**CCT College Dublin**

|  |  |
| --- | --- |
| **Modules Titles:** | Research & Professional Ethics |
| **Assessment Title:** | Research Proposal |
| **Lecture:** | Kelvin Martins |
| **Student:** | Ricardo Alves de Souza |
| **Student Number:** | 2018499 |
| **Assessment due Date:** | 05/05/2025 |
| **Date of Submission:** | 05/05/2025 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

# Table of Contents

[Proposed Title: 2](#_Toc2137924847)

[Deep Learning-Based Radiomics for Sarcoma Prognosis: A Multi-Modal Approach to Predict Progression-Free Survival 2](#_Toc1682296204)

[Research Objectives and/or Hypothesis 3](#_Toc723522617)

[Objectives 3](#_Toc1262915095)

[Hypothesis 3](#_Toc255595082)

[Literature Review 4](#_Toc790479030)

[Machine Learning for Survival Prediction 4](#_Toc1000643109)

[Radiomics in Cancer Prognosis 4](#_Toc221765502)

[Deep Learning-Based Radiomics 5](#_Toc948938522)

[Predictive Factors and Biomarkers 6](#_Toc41632031)

[Radiomics for Rare Cancers 6](#_Toc1604977690)

[High-Grade Glioma and Generalizability 6](#_Toc700112695)

[Limitations and Future Directions 6](#_Toc161439746)

[Proposed Sampling Strategy 7](#_Toc832502631)

[Proposed Primary Research Methodology 8](#_Toc837219770)

[Ethical Considerations 9](#_Toc526065318)

[References 10](#_Toc276106049)

# Proposed Title:

Deep Learning-Based Radiomics for Sarcoma Prognosis: A Multi-Modal Approach to Predict Progression-Free Survival

# Research Objectives and/or Hypothesis

## Objectives

To develop and optimize deep learning models, including Deep Learning Neural Networks (DLNNs) and Deep Learning-based Radiomics (DLR), to automatically extract and analyse high-dimensional features from existing sarcoma imaging (MRI/CT/PET-CT) and clinical datasets for predicting PFS.

To evaluate the computational efficiency and predictive performance of deep learning models compared to traditional machine learning approaches (e.g., Cox Proportional Hazards, Random Survival Forests) using metrics such as AUC and concordance index.

To validate the predictive performance of the proposed models using cross-validation and external cohorts to ensure generalizability.

To quantify the improvement in predictive accuracy by integrating radiomics and clinical data using deep learning compared to models relying solely on clinical data.

To automate risk stratification of sarcoma patients into high-and low-risk groups using deep learning predictions to demonstrate scalable, data-driven decision support.

Hypothesis

H1: Deep learning models integrating radiomics and clinical features will outperform traditional machine learning models in predicting PFS for sarcoma patients by efficiently capturing complex, non-linear patterns in high-dimensional data.

H2: Deep Learning-based Radiomics (DLR) models will achieve higher predictive accuracy and computational efficiency compared to traditional statistical methods (e.g., Cox regression) by automating feature extraction and reducing preprocessing requirements.

H3: Automated feature extraction from normalized MRI images using deep learning will improve the scalability, reproducibility, and predictive power of models for PFS in sarcoma patients.

# Literature Review

Sarcomas represent only about 1% of all cancers, and their extreme rarity and heterogeneity make progression-free survival (PFS), one of the most important endpoints in cancer clinical trials, particularly challenging. Developments in data analytics, especially ML and radiomics, represent a new opportunity to improve prognosis prediction models for rare and heterogeneous cancers such as sarcoma. This literature review summarizes principal studies that, using ML and radionics, have predicted progression-free survival (PFS) and overall survival (OS) in sarcomas and similar diseases, outlining methods, results, and disadvantages. The review examines nine strong sources to support the research they propose, primarily on the approaches these sources use to write predictive models for sarcoma patients.

## Machine Learning for Survival Prediction

Machine learning provides a valuable alternative to traditional regression methods such as the Cox Proportional Hazards (CPH) model, which assumes linearity in the relationships between the predictors and the hazard function (or log hazard). It has emerged as a popular tool for survival analysis. Matsuo et al. (2019) find a significant improvement in the prediction of PFS and OS in the primitive cervical cancer model in the comparison of the deep learning neural network (DLNN) model with CPH. The DLNN significantly outperformed dropout-based Cox proportional hazards (CPH) in terms of PFS (MAE, 29.3 vs. 316.2) by accounting for non-linear relationships between 40 clinical and tumor-related features.

This indicates that as tumor heterogeneity leads to complex feature interactions, DLNN may be helpful for application on sarcoma. There has not been an abundance of studies utilizing a concordance index to assess the prognostic capacity of clinical factors or models in sarcoma research.

Similarly, Liu et al. (2020) built an ANN model for predicting 1-year PFS in HCC patients. Through a cohort including 2890 patients, the model obtained clinical risk factors including tumor size and portal vein tumor thrombus and laboratory markers such as alpha-fetoprotein (AFP). The AUROC corresponding to ANN was as high as 0.866, which was higher than traditional staging systems such as TNM and BCLC.

The ability of the study to handle high-dimensional data highlights the promise of ANNs for sarcoma in which many clinical and imaging variables exist. In addition, the ANN scores with associated risk stratification of patients in high, medium, and low-risk groups open a horizon for personalized treatment based on risk level, one of the most prominent objectives of sarcoma prognosis.

## Radiomics in Cancer Prognosis

Radiomics has improved prognostic modelling by capturing tumor heterogeneity through high-throughput extraction of quantitative features from medical images. Zhu et al. (2024) utilized MRI-based features from 202 patients to develop a radiomics nomogram for predicting PFS in sarcoma patients. When using LASSO Cox regression, the AUCs achieved with their model, ROI-Net, were 0.947, 0.907, and 0.924 after 300, 600, and 900 days, respectively.

Adding clinical characteristics such as peritumoral edema and metastases enhanced model performance, showcasing the effectiveness of radiomics clinical integration. These results are of great scientific value, but the conclusion of the study stands in the limitations of its retrospective design and absence of external validation, which should be a requirement of further sarcoma studies to ensure the multicentre applicability of the results.

Cui et al. (2022) explored the ML models based on the imaging features (radionics) from CT images to predict PFS and OS in patients with oesophageal squamous cell carcinoma (ESCC). In the test cohort, their combined radionics-clinical model surpassed standalone models with AUCs of 0.833 and 0.768 for PFS and OS, respectively. This contribution relies on clinical variables, including TNM stage, tumour differentiation, and ECOG performance status, that also increased predictive accuracy. In sarcoma, for which clinical staging has clear relevance but is often insufficient on its own, these findings generate new opportunities for diagnostic differentiation.

Applying LASSO Cox regression for feature selection is one of the recommended approaches for managing high-dimensional radionics data and provides an established theoretical framework for sarcoma investigations.

Peeken et al. (2018) investigated treatment-related variables in soft tissue sarcoma (STS) patients receiving radiotherapy. AUCs for death, local progression, and systemic progression using a random forest-based model with integrated dose and planning target volume were 0.87, 0.88, and 0.84, respectively.

Features of response to therapy improved performance, indicating that therapy-related data may help improve prognostic models in sarcoma. The emphasis on treatment-associated variables in this study overcomes the limitations of classical prognostic models based on a staging framework alone. It provides data to inform sarcoma-embedded trials on the selection of therapy-related features.

## Deep Learning-Based Radiomics

Deep learning-based radionics (DLR) enhanced conventional radionics by outperforming feature representation automatically.

Gu et al. (2022) proposed a 3D multi-modality DLR model using the PET/CT images to predict 5-year PFS in nasopharyngeal carcinoma (NPC). It produced receiver operating characteristic curves with area under curves (AUCs) of 0.842 and 0.823 in internal and external cohorts, which were greater than conventional radionics (AUC: 0.796 and 0.782, respectively) adding the TNM stage further enhanced performance, highlighting the supplemental information provided by clinical features. Sarcoma has a wide range of subtypes, and the high dimensionality of the 3D CNNs which the study has implemented to convey the tumor heterogenicity accurately, is very favourable in this regard.

The focus on external validation strengthens the study and its generalizability to the general sarcoma research landscape.

Hussain et al. (2023) performed a systematic review of deep learning, radiomics, and radio genomics studies using digital breast tomosynthesis (DBT) and 3D mammography. Their review of 30 studies showed that DLR models have the potential to optimize early breast cancer detection by capturing both complex and straightforward imaging phenotypes. The review observed that DLR models used for DBT images surpassed the performance of traditional radionics through the utilization of three-dimensional convolutional neural networks (3D CNNs) for densely capturing spatial and textural features, demonstrating improved accuracy for disease outcome predictions. While these results are related to breast cancer, as both tumours are complex, the use of 3D imaging modalities, for example, MRI for sarcoma, which both benefit from heterogeneous tumor characterization, is also relevant for sarcoma. Parallel benchmarking of radiomics and DLR models is the focus of this study and provides a systematic framework to compare model performance within sarcoma research.

Predictive Factors and Biomarkers

De Nonneville et al. (2019) validated pre-therapeutic absolute neutrophil count (ANC) as a predictive factor of progression-free survival (PFS) in patients with soft tissue sarcomas (STS) treated with trabectedin. Patients with neopterin normal (ANC < 7.5 G/L) indeed had a median PFS of 5.78 months versus 3.22 months for those with elevated ANC ≥ 7.5 G/L (p = 0.009).

In this regard, while this study used a mathematical model to define ANC thresholds, their approach to data-driven biomarker discovery may be leveraged in future sarcoma imaging feature models. High ANC is associated with inferior PFS, which indicates inflammation may mirror tumor biology and points to the need to integrate biomarker data into sarcoma prognostic models.

## Radiomics for Rare Cancers

Destito et al. (2023) applied radiomics to predict overall survival (OS) and progression-free survival (PFS) in a rare cancer, primary central nervous system lymphoma (PCNSL).

They observed that for overall survival (OS) and progression-free survival (PFS), the AUC associated with their machine-learning (ML) models using normalized MRI images achieved a 23% enhancement as compared to clinical prognostic factors for OS and approaching a 50% enhancement for the PFS analysis.

Feature stability after Z-score normalization was improved dramatically (p<10⁻¹²), pointing out the critical role played by image preprocessing.

Sarcomas are rare cancers, so the small sample sizes in the study are in keeping with current research in sarcomas, which requires clear feature selection and normalization (Meza et al. The fact that textural features from normalized images outperformed first-order features is in line with survival prediction being driven by intra-lesion heterogeneity.

This concept is also valid for sarcoma.

## High-Grade Glioma and Generalizability

Kwiatkowska-Miernik et al. (2024) applied radiomics and ML to predict the overall PFS of high-grade glioma patients and reached a 1-MAPE of 92.27% using a random forest model.

One recent example of such a strategy is the study with 109 radiomic features and four clinical variables (sex, weight, age, tumor location) achieving higher accuracies than purely radiomic-based models, demonstrating the advantage of multi-modal feature sets.

While the small sample size and semi-automatic segmentation indicate challenges in reproducibility, which could be solved by automated segmentation additionally in better-powered cohorts for sarcoma research, these findings point to specific image processing strategies to test different hypotheses in single centres or even collaborative studies.

The strong predictive performance of the random forest model justifies its relative complacency for sarcoma, especially in the context of heterogeneous data.

# Limitations and Future Directions

Many of these studies were limited by small sample size, retrospective design, and a lack of external validation, particularly in rare cancers such as sarcoma and PCNSL.

For instance, Zhu et al. (2024) and Cui et al. (2022) pointed out the need for additional prospective research and external cohorts to confirm their findings.

As emphasized by Kwiatkowska-Miernik et al. (2024), dependence on manual or semi-automatic segmentation creates reproducibility issues. In addition, since the potential biological significance of radiomic features is uncertain, radiomic features urgently need to be combined with genomic and pathological data (Gu et al., 2022).

The black box of DLR models, advanced by Gu et al. (2022), highlights the need for interpretable AI for clinical implementation.

The current studies also make important contributions to the work proposed here by showing the effectiveness of ML and radionics in predicting PFS across a range of cancers.

Matsuo et al. (2019) and Liu et al. (2020) emphasize the strength of neural networks on complex feature interplay that also directs the selection of DLNN and ANN models for sarcoma.

Zhu et al. (2024) and Cui et al. (2022) bring support for the proposed multi-modal approach, including the fusion of radionics and clinical features. Peeken et al. (2018) and De Nonneville et al. (2019) highlight the value of treatment-related and biomarker information, leading to the addition of such variables.

Gu et al. (2022) and Destito et al. (2023) emphasize, respectively, the importance of DLR and image normalization for uncommon malignancies, Kwiatkowska-Miernik et al. (2024) demonstrate that random forests might be the answer for heterogeneous data.

In summarizing the gaps that remain, Zhu et al. (2024) lays out a clear plan for future sarcoma research involving larger, multicentre, prospective cohorts, which would increase generalizability to the whole population, while automated segmentation tools, as proposed by Destito et al. (2023) allow reproducibility and normalization methods to create feature stability.

As suggested by Gu et al. (2022), the integration of genomic data could reveal the biological meaning of the radiomic features. Furthermore, building interpretable ML models, as suggested by Gu et al. (2022), will allow translation into the clinic.

The reviewed studies suggest that machine learning (ML) and radiomics have promise to significantly improve progression-free survival (PFS) prediction for sarcoma patients.

Using combined clinical and radionics features as input to state-of-the-art DLR models while tackling the lack of reproducibility due to methodological variances via normalization, it is possible to build on these findings to ultimately define a versatile, robust predictive tool that could be easily integrated into the personalizing treatment of patients with sarcoma.

To ensure that the models are generalizable and facilitate clinical actionability, the proposal outlines a novel approach leveraging external validation and multi-modal data integration, as well as automated segmentation, to advance predictive tools to improve patient outcomes in this difficult-to-treat disease.

# Proposed Sampling Strategy

This research will utilize existing datasets from adults aged 18 years and older with a confirmed sarcoma diagnosis, sourced from established registries such as the French Sarcoma Group and the National Cancer Institute [US] SEER program.

Datasets must include pre-treatment imaging (MRI/CT/PET-CT) and clinical variables, such as age, sex, and tumor stage to support deep learning model development for predicting PFS. Patients are eligible if they have imaging data available (MRI, CT, or PET/CT scans) prior to initiating treatment, along with detailed clinical records that document tumor characteristics, treatment history, and follow-up outcomes.

The population includes patients with 2-5 years of multi-modal data, clinical and imaging, required for ML model development to predict PFS, as this data is challenging to come by. The research is targeting sarcoma as an area of need due to its heterogeneity and rarity of disease. Therefore, personalized prognostic tools are lacking in this space.

This study will take a purposive sampling approach, sampling patients from various cancer centres to ensure variation in the type, stage, and treatment regimen of sarcoma.

More specifically, a stratified sampling method will be employed to ensure representation across the following important subgroups: STS versus bone sarcoma, low-grade versus high-grade tumours, and early-stage versus metastatic disease.

A purposive sampling method will be used to acquire high-quality, multi-modal datasets from multiple cancer registries with ample variability in imaging and clinical characteristics to build generalizable deep learning models.

To achieve maximal generalizability of the model, stratification will be based on data characteristics, such as imaging modality and data completeness rather than clinical subtypes. The sample size is estimated (~300 patients, power = 0.8, alpha = 0.05) for deep learning model training considering the rare nature of sarcoma. As previously shown by Liu et al. (2020) this size achieves a balance between statistical robustness and making use of curated registry data.

By harnessing prior datasets, these approaches are scalable to multiple conditions and disease states without the need for expensive prospective data collection, addressing the existing need for such an analytics platform to provide data-driven guidelines.

The generalizability of the results is fortified here as multi-center data can overcome such limitations as noted by Cui et al. (2022) in single-center studies.

# Proposed Primary Research Methodology

The primary methodology will involve retrospective data extraction from existing sarcoma registries, such as French Sarcoma Group and SEER containing pre-treatment imaging (MRI/CT/PET-CT) and clinical variables, such as age, sex and tumor stage.

Development of deep learning models such as Deep Learning Neural Network (DLNN) and Deep Learning-based Radiomics (DLR) to mine imaging data for feature extraction with a reduced dependency on manual segmentation.

Then, automated tools, such as 3D Convolutional Neural Networks (CNNs), will be designed to generalize from ROI-Net (Zhu et al., 2024), to quantify properties of tumours in a high-throughput fashion. Imaging data will be pre-processed to obtain stable features (Destito et. al. (2023).

Clinical variables will be harmonized across datasets to provide one common high-dimensional input to deep learning models.

Using standard imaging protocols and clinical assessments to increase the recency of data.

The rationale behind these approaches is that they can create a desirable, high-quality, multi-modal dataset for training ML models.

The study of Cui et al. (2022) provides further evidence that CT-based radionics can be used to predict survival while maintaining cost-effectiveness by relying on high-quality datasets that were already available.

In comparison to semi-automatic methods cited by Kwiatkowska-Miernik et al. (2024), automated segmentation, as used by Zhu et al. (2024), decreases manual work and increases reproducibility across centers.

Data gaps are filled by clinician surveys, ensuring that aspects, such as performance status, which De Nonneville et al. (2019) identified to be significant for PFS prediction. Although both the time and costs would not justify the collection of prospective data, it may well be the case that more is needed in terms of real-world evidence, as already suggested recently by Peeken et al. (2018).

Prospective data would undoubtedly be able to make it less dated from what the standard practical treatment is concerned. All these methods can help enable the development of ML models that are both accurate and generalizable by providing aggregate data of relevant clinical and imaging features.

# Ethical Considerations

Numerous ethical and legal/regulatory issues need to be addressed to ensure such research is conducted responsibly.

To start, informed consent is paramount. In common registry-based studies, finding fully anonymized data would require an IRB's approval to justify minimal risk to participants, allowing for waiver of consent needed for retrospective data.

Written or oral informed consent will be sought, depending on local regulations, to provide information about the purpose of the study, data usage, and all relevant information about the study to the potential risk of the participants in advance to ensure that the individual agrees to participate voluntarily with the right to withdraw at any time.

Secondly, data privacy is of utmost importance, and customers must comply with regulations like the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the USA.

Identifiable patient information, such as names and dates of birth, must be anonymized, and individual patients must be issued with unique identifiers.

Third, the bias and fairness of ML models need to be mitigated as if the model automatically predicts biased treatment, such as prediction based on gender or ethnicity that members of the same demographic group would be unfairly treated.

Fourth, the security of the data is important so that no one has access or breaches; this requires secure storage and transmission protocols. Lastly, transparency in ML model development is necessary to overcome the "black box" problem that Gu et al. (2022) point out in the generation of predictive outputs, making them reliable and trustworthy for clinicians and patients alike.

In response to these considerations, the following measures will be taken.

A harmonized, multilingual, and ethics-approved consent form will be created for potential participants to facilitate informed consent.

For datasets involving retrospective data, such a waiver request will be submitted to the IRB detailing all anonymization steps taken and compliance with suitable ethical guidelines.

Anonymization will also be in accordance with ISO/IEC 27001 to ensure data privacy; a data management plan will outline processes for data handling, storage, and destruction post-study.

As per guidelines by AI ethics, bias mitigation is done using fairness-aware algorithms in feature selection.

In contrast, model performance is audited across demographic groups, such as age, sex, and ethnicity, using metrics like demographic parity.

Hosting encrypted servers, using multi-factor authentication, periodic security audits, as well as having a data breach response plan will ensure the protection of data successfully.

To increase transparency, non-sensitive study materials, such as ML model architectures and feature selection approach, will be made available in a public repository as performed by Destito et al. (2023), as they promote reproducibility and trust.

# References

Cui, Y., Li, Z., Xiang, M., Han, D., Yin, Y., & Ma, C. (2022). Machine learning models predict overall survival and progression free survival of non-surgical esophageal cancer patients with chemoradiotherapy based on CT image radiomics signatures. Radiation Oncology, 17(1), 212. <https://link.springer.com/article/10.1186/s13014-022-02186-0>

De Nonneville, A., Barbolosi, D., Andriantsoa, M., El-Cheikh, R., Duffaud, F., Bertucci, F., & Salas, S. (2019). Validation of neutrophil count as an algorithm-based predictive factor of progression-free survival in patients with metastatic soft tissue sarcomas treated with trabectedin. Cancers, 11(3), 432. <https://doi.org/10.3390/cancers11030432>

Destito, M., Marzullo, A., Leone, R., Zaffino, P., Steffanoni, S., Erbella, F., ... & Spadea, M. F. (2023). Radiomics-based machine learning model for predicting overall and progression-free survival in rare cancer: a case study for primary CNS lymphoma patients. Bioengineering, 10(3), 285. https://doi.org/10.3390/bioengineering10030285

Gu, B., Meng, M., Bi, L., Kim, J., Feng, D. D., & Song, S. (2022). Prediction of 5-year progression-free survival in advanced nasopharyngeal carcinoma with pretreatment PET/CT using multi-modality deep learning-based radiomics. Frontiers in Oncology, 12, 899351. <https://doi.org/10.3389/fonc.2022.899351>

Hussain, S., Lafarga-Osuna, Y., Ali, M., Naseem, U., Ahmed, M., & Tamez-Peña, J. G. (2023). Deep learning, radiomics and radiogenomics applications in the digital breast tomosynthesis: a systematic review. BMC bioinformatics, 24(1), 401. <https://link.springer.com/article/10.1186/s12859-023-05515-6>

Kwiatkowska-Miernik, A., Wasilewski, P. G., Mruk, B., Sklinda, K., Bujko, M., & Walecki, J. (2024). Estimating progression-free survival in patients with primary high-grade glioma using machine learning. Journal of Clinical Medicine, 13(20), 6172. <https://doi.org/10.3390/jcm13206172>

Liu, X., Hou, Y., Wang, X., Yu, L., Wang, X., Jiang, L., & Yang, Z. (2020). Machine learning-based development and validation of a scoring system for progression-free survival in liver cancer. Hepatology International, 14, 567-576. <https://link.springer.com/article/10.1007/s12072-020-10046-w>

Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R. S., Takiuchi, T., Liu, Y., & Roman, L. D. (2019). Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. American Journal of Obstetrics and Gynecology, 220(4), 381-e1. <https://doi.org/10.1016/j.ajog.2018.12.030>

Peeken, J. C., Goldberg, T., Knie, C., Komboz, B., Bernhofer, M., Pasa, F., ... & Combs, S. E. (2018). Treatment-related features improve machine learning prediction of prognosis in soft tissue sarcoma patients. Strahlentherapie und Onkologie, 1-11. DOI:10.1007/s00066-018-1294-2

Zhu, N., Niu, F., Fan, S., Meng, X., Hu, Y., Han, J., & Wang, Z. (2024). Predicting progression-free survival in sarcoma using MRI-based automatic segmentation models and radiomics nomograms: a preliminary multicenter study. Skeletal Radiology, 1-11. <https://link.springer.com/article/10.1007/s00256-024-04837-7>