 times more predictive power than the classifier without skill, with the best accurate model having an AUPRC of 0.70 (FDR = 0.05). The accuracy varied according to the drug utilized and the LM volume. According to the researchers, using radiomic characteristics to predict lesion-specific therapy responses is an innovative approach. Recognizing the biodiversity in metastatic subclones can be used to assess treatment response, potentially facilitating management options involving selective ablation of resistant clones during systemic therapy.

Despite cancer diagnosis and treatment constantly improving, there nevertheless remain diagnostic regions where it is extremely difficult to accurately detect disease development. Head and neck cancers (HNCs) are a complex group of malignancies that present significant diagnostic

challenges [105]. Skull base malignancies are notoriously difficult to collect samples and diagnose, and because they grow slowly, patients typically do not have symptoms until the final stages. Because of the skull base's distinctive and convoluted anatomical structure, surgeons are frequently unable to determine the type of tumor and the area of resection earlier than surgery. Despite the complexity of the diagnoses, scientists have lately conducted study on radiomics in HNC diagnosis. Experts can create models that more accurately describe tumor growth and development by analyzing enormous amounts of clinically relevant data, allowing patient-specific and sequential therapy. The scientists Peng Z., Wang Y., Jiang S., Fan R., Zhang H., and Jiang W. released a paper where they reviewed and established the concepts and procedures of radiomics and machine learning, along with their current uses in head and neck cancers, as well as the guidance and applications of artificial intelligence to the therapy and diagnostics of HNC. According to their findings, the scientists collected imaging features from MRI scans of 85 individuals in the training sample and demonstrated that an MRI radiomics pattern could distinguish stage III-IV squamous cell carcinoma of the head and neck from stages I-II HNC carcinoma of squamous cells. As a result, it indicates that radiomics may become an important tool for preoperative staging. Radiomics is the extraction of relevant, quantitative data from medical images that can be combined with other common predictive factors such as clinical setting, tissue molecule markers, and morphological characteristics. Several studies have shown that a model of prediction based on imaging and other data is more effective at predicting sickness and survival. Machine learning has been shown to be an effective method of statistical analysis required to correctly produce and use massive amounts of high-dimensional data. The type of data as well as the investigation's purpose dictate the modelling approach utilized. The category of data and the purpose of the inquiry dictate the modeling approach utilized. Popular machine learning algorithms include, random forests (RF), logistic regression, Bayesian models, support vector machines (SVMs), and, more recently, deep learning. According to the study's authors, the technique has been widely used in the building of many prediction models for HNC [106].

This chapter presents a detailed account of the radiomics process, from its genesis in the early 1970s to its rapid development and application in a variety of settings these days. The literature review also briefly summarizes the radiomics process's essential steps and its limitations, as well as its applications in medical physics. This section has shown that radiomics has a high potential for improving cancer diagnosis and prognosis by extracting quantitative imaging features and merging them with clinical data. Several studies have successfully used radiomics to treat several forms of cancer, including lung, kidney, breast, and soft tissue sarcomas, demonstrating its capacity to improve prognostic modelling and treatment planning. However, research indicates that HNC malignancies are especially difficult because of the complicated anatomy of the skull. Many critical components in the head and neck area are closely intertwined, including the brain, cranial nerves, major blood arteries, airways, and soft tissues, making it difficult to discern between healthy and injured tissue [107, 108]. Furthermore, HNC cancer exhibits a great degree of diversity in tumor biology, growth patterns, and response to treatment. This variation hinders the development of standardized therapeutic procedures, necessitating personalized medical strategies [108]. As a result, the anatomically complicated face and neck region makes it extremely difficult to effectively diagnose and arrange treatment for patients. Despite recent efforts, study in this anatomical subject remains restricted. Considering these limitations and the need for improved diagnostic tools, the purpose of this work is to create a radiomics model using medical imaging of the head and neck to improve the accuracy of diagnosis and treatment planning in this complex area.

2. Methods

2.1. Data set

In this study, delta radiomic features were extracted from PET/CT imaging data of the head and neck region in a cohort of 55 patients. Delta features were defined as the quantitative differences in radiomic parameters computed between two imaging time points, prior to and following chemotherapy administration.

The dataset can be divided into two separate groups of variables describing each patient. One group consists of clinical (described in Table 2). In this study, variables Treatment Type and Survival Outcome (Lived/Deceased) will be employed to demonstrate the implementation of the defined procedure for developing the prognostic model. In this case, two prognostic models will be developed to:

- identify which delta radiomic features are associated with survival outcomes.
- assess which treatment types lead to greater changes in radiomic features.

Table 2. Clinical data.

Variable	Meaning
Treatment type	A multi-class categorical variable, which was used to represent the type of chemotherapy administered, with values 0, 1, and 2 corresponding to distinct treatment regimens applied to each patient
Lived / died	A binary outcome variable, which was defined to represent patient survival status following treatment, where 0 indicates survival and 1 denotes death
Amzius_dgn	A numerical variable, which indicated at what age patient was diagnosed.

The second group of variables consists of delta radiomic features, which reflect changes in quantitative imaging biomarkers derived from head and neck PET/CT scans performed before and after chemotherapy. The delta features in the dataset can be categorized according to the type of feature (see Table 3).

Table 3. Categories of main delta features.

Feature category	Examples	Number of features
Conventional	SUVbwmean, TLG	8
Discretized Intensity	Entropy, SUVbwKurtosis	8
Shape	Volume, Sphericity	5
GLCM	Contrast, Correlation	8
GLRLM	LRE, HGRE	10
GLZLM & NGLDM	SZE, Coarseness	10

Radiomic feature extraction yielded a total of 49 features, grouped into six main categories: conventional first-order statistics (e.g. SUVmean, SUVmax, TLG), discretized intensity-based metrics (e.g., histogram entropy), shape-based 3D geometric descriptors (e.g., volume, sphericity), and texture features including gray-level co-occurrence matrix (GLCM), gray-level run-length matrix

(GLRLM), gray-level zone length matrix (GLZLM), and neighborhood gray-level difference matrix (NGLDM).

2.2. Radiomics workflow

2.2.1. Verifying data's balance

Assessing whether the target variable is balanced is an important step before starting to develop supervised ML models with a categorical target variable that has two or more classes. In other words, in imbalanced datasets, the number of observations in each class is noticeably different, and the resulting supervised machine learning model may not be accurate [109]. To determine whether the target variable is balanced, it is necessary to analyze the distribution of the target variable or class labels, otherwise known as categories, in the dataset. This can be achieved using a variety of methods. For instance, the distribution of the classes can be measured, including the calculation of the frequency of each class, and displayed in a bar or pie chart, which is an excellent way to assess the distribution of the target variable [110]. If the distribution of classes is not equal, this is not necessarily a sign that the dataset is imbalanced, as this is not only influenced by the number of classes but also by the size of the dataset. Therefore, statistical tests to check whether the target variable is balanced can be performed using *Chi-Square Goodness of Fit test*.

The Chi-Square Goodness-of-Fit test is a statistical tool for determining if the observed frequencies of categorical outcomes differ significantly from the anticipated frequencies for a given distribution. In the context of this study, the test is used to determine whether the class distribution of the target variable is significantly different from a uniform distribution, indicating imbalance in class representation [111]. In this case, hypotheses will be formulated as follows:

$$H_0: The\ target\ variable\ is\ balanced$$

$$H_a: The\ targef\ variable\ is\ imbalanced$$

In this instance, the null hypothesis H_0 indicates that the target variable in the dataset is balanced. The alternative hypothesis H_a is that the target variable in that dataset is imbalanced, which means that at least one category has a considerably different frequency than expected from a balanced distribution. The hypotheses will be tested at a significance level of 0.05 (α = 0.05), with the p-value used to assess statistical significance. For instance, if the ***p-value*** is greater than or equal to 0.05, the null hypothesis is not rejected, and the target variable can be considered as balanced [112, 113].

This method is particularly suitable for multi-class categorical target variables and is widely employed when confirming the representativeness of categorical groups in classification issues, such as those encountered in radiomics-based prognostic modelling.

2.2.2. Synthetic data generation

The imbalance of classes remains one of the most significant challenges to address issues with classification in medicine, because great accuracy must be achieved for a small number of data points. For instance, in medical statistics for diagnosis of cancer, benign tumors are far more common than dangerous tumors. When statistical models are created on extremely skewed datasets, machine

learning algorithms frequently overestimate the largest class therefore produce incorrect results, especially when the dataset's minority of instances to be classified is too small [114].

Since it is extremely difficult to gather data in medicine in a way that ensures the set is balanced, the problem of set imbalance can be tackled addressed in a variety of ways, including random oversampling. This is a simple approach for randomly reproducing a minority of class samples to balance the class distribution. The method selects minority class values at random from an existing dataset and replicates them without modification. This method can be beneficial for efficiently generating minority class observations. However, it should be noted that data duplication will result in overfitting, hence this strategy is only appropriate when the dataset is exceedingly limited, and no other method can be used [115]. On the other hand, synthetic data production is a more efficient technique to extend an existing dataset. There are several ways, for example SMOTE (synthetic minority oversampling method) and GAN (generative adversarial networks). The SMOTE approach creates synthetic samples by interpolating existing minority class samples in the feature space. The method operates by selecting a minority class sample and creating new points along a line connecting it to its *k* nearest neighbors [116]. In contrast, the GAN approach comprises of a generator that generates synthetic data and a discriminator which distinguishes between genuine and fake data. Eventually, the generator learns to produce synthetic samples which are extremely close to actual ones [117]. An overall comparison of three different methods could be seen below (see Table 4).

Table 4. Comparison of different data generating techniques [118].

	Random Oversampling	**SMOTE**	**GANs**
Advantages	• Easy to implement • Retains the original feature distribution	• Computationally efficient • Generates new synthetic samples (prevents overfitting)	• Can generate highly realistic data for radiomics and images • Can create completely new variations
Disadvantages	• High risk of overfitting • No added diversity	• Does not generate new patterns, only variations of existing ones. • Not effective for highly non-linear data (deep learning features)	• Needs sufficient training data to create meaningful samples • Challenging validation of synthetic samples

A comparison of the three techniques shows that SMOTE and GAN are better than random over-sampling, which lacks data diversity and has a high risk of overfitting. When comparing SMOTE and GAN, it is necessary to evaluate the analysis requirements and complexity of the data. The SMOTE approach is suitable for supervised machine learning methods such as regression and classification, whereas GAN is more appropriate for unsupervised machine learning models such as clustering or association analysis [119]. It is also very important to consider the size of the dataset, as GAN needs a large dataset to train the method, while the SMOTE algorithm works well on a smaller dataset.

Given the use of supervised machine learning models to build the radiomics model in this project, as well as the size of the dataset, the SMOTE method will be chosen to obtain additional data.

2.2.3. Selection of significant variables

2.2.3.1. Correlation analysis

The sets of radiomics features extracted from medical images contain a great number of variables, however not all of them are important for developing predictive models. To create an accurate and robust model, it is critical to identify and leave only the most significant features while removing duplicate or non-contributory ones. Based on extensive literature on radiomics research practices and the application of machine learning algorithms, correlation analysis is one of the first steps in the selection of significant features. It helps to remove redundant information by identifying variables that are highly correlated and providing overlapping information. [120, 121]. In the field of machine learning, correlation quantifies the statistical relationship between two continuous variables. Spearman's rank correlation coefficient measures the strength and direction of monotonic correlations while making no assumptions about linearity. In other words, this analysis looks at whether an increase in one variable is consistently associated with an increase or decrease in another variable, regardless of functional form. The coefficient ranges from -1 to 1, with extreme values indicating a completely monotonic correlation. This non-parametric measure is particularly useful for finding duplicate features with strong monotonic relationships, including those associated through nonlinear but order-preserving transformations [122, 123].

In this study, correlation analysis will be performed as a prior step before applying feature selection algorithms. Features with high positive or negative correlation ($|\rho| > 0.9$ or $|\rho| < -0.9$) will be eliminated to reduce multicollinearity [124]. This approach guarantees that only non-redundant features are fed into the machine learning pipeline, which improves model performance and interpretability [125]. This will be accomplished by using a correlation matrix. It gives a structured overview of the pairwise associations between all features, with each cell reflecting the correlation coefficient between a specific pair. This matrix provides a thorough view of feature interdependencies and serves as the foundation for intelligent feature reduction [126].

2.2.3.2. Selection of significant variables

After performing correlation analysis to eliminate redundant variables, the next step involves applying supervised machine learning algorithms to identify the most significant features to the target classification task. This process requires the data set to be appropriately split into two or more subsets [127]. Machine learning needs to split the data into multiple subsets to ensure good generalization of the model to previously unseen data and to avoid overfitting. The dataset is frequently divided into training and testing subsets, with each receiving 80% and 20% of the data, respectively. However, if hyperparameters are to be tuned, the dataset should be divided into three subsets: training, validation, and testing, with 70%, 15%, and 15% of the data allotted to each [128].

As an alternative, k-fold cross-validation could be applied to improve model evaluation. In this case, the data is first split into training (80%) and testing (20%) subsets. The training set is then subjected to k-fold cross-validation, which divides it into k equal folds. In each of the k iterations, one-fold is used as a test set, with the remaining k-1 folds serving as the training set. This procedure is repeated k times, with each fold serving as a test set once and the remaining k-1 folds being utilized for training. By the end of the process, each data point has been added to the test set once and the training set k-1 times. This method makes greater use of the data, offering a more thorough examination while lowering the risk of overfitting. Following cross-validation and model training, the model that

remains is tested on the 20% testing set to see how it performs on previously unseen data. Employing an 80/20 data split with cross-validation ensures that there is a separate from one another, unaltered testing set to assess the model's generalization ability. Cross-validation provides an even more powerful evaluation throughout training by validating different subsets of the training data along with lowering the risk of overfitting [129].

In this study, the dataset was separated into training and testing subsets, with 80% used for training and 20% for testing. Furthermore, 10-fold cross-validation was used to the training data. 10-fold cross-validation was chosen because it provides a more trustworthy assessment of model performance than a simple split into training and testing sets. Using several subgroups for both training and testing reduces the variability of performance estimates and strikes a compromise between bias and variance [130].

2.2.3.2.1. Recursive Feature Elimination (RFE)

Recursive feature elimination is a feature selection approach that seeks to estimate which features are most useful in discriminating against the classes of interest. This strategy can minimize non-significant features to obtain a final feature set that is significant for target variable prediction while not reducing the final classification accuracy. This procedure entails training a model, ranking characteristics by importance, and eliminating those that are the least relevant in predicting the target variable (see Fig. 8). This process is repeated when the desired number of features has been achieved [131]. There are various options for training models, such as Random Forest or Support Vector Machines [132]. In this scenario, Random Forest was chosen because of its capacity to effectively deal with multi-class target variables [133].

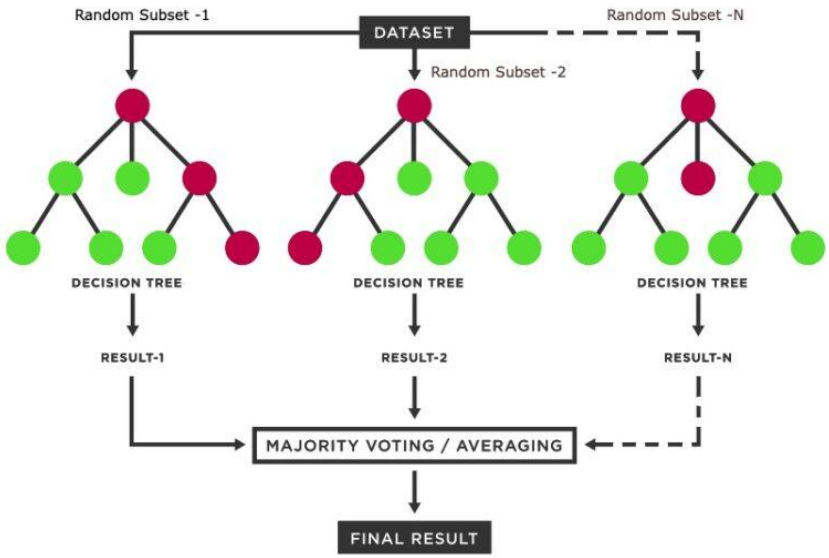


Fig. 8. Random Forest algorithm scheme [134].

The RFE algorithm principle is based on the importance of variables assessment, which is determined internally by RF classifiers and involves numerous rounds of classification. Each cycle consists of developing a new random forest classification model, measuring its accuracy using cross-validation, examining the feature importance metrics for each feature utilized, and updating the feature-set that will be used in the next round of the operation. The first classification round uses all the available features. The weakest performers are then identified using the variable of importance metric, which

the model estimates during the learning process. Several of the weakest characteristics are then removed from the feature set, and the operation is repeated. Once multiple classification models are built across different iterations, the final prediction is determined using a combination technique such as majority voting (for classification tasks) or averaging (for regression tasks). This ensemble approach ensures a more robust and generalized final model, as it aggregates the predictions from multiple models rather than relying on a single iteration. By doing so, RFE seeks to reduce any dependencies or collinearity in the input features [131]. The final dataset is the subset that yields the best-performing model.

2.2.3.2.2EXtreme Gradient Boost (XGBoost)

XGBoost, or eXtreme Gradient Boosting, is a sophisticated ensemble learning method that builds decision trees progressively, with each new tree rectifying the errors produced by prior trees. (Fig. 9) [135].

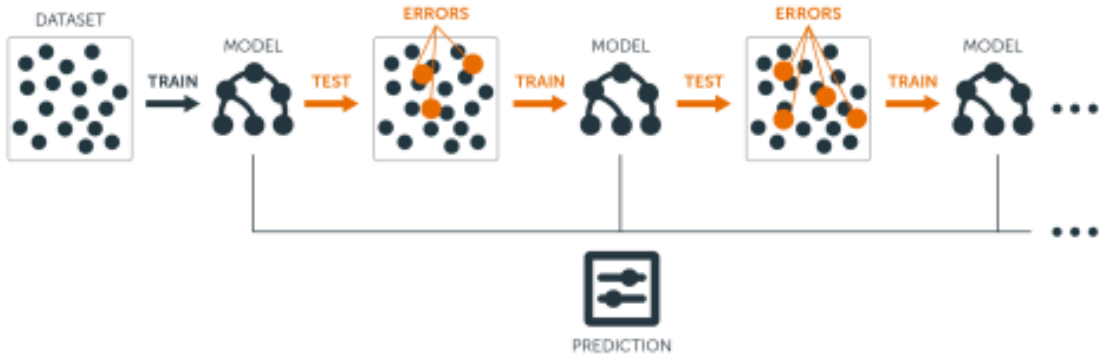


Fig. 9. Schematical representation of eXtreme Gradient Boost [136].

To improve prediction performance, this approach uses an additive training procedure in which trees are built iteratively and fitted to the loss function's negative gradient [137]. XGBoost uses regularization terms to correct model complexity, lowering the danger of overfitting by regulating the number of leaves in trees [138]. One of its significant advantages is its capacity to handle missing values by learning appropriate split directions even when some data points are missing [139]. Furthermore, XGBoost has a built-in system for feature selection, ranking features depending on their importance in decision tree development [138]. Feature importance is determined by using a range of metrics, including gain, which measures the improvement in accuracy provided by a feature, and frequency, which measures how frequently a feature is used across all trees. However, frequency alone may be misleading if a feature is widely utilized but fails to significantly add to performance [140]. XGBoost also integrates cross-validation techniques to evaluate model performance and fine-tune hyperparameters, ensuring robust generalization to unseen data. Furthermore, it supports early stopping, a strategy that stops training when the validation performance fails to improve, preventing overfitting and lowering computational cost [141, 142].

2.2.3.3. Performance evaluation

Assessing a machine learning model's performance is crucial for ensuring that it is reliable, efficient, and adaptable to new data sets. Without sufficient assessment, the model's predictions may be erroneous, biased, or deceptive, resulting in poor decision-making. [143, 144]. Overfitting, which occurs when a model identifies patterns in training data but performs poorly on unseen data, can be

identified and mitigated using performance measures. Overfitting occurs when the model over-simplifies the relevant models, resulting in poor performance on both training and testing data. [145]. To achieve robustness, different models must be compared using appropriate assessment measures to determine which one best balances accuracy, generalizability, and computing efficiency [143].

2.2.3.3.1. AUROC

The Area Under the Receiver Operating Characteristic Curve (AUROC or ROC-AUC) is a popular statistic for evaluating the effectiveness of classification models [146]. A widespread assumption is that AUROC is a combination of AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic Curve), although this is incorrect. In fact, AUROC refers only to the AUC calculated for the ROC curve. To completely appreciate its significance, it is necessary to first define the distinct notions of AUC and ROC [147].

The AUC is a scalar metric that measures a model's overall ability to differentiate between positive and negative classifications. It presents a single-value summary of the ROC curve, providing a comprehensive measure of categorization performance [148]. A higher AUC value indicates greater model performance, with greater or equal to 0.9 value signifying perfect discrimination and 0.5 implying that the model performs no better than random chance [149, 150]. Table 5 provides a full description of the AUC values.

Table 5. Interpretation of model's performance based on AUC value [149, 150].

AUC Value	Interpretation of model's performance
$0.9 \leq AUC$	Excellent
$0.8 \leq AUC < 0.9$	Good
$0.7 \leq AUC < 0.8$	Moderate
$0.6 \leq AUC < 0.7$	Poor
$0.5 \leq AUC < 0.6$	Random or failed

The ROC is a visualization tool that shows a classifier's ability to distinguish across classes. It depicts the True Positive Rate (TPR) against the False Positive Rate (FPR) over various classification thresholds, allowing for an evaluation of the compromise between specificity and sensitivity. The ROC curve is especially useful for establishing the best decision threshold for a model, making it a vital tool in performance evaluation [151].

In this study, the AUC and ROC metrics are interpreted differently than they are in common studies. Traditionally, these measures are used to evaluate binary classification tasks, which measure a model's ability to identify accurately between two distinct classes [152, 153]. However, as the research presented herein involves a multi-class classification situation, these performance metrics must be evaluated and applied differently in order to appropriately reflect the complexities of multi-class predictive modeling.

Traditional AUC and ROC studies must be adapted when dealing with multi-class target variables since they are built for binary classification problems by default. The one-vs-rest (OVR) principle is a commonly used strategy for extending ROC analysis to multi-class issues, in which each class is compared individually to the sum of all other classes. In this study, class labels are translated into

binary indicators: the target class is labelled as 1 and all other classes as 0 [154]. A visual example of one class vs all other classes classification performance is depicted in Fig. 10.

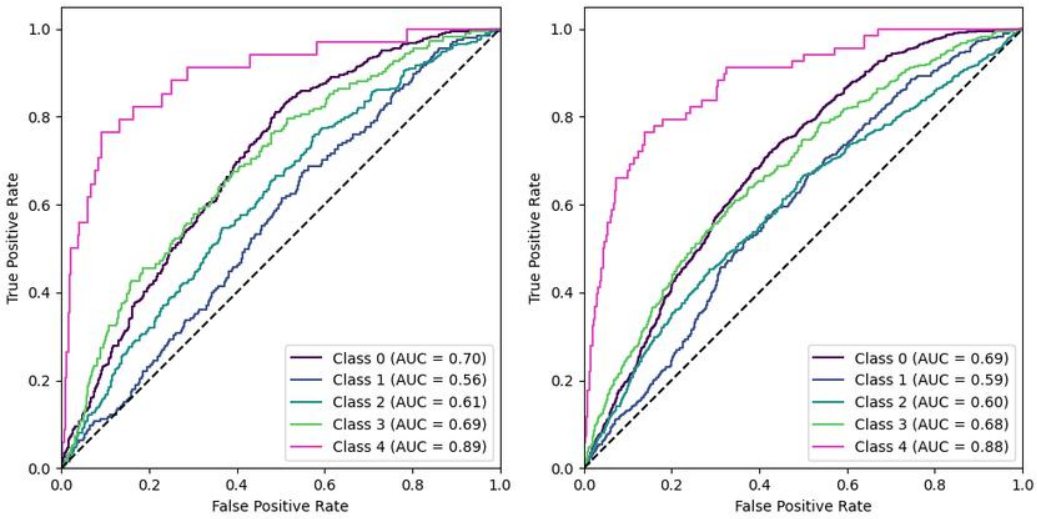


Fig. 10. Illustration of ROC curves for each class using One-vs-Rest ROC strategy [155] .

To evaluate the overall performance of the model in a multi-class environment, use the macro-level AUC. This is calculated by first calculating the AUC for each class in the binary classification task, which assesses the model's ability to distinguish the target class from the other classes. The AUC scores are then arithmetically averaged, giving each class equal weight, regardless of frequency. This helps to ensure that no single class, especially the most common class, has a disproportionate impact on the overall score if the classes are unevenly distributed. Thus, the macro-AUC is a summary statistic that reflects the performance of the model evenly across all classes, thus providing a balanced and unbiased assessment, particularly useful when all classes are equally important [156, 157].

2.2.3.3.2. Confusion matrix

The confusion matrix is crucial to evaluating the performance of various classification systems. The confusion matrix can be described as the mix of predicted and actual class instances. It enables for the specification of an extensive variety of performance criteria (such as accuracy, precision, and recall). In other words, a confusion matrix can be defined as a set performance indicator that can be used in a classification task to evaluate an algorithm or to compare the performance of different algorithms. The confusion matrix can be applied to both binary and multiclass classification problems [158, 159]. Fig. 11 shows an example of both types of confusion matrixes.

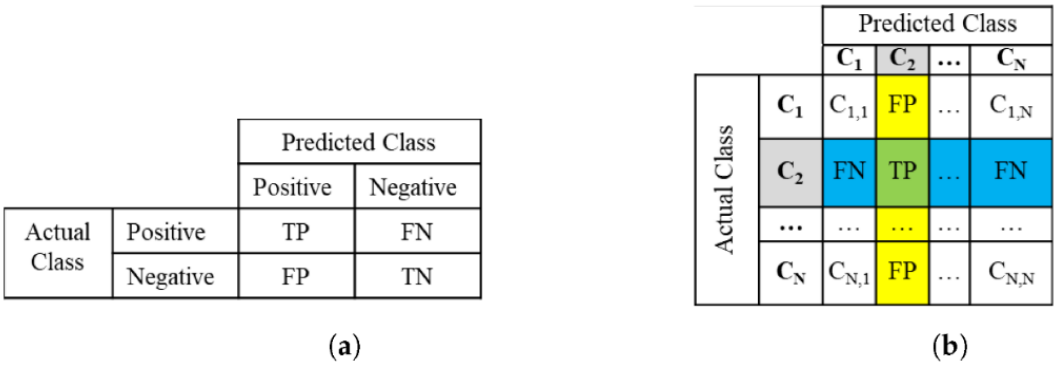


Fig. 11. Confusion matrix examples for binary and multiclass classification problems [159].

Each displayed column of the matrix indicates occurrences of a predicted class, whereas each row represents cases of an actual class. Confusion matrix can provide insight not just into the errors made by a classifier, additionally the kind of errors that result [159]. The description of the main evaluation metrics, including what they measure and how they work are displayed in Table 6.

Table 6. Evaluation metrics [158–160].

Metric	What it measures	How it works
Accuracy	Measures the proportion of correctly classified instances among all instances.	$\frac{TP}{\sum(TP+FP+FN)}$
Precision	Measure how many instances classified as a specific class were correct.	$\frac{TP}{TP_i+FP_i}$
Specificity	The proportion of instances not belonging to class *i* that were correctly classified as not being in class *i*.	$\frac{TN}{TN_i+FP_i}$
Recall	Measures how well a model identifies actual instances of class *i*.	$\frac{TP}{TP_i+FN_i}$
F1	The harmonic mean of precision and recall, balancing false positives and false negatives.	$2 \cdot \frac{Presicion_i \cdot Recall_i}{Precision_i + Recall_i}$

2.2.4. Development of prognostic model

The next step in the radiomics process will be to create a prediction model based on delta radiomics features to determine which chemotherapy types result in bigger changes in delta radiomic features and which delta characteristics can be related with better survival outcomes. Given that this is still a classification task with complex, high-dimensional data, a tree-based machine learning technique will be used. Tree-based models are especially well-suited to perform such tasks because of their high accuracy, resistance to overfitting, and ability to capture non-linear correlations and feature interactions. Standard evaluation measures will be used to verify that the model's performance is both accurate and well interpretable. Additionally, statistical tests will be utilized to validate whether the results reflect meaningful differences rather than random variation.

2.2.4.1. Categorical boosting (CatBoost)

Categorical Boosting is known as a supervised machine learning approach, from the family of gradient boosting models, which builds a series of decision trees where each new tree attempts tries to correct the errors made by the previous ones. This approach uses the gradient of the loss function

to guide the learning process and improve model performance [161]. The scheme of this approach is depicted in Fig. 12.

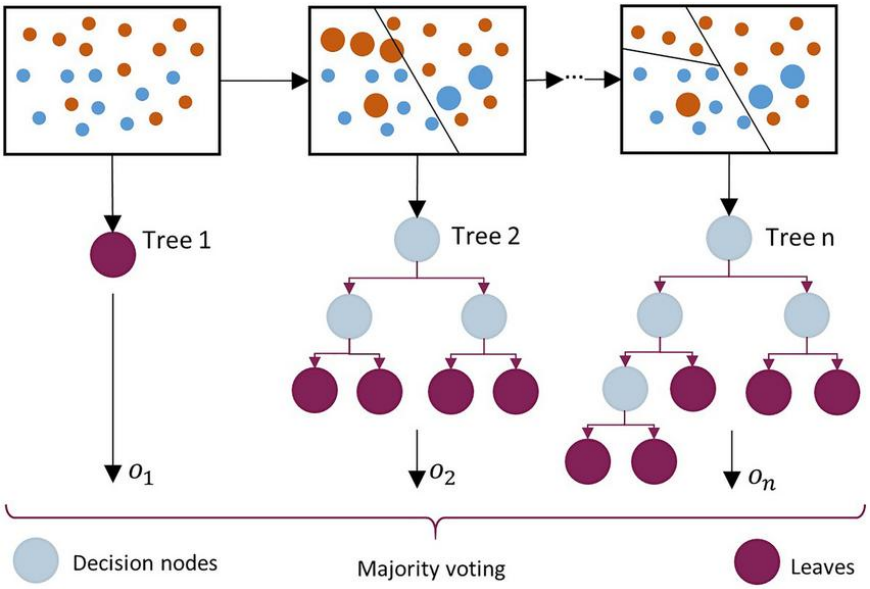


Fig. 12. Illustration of Categorical Boost approach [162].

At first glance, CatBoost appears to be similar to the XGBoost algorithm employed in this study for feature selection. However, CatBoost is frequently seen as a more advanced alternative, particularly when working with categorical data. Usually, gradient boosting algorithms require categorical variables to be transformed to numerical values using techniques such as one-hot or label encoding. These strategies can result in loss of data and overfitting [163]. CatBoost, on the other hand, handles categorical data natively by employing ordered target statistics, a method that converts categorical values into numerical representations based on the average target value while carefully avoiding data leakage. Another important contrast is that CatBoost uses ordered boosting rather than classical boosting, which uses the full dataset to train each tree. Ordered boosting imitates an online learning process by training each tree only on data that would be accessible at prediction time. This technique reduces overfitting and creates a more generalizable model [161, 164, 165].

2.2.4.2. Performance evaluation

At this point in the study, it is very important to note that binary and multiclass classification tasks require different performance evaluation methods. For the multiclass task, the evaluation was done to find out not only how well the models predict the target variable overall, but also how well the CatBoost method predicts each class of the target variable for the multiclass task. Thus, for this case, the One-vs-Rest ROC curve will be used described previously in section a 2.3.3.3.1. As the prediction of each class is very important at this stage of the study, the ROC curves will be constructed to reflect not only how well the model performs, but also how well each class is predicted compared to the others (One Class Strategy versus Others). In the binary case, a standard ROC curve will be used to evaluate the performance of the model. In addition, for both cases, the confusion matrix will be employed to evaluate the accuracy of the model and how well models are able to distinguish between positive and negative classes (see Fig. 11).

2.2.4.3. Statistical test

Statistical analysis is required in scientific research to objectively assess whether observed differences or connections in data are likely to be actual effects rather than random noise. Statistical tests can provide a formal mechanism for hypothesis testing, allowing researchers to draw accurate and reproducible conclusions from empirical data [166]. In this study, statistical tests will be used to see if delta radiomic characteristics differ significantly between clinical variables groups. In the treatment type analysis, the Kruskal-Walli's test, followed by Dunn's post hoc test, will be used to determine whether different therapies are associated with distinct patterns of change in delta radiomic features, indicating a treatment effect on imaging-derived biomarkers. In the case of survival outcome analysis, the Mann-Whitney U test will be used to determine whether delta characteristics differ substantially between patients who survived and those who did not. These results will indicate delta radiomic features potential in prognostic relevance.

2.2.4.3.1. Kruskal–Walli's test

The Kruskal-Walli's test is a nonparametric statistical test that detects statistically significant variations in the distribution of a continuous variable between two or more independent groups. Unlike ANOVA, it does not imply normality or equal variances, making it ideal for not normally distributed data or features containing outliers. In this investigation, the Kruskal-Walli's test will be employed to determine whether the distribution of delta radiomics characteristics varies significantly among treatment types, hence assessing the potential relationship between treatment and imaging-derived biological alterations [167]. In this case, hypotheses will be formulated as follows:

$$H_0: The\ distribution\ of\ delta\ feature\ X\ does\ not\ significantly\ differ\ between\ different\ treatment\ types$$

$$H_a: The\ distribution\ of\ delta\ feature\ X\ differs\ significanlty\ in\ at\ least\ one\ treatment\ group$$

This study uses a significant level of 0.05 ($\alpha = 0.05$), which corresponds to a 95% confidence interval. For example, if the hypnosis ends up resulting p-value is less than 0.05, the null hypothesis is considered as rejected, showing that the distribution of the delta radiomics feature differs significantly between treatment groups [167]. This result can indicate that the type of treatment may be related to detectable changes in imaging-derived features.

2.2.4.3.2. Post-Hoc analysis - Dunn's test

If the Kruskal-Walli's test results showed statistically significant results, then Dunn's tests can be employed to check at which exact categories the difference is significant. Dunn's test is performed pairwise, by comparing each independent group [168]. Hypotheses for this situation will be constructed for each pairwise comparison in the following manner:

$$H_0: The\ distribution\ of delta\ feature\ X\ is\ the\ same\ in\ both\ A\ and\ B\ \ treatment\ types$$

$$H_a: The\ distribution\ of\ delta\ feature\ X\ is\ different\ in\ A\ and\ B$$

Dunn's test will be performed with a significance level α = 0.05. If the p-value is less than 0.05, the null hypothesis is rejected, indicating a substantial difference in delta feature value change across treatments. In addition, to avoid inflated false positive rates and to produce more statistically trustworthy and reproducible results, the Benjamini-Hochberg (FDR Control) approach will be used [169, 170].

2.2.4.3.3. Mann–Whitney U test

Mann-Whitney U tests, commonly referred to as the Wilcoxon Rank Sum test, will be used to determine whether there is a significant difference between two groups. This test is only applicable when dealing with binary data, hence in this example, the patient lived or died following the treatment [171].

$$H_0: There\ is\ no\ significant\ difference\ in\ the\ distribution\ of\ delta\ features\ X\ between\ the\ two\ groups$$

$$H_a: The\ distribution\ of\ delta\ feature\ X\ differs\ significantly\ between\ the\ two\ groups$$

Consistent with previous statistical tests, the Mann–Whitney U test will be conducted using a significance level of α = 0.05. If the resulting p-value is less than 0.05, the null hypothesis will be rejected, indicating a statistically significant difference between the two groups.

3. Results

To demonstrate the practical use, reliability and effectiveness of the developed workflow of the radiomics process, two illustrative examples are provided below. They depict the entire step-by-step procedure of radiomics prognostic model's development, from data preparation to feature processing, model training, and evaluation. The most important methodological decisions are thoroughly explored, together with their justifications and implications for the model's overall performance and interpretation. The included examples provided a comprehensive overview of how the method could potentially be applied to real-world datasets in medical physics.

3.1. Analysis of Delta Features Changes Across Treatment Types

When implementing machine learning methodologies, it is important to understand the dataset's structure and variable distribution to ensure model robustness and avoid bias caused by class imbalance. The first step of this study was to examine the distribution of the target variable, which determines the type of chemotherapy treatment. The Chi-square goodness-of-fit test was used to determine whether the observed class frequencies deviated significantly from a uniform distribution.

Table 7. The frequencies of target variable.

0	**1**	**2**	**Chi-Square Goodness-of-Fit Test Result**
Original data set			
23	24	8	0.01
Data set after synthetic generation			
23	24	24	0.98

As seen in Table 7, the original dataset was highly imbalanced ($p = 0.01$), with treatment groups 0, 1, and 2 containing 23, 24, and 8 observations respectively. To compensate for this imbalance, synthetic oversampling was used to boost the representation of Treatment type 2 to 24 cases. After generating synthetic data, a repeated Chi-square test revealed no significant variation from a uniform distribution ($p = 0.98$). By establishing class balance prior to model training, the likelihood of bias towards more frequent treatment groups is eliminated, which is critical for developing models that can generalize effectively to new data. In clinical applications, ensuring fair representation among treatment categories helps to avoid skewed predictions, which could lead to misinterpretation of treatment outcomes. Achieving a homogeneous distribution of classes improves the reliability of the created models, increasing their potential for accurate and therapeutically relevant decision assistance. These results, shown in Table 7, confirm that the dataset is well balanced, giving a solid platform for future feature selection, model development, and clinical deployment.

To minimize feature redundancy and address multicollinearity, a pairwise correlation analysis with Spearman coefficients was performed after class balance evaluation but before machine learning-based feature selection. Reducing redundancy of features, it is critical to guarantee that the final model captures significant, non-redundant patterns in the imaging data, thus boosting diagnostic reliability and supporting clinically interpretable outcomes. Features with correlation coefficients larger than 0.9 were deemed highly collinear; in such circumstances, one trait from each associated pair was removed while keeping the one with higher variance or more clinical importance. Fig. 13 depicts this

procedure, with the left matrix showing the correlation structure of the original delta radiomic feature set, and the right matrix showing the reduced set following collinearity-based filtering.

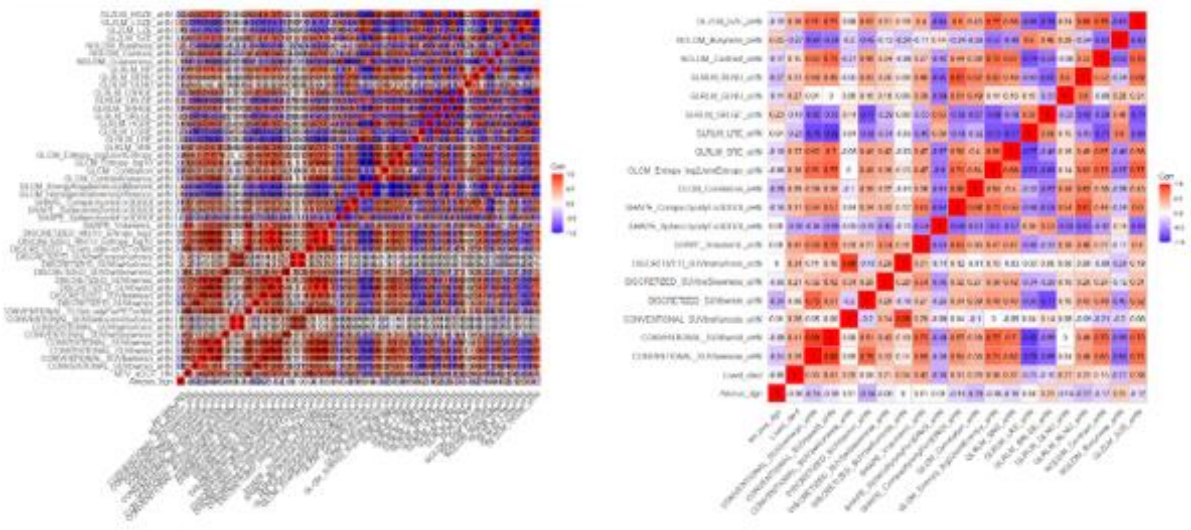


Fig. 13. Redundancy reduction using pairwise correlation analysis (The full-size correlation matrices are given in Appendix 1).

Although correlation thresholding reduces redundancy by removing highly collinear data, it may not always determine which features contain the most predictive or therapeutically important information. As a result, selecting only the most significant characteristics is critical for improving model interpretability, reducing overfitting, and ensuring that the final prediction model is statistically robust and therapeutically relevant. In this study, two machine learning algorithms, Recursive Feature Elimination with Random Forest (RFE-RF) and Extreme Gradient Boosting (XGBoost), were used to systematically identify the most relevant delta radiomic characteristics. To evaluate the effectiveness of these algorithms and decide which method selects significant features more accurately for further model building, their performance was assessed using the indicators and processes provided in Section 2.2.3.3. Table 8 summarizes the comparative performance metrics for the RFE-RF and XGBoost feature selection algorithms.

Table 8. Performance metrics of feature selection algorithms.

Metric	Recursive Feature Elimination with RF	eXtreme Gradient Boost
Accuracy	070	0.69
Precision	0.67	0.67
Specificity	0.83	0.84
Recall	0.76	0.68
F1	0.69	0.67

Table 8 shows that the overall accuracy of the feature selection approaches was similar (0.70 for RFE-RF and 0.69 for XGBoost), as was their precision (0.67). However, in a broader sense, the RFE-RF-based model had higher recall (0.76 vs. 0.68) and F1-score (0.69 vs. 0.67), indicating a more balanced capacity to properly identify and categorize treatment kinds. Although XGBoost had slightly greater specificity (0.84 vs. 0.83), the overall results indicate that the RFE-RF model, in this study case, is a more efficient feature selection strategy for this task. From the standpoint of therapy, the higher recall and F1-score acquired with RFE-RF are especially significant, implying a greater ability

to properly identify distinct treatment categories. This enhanced differentiation across therapy kinds increases the model's potential clinical application, providing more dependable guidance for treatment assessment and decision-making.

These standard classification measures helped to evaluate the model's performance, but ROC curves and AUC values provide a more thorough assessment of the discriminatory strength of all classes (treatment types). Figure 14 depicts the macro-averages of the ROC curves for both models.

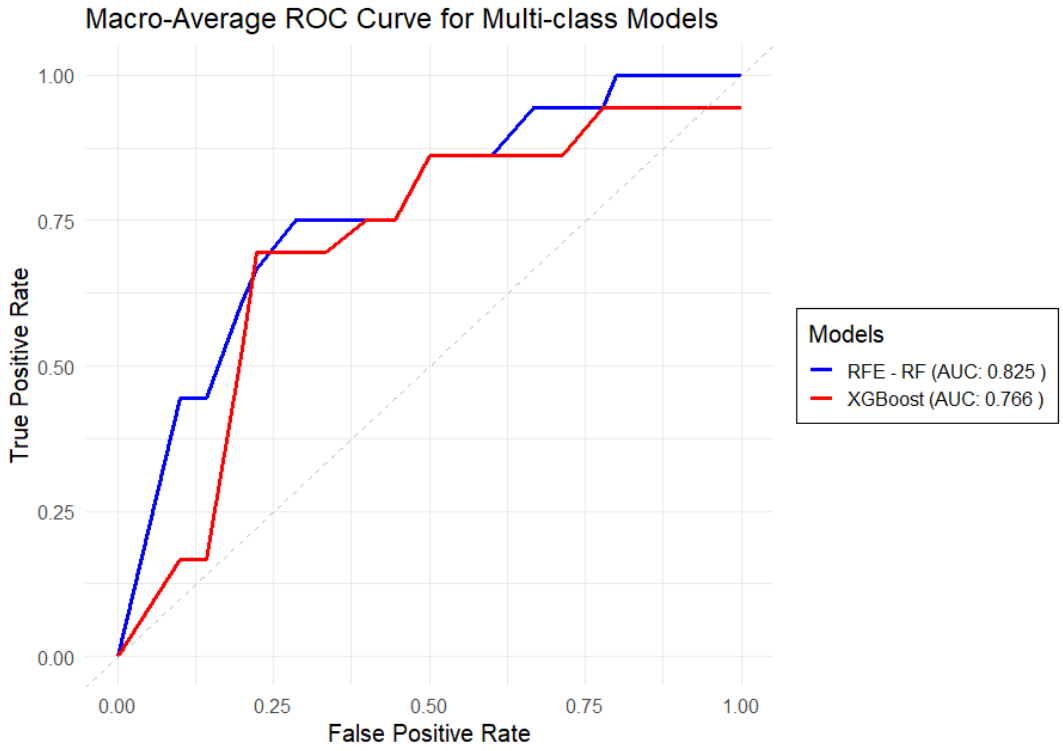


Fig. 14. Macro-average ROC curves comparing multi-class model performance using RFE-RF and XGBoost feature selection methods.

Fig. 14 shows the macro-averaged ROC curves for multi-class classification models created with two feature selection strategies: RFE-RF (blue) and XGBoost (red). As seen in the figure, the RFE-RF model had a higher macro-averaged AUC of 0.825, showing stronger overall discriminative capacity across all classes than the XGBoost-based model, which had an AUC of 0.766. RFE-RF's ROC curve consistently exceeds that of XGBoost across the majority of the False Positive Rate (FPR) range, notably in the 0.25-0.75 interval, where models are most used in clinical settings. This interval shows the delicate balance of minimizing false positives while keeping high sensitivity, which is critical for creating models for clinical decision assistance. A better trade-off in this range indicates that the RFE-RF model can correctly classify patients into distinct treatment types while decreasing the risk of misclassification. Clinically, this increased discrimination can contribute to more accurate treatment assessments, better patient stratification, and eventually more dependable support for therapeutic decision-making. Although both models perform well (AUC > 0.7), RFE-RF appears to be the more effective strategy for selecting delta radiomic features for treatment categorization in this multi-class setting. In clinical terms, a greater AUC indicates a better capacity to distinguish between different chemotherapy treatment types, which is crucial for ensuring that predictive models give accurate and reliable support for treatment evaluation. Thus, based on the higher performance of the RFE-RF-based model, the relevant delta radiomic features revealed by this method will be used for further analysis.

Based on the feature selection results, the most significant delta radiomic features obtained using the RFE-RF approach were used to build a prediction model using the CatBoost algorithm. Following model training, the ten most significant delta radiomic characteristics were determined based on their contribution to predicting the target variable - treatment type.

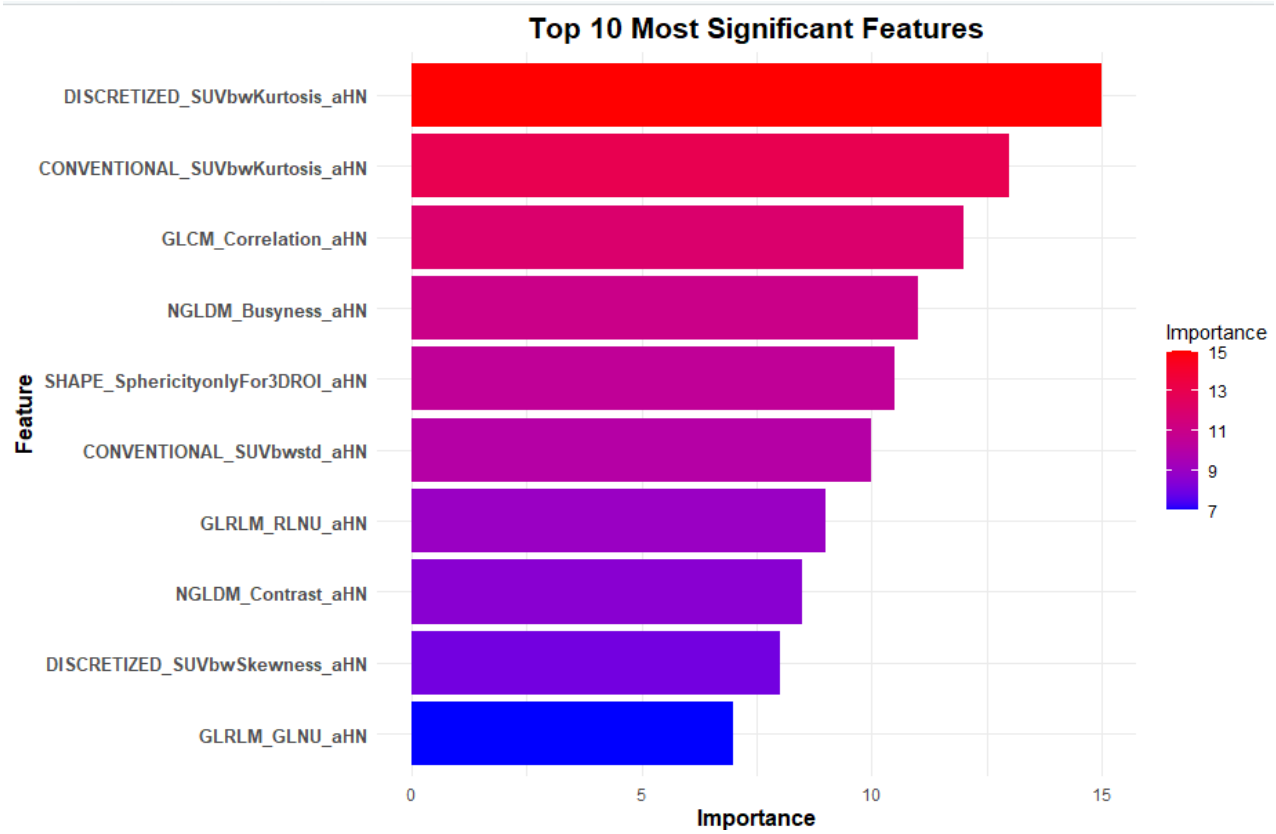


Fig. 15. Significant delta radiomic features identified by the CatBoost model for treatment type classification.

Fig. 15 depicts the most significant delta features detected by the CatBoost model, which was trained to distinguish chemotherapy treatment types. The features fall into six radiomic categories, as specified in section 2.1.: *conventional* (SUVbwstd, SUVbwKurtosis), *discretized intensity* (SUVbwKurtosis, SUVbwSkewness), *shape* (e.g., Sphericity), *GLCM* (Correlation), *GLRLM* (RLNU, GLNU), and *NGLDM* (Busyness, Contrast). Among the most significant features were DISCRETIZED_SUVbwKurtosis, CONVENTIONAL_SUVbwKurtosis, GLCM_Correlation, and NGLDM_Busyness_aHN. This suggests that intensity and textural traits were crucial in discriminating between treatment types. It's also worth noting that the method distinguishes between two SUV kurtosis metrics: discrete and conventional. This implies that changes in the peak or tails of the SUV distribution before and after treatment are substantially linked with the treatment type. The inclusion of GLCM_Correlation_aHN and NGLDM_Busyness_aHN among the top features emphasizes the significance of textural heterogeneity in PET/CT imaging, which may reflect tissue complexity and responsiveness to treatment. These qualities indicate that delta radiomics can aid in the non-invasive determination of how various types of treatment affect distinct tissue attributes. By detecting minor changes in tumor form and activity on imaging, the model can help clinicians make better treatment decisions and manage patient care more efficiently.

Once the model has determined the most important features for predicting the type of treatment, it is time to evaluate the CatBoost model's performance. Fig. 16 illustrates the discriminatory power of the ROC curves by class. As can be observed, treatment type 2 has the highest AUC, showing excellent class separability. The system also discriminates well between treatment types 0 and 1, with an AUC much higher than 0.85. All curves are significantly higher than the diagonal line (random guess), suggesting that the model consistently divides classes. From a clinical standpoint, this robust class separability indicates that delta radiomic features capture treatment-specific imaging characteristics, allowing the model to confidently discriminate between therapy types. This may help doctors better understand how different therapies alter tissue features and support the development of imaging-based instruments for planning treatment, monitoring, or stratification in precise oncology.

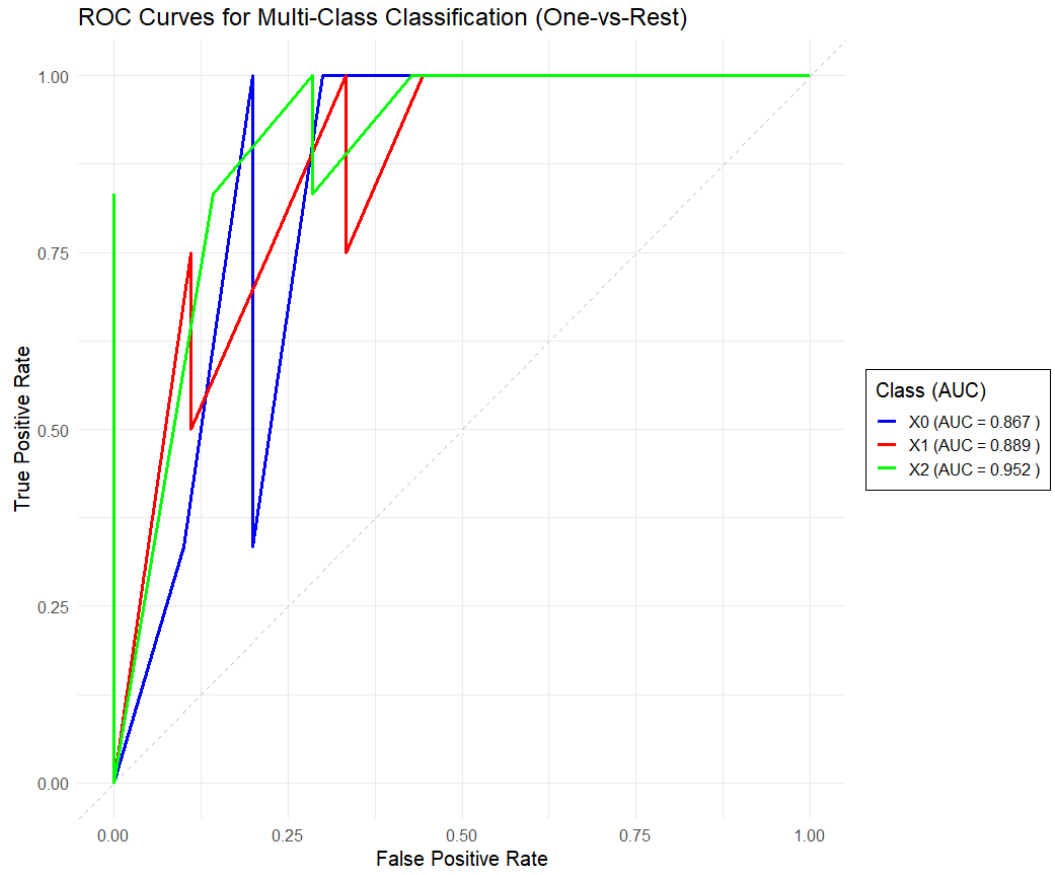


Fig. 16. Performance of the CatBoost approach.

In addition to ROC curves, it is also worth considering the overall performance metrics (see Table 9)

Table 9. Performance metrics of CatBoost algorithm.

Metric	CatBoost
Accuracy	0.71
Precision	0.64
Specificity	0.85
Recall	0.79
F1	0.65

In this case, the overall accuracy of the model is 71% and the recall is 79%, indicating that the model can perform reliably across different classes and successfully identifies the most relevant cases. The F1 rate (65%) in this case shows a moderate balance between precision and recall, while the high

specificity (85%) indicates that the model correctly identifies negative cases. Hence, the data is correctly identified. High recall and specificity are especially useful in clinical settings because they reduce the danger of misclassifying patients and ensure that different treatment types may be effectively differentiated based on imaging findings

Since the most predictive delta radiomic parameters were found to distinguish between different treatment types, and the model's performance on this task was judged to be good, it is now possible to analyze how these properties differ between treatment types. This following analysis allows for the identification of imaging biomarkers that reflect treatment-specific tissue alterations, offering additional insight into therapeutic response and tumor biology under various regimens. This part of the study focuses on the four factors that had the greatest impact on predicting the type of treatment.

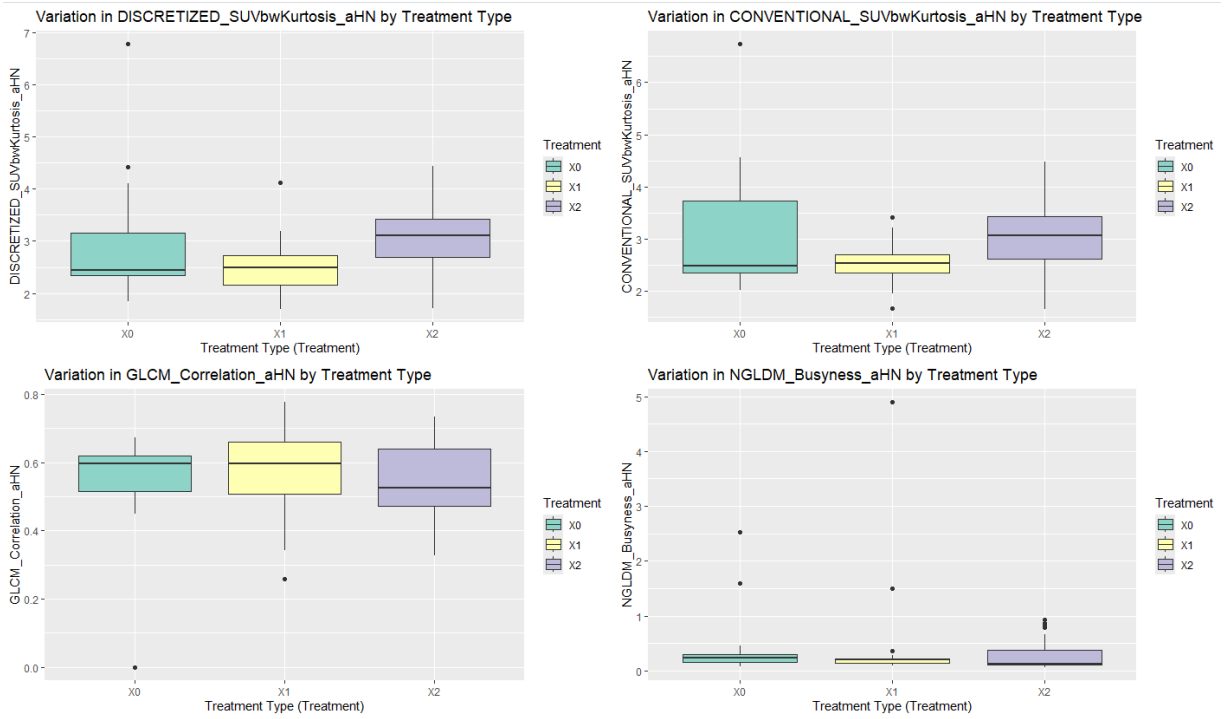


Fig. 17. Variation in significant delta radiomics features across different treatment types.

Figure 17 shows that a **DISCRETIZED_SUVbwKurtosis_aHN** delta feature has higher values in treatment type 2, showing that this form of treatment may create significant changes in the distribution's peakedness of SUV values following treatment. Type 1 has the lowest median and smallest dispersion, implying a more consistent or less variable effect on this feature. The median value for treatment type 0 was higher than for treatment type 1, and the distribution was wider on average, implying that this type of chemotherapy induces more changes in the highest point distribution of diversity features. Although less significant than in treatment type 2, variability in 0 indicates a possibly variable response among people taking this medication.

The **CONVENTIONAL_SUVbwKurtosis_aHN** delta feature follows a pattern comparable to the discretized version. 0 and 2 treatment types have larger variability and medians compared to treatment type 1. This suggests that treatment types 0 and 2 may cause more metabolic heterogeneity or alterations to tumor shape.

The median values of the **GLCM_Correlation_aHN** delta attribute, which describes changes in tissue texture, were higher for treatment type 1, which may indicate stronger correlations in local intensity induced by this treatment. Treatment type 2 is slightly more diffuse than treatment type 1, suggesting that the textural consequences are more variable, and hence this treatment produces different effects on patients. Treatment type 0 had the least spread values in this case, which means that it does not cause significant changes in tissue texture.

NGLDM_Busyness_aHN has higher values for treatment type 2 and more outliers, albeit smaller, which may indicate that it causes more complex or variable tissue responses. Treatment types 0 and 1 have extremely low median values, indicating modest changes or more consistent tissue responses. Although as can be seen from the boxplot treatment types 0 and 1 have a number of outliers with higher values, which may indicate that these textural changes are closely related to other features or reflect noise in the data. Taking together, these changes reinforce the importance of delta radiomics in capturing small but significant differences in tissue response between different types of chemotherapy treatment.

While boxplots offer an initial visual summary of how delta radiomic characteristics vary across treatment types, thorough statistical analysis is required to evaluate whether the observed variations are statistically significant. To this purpose, a Kruskal-Wallis H test was run on every one of the top-ranked traits to see if the distributions differed significantly between the three chemotherapy treatment groups. For features where the Kruskal-Walli's test revealed significant group differences, Dunn's post hoc test with multiple comparison correction was used to determine whether treatment pairs had statistically significant delta feature values.

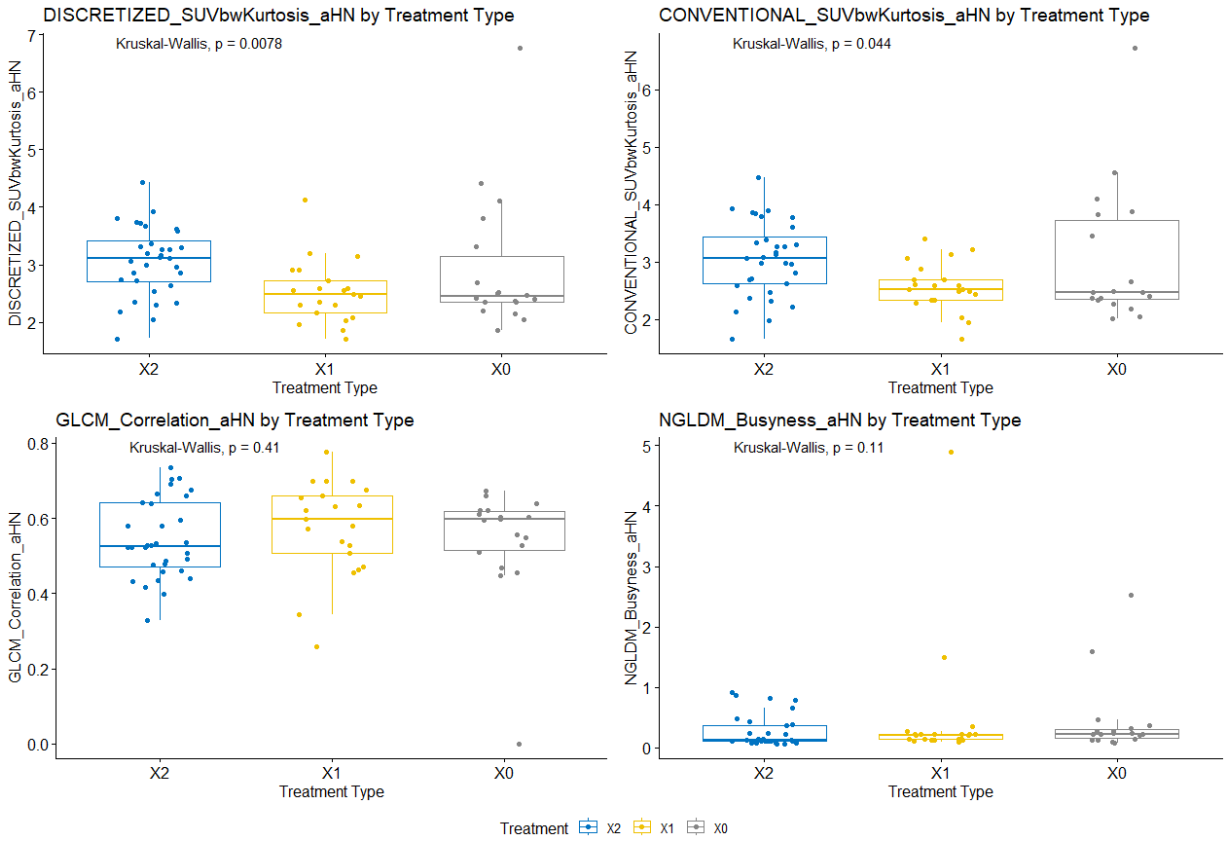


Fig. 18. The boxplot of delta radiomic features across chemotherapy treatment types with Kruskal–Walli's test results.

Statistical analysis using the Kruskal-Walli's test indicated that both top-ranked delta radiomic features: DISCRETIZED_SUVbwKurtosis (*p = 0.0078*) and CONVENTIONAL_SUVbwKurtosis (*p = 0.044*) varied significantly among chemotherapy treatment groups. Therefore, the next step is a post hoc Dunn's test with FDR correction. Results of both tests can be seen in Table 9.

Table 10. Statistical comparison of delta features across different treatment types.

Feature name	Kruskal Walli's	Dunn's test	
DISCRETIZED_SUVbwKurtosis_aHN	0.0078	0 vs 1	0.18
		0 vs 2	**0.047**
		1 vs 2	**0.004**
CONVENTIONAL_SUVbwKurtosis_aHN	0.044	0 vs 1	0.21
		0 vs 2	0.11
		1 vs 2	**0.023**
GLCM_Correlation_aHN	0.41	-	
NGLDM_Busyness_aHN	0.11	-	

Statistical analysis of the most important delta radiomic features revealed that only the kurtosis-based SUV characteristics **DISCRETIZED_SUVbwKurtosis** and **CONVENTIONAL_SUVbwKurtosis** differed significantly between chemotherapy treatment types. In simple terms, these characteristics revealed that patients getting type 2 treatment saw more significant changes in the shape of the SUV distribution than patients receiving type 1 treatment. For **DISCRETIZED_SUVbwKurtosis**, the difference between 2 and 0 was likewise borderline significant (0.047), supporting the fact that two treatments may cause bigger metabolic alterations.

Two texture-based features, **GLCM_Correlation** and **NGLDM_Busyness**, however, showed no statistically significant variations between treatment types, despite some obvious variances in box distribution. This shows that, while textural traits are crucial for overall delta radiomic analysis, they could be less sensitive to treatment-induced changes in this specific clinical situation.

In summary, the statistically significant difference between DISCRETIZED_SUVbwKurtosis and CONVENTIONAL_SUVbwKurtosis over the types of chemotherapy demonstrates their ability to detect biologically relevant, treatment-induced metabolic alterations. Although texture-based variables such as GLCM_Correlation and NGLDM_Busyness did not approach statistical significance, their variability suggests that they may have complementary value in a wider multiparametric context. The successful selection of kurtosis-based SUV measurements as discriminative features demonstrates the efficacy of the proposed delta radiomics methodology in identifying clinically useful imaging biomarkers. If these results are externally verified with independent, multi-institutional datasets, the established methodology has the potential to significantly improve non-invasive response examination, guide treatment decision-making, and enable individualized treatment options in oncology.

3.2. Delta Radiomic Features as Predictors of Survival

As in the previous case, the frequency distribution of the target variable was examined using the Chi-Square Goodness-of-Fit Test. Since the null hypothesis was rejected (*p-value is 0.04*), suggesting a severe class imbalance, a synthetic data generation procedure was used to produce more observations. Following augmentation, the Chi-Square Goodness-of-Fit Test revealed that the class distribution

was no longer statistically different (*p-value is 0.57*), indicating that the data set is now balanced and ready for further research (see Table 11).

Table 11. The frequencies of the target variable.

0	1	Chi-Square Goodness-of-Fit Test Result
Original data set		
20	35	0.04
Data set after synthetic generation		
40	35	0.57

As previously stated, a correlation analysis was performed here to eliminate redundant variables, with a threshold of 0.9. This approach, as was justified in the previous subsection, ensures that the retained characteristics give independent information for downstream modeling. The correlation matrices before (left) and after (right) the removal of redundant variables are presented in Fig. 19.

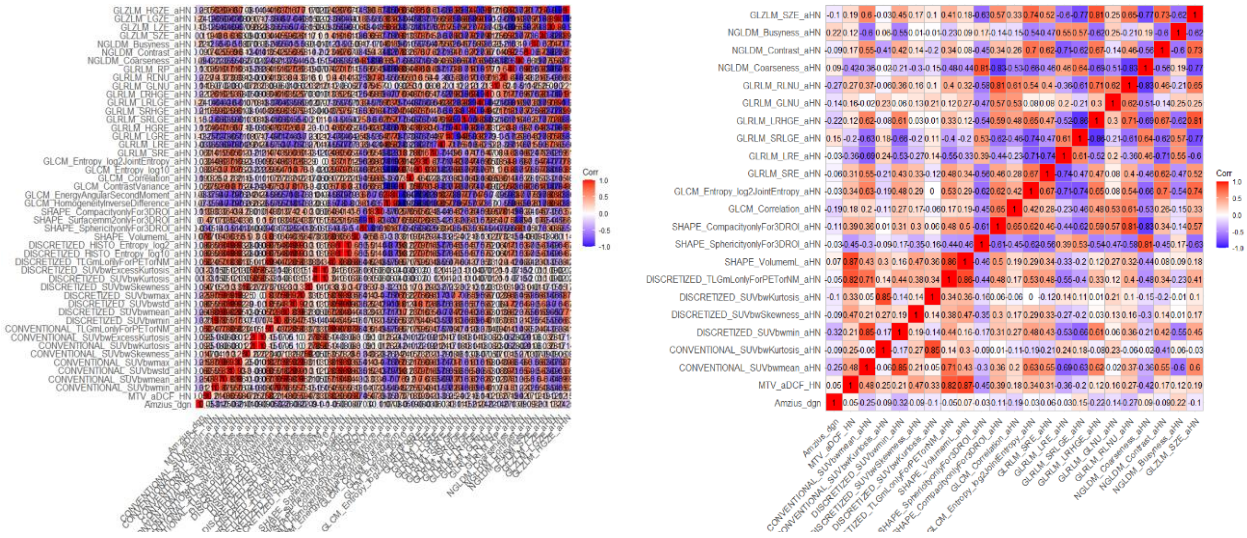


Fig. 19. Redundancy reduction by using correlation analysis (The full-size correlation matrices are given in Appendix 1).

As in the previous case, following the correlation analysis, the next step was to identify the most significant delta radiomics features. Two feature selection algorithms, Recursive Feature Elimination with Random Forest (RFE-RF) and XGBoost, were employed to compare their effectiveness in selecting informative features. The performance evaluation of both approaches is presented in Table 12, allowing for a comparative analysis of their accuracy and feature selection capability. As in the previous model, careful feature selection at this stage is critical to preserve clinically meaningful imaging biomarkers, ensuring that the produced models remain robust and translationally relevant for clinical applications.

Table 12. A comparison of performance metrics of feature selection algorithms.

Metric	Recursive Feature Elimination with RF	eXtreme Gradient Boost
Accuracy	0.86	0.78
Precision	0.95	0.90
Specificity	0.68	0.50
Recall	0.71	0.76
F1	0.64	0.68

In this analysis, RFE-RF achieved the highest overall accuracy (0.86) and precision (0.95), indicating that the model based on these delta features made highly accurate predictions with few false positives. It also demonstrated higher specificity (0.68), meaning it was better at correctly identifying negative instances compared to XGBoost. XGBoost, while slightly behind in accuracy (0.78) and precision (0.90), outperformed RFE-RF in recall (0.76) and F1 score (0.68), suggesting it captured more true positives and achieved a better balance between precision and recall. As in the previous case, both models performed quite well; however, the features selected by RFE-RF were used for further model development, as this method demonstrated better performance in correctly identifying false positive observations. From a clinical standpoint, our findings indicate that the RFE-RF-based model may give more reliable support in predicting patient survival outcomes while reducing the possibility of misclassification. High accuracy and specificity are especially crucial when using prognostic models to identify patients who are at a higher risk of poor outcomes, allowing for early intervention or more frequent clinical surveillance. This approach helps to improve the reliability of imaging-based survival forecasts, resulting in more tailored and informed patient treatment.

To further evaluate the predictive performance of the selected delta radiomics features, a comparative ROC analysis was performed for the models developed using features selected by RFE-RF and XGBoost algorithms (see Fig. 20).

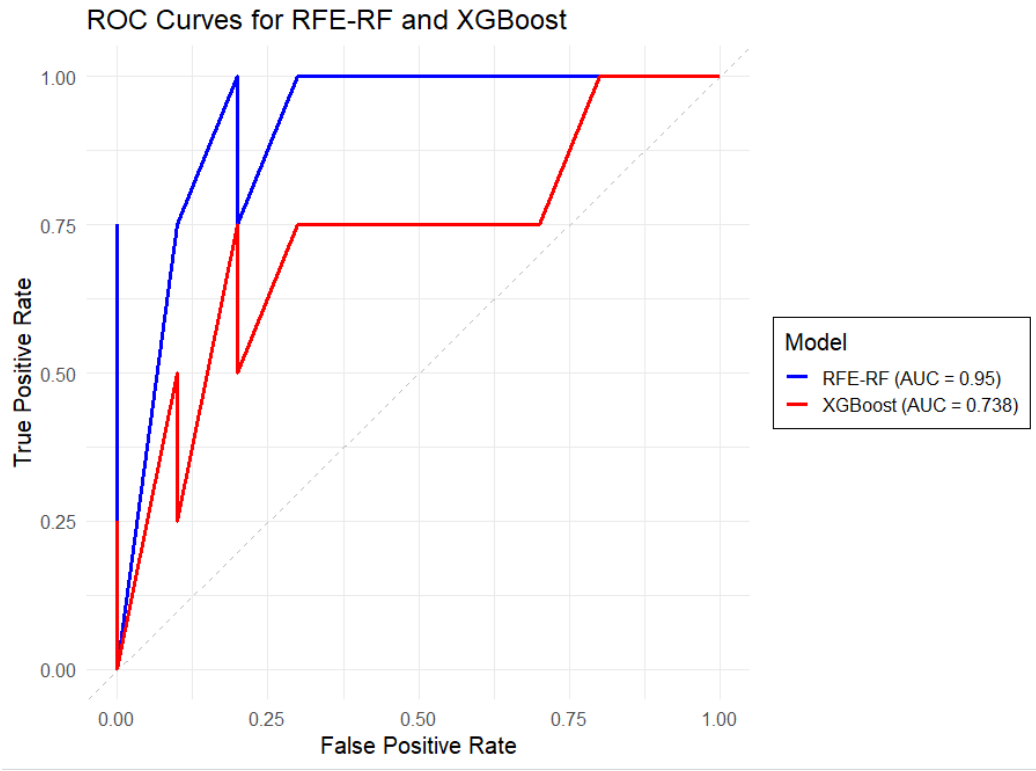


Fig. 20. Comparison of the performance of two ML approaches using ROC Curves.

The results showed that the model using RFE-RF-selected features had a significantly higher Area Under the Curve (AUC) value of 0.95, compared to 0.738 for the XGBoost-based model. This significant difference implies that the RFE-RF resulted in improved class separability and prediction performance. The ROC curve of the RFE-RF model was continuously higher over the entire range of false positive rates, indicating a better sensitivity-specificity trade-off. In clinical terms, a larger AUC

indicates that the model can more reliably predict patient survival given imaging data. This is significant because it allows clinicians to monitor the patient's reaction to treatment and determine which patients require more aggressive treatment or closer monitoring.

In the next stage, the CatBoost algorithm was applied to develop the prognostic model. During this process, the most significant features were identified (see Fig. 21).

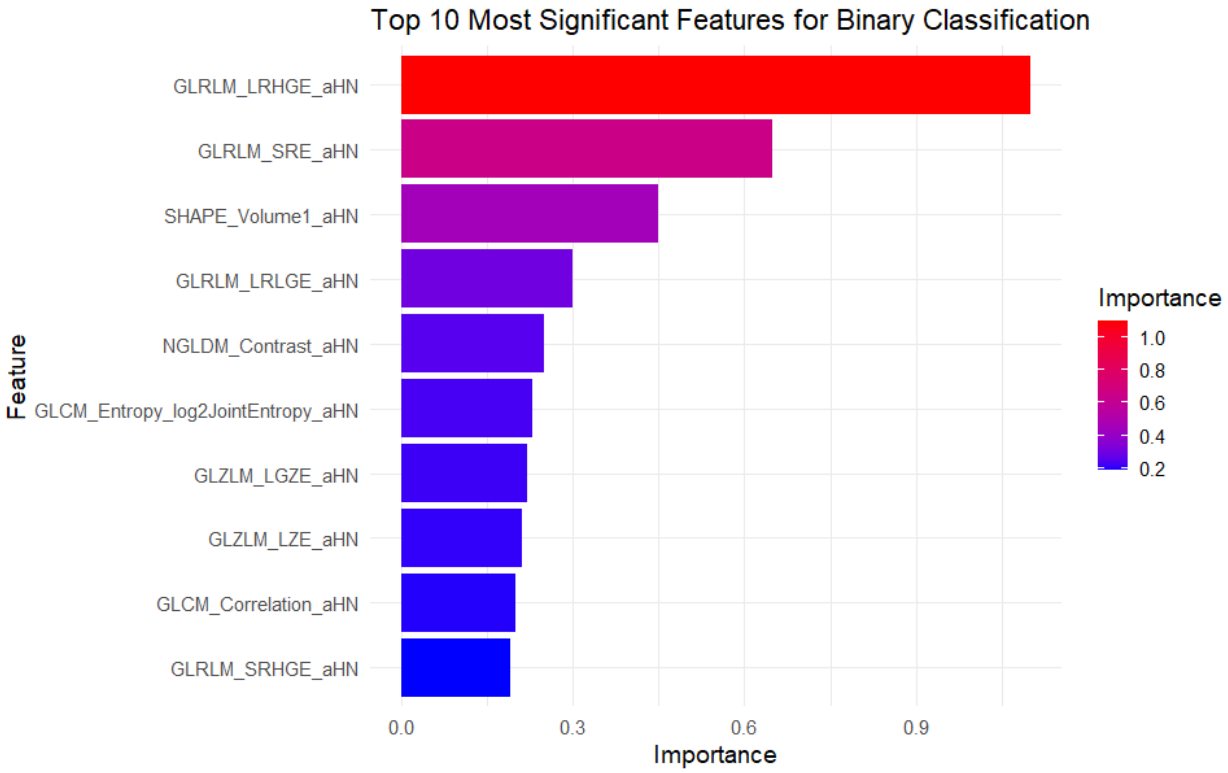


Fig. 21. The most significant features by CatBoost.

In this case, GLRLM characteristics emerged as the most significant category, indicating that changes in run-length texture patterns in survival analysis are highly predictive of the binary classification test. Shape as well as GLCM-based heterogeneity aspects also have an important impact. Interestingly, conventional along with discretized intensity features were not placed among the top ten, as they were in the prior case, when survival outcomes of effect on delta features were investigated, implying that greater complexity textural and structural alterations may be more useful for patient survival modeling in this setting. These findings imply that survival outcomes may be linked to minor changes in tumor microarchitecture and heterogeneity rather than simple intensity changes. This shows that in clinical practice, texture-based radiomic indicators may aid in capturing underlying biological aggressiveness, tissue disorganization, and response failure, all of which have a significant impact on patient prognosis. Recognizing and quantifying these complicated imaging patterns can improve risk categorization and treatment planning.

Since the model has determined the most important features for predicting the survival outcome, it is time to evaluate the CatBoost model's performance. To assess the model’s effectiveness, key performance metrics and the ROC curve were employed to measure the model’s classification ability and discriminative power (see Table 13 and Fig. 22).

Table 13. Performance metrics to evaluate the CatBoost approach.

Metric	CatBoost
Accuracy	0.69
Precision	0.80
Specificity	0.65
Recall	0.68
F1	0.59

With an accuracy of 69% and a high precision (0.80), the model reliably detects positive situations while producing few false positives. This is useful in applications where false alarms should be avoided. However, this model has a modest recall (0.68), implying that only about 68% of real positive cases are successfully detected. Notably lower than accuracy, indicating that the model misses some true cases. Additionally, the model exhibits low specificity (0.65). The capacity to correctly recognize negative situations is limited, showing opportunity for improvement in managing the negative class. A low F1 Score (0.59) indicates a moderate balance of precision and recall, but not optimal. From a clinical standpoint, these findings suggest that the model accurately predicts high-risk patients who did not survive but fails to detect a significant proportion of true positive cases. While the model's moderate specificity helps to alleviate unwarranted concern for certain patients who are likely to live, its limited recall and inadequate F1 score emphasize the potential of missing individuals who are at risk of poor outcomes. In clinical practice, it is critical to improve sensitivity to avoid missing patients who could benefit from early care, while also preserving sufficient precision to avoid overtreatment. Training on a larger dataset is anticipated to improve the model's performance, notably its capacity to consistently identify patients in danger of mortality.

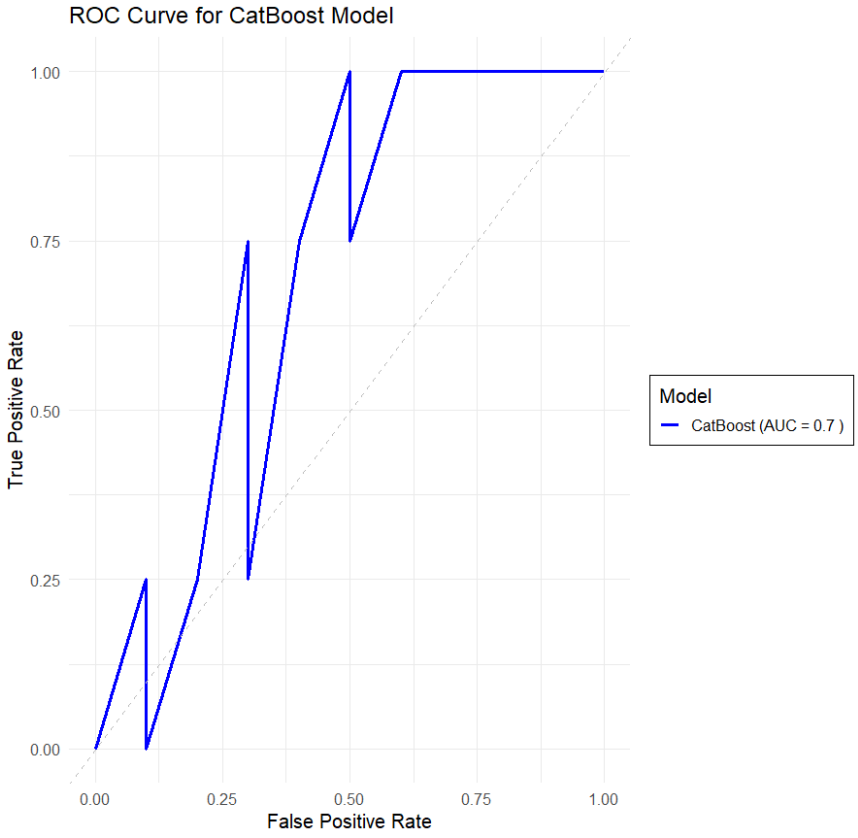


Fig. 22. The ROC curve for CatBoost model in binary classification.

The model distinguishes the two classes relatively well, with an AUC of 0.70. The ROC curve reveals some capacity to distinguish between classes above random guessing (AUC > 0.5). While the ROC curve rises above the diagonal, it does not have a consistently steep slope, implying that the model's sensitivity and specificity trade-offs vary with decision threshold. From a clinical point of view, such results may have some predictive value, but it would be worth expanding the dataset to make the model more efficient.

Given that the relevant aspects linked with the target variable (lived/died) have already been found, it is now possible to assess how each delta radiomics feature differs between patients who survived as opposed to those who did not after treatment. This technique could help predict treatment outcomes, like survival. The Fig. 23 illustrates how delta features differ between the patients who survived (0) and those who died after the treatment (1).

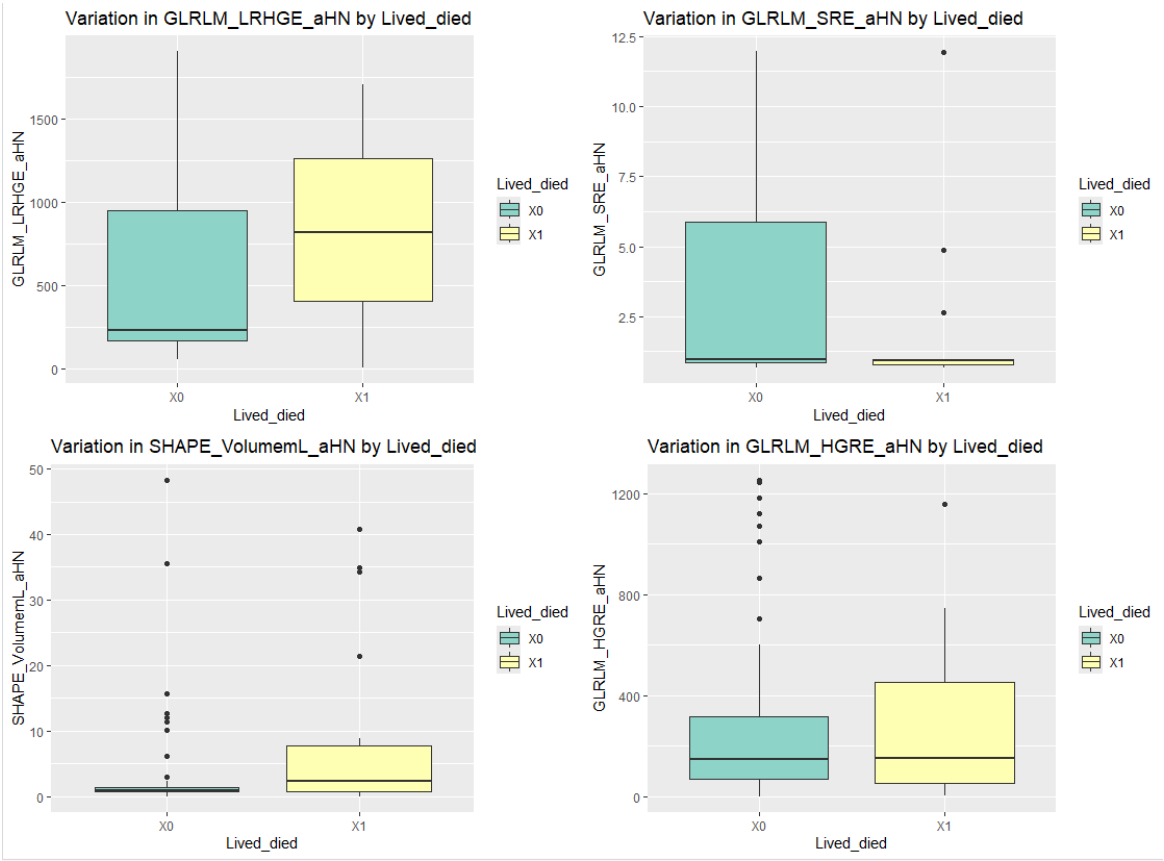


Fig. 23. Variation in significant delta radiomics features across survival outcomes.

Patients who died (1) demonstrated greater post-treatment increases in Long Run High Gray-Level Emphasis (GLRLM_LRHGE_aHN), than those who survived (0). This could indicate increased tissue heterogeneity or complexity linked with more aggressive illness. Patients that lived (0) had higher GLRLM_SRE_aHN values, which could equate to finer, more uniform texture patterns, implying a greater therapeutic response. Fig. 23 shows a significant difference, indicating substantial discriminative capability. The surviving patient (0) has relatively low volume changes (SHAPE_VolumemL_aHN), but the deceased patients (1) have greater volume increases, indicating tumor progression or inadequate shrinking post-treatment. Deceased patients (1) likely to have higher GLRLM_HGRE_aHN values, presumably indicating the presence of high-intensity, coarse textures post-treatment – a symptom of lingering aggressive tumor components. Although major differences

in delta characteristics between deceased and surviving patients are visible in Fig. 23, statistical analysis will be used to assess whether these differences are statistically significant.

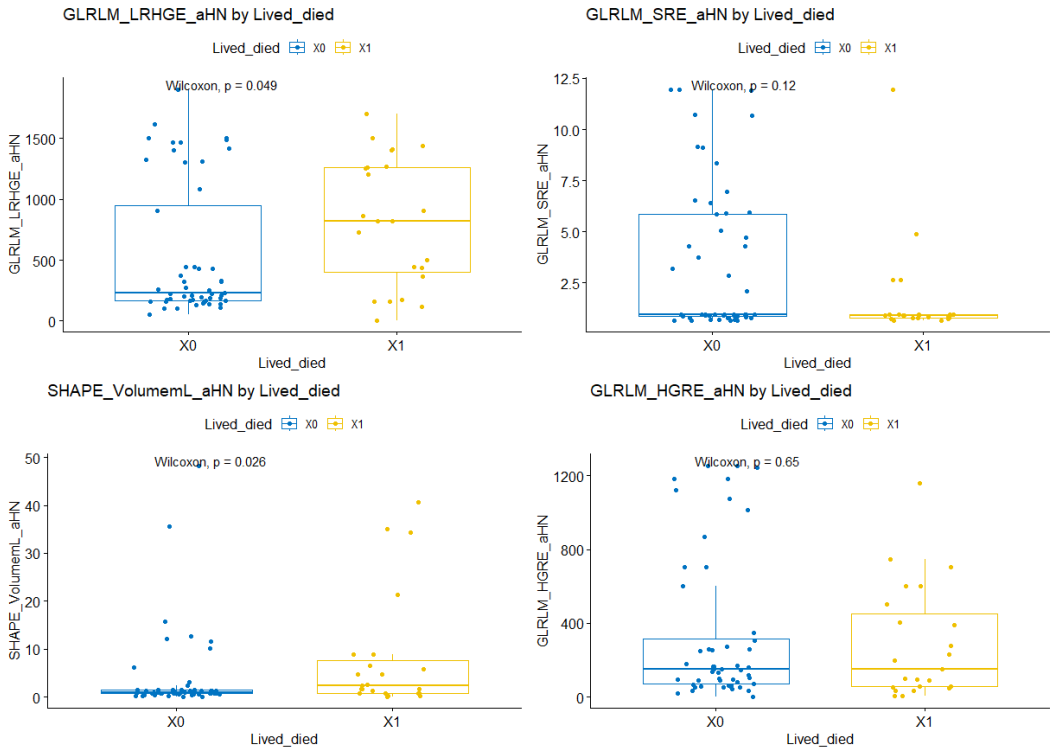


Fig. 24. Mann–Whitney U test for delta features differentiating deceased and surviving patients.

Fig. 24 depicts the distribution of the four most significant delta radiomics features, GLRLM_LRHGE_aHN, GLRLM_SRE_aHN, SHAPE_VolumemL_aHN, and GLRLM_HGRE_aHN, between two patient outcome groups: survived (0) and died (1) after therapy. To determine if the observed changes in feature distributions were statistically significant, the Mann-Whitney U test was used. This non-parametric test allows for the comparison of feature values between the groups without assuming normal distribution, giving insight into the possible prognostic value of each radiomic feature. Results can be seen in table 14.

Table 14. Statistical significance of feature differences between survived and deceased patients.

Feature	Mann–Whitney U test
SHAPE_VolumemL_aHN	**0.026**
GLRLM_LRHGE_aHN	**0.049**
GLRLM_SRE_aHN	0.12
GLRLM_HGRE_aHN	0.65

The Mann-Whitney U test found that **SHAPE_VolumemL_aHN** (*p = 0.026*) differed statistically significantly from the two outcome groups, implying that changes in tumor volume throughout therapy may be related with patient survival. Greater tumor shrinking during treatment may indicate a stronger therapeutic response, a lower tumor burden, and thus better survival prospects. In contrast, persistent or rising tumor volume may signal treatment resistance and poorer outcomes. **GLRLM_LRHGE_aHN** also neared statistical significance (*p = 0.049*), indicating a possible relationship between strong gray-level run emphasis and poor outcomes. Clinically, this may reflect a more structured, densely cellular tumor structure, possibly related to the aggressive type of tumor

and its resistance to treatment. This borderline outcome calls for additional exploration in a larger cohort. In contrast, **GLRLM_SRE_aHN** ($p = 0.12$) and **GLRLM_HGRE_aHN** ($p = 0.65$) revealed no statistically significant differences between survivors and non-survivors. This finding suggests that, in the current group fine scale intertumoral textural heterogeneity (GLRLM_SRE_aHN) and the amount of extremely dense, radiologically bright tumor areas (GLRLM_HGRE_aHN) may not independently predict survival. These textural characteristics, despite indicating microstructural complexity or fibrotic content, may not have a strong direct association with therapeutic response or biological aggression in this environment.

Lastly, the findings imply that SHAPE_VolumemL_aHN and GLRLM_LRHGE_aHN could be potential non-invasive imaging biomarkers for predicting survival outcomes. Their statistical correlation with survival outcome variables suggests that they may reflect therapy efficacy and tumor aggressiveness. Other textural characteristics were not significant, yet they may potentially contribute to a more complete tumor characterization. In general, the proposed delta radiomics workflow successfully detects prognostically significant features; nevertheless, external validation in independent, multi-institutional cohorts is required to ensure their clinical utility and to validate them as accurate non-invasive biomarkers for personalized cancer management.

It is crucial to note that the results of the developed models may only be externally validated or properly compared to similar studies if every stage of the radiomics workflow, from image acquisition to model development, are standardized and implemented employing consistent protocols, ensuring data integrity and methodological comparability.

Conclusions

1. Radiomics has advanced greatly from its beginnings in texture analysis in the 1970s, becoming an effective tool for personalized oncology. Modern radiomics uses advanced machine learning alongside medical imaging to extract high-dimensional characteristics that capture tumor-related parameters such as shape, texture, and intensity. Delta, and dynamic radiomics are key breakthroughs that have improved the capacity to measure tumor features, treatment response, and prognosis across many cancer types noninvasively. Despite its considerable promise, radiomics currently faces significant obstacles such as insufficient methodological standardization, small and varied datasets, and reproducibility issues. These challenges limit their inclusion into medical facilities and decision-making. In medical physics, radiomics is becoming increasingly important in quantitative imaging and individualized therapy planning. When integrated with clinical, genetic, and demographic data, it can offer strong predictive models that improve diagnosis accuracy and facilitate decision making. However, to realize its full potential, further progress is needed to ensure appropriate standardization of imaging and feature extraction protocols to ensure reproducibility and reliability of radiomics studies, inter-institutional availability of data, and sufficient clinical validation to integrate the use of radiomics models into the clinical setting.
2. To address the complexities of head and neck cancer, a delta radiomics methodology was created to examine temporal changes in imaging-derived features. This workflow was tailored to the anatomical and clinical features of these tumors through choosing machine learning approaches capable of handling complex, structured, high-dimensional data, as tumors in the head and neck region tend to be associated with irregular geometry and significant intratumoral heterogeneity. The developed workflow begins with the evaluation of delta features, followed by data preparation, which includes statistical evaluation of data distribution and synthetic data generation. Next follows correlation analysis, which eliminates strongly related characteristics to reduce redundancy and improve overall model accuracy, machine learning-based modeling, and statistical validation. By focusing on temporal feature changes, this strategy seeks to give a more dynamic and biologically meaningful assessment of treatment response than typical static radiomics.
3. To validate this suggested delta radiomics workflow, several machine learning models were created: one for treatment classification and other for survival result prediction. RFE-RF successfully identified discriminative delta features, with AUCs of 0.825 (treatment) and 0.95 (survival). CatBoost models trained with these features performed well in both challenges. Statistical study revealed that kurtosis-based SUV characteristics differed significantly among treatment types, whereas SHAPE_VolumemL_aHN and GLRLM_LRHGE_aHN were linked to survival. Together, these findings confirm the validity as well as reliability of the suggested methodology for assessing treatment-related changes and predicting survival outcomes. If verified externally in larger, multi-institutional cohorts, the detected delta radiomic characteristics could be transformed into non-invasive imaging biomarkers that could help in early treatment response analysis and survival risk stratification. This establishes this workflow as a possible decision-support tool for individualized cancer treatment planning in future clinical studies and clinical settings.

Recommendations

This chapter presents a systematic, evidence-based workflow for developing prediction models using delta radiomics features. Each methodological step has been carefully designed to address the specific challenges inherent in radiomics studies on high-dimensional and low-sample datasets, including class imbalance, overfitting and overfitting of features, and statistical insignificance. The proposed workflow highlights all important steps to avoid overfitting and overall model bias. To ensure transparency, reproducibility and practical applicability of the developed model for head and neck cancer, R code for each step is provided in the appendices.

1. Exploration of the dataset

Prior to any modeling, it is essential to perform an initial exploration of the dataset to understand its structure and content. This includes examining the types and distributions of clinical and radiomics variables, identifying missing data, detecting potential outliers, and summarizing key descriptive statistics. Visual inspection techniques, such as histograms, could be employed. Thorough dataset exploration ensures that subsequent preprocessing and modeling steps are appropriately tailored to the data characteristics. *A practical code example is provided in Appendix 2.*

2. Assessment and Correction of Target Variable Imbalance.

Before developing any predictive models, it is critical to determine whether the distribution of outcome variables (such as survival outcome) is balanced across all classes. Imbalanced datasets can skew machine learning algorithms toward majority classes, producing misleading results. To assess statistical imbalance, utilize the Chi-Square Goodness-of-Fit test. If an imbalance in the dataset is found, use synthetic data generation techniques such as SMOTE (Synthetic Minority Over-sampling methodology) or another methodology appropriate for the dataset. *A practical code example is provided in Appendix 2.*

3. Correlation Analysis for Redundancy Reduction

Radiomics datasets often contain a huge number of features, many of which may be strongly associated. In many circumstances, such information will not provide any more relevant information, and including such duplicate features may result in multicollinearity, which hampers model interpretation and performance. Spearman's rank correlation can discover and remove highly associated characteristics by imposing a correlation coefficient cutoff ($\rho > 0.9$). This stage not only decreases computing complexity and helps to speed up the process, but it also improves the model's reliability and interpretability by guaranteeing that only unique, non-redundant features are input into it. *Code examples are detailed in Appendix 3.*

4. Model Training and Cross-Validation Strategy

The default strategy for building a machine learning model is to divide the dataset into two subsets: 80% for training the algorithm and 20% for validation. To test the ability of the model to generalize over new data, it is recommended to use 10-fold cross-validation, repeating it in the training and testing phases on different subsets of the data. This approach is particularly important for small datasets to avoid over-optimistic performance estimates. *Code implementation is illustrated in Appendix 4.*

5. Feature Selection Based on Predictive Relevance and development of the final model

Following correlation analysis, supervised machine learning-based feature selection should be used to determine the most useful features for the classification task. Several algorithms can be used for this reason, including Recursive Feature Elimination (RFE) with Random Forest, eXtreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM), among others. These methods are especially useful in high-dimensional datasets, such as those used in radiomics, because they make it easier to identify the features that contribute the most significantly to the model's predictive performance. It is important to realize that not all algorithms are naturally suited for multi-class classification tasks; thus, careful selection according to the characteristics of the dataset and the classification target is required to guarantee methodological appropriateness and model correctness.

Once the attributes have been selected, prediction models should be built using algorithms that fit the data. In this case, the data required an algorithm that supports multi-class classification, so CatBoost was chosen as the ideal algorithm for small datasets. Moreover, it is robust to overfit due to the use of ordered boosting. To test the ability of the model to generalize over new data, it is recommended to use 10-fold cross-validation, repeating it in the training and testing phases on different subsets of the data. This approach is particularly important for small datasets to avoid over-optimistic performance estimates. *Examples are provided in Appendix 5.*

6. Comprehensive Performance Evaluation

For both binary and multiclass classification problems, model evaluation should extend above the use of confusion matrices alone. Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) is a popular and informative performance statistic. While ROC-AUC is commonly used for binary classification, it can be extended for multiclass issues utilizing methodologies such as the one-vs-rest (OvR) approach. In this context, macro-averaged AUC is widely used, which computes the AUC for each class independently before averaging the results, treating all classes equally, regardless of frequency. ROC-AUC measures the model's ability to discriminate between classes, but it does not consider class imbalances or specific types of misclassifications. As a result, it is critical to supplement it with other evaluation metrics generated from the confusion matrix, such as recall, precision, and F1 score. Together, these measures provide a more thorough images of the model's classification performance, especially when dealing with imbalanced data or analyzing performance across numerous classes. *Metric calculation examples are available in Appendix 6.*

7. Statistical Validation of Radiomics Features

Include statistical hypothesis testing to see if changes in radiomics correlate with outcomes. As in this study, Kruskal-Wallis and Dunn post-hoc tests were employed to evaluate the delta characteristics of the various treatment types and discover variations between groups. For binary outcomes like survival status, the Mann-Whitney U test is appropriate for detecting significant variations in delta features. To account for multiple testing errors, false discovery rate (FDR) correction (e.g., Benjamini-Hochberg) should be applied to all tests. *Code examples are shown in Appendix 7.*

References

1. What Is Cancer? - NCI. [online]. [Accessed 25 March 2025]. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer
2. What Is Cancer? | Cancer Basics | American Cancer Society. [online]. [Accessed 25 March 2025]. Available from: https://www.cancer.org/cancer/understanding-cancer/what-is-cancer.html
3. BRAY BSC, Freddie, LAVERSANNE, | Mathieu, HYUNA, |, PHD, Sung, FERLAY, Jacques, SIEGEL MPH, Rebecca L, SOERJOMATARAM, Isabelle, AHMEDIN, | and DVM, Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* [online]. 1 May 2024. Vol. 74, no. 3, p. 229–263. [Accessed 25 March 2025]. DOI 10.3322/CAAC.21834. Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21834
4. Global cancer burden growing, amidst mounting need for services. [online]. [Accessed 25 March 2025]. Available from: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services
5. Precision Oncology Program | FDA. [online]. [Accessed 25 March 2025]. Available from: https://www.fda.gov/about-fda/oncology-center-excellence/precision-oncology-program
6. Precision Oncology: Where are we in 2024? [online]. [Accessed 25 March 2025]. Available from: https://na.geneseeq.com/precision-oncology/
7. MIN, Ningning, WEI, Yufan, ZHENG, Yiqiong and LI, Xiru. Advancement of prognostic models in breast cancer: a narrative review. *Gland Surgery* [online]. 2021. Vol. 10, no. 9, p. 2815. [Accessed 25 March 2025]. DOI 10.21037/GS-21-441. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8514300/
8. PHUNG, Minh Tung, TIN TIN, Sandar and ELWOOD, J. Mark. Prognostic models for breast cancer: a systematic review. *BMC Cancer* [online]. 14 March 2019. Vol. 19, no. 1, p. 230. [Accessed 25 March 2025]. DOI 10.1186/S12885-019-5442-6. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6419427/
9. LÜÖND, Fabiana, TIEDE, Stefanie and CHRISTOFORI, Gerhard. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer 2021 125:2* [online]. 6 April 2021. Vol. 125, no. 2, p. 164–175. [Accessed 25 March 2025]. DOI 10.1038/s41416-021-01328-7. Available from: https://www.nature.com/articles/s41416-021-01328-7
10. KARMAZANOVSKY, Grigory, GRUZDEV, Ivan, TIKHONOVA, Valeriya, KONDRATYEV, Evgeny and REVISHVILI, Amiran. Computed tomography-based radiomics approach in pancreatic tumors characterization. *Radiologia Medica* [online]. 1 November 2021. Vol. 126, no. 11, p. 1388–1395. [Accessed 25 March 2025]. DOI 10.1007/S11547-021-01405-0/FIGURES/1. Available from: https://link.springer.com/article/10.1007/s11547-021-01405-0
11. MAYERHOEFER, Marius E., MATERKA, Andrzej, LANGS, Georg, HÄGGSTRÖM, Ida, SZCZYPIŃSKI, Piotr, GIBBS, Peter and COOK, Gary. Introduction to Radiomics. *Journal of Nuclear Medicine* [online]. 1 April 2020. Vol. 61, no. 4, p. 488–495. [Accessed 25 March 2025]. DOI 10.2967/JNUMED.118.222893. Available from: https://jnm.snmjournals.org/content/61/4/488
12. SCAPICCHIO, Camilla, GABELLONI, Michela, BARUCCI, Andrea, CIONI, Dania, SABA, Luca and NERI, Emanuele. A deep look into radiomics. *La Radiologia Medica* [online]. 1 October 2021. Vol. 126, no. 10, p. 1296. [Accessed 25 March 2025]. DOI 10.1007/S11547-021-01389-X. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8520512/

13. ZHAO, Binsheng. Understanding Sources of Variation to Improve the Reproducibility of Radiomics. *Frontiers in Oncology* [online]. 29 March 2021. Vol. 11, p. 633176. [Accessed 25 March 2025]. DOI 10.3389/FONC.2021.633176/FULL. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8039446/
14. REIAZI, Reza, ABBAS, Engy, FAMIYEH, Petra, REZAIE, Aria, KWAN, Jennifer Y.Y., PATEL, Tirth, BRATMAN, Scott V., TADIC, Tony, LIU, Fei Fei and HAIBE-KAINS, Benjamin. The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review. *Computers in Biology and Medicine*. 1 June 2021. Vol. 133, p. 104400. DOI 10.1016/J.COMPBIOMED.2021.104400.
15. LI, Xiao Tian and HUANG, Raymond Y. Standardization of imaging methods for machine learning in neuro-oncology. *Neuro-Oncology Advances* [online]. 31 December 2020. Vol. 2, no. Supplement_4, p. iv49–iv55. [Accessed 25 March 2025]. DOI 10.1093/NOAJNL/VDAA054. Available from: https://dx.doi.org/10.1093/noajnl/vdaa054
16. MARCU, David C., GRAVA, Cristian and MARCU, Loredana G. Current Role of Delta Radiomics in Head and Neck Oncology. *International journal of molecular sciences* [online]. 1 February 2023. Vol. 24, no. 3. [Accessed 22 April 2025]. DOI 10.3390/IJMS24032214. Available from: https://pubmed.ncbi.nlm.nih.gov/36768535/
17. DING, Haoran, WU, Chenzhou, LIAO, Nailin, ZHAN, Qi, SUN, Weize, HUANG, Yingzhao, JIANG, Zhou and LI, Yi. Radiomics in Oncology: A 10-Year Bibliometric Analysis. *Frontiers in Oncology* [online]. 20 September 2021. Vol. 11, p. 689802. [Accessed 16 October 2024]. DOI 10.3389/FONC.2021.689802/BIBTEX. Available from: www.frontiersin.org
18. ALDERSON, Philip O. and SUMMERS, Ronald M. The Evolving Status of Radiomics. *JNCI Journal of the National Cancer Institute* [online]. 1 September 2020. Vol. 112, no. 9, p. 869. [Accessed 16 October 2024]. DOI 10.1093/JNCI/DJAA018. Available from: /pmc/articles/PMC7492766/
19. HUANG, Chao, CINTRA, Murilo, BRENNAN, Kevin, ZHOU, Mu, COLEVAS, A. Dimitrios, FISCHBEIN, Nancy, ZHU, Shankuan and GEVAERT, Olivier. Development and validation of radiomic signatures of head and neck squamous cell carcinoma molecular features and subtypes. *EBioMedicine* [online]. 1 July 2019. Vol. 45, p. 70–80. [Accessed 16 October 2024]. DOI 10.1016/J.EBIOM.2019.06.034. Available from: https://pubmed.ncbi.nlm.nih.gov/31255659/
20. AERTS, Hugo J.W.L., VELAZQUEZ, Emmanuel Rios, LEIJENAAR, Ralph T.H., PARMAR, Chintan, GROSSMANN, Patrick, CAVALHO, Sara, BUSSINK, Johan, MONSHOUWER, René, HAIBE-KAINS, Benjamin, RIETVELD, Derek, HOEBERS, Frank, RIETBERGEN, Michelle M., LEEMANS, C. René, DEKKER, Andre, QUACKENBUSH, John, GILLIES, Robert J. and LAMBIN, Philippe. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications 2014 5:1* [online]. 3 June 2014. Vol. 5, no. 1, p. 1–9. [Accessed 16 October 2024]. DOI 10.1038/ncomms5006. Available from: https://www.nature.com/articles/ncomms5006
21. PAREKH, Vishwa S and JACOBS, Michael A. Multiparametric radiomics methods for breast cancer tissue characterization using radiological imaging. [online]. 2020. Vol. 180, p. 407–421. [Accessed 18 January 2025]. DOI 10.1007/s10549-020-05533-5. Available from: https://doi.org/10.1007/s10549-020-05533-5
22. GUO, Liangcun, DU, Siyao, GAO, Si, ZHAO, Ruimeng, HUANG, Guoliang, JIN, Feng, TENG, Yuee and ZHANG, Lina. Delta-Radiomics Based on Dynamic Contrast-Enhanced MRI Predicts Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy. *Cancers*. 2022. Vol. 2022, p. 3515. DOI 10.3390/cancers14143515.

23. MA, Y, MA, W, XU, X and CAO, F. How Does the Delta-Radiomics Better Differentiate Pre-Invasive GGNs From Invasive GGNs? Front. *Oncol* [online]. 2020. Vol. 10, p. 1017. [Accessed 21 January 2025]. DOI 10.3389/fonc.2020.01017. Available from: www.frontiersin.org
24. NASIEF, Haidy, ZHENG, Cheng, SCHOTT, Diane, HALL, William, TSAI, Susan, ERICKSON, Beth and LI, X Allen. ARTICLE OPEN A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. [online]. [Accessed 21 January 2025]. DOI 10.1038/s41698-019-0096-z. Available from: https://doi.org/10.1038/s41698-019-0096-z
25. QU, Hui, SHI, Ruichuan, LI, Shuqin, CHE, Fengying, WU, Jian, LI, Haoran, CHEN, Weixing, ZHANG, Hao, LI, Zhi and CUI, · Xiaoyu. Dynamic radiomics: A new methodology to extract quantitative time-related features from tomographic images. *Applied Intelligence* [online]. Vol. 1, p. 3. [Accessed 28 January 2025]. DOI 10.1007/s10489-021-03053-3. Available from: https://doi.org/10.1007/s10489-021-03053-3
26. MENG, Yiming, SUN, Jing, QU, Na, ZHANG, Guirong, YU, Tao and PIAO, Haozhe. Application of Radiomics for Personalized Treatment of Cancer Patients. *Cancer Management and Research* [online]. 2019. Vol. 11, p. 10851. [Accessed 4 February 2025]. DOI 10.2147/CMAR.S232473. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6941598/
27. PARMAR, Chintan, GROSSMANN, Patrick, BUSSINK, Johan, LAMBIN, Philippe and AERTS, Hugo J.W.L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific reports* [online]. 17 August 2015. Vol. 5. [Accessed 4 February 2025]. DOI 10.1038/SREP13087. Available from: https://pubmed.ncbi.nlm.nih.gov/26278466/
28. LAMBIN, Philippe, LEIJENAAR, Ralph T.H., DEIST, Timo M., PEERLINGS, Jurgen, DE JONG, Evelyn E.C., VAN TIMMEREN, Janita, SANDULEANU, Sebastian, LARUE, Ruben T.H.M., EVEN, Aniek J.G., JOCHEMS, Arthur, VAN WIJK, Yvonka, WOODRUFF, Henry, VAN SOEST, Johan, LUSTBERG, Tim, ROELOFS, Erik, VAN ELMPT, Wouter, DEKKER, Andre, MOTTAGHY, Felix M., WILDBERGER, Joachim E. and WALSH, Sean. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews. Clinical oncology* [online]. 1 December 2017. Vol. 14, no. 12, p. 749–762. [Accessed 4 February 2025]. DOI 10.1038/NRCLINONC.2017.141. Available from: https://pubmed.ncbi.nlm.nih.gov/28975929/
29. VIA, Luigi La, SANGIORGIO, Giuseppe, STEFANI, Stefania, MARINO, Andrea, NUNNARI, Giuseppe, COCUZZA, Salvatore, MANTIA, Ignazio La, CACOPARDO, Bruno, STRACQUADANIO, Stefano, SPAMPINATO, Serena, LAVALLE, Salvatore and MANIACI, Antonino. The Global Burden of Sepsis and Septic Shock. *Epidemiologia 2024, Vol. 5, Pages 456-478* [online]. 25 July 2024. Vol. 5, no. 3, p. 456–478. [Accessed 4 February 2025]. DOI 10.3390/EPIDEMIOLOGIA5030032. Available from: https://www.mdpi.com/2673-3986/5/3/32/htm
30. LITJENS, Geert, KOOI, Thijs, BEJNORDI, Babak Ehteshami, SETIO, Arnaud Arindra Adiyoso, CIOMPI, Francesco, GHAFOORIAN, Mohsen, VAN DER LAAK, Jeroen A.W.M., VAN GINNEKEN, Bram and SÁNCHEZ, Clara I. A survey on deep learning in medical image analysis. *Medical image analysis* [online]. 1 December 2017. Vol. 42, p. 60–88. [Accessed 4 February 2025]. DOI 10.1016/J.MEDIA.2017.07.005. Available from: https://pubmed.ncbi.nlm.nih.gov/28778026/
31. LAO, Jiangwei, CHEN, Yinsheng, LI, Zhi Cheng, LI, Qihua, ZHANG, Ji, LIU, Jing and ZHAI, Guangtao. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific reports* [online]. 1 December 2017. Vol. 7, no. 1. [Accessed 4 February 2025]. DOI 10.1038/S41598-017-10649-8. Available from: https://pubmed.ncbi.nlm.nih.gov/28871110/

32. KOÇAK, Burak, PONSIGLIONE, Andrea, STANZIONE, Arnaldo, BLUETHGEN, Christian, SANTINHA, João, UGGA, Lorenzo, HUISMAN, Merel, KLONTZAS, Michail E., CANNELLA, Roberto and CUOCOLO, Renato. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*. 2 July 2024. DOI 10.4274/DIR.2024.242854.
33. VAROQUAUX, Gaël and CHEPLYGINA, Veronika. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine 2022 5:1* [online]. 12 April 2022. Vol. 5, no. 1, p. 1–8. [Accessed 4 February 2025]. DOI 10.1038/s41746-022-00592-y. Available from: https://www.nature.com/articles/s41746-022-00592-y
34. MANIACI, Antonino, LAVALLE, Salvatore, GAGLIANO, Caterina, LENTINI, Mario, MASIELLO, Edoardo, PARISI, Federica, IANNELLA, Giannicola, CILIA, Nicole Dalia, SALERNO, Valerio, CUSUMANO, Giacomo and LA VIA, Luigi. The Integration of Radiomics and Artificial Intelligence in Modern Medicine. *Life 2024, Vol. 14, Page 1248* [online]. 1 October 2024. Vol. 14, no. 10, p. 1248. [Accessed 4 February 2025]. DOI 10.3390/LIFE14101248. Available from: https://www.mdpi.com/2075-1729/14/10/1248/htm
35. PINTO-COELHO, Luís. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. *Bioengineering* [online]. 1 December 2023. Vol. 10, no. 12, p. 1435. [Accessed 4 February 2025]. DOI 10.3390/BIOENGINEERING10121435. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10740686/
36. The Role of AI in Medical Imaging | BGO Software. [online]. [Accessed 4 February 2025]. Available from: https://www.bgosoftware.com/blog/the-role-of-ai-in-medical-imaging/
37. COLLIN, Catherine Bjerre, GEBHARDT, Tom, GOLEBIEWSKI, Martin, KARADERI, Tugce, HILLEMANNS, Maximilian, KHAN, Faiz Muhammad, SALEHZADEH-YAZDI, Ali, KIRSCHNER, Marc, KROBITSCH, Sylvia and KUEPFER, Lars. Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation. *Journal of Personalized Medicine* [online]. 1 February 2022. Vol. 12, no. 2, p. 166. [Accessed 4 February 2025]. DOI 10.3390/JPM12020166. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8879572/
38. LE, Matthieu, DELINGETTE, Herve, KALPATHY-CRAMER, Jayashree, GERSTNER, Elizabeth R., BATCHELOR, Tracy, UNKELBACH, Jan and AYACHE, Nicholas. Personalized Radiotherapy Planning Based on a Computational Tumor Growth Model. *IEEE transactions on medical imaging* [online]. 1 March 2017. Vol. 36, no. 3, p. 815–825. [Accessed 4 February 2025]. DOI 10.1109/TMI.2016.2626443. Available from: https://pubmed.ncbi.nlm.nih.gov/28113925/
39. VAN TIMMEREN, Janita E., CESTER, Davide, TANADINI-LANG, Stephanie, ALKADHI, Hatem and BAESSLER, Bettina. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging* [online]. 1 December 2020. Vol. 11, no. 1, p. 1–16. [Accessed 26 February 2025]. DOI 10.1186/S13244-020-00887-2/TABLES/3. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-020-00887-2
40. SCAPICCHIO, Camilla, GABELLONI, Michela, BARUCCI, Andrea, CIONI, Dania, SABA, Luca and NERI, Emanuele. A deep look into radiomics. *La Radiologia Medica* [online]. 1 October 2021. Vol. 126, no. 10, p. 1296. [Accessed 26 February 2025]. DOI 10.1007/S11547-021-01389-X. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8520512/
41. Medical Imaging: Modalities & Types of Equipment. [online]. [Accessed 26 February 2025]. Available from: https://www.excedr.com/blog/medical-imaging-and-radiology-overview
42. PERNICIANO, Alessandra, LODDO, Andrea, DI RUBERTO, Cecilia and PES, Barbara. Insights into radiomics: impact of feature selection and classification. *Multimedia Tools and*

Applications 2024 [online]. 15 November 2024. P. 1–27. [Accessed 26 February 2025]. DOI 10.1007/S11042-024-20388-4. Available from: https://link.springer.com/article/10.1007/s11042-024-20388-4

43. BECKMANN, Matthias, NICKEL, Judith, BECKMANN, Matthias and NICKEL, Judith. Optimized filter functions for filtered back projection reconstructions. *Inverse Problems and Imaging* [online]. 2025. Vol. 0, no. 0, p. 0–0. [Accessed 26 February 2025]. DOI 10.3934/IPI.2025003. Available from: https://www.aimsciences.org/en/article/doi/10.3934/ipi.2025003

44. GOTHWAL, Ritu, TIWARI, Shailendra and SHIVANI, Shivendra. Computational Medical Image Reconstruction Techniques: A Comprehensive Review. *Archives of Computational Methods in Engineering 2022 29:7* [online]. 25 July 2022. Vol. 29, no. 7, p. 5635–5662. [Accessed 26 February 2025]. DOI 10.1007/S11831-022-09785-W. Available from: https://link.springer.com/article/10.1007/s11831-022-09785-w

45. KOETZIER, Lennart R., MASTRODICASA, Domenico, SZCZYKUTOWICZ, Timothy P., VAN DER WERF, Niels R., WANG, Adam S., SANDFORT, Veit, VAN DER MOLEN, Aart J., FLEISCHMANN, Dominik and WILLEMINK, Martin J. Deep Learning Image Reconstruction for CT: Technical Principles and Clinical Prospects. *Radiology* [online]. 1 March 2023. Vol. 306, no. 3. [Accessed 26 February 2025].
DOI 10.1148/RADIOL.221257/ASSET/IMAGES/LARGE/RADIOL.221257.FIG11.JPEG.
Available from: https://pubs.rsna.org/doi/10.1148/radiol.221257

46. CHEN, Yutong, SCHONLIEB, Carola Bibiane, LIO, Pietro, LEINER, Tim, DRAGOTTI, Pier Luigi, WANG, Ge, RUECKERT, Daniel, FIRMIN, David and YANG, Guang. AI-Based Reconstruction for Fast MRI-A Systematic Review and Meta-Analysis. *Proceedings of the IEEE*. 1 February 2022. Vol. 110, no. 2, p. 224–245. DOI 10.1109/JPROC.2022.3141367.

47. Image Segmentation — A Beginner's Guide | Medium. [online]. [Accessed 27 February 2025]. Available from: https://medium.com/@raj.pulapakura/image-segmentation-a-beginners-guide-0ede91052db7

48. What Is Image Segmentation? | IBM. [online]. [Accessed 27 February 2025]. Available from: https://www.ibm.com/think/topics/image-segmentation

49. MA, Jun, HE, Yuting, LI, Feifei, HAN, Lin, YOU, Chenyu and WANG, Bo. Segment anything in medical images. *Nature Communications* [online]. 1 December 2024. Vol. 15, no. 1, p. 654. [Accessed 27 February 2025]. DOI 10.1038/S41467-024-44824-Z. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10803759/

50. Revolutionizing Medical Imaging with Semantic Segmentation | Keymakr. [online]. [Accessed 27 February 2025]. Available from: https://keymakr.com/blog/revolutionizing-medical-imaging-with-semantic-segmentation/

51. VAASSEN, Femke, HAZELAAR, Colien, VANIQUI, Ana, GOODING, Mark, VAN DER HEYDEN, Brent, CANTERS, Richard and VAN ELMPT, Wouter. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*. 1 January 2020. Vol. 13, p. 1–6. DOI 10.1016/J.PHRO.2019.12.001.

52. MCGRATH, Hari, LI, Peichao, DORENT, Reuben, BRADFORD, Robert, SAEED, Shakeel, BISDAS, Sotirios, OURSELIN, Sebastien, SHAPEY, Jonathan and VERCAUTEREN, Tom. Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI. *International Journal of Computer Assisted Radiology and Surgery* [online]. 1 September 2020. Vol. 15, no. 9, p. 1445. [Accessed 27 February 2025]. DOI 10.1007/S11548-020-02222-Y. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7419453/

53. XU, Yan, QUAN, Rixiang, XU, Weiting, HUANG, Yi, CHEN, Xiaolong and LIU, Fengyuan. Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. *Bioengineering 2024, Vol. 11, Page 1034* [online]. 16 October 2024. Vol. 11, no. 10, p. 1034. [Accessed 27 February 2025]. DOI 10.3390/BIOENGINEERING11101034. Available from: https://www.mdpi.com/2306-5354/11/10/1034/htm
54. CUI, Yunfeng and YIN, Fang Fang. Impact of image quality on radiomics applications. *Physics in Medicine & Biology* [online]. 22 July 2022. Vol. 67, no. 15, p. 15TR03. [Accessed 26 March 2025]. DOI 10.1088/1361-6560/AC7FD7. Available from: https://iopscience.iop.org/article/10.1088/1361-6560/ac7fd7
55. ROGERS, William, SEETHA, Sithin Thulasi, REFAEE, Turkey A.G., LIEVERSE, Relinde I.Y., GRANZIER, Renée W.Y., IBRAHIM, Abdalla, KEEK, Simon A., SANDULEANU, Sebastian, PRIMAKOV, Sergey P., BEUQUE, Manon P.L., MARCUS, Damiënne, VAN DER WIEL, Alexander M.A., ZERKA, Fadila, OBERIJE, Cary J.G., VAN TIMMEREN, Janitae, WOODRUFF, Henry C. and LAMBIN, PHILIPPE. Radiomics: from qualitative to quantitative imaging. *The British Journal of Radiology* [online]. 1 April 2020. Vol. 93, no. 1108, p. 20190948. [Accessed 26 March 2025]. DOI 10.1259/BJR.20190948. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7362913/
56. AL-THELAYA, Khaled, GILAL, Nauman Ullah, ALZUBAIDI, Mahmood, MAJEED, Fahad, AGUS, Marco, SCHNEIDER, Jens and HOUSEH, Mowafa. Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: A survey. *Journal of Pathology Informatics* [online]. 1 January 2023. Vol. 14, p. 100335. [Accessed 27 February 2025]. DOI 10.1016/J.JPI.2023.100335. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10622844/
57. ZHANG, W ;, GUO, Y ;, JIN, Q, ZHANG, Wenchao, GUO, Yu and JIN, Qiyu. Radiomics and Its Feature Selection: A Review. *Symmetry 2023, Vol. 15, Page 1834* [online]. 27 September 2023. Vol. 15, no. 10, p. 1834. [Accessed 6 April 2024]. DOI 10.3390/SYM15101834. Available from: https://www.mdpi.com/2073-8994/15/10/1834/htm
58. ZHANG, W ;, GUO, Y ;, JIN, Q, ZHANG, Wenchao, GUO, Yu and JIN, Qiyu. Radiomics and Its Feature Selection: A Review. *Symmetry 2023, Vol. 15, Page 1834* [online]. 27 September 2023. Vol. 15, no. 10, p. 1834. [Accessed 27 February 2025]. DOI 10.3390/SYM15101834. Available from: https://www.mdpi.com/2073-8994/15/10/1834/htm
59. VIJITHANANDA, Sahan M., JAYATILAKE, Mohan L., HEWAVITHANA, Badra, GONÇALVES, Teresa, RATO, Luis M., WEERAKOON, Bimali S., KALUPAHANA, Tharindu D., SILVA, Anil D. and DISSANAYAKE, Karuna D. Feature extraction from MRI ADC images for brain tumor classification using machine learning techniques. *BioMedical Engineering Online* [online]. 1 December 2022. Vol. 21, no. 1, p. 1–21. [Accessed 26 March 2025]. DOI 10.1186/S12938-022-01022-6/FIGURES/6. Available from: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-022-01022-6
60. KOÇAK, Burak, ŞEBNEM, Emine, ECE, Durmaz, ÖZGÜR KILIÇKESMEZ, Ateş and İSTANBUL, Ö K). Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol* [online]. 2019. Vol. 25, p. 485–495. [Accessed 27 February 2025]. DOI 10.5152/dir.2019.19321. Available from: https://deepcognition.ai/
61. HUANG, Fangliang and MARIANO, Vladimir Y. Application of MRI Radiomics Combined With Deep Learning Technology in Glioma Segmentation and Survival Prognosis. *Journal of*

Artificial Intelligence and Technology [online]. 22 February 2025. [Accessed 27 February 2025]. DOI 10.37965/JAIT.2025.0681. Available from: https://ojs.istp-press.com/jait/article/view/681

62. VIAL, Alanna, STIRLING, David, FIELD, Matthew, ROS, Montserrat, RITZ, Christian, CAROLAN, Martin, HOLLOWAY, Lois and MILLER, Alexis A. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. *Translational Cancer Research* [online]. 1 June 2018. Vol. 7, no. 3, p. 803–816. [Accessed 27 February 2025]. DOI 10.21037/TCR.2018.05.02. Available from: https://tcr.amegroups.org/article/view/21823/html

63. YE, Jing Yuan, FANG, Peng, PENG, Zhen Peng, HUANG, Xi Tai, XIE, Jin Zhao and YIN, Xiao Yu. A radiomics-based interpretable model to predict the pathological grade of pancreatic neuroendocrine tumors. *European Radiology* [online]. 1 March 2024. Vol. 34, no. 3, p. 1994–2005. [Accessed 27 February 2025]. DOI 10.1007/S00330-023-10186-1/FIGURES/6. Available from: https://link.springer.com/article/10.1007/s00330-023-10186-1

64. LIU, Meng Wen, ZHANG, Xue, WANG, Yan Mei, JIANG, Xu, JIANG, Jiu Ming, LI, Meng and ZHANG, Li. A comparison of machine learning methods for radiomics modeling in prediction of occult lymph node metastasis in clinical stage IA lung adenocarcinoma patients. *Journal of Thoracic Disease* [online]. 29 March 2024. Vol. 16, no. 3, p. 1765–1776. [Accessed 27 February 2025]. DOI 10.21037/JTD-23-1578/COIF. Available from: https://jtd.amegroups.org/article/view/84589/html

65. MOHD HANIFF, Nurin Syazwina, NG, Kwan Hoong, KAMAL, Izdihar, MOHD ZAIN, Norhayati and ABDUL KARIM, Muhammad Khalis. Systematic review and meta-analysis on the classification metrics of machine learning algorithm based radiomics in hepatocellular carcinoma diagnosis. *Heliyon*. 30 August 2024. Vol. 10, no. 16, p. e36313. DOI 10.1016/J.HELIYON.2024.E36313.

66. WEN, Haoxiang, LIANG, Ruiming, LIU, Xiaofei, YU, Yang, LIN, Shuirong, SONG, Zimin, HUANG, Yihao, YU, Xi, CHEN, Shuling, CHEN, Lili, QIAN, Baifeng, SHEN, Jingxian, XIAO, Han and SHEN, Shunli. Predicting Pathological Response of Neoadjuvant Conversion Therapy for Hepatocellular Carcinoma Patients Using CT-Based Radiomics Model. *Journal of hepatocellular carcinoma* [online]. 2024. Vol. 11, p. 2145–2157. [Accessed 27 February 2025]. DOI 10.2147/JHC.S487370. Available from: https://pubmed.ncbi.nlm.nih.gov/39502744/

67. ETEHADTAVAKOL, Mehdi, ETEHADTAVAKOL, Mahnaz, MOALLEM, Golnaz and NG, Eddie Y.K. Evaluating Radiomics Feature Reduction for Thyroid Nodule Segmentation in Thermal Imaging. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [online]. 2025. Vol. 15279 LNCS, p. 69–87. [Accessed 27 February 2025]. DOI 10.1007/978-3-031-76584-1_7/TABLES/10. Available from: https://link.springer.com/chapter/10.1007/978-3-031-76584-1_7

68. COBO, Miriam, MENÉNDEZ FERNÁNDEZ-MIRANDA, Pablo, BASTARRIKA, Gorka and LLORET IGLESIAS, Lara. Enhancing radiomics and Deep Learning systems through the standardization of medical imaging workflows. *Scientific Data* [online]. 1 December 2023. Vol. 10, no. 1, p. 732. [Accessed 21 October 2024]. DOI 10.1038/S41597-023-02641-X. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10590396/

69. LI, Xiao Tian and HUANG, Raymond Y. Standardization of imaging methods for machine learning in neuro-oncology. *Neuro-Oncology Advances* [online]. 31 December 2020. Vol. 2, no. Supplement_4, p. iv49–iv55. [Accessed 21 October 2024]. DOI 10.1093/NOAJNL/VDAA054. Available from: https://dx.doi.org/10.1093/noajnl/vdaa054

70. RATHORE, Saima, BAKAS, Spyridon, PATI, Sarthak, AKBARI, Hamed, KALAROT, Ratheesh, SRIDHARAN, Patmaa, ROZYCKI, Martin, BERGMAN, Mark, TUNC, Birkan, VERMA,

Ragini, BILELLO, Michel and DAVATZIKOS, Christos. Brain Cancer Imaging Phenomics Toolkit (brain-CaPTk): An Interactive Platform for Quantitative Analysis of Glioblastoma. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [online]. 2018. Vol. 10670 LNCS, p. 133–145. [Accessed 21 October 2024]. DOI 10.1007/978-3-319-75238-9_12. Available from: https://link.springer.com/chapter/10.1007/978-3-319-75238-9_12

71. LAMBIN, Philippe, ZINDLER, Jaap, VANNESTE, Ben G.L., DE VOORDE, Lien Van, EEKERS, Daniëlle, COMPTER, Inge, PANTH, Kranthi Marella, PEERLINGS, Jurgen, LARUE, Ruben T.H.M., DEIST, Timo M., JOCHEMS, Arthur, LUSTBERG, Tim, VAN SOEST, Johan, DE JONG, Evelyn E.C., EVEN, Aniek J.G., REYMEN, Bart, REKERS, Nicolle, VAN GISBERGEN, Marike, ROELOFS, Erik, CARVALHO, Sara, LEIJENAAR, Ralph T.H., ZEGERS, Catharina M.L., JACOBS, Maria, VAN TIMMEREN, Janita, BROUWERS, Patricia, LAL, Jonathan A., DUBOIS, Ludwig, YAROMINA, Ala, VAN LIMBERGEN, Evert Jan, BERBEE, Maaike, VAN ELMPT, Wouter, OBERIJE, Cary, RAMAEKERS, Bram, DEKKER, Andre, BOERSMA, Liesbeth J., HOEBERS, Frank, SMITS, Kim M., BERLANGA, Adriana J. and WALSH, Sean. Decision support systems for personalized and participative radiation oncology. *Advanced Drug Delivery Reviews*. 15 January 2017. Vol. 109, p. 131–153. DOI 10.1016/J.ADDR.2016.01.006.

72. LAMBIN, Philippe, VAN STIPHOUT, Ruud G.P.M., STARMANS, Maud H.W., RIOS-VELAZQUEZ, Emmanuel, NALBANTOV, Georgi, AERTS, Hugo J.W.L., ROELOFS, Erik, VAN ELMPT, Wouter, BOUTROS, Paul C., GRANONE, Pierluigi, VALENTINI, Vincenzo, BEGG, Adrian C., DE RUYSSCHER, Dirk and DEKKER, Andre. Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nature Reviews Clinical Oncology 2012 10:1* [online]. 20 November 2012. Vol. 10, no. 1, p. 27–40. [Accessed 31 January 2025]. DOI 10.1038/nrclinonc.2012.196. Available from: https://www.nature.com/articles/nrclinonc.2012.196

73. MILES, Kenneth. Radiomics for personalised medicine: the long road ahead. *British Journal of Cancer 2020 122:7* [online]. 15 January 2020. Vol. 122, no. 7, p. 929–930. [Accessed 31 January 2025]. DOI 10.1038/s41416-019-0699-8. Available from: https://www.nature.com/articles/s41416-019-0699-8

74. ZHANG, Yucheng, LOBO-MUELLER, Edrise M., KARANICOLAS, Paul, GALLINGER, Steven, HAIDER, Masoom A. and KHALVATI, Farzad. Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images. *Scientific Reports 2021 11:1* [online]. 14 January 2021. Vol. 11, no. 1, p. 1–11. [Accessed 31 January 2025]. DOI 10.1038/s41598-021-80998-y. Available from: https://www.nature.com/articles/s41598-021-80998-y

75. VAN SOEST, Johan, MELDOLESI, Elisa, VAN STIPHOUT, Ruud, GATTA, Roberto, DAMIANI, Andrea, VALENTINI, Vincenzo, LAMBIN, Philippe and DEKKER, Andre. Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility. *Medical Physics* [online]. 1 September 2017. Vol. 44, no. 9, p. 4961–4967. [Accessed 31 January 2025]. DOI 10.1002/MP.12423. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.12423

76. VAN TIMMEREN, Janna E., CARVALHO, Sara, LEIJENAAR, Ralph T.H., TROOST, Esther G.C., VAN ELMPT, Wouter, DE RUYSSCHER, Dirk, MURATET, Jean Pierre, DENIS, Fabrice, SCHIMEK-JASCH, Tanja, NESTLE, Ursula, JOCHEMS, Arthur, WOODRUFF, Henry C., OBERIJE, Cary and LAMBIN, Philippe. Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-

PET/CT radiomics. *PloS one* [online]. 1 June 2019. Vol. 14, no. 6. [Accessed 31 January 2025]. DOI 10.1371/JOURNAL.PONE.0217536. Available from: https://pubmed.ncbi.nlm.nih.gov/31158263/

77. GOCERI, Evgin. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review* [online]. 1 November 2023. Vol. 56, no. 11, p. 12561–12605. [Accessed 4 February 2025]. DOI 10.1007/S10462-023-10453-Z/TABLES/10. Available from: https://link.springer.com/article/10.1007/s10462-023-10453-z

78. LO IACONO, F, MARAGNA, R., PONTONE, G. and CORINO, V. D. A. A Novel Data Augmentation Method for Radiomics Analysis Using Image Perturbations. *Journal of Imaging Informatics in Medicine 2024 37:5* [online]. 6 May 2024. Vol. 37, no. 5, p. 2401–2414. [Accessed 4 February 2025]. DOI 10.1007/S10278-024-01013-0. Available from: https://link.springer.com/article/10.1007/s10278-024-01013-0

79. YASAKA, Koichiro, AKAI, Hiroyuki, KUNIMATSU, Akira, KIRYU, Shigeru and ABE, Osamu. Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology* [online]. 1 April 2018. Vol. 36, no. 4, p. 257–272. [Accessed 4 February 2025]. DOI 10.1007/S11604-018-0726-3/FIGURES/11. Available from: https://link.springer.com/article/10.1007/s11604-018-0726-3

80. JHA, A. K., MITHUN, S., JAISWAR, V., SHERKHANE, U. B., PURANDARE, N. C., PRABHASH, K., RANGARAJAN, V., DEKKER, A., WEE, L. and TRAVERSO, A. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Scientific Reports 2021 11:1* [online]. 21 January 2021. Vol. 11, no. 1, p. 1–12. [Accessed 19 February 2025]. DOI 10.1038/s41598-021-81526-8. Available from: https://www.nature.com/articles/s41598-021-81526-8

81. PARK, Ji Eun, PARK, Seo Young, KIM, Hwa Jung and KIM, Ho Sung. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean Journal of Radiology* [online]. 1 July 2019. Vol. 20, no. 7, p. 1124. [Accessed 19 February 2025]. DOI 10.3348/KJR.2018.0070. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6609433/

82. PARK, Ji Eun, PARK, Seo Young, KIM, Hwa Jung and KIM, Ho Sung. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean Journal of Radiology* [online]. 1 July 2019. Vol. 20, no. 7, p. 1124–1137. [Accessed 21 February 2025]. DOI 10.3348/KJR.2018.0070. Available from: https://doi.org/10.3348/kjr.2018.0070

83. THOMAS, Hannah Mary T., WANG, Helen Y.C., VARGHESE, Amal Joseph, DONOVAN, Ellen M., SOUTH, Chris P., SAXBY, Helen, NISBET, Andrew, PRAKASH, Vineet, SASIDHARAN, Balu Krishna, PAVAMANI, Simon Pradeep, DEVADHAS, Devakumar, MATHEW, Manu, ISIAH, Rajesh Gunasingam and EVANS, Philip M. Reproducibility in Radiomics: A Comparison of Feature Extraction Methods and Two Independent Datasets. *Applied Sciences 2023, Vol. 13, Page 7291* [online]. 19 June 2023. Vol. 13, no. 12, p. 7291. [Accessed 19 February 2025]. DOI 10.3390/APP13127291. Available from: https://www.mdpi.com/2076-3417/13/12/7291/htm

84. TOMASZEWSKI, Michal R. and GILLIES, Robert J. The Biological Meaning of Radiomic Features. *Radiology* [online]. 1 March 2021. Vol. 298, no. 3, p. 505. [Accessed 27 October 2024]. DOI 10.1148/RADIOL.2021202553. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7924519/

85. MAJUMDER, Shweta, KATZ, Sharyn, KONTOS, Despina and ROSHKOVAN, Leonid. State of the art: radiomics and radiomics-related artificial intelligence on the road to clinical translation. *BJR|Open* [online]. 12 December 2023. Vol. 6, no. 1. [Accessed 27 October 2024]. DOI 10.1093/BJRO/TZAD004. Available from: https://dx.doi.org/10.1093/bjro/tzad004
86. DECOUX, Antoine, DURON, Loic, HABERT, Paul, ROBLOT, Victoire, ARSOVIC, Emina, CHASSAGNON, Guillaume, ARNOUX, Armelle and FOURNIER, Laure. Comparative performances of machine learning algorithms in radiomics and impacting factors. *Scientific Reports 2023 13:1* [online]. 28 August 2023. Vol. 13, no. 1, p. 1–10. [Accessed 30 October 2024]. DOI 10.1038/s41598-023-39738-7. Available from: https://www.nature.com/articles/s41598-023-39738-7
87. LISSON, Catharina Silvia, LISSON, Christoph Gerhard, MEZGER, Marc Fabian, WOLF, Daniel, SCHMIDT, Stefan Andreas, THAISS, Wolfgang M., TAUSCH, Eugen, BEER, Ambros J., STILGENBAUER, Stephan, BEER, Meinrad and GOETZ, Michael. Deep Neural Networks and Machine Learning Radiomics Modelling for Prediction of Relapse in Mantle Cell Lymphoma. *Cancers* [online]. 2 April 2022. Vol. 14, no. 8, p. 2008. [Accessed 30 October 2024]. DOI 10.3390/CANCERS14082008. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9028737/
88. WAGNER, Matthias W., NAMDAR, Khashayar, BISWAS, Asthik, MONAH, Suranna, KHALVATI, Farzad and ERTL-WAGNER, Birgit B. Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology* [online]. 1 December 2021. Vol. 63, no. 12, p. 1957. [Accessed 30 October 2024]. DOI 10.1007/S00234-021-02813-9. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8449698/
89. DECOUX, Antoine, DURON, Loic, HABERT, Paul, ROBLOT, Victoire, ARSOVIC, Emina, CHASSAGNON, Guillaume, ARNOUX, Armelle and FOURNIER, Laure. Comparative performances of machine learning algorithms in radiomics and impacting factors. *Scientific Reports 2023 13:1* [online]. 28 August 2023. Vol. 13, no. 1, p. 1–10. [Accessed 30 October 2024]. DOI 10.1038/s41598-023-39738-7. Available from: https://www.nature.com/articles/s41598-023-39738-7
90. MAYERHOEFER, Marius E., MATERKA, Andrzej, LANGS, Georg, HÄGGSTRÖM, Ida, SZCZYPIŃSKI, Piotr, GIBBS, Peter and COOK, Gary. Introduction to radiomics. *Journal of Nuclear Medicine*. 1 April 2020. Vol. 61, no. 4, p. 488–495. DOI 10.2967/JNUMED.118.222893.
91. Introduction to Radiomics - PMC. [online]. [Accessed 7 February 2025]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9374044/
92. ORLHAC, Fanny, NIOCHE, Christophe, KLYUZHIN, Ivan, RAHMIM, Arman and BUVAT, Irène. Radiomics in PET imaging: a practical guide for newcomers. *PET Clinics* [online]. Vol. 2021, no. 4. [Accessed 7 February 2025]. Available from: https://hal.science/hal-03320546v1
93. WAGNER, Matthias W, KHASHAYAR NAMDAR, ·, BISWAS, · Asthik, MONAH, Suranna, FARZAD KHALVATI, · and ERTL-WAGNER, Birgit B. Radiomics, machine learning, and artificial intelligence-what the neuroradiologist needs to know. [online]. Vol. 1, p. 3. [Accessed 28 January 2025]. DOI 10.1007/s00234-021-02813-9. Available from: https://doi.org/10.1007/s00234-021-02813-9
94. Introduction to Radiomics. [online]. 2020. [Accessed 18 February 2025]. DOI 10.2967/jnumed.118.222893. Available from: http://www.snmmilearningcenter.org
95. JIN, Dan, NI, Xiaoqiong, TAN, Yanhuan, YIN, Hongkun and FAN, Guohua. Radiomics based on dual-layer spectral detector CT for predicting EGFR mutation status in non-small cell lung cancer. *J Appl Clin Med Phys*. 2025. P. 26. DOI 10.1002/acm2.14616.

96. Survival Rates for Kidney Cancer | American Cancer Society. [online]. [Accessed 28 January 2025]. Available from: https://www.cancer.org/cancer/types/kidney-cancer/detection-diagnosis-staging/survival-rates.html
97. SIEGEL MPH, Rebecca L, GIAQUINTO, Angela N, AHMEDIN, |, DVM, Jemal and SIEGEL, Rebecca L. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians* [online]. 1 January 2024. Vol. 74, no. 1, p. 12–49. [Accessed 5 February 2025]. DOI 10.3322/CAAC.21820. Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21820
98. Invasive Breast Cancer (IDC/ILC) | American Cancer Society. [online]. [Accessed 5 February 2025]. Available from: https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/invasive-breast-cancer.html
99. Invasive Ductal Carcinoma (IDC): Overview, Treatment & Prognosis. [online]. [Accessed 5 February 2025]. Available from: https://www.nationalbreastcancer.org/invasive-ductal-carcinoma/
100. SHEA, Eric Ka Ho, KOH, Valerie Cui Yun and TAN, Puay Hoon. Invasive breast cancer: Current perspectives and emerging views. *Pathology International* [online]. 1 May 2020. Vol. 70, no. 5, p. 242–252. [Accessed 5 February 2025]. DOI 10.1111/PIN.12910. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/pin.12910
101. LIU, Hanqin, XIA, Han, YIN, Xiaoxiao, QIN, Aiping, ZHANG, Wen, FENG, Shuang and JIN, Jing. Study on the differentiation of infiltrating breast cancer molecular subtypes based on ultrasound radiomics. *Clinical Breast Cancer* [online]. January 2025. Vol. 0, no. 0. [Accessed 5 February 2025]. DOI 10.1016/J.CLBC.2025.01.005. Available from: http://www.clinical-breast-cancer.com/article/S1526820925000163/fulltext
102. BRAY, Freddie, LAVERSANNE, Mathieu, WEIDERPASS, Elisabete and SOERJOMATARAM, Isabelle. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* [online]. 15 August 2021. Vol. 127, no. 16, p. 3029–3030. [Accessed 17 February 2025]. DOI 10.1002/CNCR.33587. Available from: https://pubmed.ncbi.nlm.nih.gov/34086348/
103. Soft Tissue Sarcoma: Symptoms, Treatment & Prognosis. [online]. [Accessed 17 February 2025]. Available from: https://my.clevelandclinic.org/health/diseases/21732-soft-tissue-sarcoma
104. GEADY, Caryn, ABBAS-AGHABABAZADEH, Farnoosh, KOHAN, Andres, SCHUETZE, Scott, SHULTZ, David and HAIBE-KAINS, Benjamin. Radiomic-Based Prediction of Lesion-Specific Systemic Treatment Response in Metastatic Disease. *medRxiv* [online]. 13 August 2024. P. 2023.09.22.23294942. [Accessed 17 February 2025]. DOI 10.1101/2023.09.22.23294942. Available from: https://www.medrxiv.org/content/10.1101/2023.09.22.23294942v3
105. BHAT, Gh Rasool, HYOLE, Rosalie G. and LI, Jiong. Head and neck cancer: Current challenges and future perspectives. *Advances in cancer research* [online]. 1 January 2021. Vol. 152, p. 67–102. [Accessed 28 January 2025]. DOI 10.1016/BS.ACR.2021.05.002. Available from: https://pubmed.ncbi.nlm.nih.gov/34353444/
106. PENG, Zhouying, WANG, Yumin, WANG, Yaxuan, JIANG, Sijie, FAN, Ruohao, ZHANG, Hua and JIANG, Weihong. Application of radiomics and machine learning in head and neck cancers. *International Journal of Biological Sciences* [online]. 2021. Vol. 17, no. 2, p. 475–486. [Accessed 28 January 2025]. DOI 10.7150/IJBS.55716. Available from: http://www.ijbs.com://creativecommons.org/licenses/by/4.0/
107. TORTORA, Mario, GEMINI, Laura, SCARAVILLI, Alessandra, UGGA, Lorenzo, PONSIGLIONE, Andrea, STANZIONE, Arnaldo, D'ARCO, Felice, D'ANNA, Gennaro and

CUOCOLO, Renato. Radiomics Applications in Head and Neck Tumor Imaging: A Narrative Review. *Cancers* [online]. 1 February 2023. Vol. 15, no. 4, p. 1174. [Accessed 3 March 2025]. DOI 10.3390/CANCERS15041174. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9954362/

108. DIJK, Lisanne V. van and FULLER, Clifton D. Artificial Intelligence and Radiomics in Head and Neck Cancer Care: Opportunities, Mechanics, and Challenges. *American Society of Clinical Oncology Educational Book* [online]. June 2021. No. 41, p. e225–e235. [Accessed 3 March 2025]. DOI 10.1200/EDBK_320951. Available from: https://ascopubs.org/doi/10.1200/EDBK_320951

109. How do you know if your data is imbalanced? [online]. [Accessed 3 March 2025]. Available from: https://www.deepchecks.com/question/how-do-you-know-if-your-data-is-imbalanced/

110. How can I determine if my data is balanced or imbalanced? | by Faheem Siddiqi | Medium. [online]. [Accessed 3 March 2025]. Available from: https://medium.com/@faheemsiddiqi789/how-can-i-determine-if-my-data-is-balanced-or-imbalanced-080819af408c

111. Chi-Square Goodness of Fit Test | Formula, Guide & Examples. [online]. [Accessed 11 April 2025]. Available from: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

112. Chi-Square Goodness of Fit Test Introduction to Statistics | JMP. [online]. [Accessed 11 April 2025]. Available from: https://www.jmp.com/en/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test

113. 11.2 - Goodness of Fit Test | STAT 200. [online]. [Accessed 11 April 2025]. Available from: https://online.stat.psu.edu/stat200/lesson/11/11.2

114. MUKHERJEE, Mimi and KHUSHI, Matloob. SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Applied System Innovation 2021, Vol. 4, Page 18* [online]. 2 March 2021. Vol. 4, no. 1, p. 18. [Accessed 3 March 2025]. DOI 10.3390/ASI4010018. Available from: https://www.mdpi.com/2571-5577/4/1/18/htm

115. Oversampling—Handling Imbalanced Data | by Abdallah Ashraf | Medium. [online]. [Accessed 3 March 2025]. Available from: https://medium.com/@abdallahashraf90x/oversampling-for-better-machine-learning-with-imbalanced-data-68f9b5ac2696

116. TIAN, L., ZHANG, D., BAO, S., NIE, P., HAO, D., LIU, Y., ZHANG, J. and WANG, H. Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas. *Clinical Radiology*. 1 February 2021. Vol. 76, no. 2, p. 158.e19-158.e25. DOI 10.1016/J.CRAD.2020.08.038.

117. PAN, Shaoyan, FLORES, Jessica, LIN, Cheng Ting, STAYMAN, J Webster and GANG, Grace J. Generative Adversarial Networks and Radiomics Supervision for Lung Lesion Synthesis. *Proceedings of SPIE--the International Society for Optical Engineering* [online]. 13 February 2021. Vol. 11595, p. 115950O. [Accessed 3 March 2025]. DOI 10.1117/12.2582151. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8516144/

118. MICHELUCCI, Umberto. Unbalanced Datasets and Machine Learning Metrics. *Fundamental Mathematical Concepts for Machine Learning in Science* [online]. 2024. P. 185–212. [Accessed 4 March 2025]. DOI 10.1007/978-3-031-56431-4_8. Available from: https://link.springer.com/chapter/10.1007/978-3-031-56431-4_8

119. Introduction to Machine Learning in R - GeeksforGeeks. [online]. [Accessed 3 March 2025]. Available from: https://www.geeksforgeeks.org/introduction-to-machine-learning-in-r/

120. ZHANG, W ;, GUO, Y ;, JIN, Q, ZHANG, Wenchao, GUO, Yu and JIN, Qiyu. Radiomics and Its Feature Selection: A Review. *Symmetry 2023, Vol. 15, Page 1834* [online]. 27 September 2023.

Vol. 15, no. 10, p. 1834. [Accessed 3 April 2025]. DOI 10.3390/SYM15101834. Available from: https://www.mdpi.com/2073-8994/15/10/1834/htm
121. HONG, Sungsoo, HONG, Sungjun, OH, Eunsun, LEE, Won Jae, JEONG, Woo Kyoung and KIM, Kyunga. Development of a flexible feature selection framework in radiomics-based prediction modeling: Assessment with four real-world datasets. *Scientific Reports 2024 14:1* [online]. 26 November 2024. Vol. 14, no. 1, p. 1–9. [Accessed 3 April 2025]. DOI 10.1038/s41598-024-80863-8. Available from: https://www.nature.com/articles/s41598-024-80863-8
122. 11. Correlation and regression. [online]. [Accessed 10 March 2025]. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression
123. SCHOBER, Patrick and SCHWARTE, Lothar A. Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia* [online]. 1 May 2018. Vol. 126, no. 5, p. 1763–1768. [Accessed 4 April 2025]. DOI 10.1213/ANE.0000000000002864. Available from: https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients__appropriate_use_and.50.aspx
124. NASIEF, Haidy, ZHENG, Cheng, SCHOTT, Diane, HALL, William, TSAI, Susan, ERICKSON, Beth and ALLEN LI, X. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. *npj Precision Oncology 2019 3:1* [online]. 4 October 2019. Vol. 3, no. 1, p. 1–10. [Accessed 4 April 2025]. DOI 10.1038/s41698-019-0096-z. Available from: https://www.nature.com/articles/s41698-019-0096-z
125. Correlation in machine learning—All you need to know | by Abdallah Ashraf | Medium. [online]. [Accessed 10 March 2025]. Available from: https://medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9
126. Correlation Matrix: What It Is, Why It's Used & How It's Created. [online]. [Accessed 10 March 2025]. Available from: https://www.displayr.com/what-is-a-correlation-matrix/
127. ELIANE BIRBA, Delwende. A Comparative study of data splitting algorithms for machine learning model selection. *DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING*.
128. A Guide to Data Splitting in Machine Learning | by Data Science Wizards | Medium. [online]. [Accessed 5 March 2025]. Available from: https://medium.com/@datasciencewizards/a-guide-to-data-splitting-in-machine-learning-49a959c95fa1
129. PURBA, Mariana, ERMATITA, Ermatita, ABDIANSAH, Abdiansah, NOPRISSON, Handrie, AYUMI, Vina, SALAMAH, Umniy, SETIAWAN, Hadiguna and YADI, Yadi. Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach. *IJACSA) International Journal of Advanced Computer Science and Applications* [online]. Vol. 13, no. 9, p. 2022. [Accessed 5 March 2025]. Available from: www.ijacsa.thesai.org
130. VARSHNEY, Nihir and SINGH, Sofia. Enhancing Diabetes Prediction: A comparative analysis of Train-Test Split and Stratified 10-Fold Cross-Validation with SMOTE Integration. *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024*. 2024. P. 1345–1351. DOI 10.1109/IC3I61595.2024.10829095.
131. DEMARCHI, Luca, KANIA, Adam, CIEZKOWSKI, Wojciech, PIÓRKOWSKI, Hubert, OŚWIECIMSKA-PIASKO, Zuzanna and CHORMAŃSKI, Jarosław. Recursive Feature Elimination and Random Forest Classification of Natura 2000 Grasslands in Lowland River Valleys of Poland Based on Airborne Hyperspectral and LiDAR Data Fusion. *Remote Sensing 2020, Vol. 12, Page 1842* [online]. 6 June 2020. Vol. 12, no. 11, p. 1842. [Accessed 5 March 2025]. DOI 10.3390/RS12111842. Available from: https://www.mdpi.com/2072-4292/12/11/1842/htm

132. DU, Yi, SHI, Haipeng, YANG, Xiaojing and WU, Weidong. Machine learning for infection risk prediction in postoperative patients with non-mechanical ventilation and intravenous neurotargeted drugs. *Frontiers in Neurology*. 1 August 2022. Vol. 13. DOI 10.3389/FNEUR.2022.942023.
133. Multi-class Classification on Imbalanced Data using Random Forest Algorithm in Spark | by Burak Özen | Medium. [online]. [Accessed 5 March 2025]. Available from: https://burakozen.medium.com/multi-class-classification-on-imbalanced-data-using-random-forest-algorithm-in-spark-5b3d0af9b93f
134. Random Forest - Regression and Classification - Explained using Sklearn - Python | I N F O A R Y A N. [online]. [Accessed 12 March 2025]. Available from: https://infoaryan.com/blog/random-forest-algorithm-explained-using-python-sklearn/
135. DERANGULA, Anusha, EDARA, SrinivasaReddy and KUMAR KARRI, Praveen. European Journal of Molecular & Clinical Medicine Feature Selection of Breast Cancer Data Using Gradient Boosting Techniques of Machine Learning. .
136. NATEKIN, Alexey and KNOLL, Alois. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. 2013. Vol. 7, no. DEC. DOI 10.3389/FNBOT.2013.00021/FULL.
137. What is XGBoost? | IBM. [online]. [Accessed 6 March 2025]. Available from: https://www.ibm.com/think/topics/xgboost
138. LAM, Luu Ho Thanh, CHU, Ngan Thy, TRAN, Thi Oanh, DO, Duyen Thi and LE, Nguyen Quoc Khanh. A Radiomics-Based Machine Learning Model for Prediction of Tumor Mutational Burden in Lower-Grade Gliomas. *Cancers 2022, Vol. 14, Page 3492* [online]. 18 July 2022. Vol. 14, no. 14, p. 3492. [Accessed 6 March 2025]. DOI 10.3390/CANCERS14143492. Available from: https://www.mdpi.com/2072-6694/14/14/3492/htm
139. How XGBoost Handles Missing Values | by 林承慶 ChengChing Lin | Medium. [online]. [Accessed 6 March 2025]. Available from: https://medium.com/@ar851060/how-xgboost-handles-missing-values-0e645fa76837
140. XGBoost Best Feature Importance Score | XGBoosting. [online]. [Accessed 6 March 2025]. Available from: https://xgboosting.com/xgboost-best-feature-importance-score/
141. Cross Validation and Tuning – Machine Learning for Tabular Data in R. [online]. [Accessed 6 March 2025]. Available from: https://carpentries-incubator.github.io/r-ml-tabular-data/06-Exploration/index.html
142. Avoid Overfitting By Early Stopping With XGBoost In Python - MachineLearningMastery.com. [online]. [Accessed 6 March 2025]. Available from: https://machinelearningmastery.com/avoid-overfitting-by-early-stopping-with-xgboost-in-python/
143. RAINIO, Oona, TEUHO, Jarmo and KLÉN, Riku. Evaluation metrics and statistical tests for machine learning. *Scientific Reports 2024 14:1* [online]. 13 March 2024. Vol. 14, no. 1, p. 1–14. [Accessed 10 March 2025]. DOI 10.1038/s41598-024-56706-x. Available from: https://www.nature.com/articles/s41598-024-56706-x
144. Model Evaluation in Machine Learning. [online]. [Accessed 10 March 2025]. Available from: https://www.appliedaicourse.com/blog/model-evaluation-in-machine-learning/
145. LÓPEZ, Osval Antonio Montesinos, LÓPEZ, Abelardo Montesinos and CROSSA, Dr. Jose. Overfitting, Model Tuning, and Evaluation of Prediction Performance. *Multivariate Statistical Machine Learning Methods for Genomic Prediction* [online]. 14 January 2022. P. 109–139. [Accessed 10 March 2025]. DOI 10.1007/978-3-030-89010-0_4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583970/

146. MILLER, Catriona, PORTLOCK, Theo, NYAGA, Denis M. and O'SULLIVAN, Justin M. A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics* [online]. 10 September 2024. Vol. 4, p. 1457619. [Accessed 10 March 2025]. DOI 10.3389/FBINF.2024.1457619. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11420621/
147. AUC ROC Curve in Machine Learning - Analytics Vidhya. [online]. [Accessed 10 March 2025]. Available from: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
148. Understanding the ROC Curve and AUC | Towards Data Science. [online]. [Accessed 10 March 2025]. Available from: https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb/
149. ÇORBACIOĞLU, Şeref Kerem and AKSEL, Gökhan. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine* [online]. 1 October 2023. Vol. 23, no. 4, p. 195. [Accessed 10 March 2025]. DOI 10.4103/TJEM.TJEM_182_23. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10664195/
150. DE HOND, Anne A.H., STEYERBERG, Ewout W. and VAN CALSTER, Ben. Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health* [online]. 1 December 2022. Vol. 4, no. 12, p. e853–e855. [Accessed 10 March 2025]. DOI 10.1016/S2589-7500(22)00188-1. Available from: https://www.thelancet.com/action/showFullText?pii=S2589750022001881
151. ROC Curve and AUC in Machine Learning: A Complete Guide! [online]. [Accessed 10 March 2025]. Available from: https://www.simplilearn.com/what-is-a-roc-curve-and-how-to-use-it-in-performance-modeling-article
152. Understanding the ROC Curve: When and How to Use It in Binary Classification | by Sanjay Dutta | Medium. [online]. [Accessed 8 April 2025]. Available from: https://medium.com/@sanjay_dutta/understanding-the-roc-curve-when-and-how-to-use-it-in-binary-classification-724b97f641f4
153. WANG, Liang and CARVALHO, Luis. Multiclass ROC. .
154. How to Use the AUC ROC Curve for the Multi-class Model? [online]. [Accessed 8 April 2025]. Available from: https://www.deepchecks.com/question/how-to-use-the-auc-roc-curve-for-the-multi-class-model/
155. PREZJA, Fabi, ANNALA, Leevi, KIISKINEN, Sampsa, LAHTINEN, Suvi, OJALA, Timo, RUUSUVUORI, Pekka and KUOPIO, Teijo. Improving Performance in Colorectal Cancer Histology Decomposition using Deep and Ensemble Machine Learning. [online]. 25 October 2023. [Accessed 8 April 2025]. DOI 10.1016/j.heliyon.2024.e37561. Available from: http://arxiv.org/abs/2310.16954
156. Multiclass averaging. [online]. [Accessed 8 April 2025]. Available from: https://cran.r-project.org/web/packages/yardstick/vignettes/multiclass.html
157. Accuracy, precision, and recall in multi-class classification. [online]. [Accessed 8 April 2025]. Available from: https://www.evidentlyai.com/classification-metrics/multi-class-metrics
158. What is a Confusion Matrix in Machine Learning? [online]. [Accessed 10 March 2025]. Available from: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning
159. MARKOULIDAKIS, Ioannis, RALLIS, Ioannis, GEORGOULAS, Ioannis, KOPSIAFTIS, George, DOULAMIS, Anastasios and DOULAMIS, Nikolaos. Multiclass Confusion Matrix

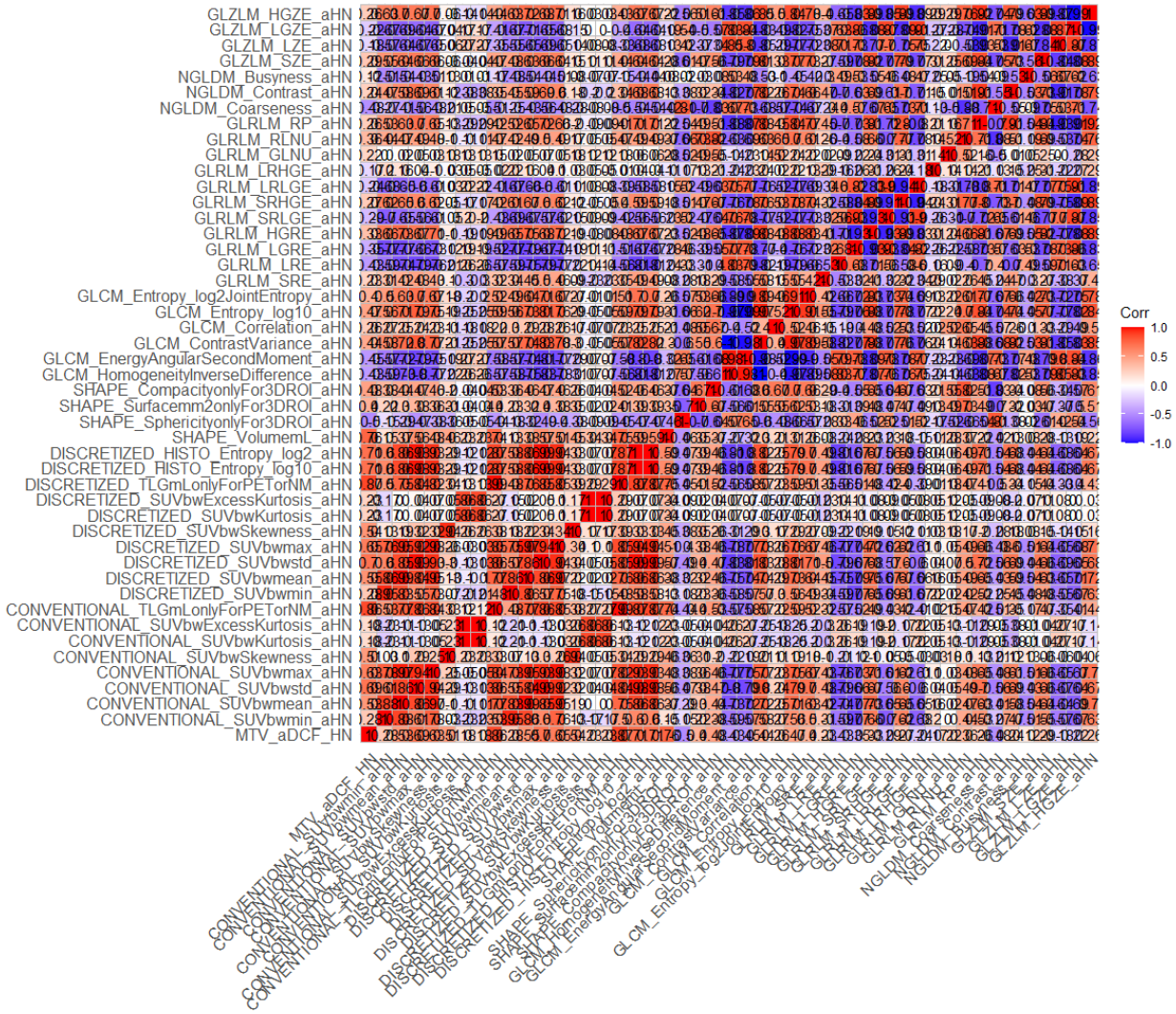
Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies 2021, Vol. 9, Page 81* [online]. 2 November 2021. Vol. 9, no. 4, p. 81. [Accessed 10 March 2025]. DOI 10.3390/TECHNOLOGIES9040081. Available from: https://www.mdpi.com/2227-7080/9/4/81/htm
160. Confusion Matrix: How To Use It & Interpret Results [Examples]. [online]. [Accessed 10 March 2025]. Available from: https://www.v7labs.com/blog/confusion-matrix-guide
161. VRETTOS, Konstantinos, TRIANTAFYLLOU, Matthaios, MARIAS, Kostas, KARANTANAS, Apostolos H and KLONTZAS, Michail E. Artificial intelligence-driven radiomics: developing valuable radiomics signatures with the use of artificial intelligence. *BJR|Artificial Intelligence* [online]. 4 March 2024. Vol. 1, no. 1. [Accessed 11 April 2025]. DOI 10.1093/BJRAI/UBAE011. Available from: https://dx.doi.org/10.1093/bjrai/ubae011
162. YOUSEFZADEH, Reza, KAZEMI, Alireza and AL-MAAMARI, Rashid S. Application of power-law committee machine to combine five machine learning algorithms for enhanced oil recovery screening. *Scientific Reports* [online]. 1 December 2024. Vol. 14, no. 1. [Accessed 11 April 2025]. DOI 10.1038/S41598-024-59387-8. Available from: https://www.researchgate.net/publication/380002791_Application_of_power-law_committee_machine_to_combine_five_machine_learning_algorithms_for_enhanced_oil_recovery_screening
163. Understanding CatBoost: The Gradient Boosting Algorithm for Categorical Data | by Aravind Kolli | Medium. [online]. [Accessed 11 April 2025]. Available from: https://aravindkolli.medium.com/understanding-catboost-the-gradient-boosting-algorithm-for-categorical-data-73ddb200895d
164. ISAKSSON, Lars Johannes, REPETTO, Marco, SUMMERS, Paul Eugene, PEPA, Matteo, ZAFFARONI, Mattia, VINCINI, Maria Giulia, CORRAO, Giulia, MAZZOLA, Giovanni Carlo, ROTONDI, Marco, BELLERBA, Federica, RAIMONDI, Sara, HARON, Zaharudin, ALESSI, Sarah, PRICOLO, Paula, MISTRETTA, Francesco Alessandro, LUZZAGO, Stefano, CATTANI, Federica, MUSI, Gennaro, DE COBELLI, Ottavio, CREMONESI, Marta, ORECCHIA, Roberto, LA TORRE, Davide, MARVASO, Giulia, PETRALIA, Giuseppe and JERECZEK-FOSSA, Barbara Alicja. High-performance prediction models for prostate cancer radiomics. *Informatics in Medicine Unlocked*. 1 January 2023. Vol. 37, p. 101161. DOI 10.1016/J.IMU.2023.101161.
165. MA, Yongfeng, XIE, Zhuopeng, LI, Wenlu and CHEN, Shuyan. Modeling driving styles of online ride-hailing drivers with model identifiability and interpretability. *Travel Behaviour and Society* [online]. 1 October 2023. Vol. 33. [Accessed 11 April 2025]. DOI 10.1016/j.tbs.2023.100645. Available from: https://www.datacamp.com/tutorial/catboost
166. RANGANATHAN, Priya. An Introduction to Statistics: Choosing the Correct Statistical Test. *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* [online]. 2021. Vol. 25, no. Suppl 2, p. S184. [Accessed 14 April 2025]. DOI 10.5005/JP-JOURNALS-10071-23815. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8327789/
167. Kruskal-Wallis H Test in SPSS Statistics | Procedure, output and interpretation of the output using a relevant example. [online]. [Accessed 14 April 2025]. Available from: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php
168. DINNO, Alexis. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal*. 2015. Vol. 15, no. 1, p. 292–300.
169. MAINTAINER, Alexis Dinno and DINNO, Alexis. Title Dunn's Test of Multiple Comparisons Using Rank Sums. . 2024.

170. WANG, Ruodu. Elementary proofs of several results on false discovery rate. . 2022.
171. Mann-Whitney U Test: Assumptions and Example | Technology Networks. [online]. [Accessed 20 April 2025]. Available from: https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

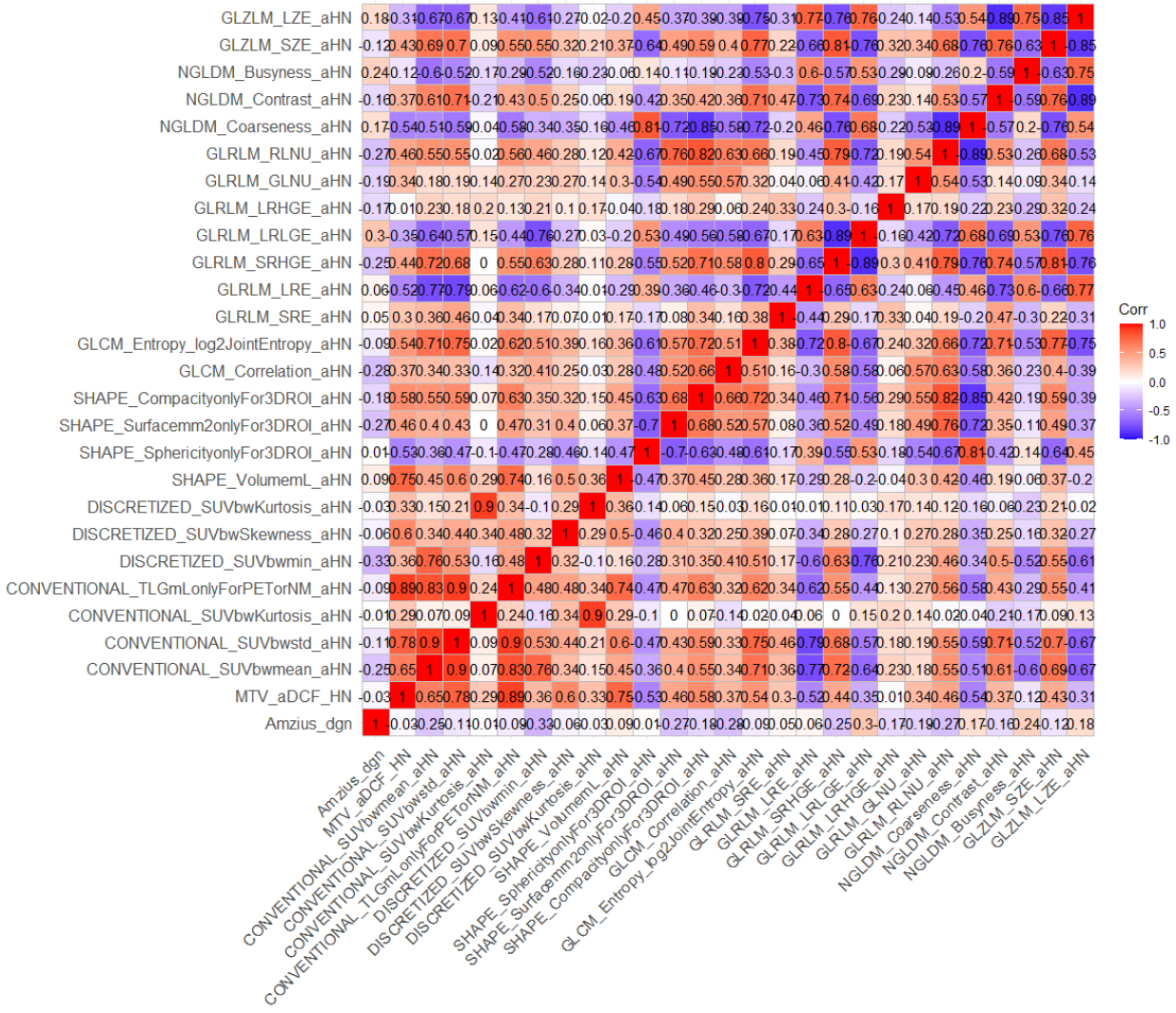
Appendices

Appendix 1. Correlation matrix

Correlation matrix of the original dataset:



Correlation matrix, after elimination of redundant features by applying a correlation coefficient threshold of 0.9.



Appendix 2. Exploration of the dataset and Assessment and Correction of Target Variable Imbalance.

The dataset was initially explored using common R functions. To remove variables with missing values, the function *data %>% select_if(~ all(!is.na(.)))* was used. The structure and types of variables were assessed using *str(data)*. If the target variable or any other variable was incorrectly classified (e.g., as integer instead of factor), the appropriate type conversion was performed using *factor(data$variable)*, as factors in R are used to represent categorical variables.

The distribution of the target variable was evaluated using the *table(data$target_variable)* function. To formally assess whether the target variable was balanced across classes, the Chi-Square Goodness-of-Fit test was performed using *chisq.test()*

In cases where the Chi-Square test indicated significant imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic examples for the minority class, thereby improving class balance. The SMOTE function used was *SMOTE ()* where, *K* defines the number of nearest neighbors considered, and *dup_size* determines the number of synthetic samples generated per minority sample.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_1%20Data%20exploration%20and%20investigation%20of%20distribution%20of%20target%20variable.R

Appendix 3. Correlation Analysis for Redundancy Reduction

Correlation analysis was used to eliminate redundant variables and reduce multicollinearity in the dataset. A correlation coefficient criterion of 0.9 was used, which indicated that pairs of variables with a Spearman correlation greater than 0.9 were highly associated. The correlation matrix was generated with the function *cor().*

To detect and remove highly correlated variables, the function *findCorrelation()* was used. This method deliberately removes the fewest variables necessary to ensure that no pair of remaining variables exceeds the given correlation threshold.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_2%20Correlation%20analysis.R

Appendix 4. Model Training and Cross-Validation Strategy

Before training, the dataset was divided into two subsets: 80% was utilized for model training, and the remaining 20% was set aside for validation.

The small sample size in this study increased the chance of overfitting. To mitigate this and achieve more stable model performance, k-fold cross-validation was used during the training process.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_3%20Data%20set%20separation.R

Appendix 5. Feature Selection Based on Predictive Relevance

Feature selection was done with two tree-based machine learning algorithms: Recursive Feature Elimination with Random Forest (RFE-RF) and eXtreme Gradient Boosting (XGBoost). These strategies were used to discover the most essential traits that help forecast the target variable.

Given the modest data size, 10-fold cross-validation was used during model training to provide robust feature selection while minimizing overfitting. Cross-validation enabled a more trustworthy estimate of model performance across diverse groups of data.

For RFE-RF: *rfeControl(), rfe()* functions were used.

For XGBoost: *trainControl()***,** *train (..., method = "xgbTree")* were used.

In both cases, model performance was evaluated using cross-validation, and feature importance was assessed based on the resulting trained models. The selection of hyperparameters for XGBoost (such as max_depth, eta, and colsample_bytree) was guided by empirical tuning to balance model complexity and performance.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_4%20Feature%20selection%20RFE-RF%20and%20XGBoost.R

Final model's development:

In this study, a multiclass classification model was created using the CatBoost algorithm. CatBoost is a gradient boosting framework known for its efficient handling of categorical features and strong performance. To confirm the model's generalizability and reduce overfitting, 10-fold cross-validation was used during the training procedure. The model was trained using the *catboost.train()* function,

The training dataset was encapsulated into a CatBoost Pool object, optimized for handling both numerical and categorical data. The 10-fold cross-validation was implemented to evaluate the model's performance across different subsets of the data, providing a robust assessment of its predictive capabilities. This approach allowed for the effective modelling of complex relationships within the data, leveraging CatBoost's capabilities to handle categorical variables and its robustness against overfitting.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_6%20CatBoost%20for%20both%20binary%20and%20multi-class%20classification.R

Appendix 6. Comprehensive Performance Evaluation

In this study, the performance of all machine learning algorithms was assessed using the most used methodologies. The confusion matrix showed a detailed breakdown of model predictions, including the number of correct and incorrect classifications for each class. This method is especially effective at detecting specific types of misclassifications. Confusion matrix can be calculated by calling the function *confusionMatrix().*

The Receiver Operating Characteristic (ROC) curve improved the diagnostic ability of the binary classifier (in the multi-class scenario, macro-ROC was utilized) system as the discrimination threshold was adjusted. The ROC curve in the binary case was calculated using the *roc()* function, whereas in the multi-class scenario, more involved code was used, which is available in the link below. The Area Under the Curve (AUC) measures the model's overall ability to differentiate between classes. The function *auc()* was used in this case to compute AUC values for both binary and multiclass scenarios. The *MultiClassSummary()* function was employed to calculate the recall, F1, and precision scores in the multiclass example.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_5%20Performance%20evaluation%20for%20both%20multi-class%20and%20binary%20classification.R

Appendix 7. Statistical Validation of Radiomics Features

A variety of nonparametric tests were used to determine the statistical significance of changes in delta radiomic characteristics between clinical groups. These tests are designed specifically for data that does not match normality assumptions.

The Mann-Whitney U Test, known as Wilcox rank sum test, was employed to compare the delta radiomics properties of two independent groups. This test is only applicable for binary classification, when the data is continuous and not regularly distributed. Function for this test - *wilcox.test()*.

Kruskal-Wallis H test was used in the multi-class classification task, which involved analyzing more than two independent groups. This test is designed for nonparametric data. It determines whether there are statistically significant differences in medians across three or more groups. In R, it can be used with the *kruskal.test()* function.

Dunn's Post Hoc Test with Benjamini-Hochberg correction was employed after the Kruskal-Wallis Test analysis, if it indicated significant differences. Then, Dunn's test was conducted to identify which specific groups differ from each other. To control false discovery rate due to multiple comparisons, the Benjamini-Hochber procedure was applied to adjust the p-values. Function for this analysis – *dunnTest()*.

The full code is available at the following link: https://github.com/blcugn3/Radiomics-workflow-for-HNC/blob/main/Step_7%20Statistical%20validation%20for%20both%20multi-class%20and%20binary%20case.lnk