# Predicting locoregional recurrence in breast cancer following breast-conserving therapy using learning-based models with multi-institutional registries

Sang Kyun Yoo

The Graduate School
Yonsei University
Department of Medicine

# Predicting locoregional recurrence in breast cancer following breast-conserving therapy using learning-based models with multi-institutional registries

A Dissertation Submitted
to the Department of Medicine
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Science

Sang Kyun Yoo

December 2024

**This certifies that the Dissertation of Sang Kyun Yoo is approved**

| | |
|---|---|
| Thesis Supervisor | Jin Sung Kim |
| Thesis Committee Member | Hojin Kim |
| Thesis Committee Member | Yong Bae Kim |
| Thesis Committee Member | Yu Rang Park |
| Thesis Committee Member | Kyu-Hwan Jung |

**The Graduate School**
**Yonsei University**
**December 2024**

# ACKNOWLEDGEMENTS

Throughout the writing of this dissertation, I have received immense support and assistance from many individuals.

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Jin Sung Kim and Professor Hojin Kim. Their patience, generous support, and invaluable guidance were instrumental in shaping my research and supporting me through challenges. Without their unwavering mentorship, this dissertation would not have been possible.

I am deeply grateful to Professor Yong Bae Kim, one of my dissertation committee members, whose expertise in radiation oncology provided invaluable perspectives and helped refine my research topic.

I also extend my heartfelt thanks to my other committee members, Professor Yu Rang Park and Professor Kyu-Hwan Jung, whose insightful feedback and constructive criticism significantly enriched the quality and depth of my research.

I would like to acknowledge Professor Joongyo Lee, Professor Nari Kim, and Professor Jin-Hwa Choi for their contributions as collaborators in this study.

I would also like to extend my sincere gratitude to Professor James J. Sohn. Through the opportunities provided by Professor Sohn, I was able to expand my perspective and gain invaluable experiences in the global academic community.

I am deeply thankful to my colleagues in MPBEL. Please know that I am grateful for all the shared experiences, even if I cannot name each of you individually. I sincerely wish everyone continued success in their academic journey and the best of luck as you complete your studies.

To my parents and younger brother, thank you for your unconditional love and support. Your belief in me sustained me even when I doubted myself. You have been my unwavering source of strength and inspiration.

Lastly, completing this dissertation has been both challenging and rewarding, and I am thankful for the personal growth I have experienced along the way.

To all who have helped me in any way, I extend my heartfelt thanks.

Sang Kyun Yoo

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

## Predicting locoregional recurrence in breast cancer following breast-conserving therapy using learning-based models with multi-institutional registries

**Purpose:** Radiotherapy (RT), alongside surgery, is an essential component that consists of breast-conserving therapy. However, in a small percentage of patients, locoregional recurrence (LRR) may occur, leading to achieving the purpose of treatment. This study aims to develop and validate a machine learning (ML) model that incorporates radiomics features from multi-institutional registries to predict the risk of LRR in breast cancer patients. By utilizing a single magnetic resonance imaging (MRI) sequence (T2-weighted with fat suppression) and identifying the key features associated with risk of LRR, this study seeks to enhance the robustness and clinical applicability of LRR risk predictive models for personalized treatment planning.

**Methods:** A multi-institutional registry of 352 breast cancer patients was retrospectively collected and analyzed. The dataset comprised diagnostic T2-weighted MRI scans with fat suppression, manually delineated primary breast tumors, and clinical factors such as age at diagnosis, tumor size, pathology, and molecular subtypes. The delineation was performed and confirmed by board-certified radiation oncologists at each institution. To address class imbalance, various data sampling methods, including oversampling techniques, were explored and evaluated. Ultimately, a balanced subset was randomly selected to address class imbalance and ensure equal representation of LRR and non-LRR cases during model development. Radiomics features, including shape, first-order statistics, and texture, were extracted from manually contoured regions of interest (ROIs). During feature extraction, the impact of MRI scan normalization on model performance was also assessed. A machine learning model was developed using feature selection techniques and principal component analysis (PCA), with logistic regression as the classifier. A domain adaptation technique was employed to improve model performance. Additionally, a model incorporating both radiomics features and clinical factors known to be associated with the risk of LRR was developed to evaluate

the added predictive value of combining different data types. The model's performance was evaluated using five-fold cross-validation and an independent test dataset, with calibration applied to improve the accuracy of probability estimates.

**Results:** The model achieved the best performance when MRI scan normalization was applied, feature selection was performed using a wrapper method (Recursive Feature Elimination, RFE), and both radiomics features and clinical factors were included as inputs. Under these conditions, the model achieved an average AUC of 0.757 (95% confidence interval, 0.715-0.799) for cross-validation and 0.762 for the independent test dataset.

**Conclusion:** In this study, a predictive model for the risk of LRR in breast cancer patients was developed by integrating radiomics features with clinical factors known to be associated with LRR risk. The findings suggest that radiomics, as a non-invasive biomarker, could contribute to enhancing personalized risk assessment when integrated with clinical factors. To further validate the proposed model's predictive power, prospective datasets should be analyzed in future studies.

---

Key words : locoregional recurrence, breast cancer, breast-conserving therapy, machine learning, radiomics

# 1. INTRODUCTION

## 1.1. Radiotherapy

Radiotherapy (RT) is one of essential modalities in the treatment of a range of cancers, as well as surgery and chemotherapy[1] The fundamental principle of RT is to induce DNA damage within malignant tissues by ionizing radiation, thereby inhibiting the ability of these cells to proliferate and ultimately leading to their elimination.[2,3] To achieve this, RT relies on the precise delivery of ionizing radiation directed to the tumor, minimizing exposure to surrounding normal tissues.

Advancements in RT techniques have significantly enhanced its precision and efficacy. These advancements include techniques such as Intensity-Modulated Radiotherapy (IMRT), Image-Guided Radiotherapy (IGRT), and Stereotactic Body Radiotherapy (SBRT), which enable highly targeted radiation delivery, effectively sparing adjacent normal tissues and minimizing side effects.[4,5] These advancements have enabled RT to effectively manage localized tumors while also addressing the challenges of advanced-stage cancers, improving treatment outcomes across diverse cancer types.

RT is often combined with other treatments to improve overall efficacy. For instance, recent advancements have highlighted the potential of combining RT with immunotherapy, leveraging the immune-activating effects of radiation to enhance systemic tumor control.[6,7] Furthermore, Chemoradiotherapy (CRT), which refers to the combined use of RT and chemotherapy, has demonstrated synergistic effects in controlling cervical, head and neck, and gastrointestinal malignancies, significantly improving survival outcomes compared to either treatment alone.[8,9] In breast cancer, RT is a key component of Breast-Conserving Therapy (BCT), targeting residual cancer cells after surgery to minimize recurrence.[10] These combinations highlight the adaptability of RT, enabling it to work effectively alongside other treatments in comprehensive cancer therapy.

## 1.2. Predictive modeling

Advancements in predictive modeling within radiation oncology have shifted the focus from generalized, population-based methods to patient-centered approaches. Traditionally, models such as the linear-quadratic model have provided foundational insights into radiation-induced biological effects, including dose-response relationships, serving as a foundation for understanding and optimizing treatment outcomes.[11] However, these models were limited in their ability to account for the intricacies of patient-specific characteristics.

Artificial Intelligence (AI) has emerged as a powerful tool for medical imaging, enabling the automation of complex tasks such as segmentation, classification, and outcome prediction that have traditionally been labor-intensive.[12-14] Among these applications, AI has demonstrated significant potential in improving prediction models through its ability to efficiently analyze complex, high-dimensional datasets. One of its key strengths comes from identifying noninvasive patterns within radiological images, providing valuable insights. By leveraging these patterns, AI-based models improve our understanding of patient-specific characteristics, ultimately enhancing more precise and personalized treatment strategies.

## 1.3. Artificial intelligence and radiomics in predictive modeling

AI-based methods, particularly those using Deep Learning (DL), have demonstrated exceptional capabilities in automating tasks such as tumor segmentation[12] from Magnetic Resonance Imaging (MRI) scans and predicting clinical outcomes (Figure 1 (A)).[13] However, deep learning models often require extensive labeled datasets and substantial computational resources, making their application challenging in situations with limited data availability. In contrast, Machine Learning (ML), a subset of AI, is particularly well-suited for handling smaller datasets, making it more feasible in medical imaging applications where labeled data is often limited.[15,16] Unlike DL models, which often lack transparency and provide limited interpretability regarding how they make predictions, ML techniques are generally more interpretable, offering a clearer understanding of the relationships between features and predicted outcomes, which is essential for decision-making.[17-19]

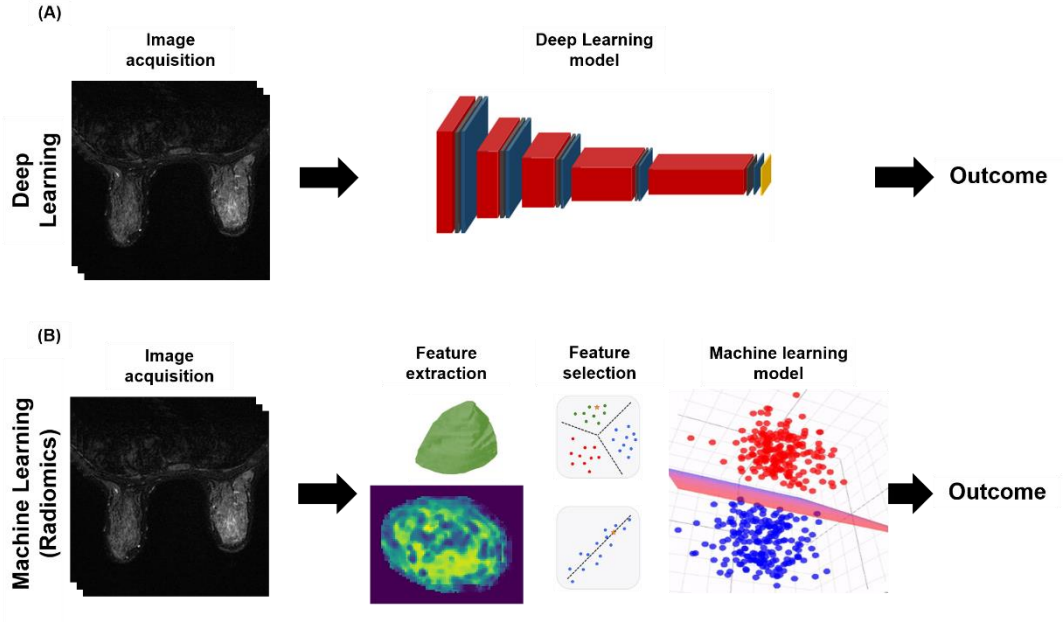


**Figure 1.** Workflow of predictive modeling in medical imaging, highlighting the automated feature extraction in DL (A) compared to the manual feature extraction and selection process in ML (B).

Radiomics has demonstrated that image-based features can capture subtle variations in tumor biology that are not easily discernible through traditional clinical assessments.[20] These radiomics features, extracted from defined Regions of Interest (ROIs) in radiological images like MRI scans, provide quantitative data that describe tumor characteristics such as intensity, shape, and texture information that may not be visible to the human perspective. Radiomics offers a way to convert standard imaging data into mineable information, potentially serving as a non-invasive prognostic biomarker.[21,22] When analyzed using learning-based algorithms, these features have shown significant potential in predicting outcomes such as risk of recurrence. In the predictive modeling process (Figure 1 (B)), radiomics features are first preprocessed to ensure consistent scales, followed by feature selection to retain only the most informative features.[23] The processed features are then used in ML models, like logistic regression, to predict clinical outcomes. This workflow of feature extraction, feature selection, and modeling allows radiomics to be effectively used for outcome prediction in clinical settings, providing a non-invasive method for understanding tumor characteristics.

## 1.4. Locoregional recurrence in breast cancer

Locoregional Recurrence (LRR), which refers to the recurrence of cancer in the originally treated breast, chest wall, or nearby lymph nodes.[24,25] In a prospective study with a median follow-up time of 3.5 years, the risk of LRR was reported as 7.0% after mastectomy and 5.4% following Breast-Conserving Surgery (BCS).[26]

RT is a fundamental treatment for managing breast cancer, which is one of the most commonly diagnosed cancers among women worldwide and has shown a steady increase over the past two decades, particularly in early-stage cases.[27] In Korea, breast cancer is the most common cancer among women, following thyroid cancer, and the incidence of early breast cancer has rapidly risen as a result of advancements in screening programs and increased public awareness.[28] Consequently, BCT, which includes partial mastectomy and RT, has become the standard treatment for early breast cancer since the 1990s, after studies demonstrated comparable outcomes to total mastectomy.[29]

Evidence from the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) meta-analysis further underscores the critical role of RT in reducing LRR.[30] The analysis demonstrated that reducing four LRR cases through RT could prevent one breast cancer-related death, highlighting the importance of minimizing recurrence to improve long-term outcomes. Moreover, LRR often necessitates total mastectomy, making recurrence prevention essential for preserving breast-conserving strategies and improving patient quality of life.

While RT significantly reduces the risk of LRR, certain patient- and tumor-related factors may still increase the likelihood of recurrence. Understanding these factors is crucial in optimizing treatment strategies and developing personalized treatment plans for patients at higher risk of recurrence. A number of factors have been shown to influence the risk of LRR, including patient age, tumor characteristics, and molecular subtypes.[31-35] Age at diagnosis has been found to be a significant factor, with younger patients more likely to experience recurrence, which may be attributed to more aggressive tumor biology in this demographic.[31,32] Tumor size is also crucial, larger tumors are generally associated with higher recurrence rates, as they are correlated with more advanced stages of the disease.[33] Additionally, pathology and nodal involvement are significant

indicators of a tumor's aggressiveness.[34] Molecular subtypes are defined based on hormone receptor status and other biological markers, which influence the risk of LRR.[35] Traditionally, luminal-type breast cancer has been associated with a low risk of LRR, while HER2-positive subtypes historically presented a higher risk.[36] However, with the advent of anti-HER2 therapies, such as trastuzumab, the LRR rates in HER2-positive patients have significantly decreased.[37] On the other hand, triple-negative breast cancer (TNBC) has been identified as having the highest risk of LRR due to its aggressive nature and limited treatment options.[38] Recently, the use of immuno-oncologic agents, such as pembrolizumab, in TNBC has shown a trend toward reduced LRR, reflecting advancements in targeted treatment approaches.[39]

## 1.5. Previous studies and current research goals

Several recent studies have explored the potential of using image-derived features from primary breast tumors to predict clinical outcomes.[40,41] Kim, JH. et al.[40] conducted a study on breast cancer heterogeneity using MRI texture analysis, which demonstrated that texture features, such as lesion heterogeneity, could serve as independent prognostic markers. This study suggested that quantifying tumor heterogeneity through texture analysis offers insights into variations in tumor biology that may not be visible through conventional imaging, ultimately helping to predict survival outcomes. Other studies conducted by Park, H. et al.[41] highlighted the utility of radiomics features for predicting disease-free survival (DFS) in patients with invasive breast cancer. They developed a radiomics nomogram that incorporated texture features and clinicopathological variables. The study demonstrated that radiomics features were highly associated with DFS.

This current study aims to develop and validate a predictive model for the risk of LRR in breast cancer patients. Specific ML techniques, such as domain adaptation, are employed to improve feature alignment and ensure consistent model performance. Additionally, clinical factors were integrated with radiomics features to provide a more comprehensive input for the ML model, enhancing the understanding of LRR risk. These advanced ML algorithms are employed to identify key features associated with the risk of LRR.

# 2. Baseline predictive model development

## 2.1. Introduction

Locoregional recurrence remains a critical challenge in the management of breast cancer. Despite the effectiveness of BCT including radiotherapy, it still fails to prevent LRR in some patients, which negatively impacts overall survival and quality of life. These challenges underline the need for accurate predictive models for risk of LRR in breast cancer patients.

The development of a predictive model for risk of LRR in breast cancer patients represents a significant step forward. Predictive models, particularly those utilizing radiomics features derived from imaging data, appear to hold potential in capturing tumor heterogeneity and identifying factors associated with recurrence.

In our previous study[42], a radiomics-based ML model was developed to predict LRR risk using data from a single registry and multiple MRI sequences. While this model demonstrated potential in predicting LRR in breast cancer patients, it was limited by its reliance on a single registry, reducing generalizability, and by the need for multiple imaging sequences, which increased the complexity of data acquisition in clinical practice. These limitations underscore the need for a model that is both more generalizable and clinically practical.

Expanding on our previous work, this chapter aims to integrate multi-institutional patient registries to develop the baseline predictive model. Radiomics features derived from a single MRI sequence, specifically T2-weighted fat-suppressed images, were utilized to improve and optimize the prediction process.

## 2.2. Multi-institutional patient registries

This study was conducted as part of the Korean Radiation Oncology Group (KROG 22-06) clinical research project, approved by the Korean Society of Radiation Oncology. A total of 455 patients with breast cancer through diagnostic breast MRI scans were initially retrospectively collected. Of these, 352 patients from registries of four different institutions were ultimately included in the analysis after excluding cases for the following reasons: (1) the lack of access to raw MRI scans, which were required for consistent extraction and preprocessing of radiomics features, and (2) incompatible MRI sequences that did not meet the study requirements. Specifically, Institution 1 contributed 114 patients (28 LRR / 86 non-LRR), Institution 2 contributed 100 patients (18 LRR / 82 non-LRR), Institution 3 contributed 88 patients (37 LRR / 51 non-LRR), and Institution 4 contributed 50 patients (1 LRR / 49 non-LRR). The study was approved by the Institutional Review Board of Severance Hospital (4-2024-1225), Yonsei University College of Medicine, Seoul, Korea. The requirement for informed consent was waived due to the retrospective nature of the study.

The entire dataset consisted of 84 LRR and 268 non-LRR cases. As illustrated in Figure 2, two different approaches were utilized to evaluate the model’s predictive performance. In the Figure 2 (A) approach, Institution 2 was designated as an independent test set, while the remaining data from Institutions 1, 3, and 4 were used for model training. This setup allowed for the evaluation of the model's ability to generalize to data from an unseen institution. In the Figure 2 (B) approach, a 5-fold cross-validation was performed using the entire dataset, ensuring that each fold served as a test set while the remaining folds were used for training. This method provided an average evaluation of the model's performance across all data.

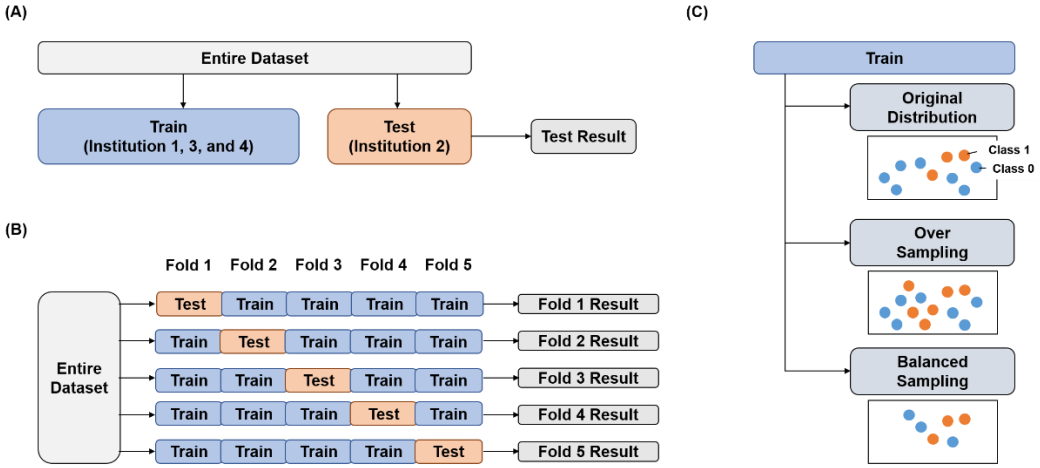


**Figure 2.** Overview of dataset composition and evaluation approaches for predicting the risk of LRR with multi-institutional patient registries.

As shown in Figure 2 (C), different sampling methods were applied to handle the class imbalance, including original distribution, over-sampling, and balanced-sampling techniques. These sampling methods were applied only to the training set in all evaluation approaches and were assessed to determine their impact on the model's predictive performance. For over-sampling, synthetic samples were generated using the SMOTE (Synthetic Minority Oversampling Technique)[43] method to increase the number of LRR cases. As another approach, non-LRR cases were randomly selected to match the number of LRR cases, resulting in a balanced dataset.

The entire dataset included three key types of information, which were diagnostic breast MRI scans utilizing the T2-weighted with fat-suppressed sequence, the manual delineation of ROIs for primary breast tumors, and several clinical factors, including age at diagnosis, tumor size, pathology, and molecular subtypes. The manual delineation was performed by one board-certified radiation oncologist at each institution, and the contours were confirmed by an experienced breast radiation oncologist at the respective institutions.

## 2.3. Radiomics features from multi-institutional data

Radiomics features were extracted from MRI scans to quantify the characteristics of breast tissue relevant to predicting LRR. To account for variations in imaging protocols across different institutions, a standardized preprocessing pipeline was employed. This included normalizing the image intensities to achieve a consistent distribution across all datasets and resampling the images to a uniform voxel spacing of 1x1x1 $mm^3$, thereby ensuring spatial consistency. The impact of preprocessing on model performance was assessed to determine its effect on improving predictive accuracy by comparing the model's performance with and without the preprocessing steps.

Manually contoured ROIs were used to define the areas of the breast from which features were extracted. The ROIs were processed to create binary masks, which were then applied to the normalized and resampled images. From the processed images, a comprehensive set of 107 radiomics features was extracted for each case, including shape, first-order statistics, and texture features. Shape features described the structural characteristics of the ROIs, while first-order statistics provided insights into the overall intensity distribution within the ROIs. Texture features captured the spatial arrangement and intensity patterns of voxels within the ROIs, focusing on the relationships and variations in gray levels, which represent the brightness of the pixels.

## 2.4. Development of baseline model

With the limited number of samples with known outcomes, ML was chosen over deep learning for predicting the risk of LRR in breast cancer patients. DL models require large datasets due to their complexity and risk of overfitting, and often exhibit higher variability in performance when data is limited. In contrast, ML models like logistic regression are computationally efficient, less prone to overfitting, and provide greater interpretability, especially in identifying key features associated with LRR risk, making them more suitable for this study.[15,16]

The ML model for predicting the risk of LRR in breast cancer patients was constructed using radiomics features extracted from MRI scans (Figure 3). The baseline model was developed through a series of steps, including preprocessing, feature normalization, feature selection using the wrapper method[44], dimensionality reduction, and applying a calibration step for the classifier. For dimensionality reduction, Principal Component Analysis (PCA) was applied to capture the most

informative components while transforming the overall set of features into a lower-dimensional representation.[45] PCA works by transforming the original features into a new set of orthogonal components that maximize variance, effectively summarizing the data with fewer dimensions while retaining the most critical information. The processed features were used to train a logistic regression model, which was subsequently calibrated using Platt scaling to improve the reliability of the predicted probabilities.[46] In the process of developing the baseline model, the impact of preprocessing and various data sampling methods, including original distribution, over-sampling, and balanced-sampling, were evaluated to ensure optimal performance.

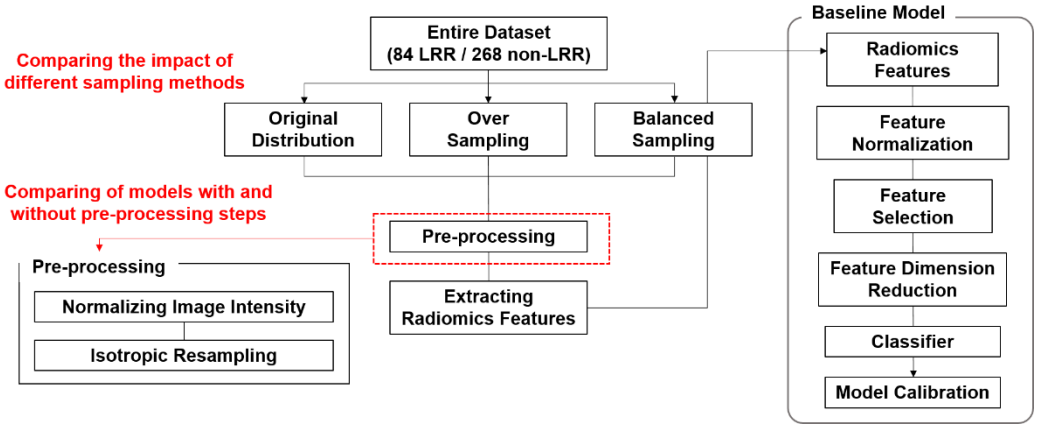


**Figure 3.** The overview for the development of the baseline predictive model, comparing the application of different sampling methods and preprocessing steps.

## 2.5. Evaluation

To evaluate the model's performance, independent testing and 5-fold cross-validation were applied as evaluation strategies. For the independent testing, the training set consisted of 252 cases, including 66 LRR and 186 non-LRR cases, while the independent test set from Institution 2 included 18 LRR and 82 non-LRR cases. For the cross-validation, the entire dataset was divided into five folds using a stratified split to ensure balanced representation of LRR and non-LRR cases across the folds. Under the original distribution, folds 1 and 2 consisted of 17 LRR and 54 non-LRR cases. Fold 3 included 16 LRR and 54 non-LRR cases, while folds 4 and 5 each included 17 LRR and 53 non-LRR cases. For over-sampling, synthetic samples were generated using the SMOTE to increase the number of LRR cases. In this scenario, folds 1 and 2 contained 64 LRR and 64 non-LRR cases, and folds 3 and 4 consisted of 63 LRR and 63 non-LRR cases, ensuring an equal representation of each class within the folds. For balanced sampling, folds 1 and 2 consisted of 17 LRR and 17 non-LRR cases, while folds 3, 4, and 5 each included 16 LRR and 16 non-LRR cases.

The performance of the predictive model was assessed using multiple metrics, including accuracy, sensitivity, specificity, and the Area Under the Receiver Operating Characteristic curve (AUC)[42], as defined in Equations (2) through (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Sensitivity = \frac{TN}{TN + FP} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{5}$$

Where, TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. TPR (True Positive Rate) is the proportion of actual positives correctly

identified by the model, and FPR (False Positive Rate) is the proportion of actual negatives that are incorrectly identified as positives (calculated as 1 - specificity).

The radiomics features were extracted using Pyradiomics v.3.0[47] in Python. The ML models utilized scikit-learn v.0.23.2[48] in Python for domain adaptation, normalization, feature selection, PCA, logistic regression, and calibration. The feature selection process employed an optimal number of features to select, resulting in the selection of 14 to 18 features out of 107. The number of components to be retained for PCA was set to ten.

## 2.6. Results

### 2.6.1. Patient characteristics

Table 1 presents the clinical characteristics of the multi-institutional patient registries included in this study. The four institutions contributed diverse patient populations, with median ages ranging from 48 to 52 years. Invasive ductal carcinoma (IDC) was the predominant pathology across all institutions, accounting for 75.4% to 98% of cases. Ductal carcinoma in situ (DCIS) and other less common pathology constituted a smaller proportion, with Institution 1 reporting the highest percentage of non-IDC cases (14.9%). Regarding tumor size, T1 stage tumors were the most prevalent across all institutions, with the highest proportion observed in Institution 2 (69%) and Institution 4 (66%). In contrast, T3 stage tumors were relatively less frequent, with Institution 1 reporting the largest proportion (5.3%) of these larger, more advanced tumors. Nodal involvement also varied between the institutions. Institution 2 had the highest proportion of patients with N0 status (78%), indicating no regional lymph node involvement, while Institution 1 had the largest percentage of patients with more advanced nodal involvement (N2 and N3 stages combined, 21.1%). The distribution of molecular subtypes also showed variation. Luminal A type, which are generally associated with a more favorable prognosis, were most frequent at Institution 4 (74%), while Institution 1 had the highest proportion of basal-like tumors (55.3%), a subtype associated with more aggressive disease. Luminal B, HER2-enriched, and basal-like subtypes were less common across all institutions, with significant variation between them.

**Table 1.** Baseline characteristics of the multi-institutional patient registries

| Characteristics | Institution 1 (N=114) | Institution 2 (N=100) | Institution 3 (N=88) | Institution 4 (N=50) |
|---|---|---|---|---|
| Age (years, median [range]) | 49 (23-75) | 52 (21-86) | 48 (44-69) | 50 (41-76) |
| Pathology, n (%) | | | | |
| IDC | 86 (75.4) | 84 (84) | 82 (93.2) | 49 (98) |
| DCIS | 11 (9.7) | 3 (3) | 1 (1.1) | 0 (0) |
| Others | 17 (14.9) | 13 (13) | 5 (5.7) | 1 (2) |
| T stage, n (%) | | | | |
| Tis | 25 (21.9) | 7 (7) | 6 (6.8) | 0 (0) |
| T1 | 51 (44.7) | 69 (69) | 57 (64.8) | 33 (66) |
| T2 | 32 (28.1) | 23 (23) | 22 (25) | 16 (32) |
| T3 | 6 (5.3) | 1 (1) | 3 (3.4) | 1 (2) |
| N stage, n (%) | | | | |
| N0 | 66 (57.9) | 78 (78) | 48 (54.5) | 34 (68) |
| N1 | 24 (21.0) | 18 (18) | 22 (25) | 13 (26) |
| N2 | 14 (12.3) | 3 (3) | 10 (11.4) | 3 (6) |
| N3 | 10 (8.8) | 1 (1) | 8 (9.1) | 0 (0) |
| Luminal type, n (%) | | | | |
| A | 15 (13.2) | 39 (39) | 32 (36.3) | 37 (74) |
| B | 16 (14.0) | 31 (31) | 22 (25) | 7 (14) |
| HER2-enriched | 20 (17.5) | 11 (11) | 13 (14.8) | 2 (4) |
| Basal-like | 63 (55.3) | 19 (19) | 21 (23.9) | 4 (8) |

***Abbreviations:*** LRR, locoregional recurrence; SMD, standardized mean difference; IDC, invasive ductal carcinoma; DCIS, ductal carcinoma in situ; HER2, human epidermal growth factor receptor 2.

### 2.6.2. Impact of preprocessing during radiomics feature extraction

Table 2 lists the comparison of the predictive model with and without preprocessing during radiomics feature extraction. The model with preprocessing during radiomics feature extraction showed a significant improvement in performance compared to the model without preprocessing. Specifically, the model with preprocessing achieved an average cross-validation accuracy of 71.8% ± 4.4% (95% CI: 65.9-77.7) and an AUC of 0.705 ± 0.055 (95% CI: 0.631-0.779). For the independent test dataset, the model achieved an accuracy of 73.0% and an AUC of 0.709. In contrast, the model without preprocessing achieved a lower average cross-validation accuracy of 68.1% ± 3.7% (95% CI: 66.2-70.6) and an AUC of 0.679 ± 0.056 (95% CI: 0.603-0.755). For the independent test dataset, the model without preprocessing achieved an accuracy of 78.0% and an AUC of 0.663. Its sensitivity, however, was extremely low at only 16.7%, indicating that the model performed poorly in identifying positive cases effectively.

**Table 2.** Baseline model evaluation: impact of preprocessing during radiomics feature extraction. The highest AUC for each evaluation method is highlighted in bold.

| Model | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Without processing | CV | 68.1±3.7 (66.2-70.6) | 63.5±4.4 (57.6-69.4) | 70.0±1.1 (68.5-71.5) | 0.679±0.056 (0.603-0.755) |
| | IND | 78.0 | 16.7 | 91.5 | 0.663 |
| With processing | CV | 71.8±2.3 (68.6-74.9) | 71.8±4.4 (65.9-77.7) | 71.7±2.0 (69.1-74.3) | **0.705±0.055 (0.631-0.779)** |
| | IND | 73.0 | 55.6 | 76.8 | **0.709** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

### 2.6.3. Impact of data sampling methods

Table 3 demonstrates the impact of different data sampling methods, including original distribution, over-sampling, and balanced-sampling, on model performance in predicting the risk of LRR. When using original distribution, the model achieved an average cross-validation accuracy of 74.8% ± 4.6% (95% CI: 68.6-81.0), with a sensitivity of only 25.3% ± 8.7% (95% CI: 13.6-37.0) and a specificity of 93.0% ± 3.7% (95% CI: 87.9-98.0). The AUC for original distribution across cross-validation was 0.642 ± 0.029 (95% CI: 0.603-0.682). The application of over-sampling improved sensitivity to 52.9% ± 3.2% (95% CI: 48.6-57.3), but resulted in a decrease in specificity to 70.5% ± 4.7% (95% CI: 64.2-76.8). The AUC in this case was 0.687 ± 0.003 (95% CI: 0.647-0.727). Balanced-sampling provided the most balanced performance among the three methods, achieving an average cross-validation accuracy of 71.8% ± 2.3% (95% CI: 68.6-74.9), sensitivity of 71.8% ± 4.4% (95% CI: 65.9-77.7), and specificity of 71.7% ± 2.0% (95% CI: 69.1-74.3). The AUC for under-sampling was 0.705 ± 0.055 (95% CI: 0.631-0.779). For the independent test dataset, balanced-sampling resulted in the highest AUC of 0.709.

**Table 3.** Baseline model evaluation: impact of data sampling methods. The highest AUC for each evaluation method is highlighted in bold.

| Model | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Original distribution | CV | 74.8±4.6 (68.6-81.0) | 25.3±8.7 (13.6-37.0) | 93.0±3.7 (87.9-98.0) | 0.642±0.029 (0.603-0.682) |
| | IND | 81.0 | 11.1 | 97.6 | 0.633 |
| Over-sampling | CV | 65.8±3.7 (60.8-70.8) | 52.9±3.2 (48.6-57.3) | 70.5±4.7 (64.2-76.8) | 0.687±0.003 (0.647-0.727) |
| | IND | 74.0 | 44.4 | 80.5 | 0.679 |
| Balanced-sampling | CV | 71.8±2.3 (68.6-74.9) | 71.8±4.4 (65.9-77.7) | 71.7±2.0 (69.1-74.3) | **0.705±0.055 (0.631-0.779)** |
| | IND | 73.0 | 55.6 | 76.8 | **0.709** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

***Note:*** This model includes preprocessing during radiomics feature extraction.

## 2.7. Discussion and conclusion

The purpose of this study was to develop and validate a predictive model for the risk of LRR in breast cancer patients using radiomics features. Building on our previous work[42], which relied on a single-institution registry and required multiple MRI sequences to predict LRR, the current study constructed a predictive model that leverages radiomics features extracted from a single MRI sequence, specifically T2-weighted images with fat suppression, and incorporates multi-institutional patient registries. In this chapter, the aim is to develop a baseline predictive model that addresses challenges such as multi-institutional variability and class imbalance, providing a foundation for further advancements in predicting LRR risk.

In developing the baseline predictive model for predicting the risk of LRR, several considerations were made to optimize the model's predictive performance. One key aspect was preprocessing during the extraction of radiomics features. This step was essential to mitigate variability resulting from differing imaging protocols, such as variations in image resolution and intensity across institutions, thereby ensuring that the radiomics features remained consistent and comparable across all datasets. Another key consideration was comparing different data sampling techniques to address the class imbalance between LRR and non-LRR cases. Sampling approaches, including balanced sampling, over-sampling using SMOTE, and retaining the original distribution, were employed to assess their impact on the model’s predictive performance. Ultimately, the balanced sampling approach was selected, as it achieved an average AUC of 0.705 for the five-fold cross-validation and an AUC of 0.709 for the independent test dataset. In contrast, the original distribution method yielded lower average AUCs of 0.642 and 0.633, while the over-sampling method achieved 0.687 and 0.679 for the five-fold cross-validation and independent test dataset, respectively. The original distribution, which retained the inherent class imbalance, likely resulted in the model being biased towards the majority class, leading to lower sensitivity in predicting LRR cases. Meanwhile, the over-sampling method using SMOTE demonstrated inherent limitations that may have contributed to its suboptimal performance. First, SMOTE generates synthetic samples by interpolating between minority class samples, which can increase existing noise in the original data and reduce the model's reliability. Second, when the number of minority samples is limited, SMOTE can lead to overfitting. The generated synthetic data may struggle to capture the intricate underlying data distribution, thereby limiting the model’s ability to generalize effectively to unseen datasets.

Despite the development of a foundational baseline predictive model in this chapter, its predictive performance remains suboptimal, highlighting the need for further enhancement. To improve the suboptimal performance of the baseline predictive model, it is essential to concentrate not only on enhancing its performance but also on identifying the key features contributing to the model's predictive outcomes.

In conclusion, this chapter establishes a foundational baseline predictive model for LRR risk in breast cancer patients. By taking into account addressing key challenges such as multi-institutional variability and class imbalance, the model establishes a foundation for future improvements.

# 3. Enhancement strategies for predictive models

## 3.1. Introduction

Radiotherapy, which effectively reduces recurrence by damaging cancer cells and improving patient outcomes, is a fundamental component of BCT alongside surgery.[27,29] However, some patients may experience LRR, involving the recurrence of cancer in treated regions, which undermines the effectiveness of BCT.[24,25]

Accurate prediction of LRR in breast cancer patients is essential for enabling personalized treatment strategies that improve clinical outcomes. However, prediction models often encounter challenges such as multi-institutional variability, class imbalance, and the limited ability of features to capture comprehensive representations. The baseline predictive model developed in Chapter 2 provides a fundamental framework for risk prediction and demonstrates an initial attempt to address these challenges. However, its suboptimal performance highlights the need for further enhancement.

Enhancing predictive models requires both performance improvement and the identification of key features contributing to model decisions. Such enhancements are critical for increasing the interpretability and clinical relevance of the models, as it provides valuable insights into the clinical determinants of LRR risk and supports interpretability in clinical applications. Furthermore, predictive models must produce reliable probability estimates to be effectively integrated into clinical decision-making.

This chapter introduces and evaluates several enhancement strategies aimed at improving the robustness, performance, and interpretability of LRR prediction models. Through these strategies, this chapter focuses on enhancing the baseline model, making it more accurate and generalizable across diverse datasets. These improvements are important for enabling predictive models to be effectively applied in clinical practice.

## 3.2. Multi-institutional patient registries

The multi-institutional patient registries detailed in Section 2.2 were utilized not only for the baseline predictive model but also for the development and evaluation of enhanced predictive models. As previously described, the dataset comprised 84 LRR cases and 268 non-LRR cases from four institutions, providing a diverse and comprehensive representation of populations.

During the development of the baseline model, various conditions were compared and evaluated to address data diversity and class imbalance. The most effective approach was implemented in the final baseline model. For this chapter, the training set adjusted through balanced sampling to mitigate class imbalance was utilized without further modification.

## 3.3. Radiomics features and clinical factors for model development

Radiomics features extracted from multi-institutional MRI scans, as described in Section 2.3, were utilized for the development of enhanced predictive models. To ensure consistency across datasets from different institutions, a standardized preprocessing pipeline, including intensity normalization and resampling to a uniform voxel spacing of 1x1x1 $mm^3$, was implemented during the development of the final baseline model. This preprocessing approach was applied in this chapter without further modification.

To provide a more comprehensive input for the ML model and enhance the understanding of LRR risk, clinical factors were integrated with radiomics features. These clinical factors, including age at diagnosis, tumor size, pathology, and molecular subtypes, were extracted from multi-institutional medical records. They were selected based on evidence from previous studies showing their significant role in LRR risk.[31-35]

## 3.4. Model enhancement strategies

As illustrated in Figure 4, the evaluation focused on assessing the effectiveness of several enhancement strategies to improving the predictive performance of the model for LRR risk. This included comparisons between models with and without domain adaptation, analyzing different feature selection methods to identify the most effective approach, and evaluating the impact of integrating clinical factors. Additionally, the role of model calibration in improving the reliability of probability estimates and performance was assessed.

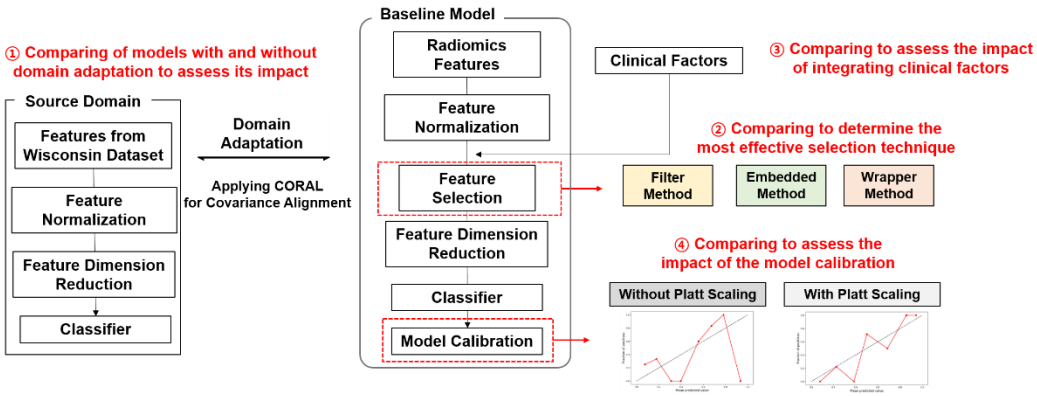


**Figure 4.** Several enhancement strategies were implemented and assessed against the baseline model, including domain adaptation, feature selection techniques, the integration of clinical factors, and model calibration.

### 3.4.1. Domain adaptation

Domain adaptation was employed to leverage the predictive capabilities of a well-trained source model to improve performance on the target dataset, which consisted of radiomics features extracted from MRI scans of breast cancer patients. The adaptation process involved aligning the source model to the target domain using a domain adaptation technique called Correlation Alignment (CORAL).[49] CORAL aims to minimize the domain shift by aligning the covariance between the source and target datasets. By reducing differences in data distribution, CORAL contributes ensure that the trained features from the source domain are more effectively transferred to the target domain.

The source model was developed using the Breast Cancer Wisconsin dataset, which collected to classify tumors as either benign or malignant.[50] This dataset comprises 569 instances, each with 30 features computed from a digitized image of a fine needle aspirate of a breast mass, including various clinical features that are biologically relevant and similar to the radiomics features used in this study. Due to these similarities, it was expected that knowledge gained from the source model could effectively enhance the predictive performance for the target dataset. The impact of domain adaptation on model performance was evaluated by comparing the performance of models developed with and without domain adaptation.

### 3.4.1. Feature selection

Feature selection is a crucial step in ML that contributes to reduce model complexity, improve computational efficiency, and enhance predictive performance by selecting relevant features. There are three common types of feature selection approaches: filter, embedded, and wrapper methods.[51] Filter methods rank features based on statistical metrics independent of the model, providing a fast way to eliminate irrelevant features. Embedded methods, on the other hand, select features during the model training process, automatically identifying important features. Wrapper methods use a model to evaluate and iteratively remove subsets of features that contribute the least to performance. In this study, these three feature selection approaches were used to determine the optimal feature set for predicting the risk of LRR in breast cancer patients.

The filter method, specifically employing the Select K Best algorithm, utilized to identify features most correlated with the risk of LRR.[52] Select K Best ranks all features based on Analysis of Variance (ANOVA) F-statistics and selects the top K features that have the highest correlation with the outcome.

The embedded method employed Least Absolute Shrinkage and Selection Operator (LASSO) to identify the most predictive features based on the inherent feature importance scores of the model.[53] LASSO incorporates L1 regularization into the logistic regression, which adds a penalty equivalent to the absolute value of the magnitude of the coefficients. This penalty causes less important feature coefficients to shrink to zero during training, effectively performing feature selection by removing irrelevant features and retaining only those that contribute the most to the prediction.

The wrapper method applied Recursive Feature Elimination (RFE) with logistic regression as the estimator.[44] RFE works by recursively fitting the model and eliminating the least important features, based on the weights of the logistic regression, in each iteration. This process continues until the optimal subset of features is identified, ensuring that only the most predictive features are retained.

The baseline model with incorporating domain adaptation was trained separately with features selected by each of these methods and subsequently compared to identify the most effective feature selection technique for predicting the risk of LRR.

### 3.4.3. Integration of clinical factors

To effectively integrate clinical factors with radiomics features for predictive modeling, proper normalization was required. Clinical factors and radiomics features were normalized independently due to their differing data characteristics. Radiomics features, derived from imaging data, are continuous, while clinical factors include both continuous and categorical data. Continuous factors, such as age at diagnosis and tumor size, were normalized using a Standard Scaler[54], while categorical factors, such as pathology and molecular subtypes, were transformed using Label Encoding[55] to convert them into a suitable format.

The normalized clinical factors and radiomics features were concatenated to form a unified feature set, which was then subjected to a feature selection algorithm to predict the risk of LRR. After comparing three feature selection methods—filter, embedded, and wrapper approaches—the most effective technique was determined and subsequently applied to the unified feature set. The impact of integrating clinical factors with radiomics features on model performance was evaluated by comparing the performance of models developed with and without the integration, using the baseline model that incorporated domain adaptation and the effective feature selection technique.

### 3.4.4. Model calibration

To improve the reliability of probability estimates from the model, the baseline model in Chapter 2 included a calibration process using Platt scaling.[46] This calibration process involved fitting a sigmoid function to the model's predicted probabilities, effectively transforming the uncalibrated outputs into calibrated probability estimates. Specifically, the Platt scaling process works by training an additional logistic regression model, where the inputs are the raw predicted scores from the original model, and the outputs are the true binary outcomes. This logistic regression fits a sigmoid curve to the predicted values, effectively adjusting the raw scores to fall within a probability range that more accurately reflects the actual likelihood of the risk of LRR. The sigmoid function maps the predicted values onto a [0, 1] scale, making the output more interpretable as a probability of the risk of LRR.

In addition to Platt scaling, other calibration methods such as Isotonic Regression[56] were considered. Isotonic Regression is a non-parametric calibration method that is more flexible compared to Platt scaling, which assumes a sigmoid relationship. However, due to the relatively small sample size and the potential risk of overfitting, Platt scaling was chosen for its simplicity and robustness. Isotonic Regression can be prone to overfitting, particularly when dealing with smaller datasets, as it tries to fit the calibration curve as closely as possible to the given data. This can lead to a model that captures noise rather than general patterns. In contrast, Platt scaling applies a parametric approach with fewer degrees of freedom, which prevents overfitting by avoiding overly complex calibration curves, making it more suitable for smaller sample sizes.

## 3.5. Evaluation and identification of key features

To evaluate the model's performance, independent testing and five-fold cross-validation were applied as detailed in Chapter 2. Metrics such as accuracy, sensitivity, specificity, and the AUC were utilized to assess the performance of the predictive models. Additionally, Expected Calibration Error (ECE)[57] was also calculated to assess the calibration performance of the predictive models. ECE quantifies the difference between the predicted probabilities and the observed outcomes, providing an indication of how well the predicted probabilities align with the actual likelihoods of positive outcomes. ECE is calculated by dividing the predicted probabilities into 10 bins, calculating the average predicted probability and the fraction of positives within each bin, and then taking a

weighted average of the absolute differences between these two values. A lower ECE indicates better calibration.

To identify the key features contributing to the model's predictive outcomes, a frequency analysis and coefficient analysis were performed. The frequency analysis was conducted to determine how often each feature is selected across cross-validation model. For the coefficient analysis, the ML model generates a coefficient map for both cross-validation and the independent test, which quantifies the contribution of each feature by analyzing the coefficient values. In the case of cross-validation, the coefficient values from the five folds were averaged to obtain a representative value for each feature.

The process of calculating coefficients for the original features involves transforming the coefficients obtained from the logistic regression model, which was trained on principal components derived through PCA, back into the original feature space. To map these coefficients back to the original feature space, we apply the following transformation:

$$w_{orig} = cP^T \quad (1)$$

Where P represents the matrix of eigenvectors corresponding to the principal components. The coefficients C obtained from the logistic regression indicate the importance of these principal components. Through this transformation, the resulting $w_{orig}$ are mapped back to the selected features.

## 3.6. Results

### 3.6.1. Impact of domain adaptation

Table 4 presents the performance of incorporating domain adaptation on the baseline model. Without domain adaptation, the model achieved an average cross-validation accuracy of 71.8% ± 2.3% (95% CI: 68.6-74.9) and an AUC of 0.705 ± 0.055 (95% CI: 0.631-0.779). For the independent test dataset, the accuracy and AUC were 73.0% and 0.709, respectively. Incorporating domain adaptation led to improvements, with a cross-validation accuracy of 73.3% ± 3.5% (95% CI: 68.7-78.0) and an AUC of 0.737 ± 0.035 (95% CI: 0.691-0.784). For the independent test dataset, the accuracy was 69.0%, and the AUC increased to 0.734.

To further demonstrate the effect of domain adaptation using CORAL, covariance alignment was assessed between the source and target datasets before and after applying CORAL. Before applying CORAL, the average covariance values across five folds were calculated as 7.229 ± 1.199. Specifically, with significant inconsistencies in Fold 1 (8.715) and Fold 3 (8.227). After applying CORAL, the average covariance across all folds was reduced to 0.868 ± 0.093. The independent dataset also showed a significant reduction, with the covariance value decreasing from 7.937 to 0.835.

**Table 4.** Impact of domain adaptation on model performance in predicting the risk of LRR. The highest AUC for each evaluation method is highlighted in bold.

| Models | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Without adaptation | CV | 71.8±2.3 (68.6-74.9) | 71.8±4.4 (65.9-77.7) | 71.7±2.0 (69.1-74.3) | 0.705±0.055 (0.631-0.779) |
| | IND | 73.0 | 55.6 | 76.8 | 0.709 |
| With adaptation | CV | 73.3±3.5 (68.7-78.0) | 72.6±2.6 (69.2-76.1) | 73.5±4.5 (67.5-79.6) | **0.737±0.035 (0.691-0.784)** |
| | IND | 69.0 | 66.7 | 69.5 | **0.734** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

### 3.6.2. Comparison of different feature selection techniques

Table 5 demonstrates the impact of different feature selection techniques, including filter, embedded, and wrapper methods, on the baseline model incorporating domain adaptation. The wrapper method, specifically RFE, demonstrated the best performance among the three feature selection methods.

For the cross-validation results, the model with the wrapper method achieved an average accuracy of 73.3% ± 3.5% (95% CI: 68.7-78.0) and an AUC of 0.737 ± 0.035 (95% CI: 0.691-0.784). For the independent test dataset, the wrapper method achieved the highest AUC of 0.734, although its accuracy of 69.0% was not the highest among the three methods.

The filter method, using the Select K Best approach, achieved an average cross-validation accuracy of 66.9% ± 4.1% (95% CI: 61.5-72.4) and an AUC of 0.684 ± 0.032 (95% CI: 0.642-0.727), while the embedded method, utilizing LASSO, achieved an average cross-validation accuracy of 64.4% ± 2.6% (95% CI: 64.0-70.9) and an AUC of 0.666 ± 0.020 (95% CI: 0.639-0.693). For the independent test dataset, the filter method achieved an accuracy of 72.0% and an AUC of 0.673, while the embedded method produced an accuracy of 78.0% and an AUC of 0.692. However, both methods demonstrated significantly lower sensitivity compared to the wrapper method and resulted in lower AUC.

**Table 5.** Comparison of different feature selection techniques in predicting the risk of LRR. The highest AUC for each evaluation method is highlighted in bold.

| Models | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Filter | CV | 66.9±4.1 (61.5-72.4) | 58.8±10.5 (44.7-72.9) | 70.0±7.6 (59.8-80.2) | 0.684±0.032 (0.642-0.727) |
| | IND | 72.0 | 38.9 | 79.3 | 0.673 |
| Embedded | CV | 67.4±2.6 (64.0-70.9) | 55.3±8.0 (44.6-66.0) | 71.9±5.4 (64.6-79.1) | 0.666±0.020 (0.639-0.693) |
| | IND | 78.0 | 50.4 | 84.1 | 0.692 |
| Wrapper | CV | 73.3±3.5 (68.7-78.0) | 72.6±2.6 (69.2-76.1) | 73.5±4.5 (67.5-79.6) | **0.737±0.035 (0.691-0.784)** |
| | IND | 69.0 | 66.7 | 69.5 | **0.734** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

***Note:*** All experiments in this table were conducted on the baseline model incorporating domain adaptation.

### 3.6.3. Impact of integration of clinical factors

Table 6 compares the performance of models trained using radiomics features only with those incorporating both radiomics and clinical factors for predicting the risk of LRR. The baseline model incorporating domain adaptation with RFE process using only radiomics features achieved an average cross-validation accuracy of 73.3% ± 3.5% (95% CI: 68.7-78.0) and an AUC of 0.737 ± 0.035 (95% CI: 0.691-0.784). For the independent test dataset, the radiomics-only model achieved an accuracy of 69.0% and an AUC of 0.734.

When clinical factors were added to the radiomics features, the model showed a slight improvement in predictive performance. The integrated model achieved an average cross-validation accuracy of 75.6% ± 2.9% (95% CI: 71.7-79.4) and an AUC of 0.757 ± 0.031 (95% CI: 0.715-0.799). For the independent test dataset, the integrated model demonstrated an accuracy of 77.0%, with an AUC of 0.762, which was slightly higher than the radiomics-only model.

**Table 6.** The numerical performance of radiomics only model and those integrating both radiomics and clinical factors for predicting the risk of LRR. The highest AUC for each evaluation method is highlighted in bold.

| Models | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Radiomics Only | CV | 73.3±3.5 (68.7-78.0) | 72.6±2.6 (69.2-76.1) | 73.5±4.5 (67.5-79.6) | 0.737±0.035 (0.691-0.784) |
| | IND | 69.0 | 66.7 | 69.5 | 0.734 |
| Radiomics + Clinical factors | CV | 75.6±2.9 (71.7-79.4) | 73.8±2.7 (70.2-77.4) | 76.1±3.1 (72.0-80.3) | **0.757±0.031 (0.715-0.799)** |
| | IND | 77.0 | 66.7 | 79.3 | **0.762** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

***Note:*** All experiments in this table were conducted on the baseline model incorporating domain adaptation with RFE process.

Figure 5 and Figure 6 illustrate the calibration curves for the models trained with radiomics features only (Figure 5) and those integrating both radiomics and clinical factors (Figure 6), respectively. For the cross-validation results, the radiomics-only model showed substantial variability among different folds (ECE values ranging from 0.067 to 0.150), especially at higher predicted probability ranges. In contrast, the model integrating both radiomics and clinical factors exhibited better calibration consistency across all folds (ECE values ranging from 0.051 to 0.109), suggesting a more robust performance across different subsets of data. For the independent test dataset, the radiomics-only model demonstrated better overall calibration, as indicated by a lower ECE value compared to the model integrating both radiomics and clinical factors. However, the radiomics-only model exhibited slight miscalibration in the predicted probability range of 0.5 to 0.8, where it showed to overestimate the actual risk of LRR. In contrast, the model integrating both radiomics and clinical factors, despite having a slightly higher overall ECE, showed improved calibration in the higher predicted probability range (0.8–1.0).

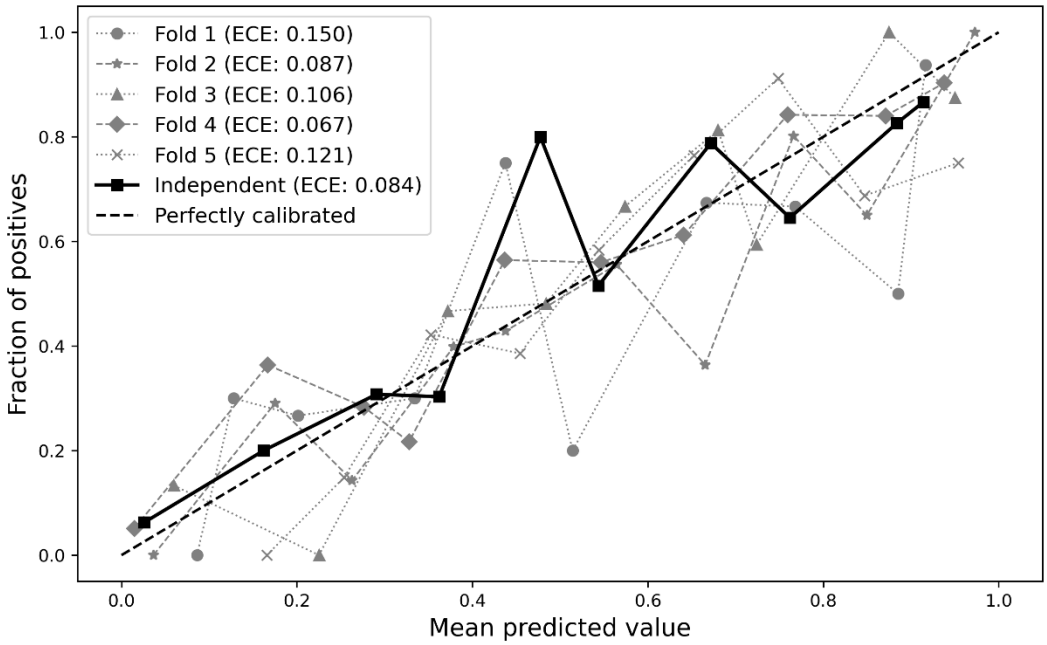


**Figure 5.** Calibration curves for the predictive model using radiomics features only, evaluated with cross-validation and independent test datasets. The dashed 45-degree line represents perfect calibration, where predicted probabilities match observed outcomes. Calibration curves for the five cross-validation folds (Fold 1-5) are shown with different markers (circle, star, triangle, diamond, and cross), while the solid black line with square markers represents the calibration curve for the independent test dataset.

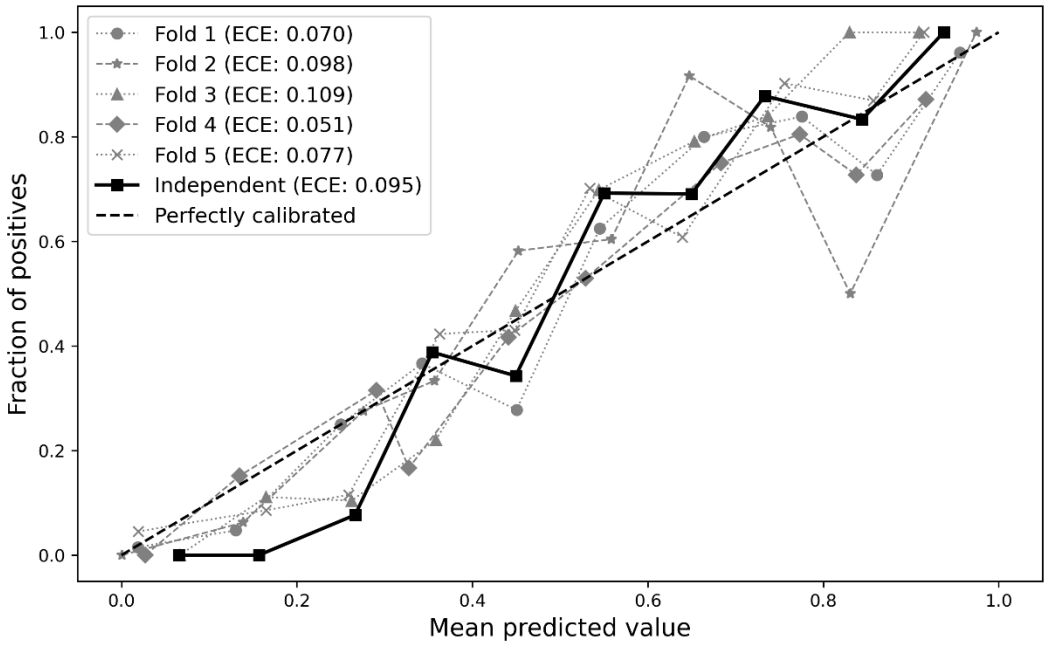


**Figure 6.** Calibration curves for the predictive model integrating both radiomics features and clinical factors, evaluated using cross-validation and independent test datasets. Calibration curves for the five cross-validation folds (Fold 1-5) are shown with different markers (circle, star, triangle, diamond, and cross), while the solid black line with square markers represents the calibration curve for the independent test dataset.

### 3.6.4. Impact of model calibration

Table 7 presents the performance of the models trained with radiomics features only and those integrating radiomics with clinical factors, without calibration for predicting the risk of locoregional recurrence (LRR). These results were compared against the models with calibration, as presented in Table 6, to highlight the impact of model calibration on predictive performance.

For the radiomics-only model without calibration, the average cross-validation accuracy was 71.0% ± 1.5% (95% CI: 69.1-73.0), with an AUC of 0.725 ± 0.045 (95% CI: 0.664-0.786). In comparison, the calibrated radiomics-only model (Table 6) achieved a higher accuracy of 73.3% ± 3.5% (95% CI: 68.7-78.0) and an AUC of 0.737 ± 0.035 (95% CI: 0.691-0.784) during cross-validation. For the independent test dataset, the uncalibrated radiomics-only model achieved an accuracy of 77.0% and an AUC of 0.680, which were lower compared to the calibrated model's performance of 69.0% accuracy and an AUC of 0.734.

For the integrated model that integrated both radiomics and clinical factors, calibration positively impacted model performance. The uncalibrated model's average cross-validation accuracy was 74.7% ± 3.2% (95% CI: 70.4-79.0), with an AUC of 0.734 ± 0.050 (95% CI: 0.667-0.801). In comparison, the calibrated integrated model (Table 6) achieved a cross-validation accuracy of 75.6% ± 2.9% (95% CI: 71.7-79.4) and an AUC of 0.757 ± 0.031 (95% CI: 0.715-0.799). For the independent test dataset, the uncalibrated integrated model had an accuracy of 78.0% and an AUC of 0.698, whereas the calibrated model improved to an accuracy of 77.0% and an AUC of 0.762.

The p-values for the differences in AUC between the models with and without calibration indicate no statistically significant differences under cross-validation ($p = 0.381$ for the radiomics-only model and $p = 0.361$ for the integrated model). Although the p-values did not indicate statistical significance, calibration appeared to slightly improve the AUC and accuracy for both models.

**Table 7.** The numerical performance of radiomics only model those integrating both radiomics and clinical factors without calibration for predicting the risk of LRR. The highest AUC for each evaluation method is highlighted in bold.

| Models without calibration | | Accuracy [%] (95% CI) | Sensitivity [%] (95% CI) | Specificity [%] (95% CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| Radiomics Only | CV | 71.0±1.5 (69.1-73.0) | 59.5±4.7 (53.2-65.7) | 74.6±1.6 (72.5-76.8) | 0.725±0.045 (0.664-0.786)[1] |
| | IND | 77.0 | 38.9 | 85.4 | 0.680 |
| Radiomics + Clinical factors | CV | 74.7±3.2 (70.4-79.0) | 61.9±7.9 (51.3-72.5) | 78.7±2.5 (75.3-82.1) | **0.734±0.050 (0.667-0.801)[2]** |
| | IND | 78.0 | 44.4 | 85.4 | **0.698** |

***Abbreviations:*** 95% CI, 95% confidence interval; AUC, area under the receiver operating characteristic curve; CV, cross-validation; IND, independent test

***Note:*** All experiments in this table were conducted on the baseline model incorporating domain adaptation with the RFE process, but without calibration.

[1, 2] p-values (0.381 and 0.361, respectively) represent the statistical significance of the differences in AUC between the models with calibration and those without calibration.

Figure 7 and Figure 8 illustrate the calibration curves for the models trained with radiomics features only and those integrating both radiomics and clinical factors, respectively, for uncalibrated models. Compared to the calibrated models depicted in Figure 5 and Figure 6, the uncalibrated models show greater deviations from the perfectly calibrated line.

For the uncalibrated radiomics-only model (Figure 7), ECE values ranged from 0.094 to 0.227 across cross-validation folds, with 0.218 for the independent test dataset. For the uncalibrated model integrating both radiomics and clinical factors (Figure 8), ECE values ranged from 0.095 to 0.223 across folds and 0.194 for the independent test dataset. These values indicate poorer calibration compared to the calibrated models shown in Figure 5 and Figure 6, reflecting greater deviations from the perfectly calibrated line.

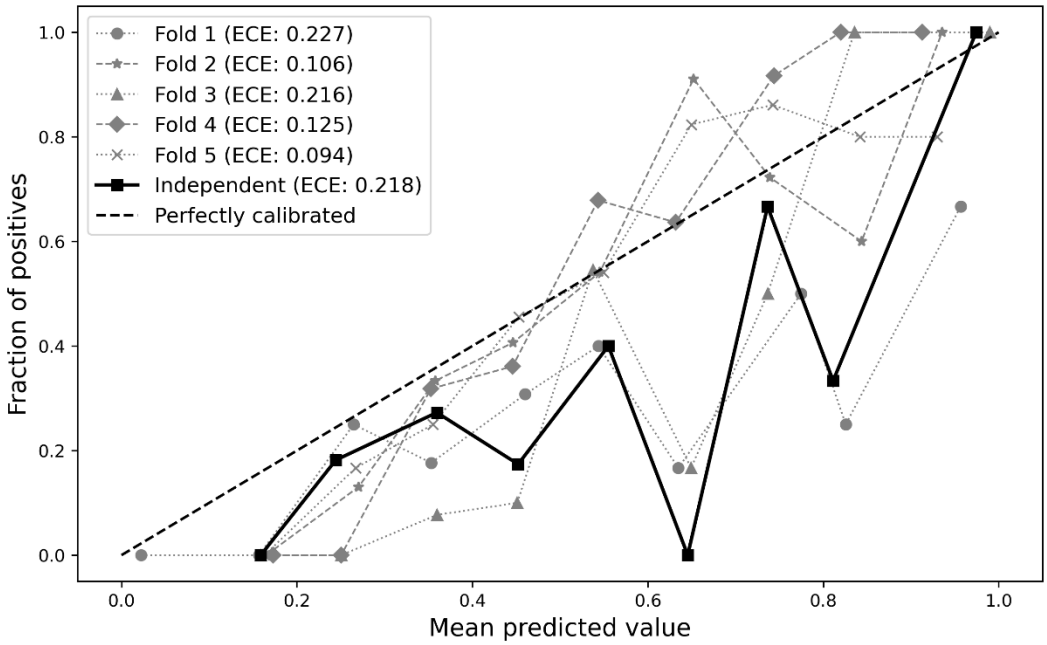


**Figure 7.** Calibration curves for the predictive model using radiomics features only, evaluated using cross-validation and independent test datasets, without calibration applied. Calibration curves for the five cross-validation folds (Fold 1-5) are shown with different markers (circle, star, triangle, diamond, and cross), while the solid black line with square markers represents the calibration curve for the independent test dataset.

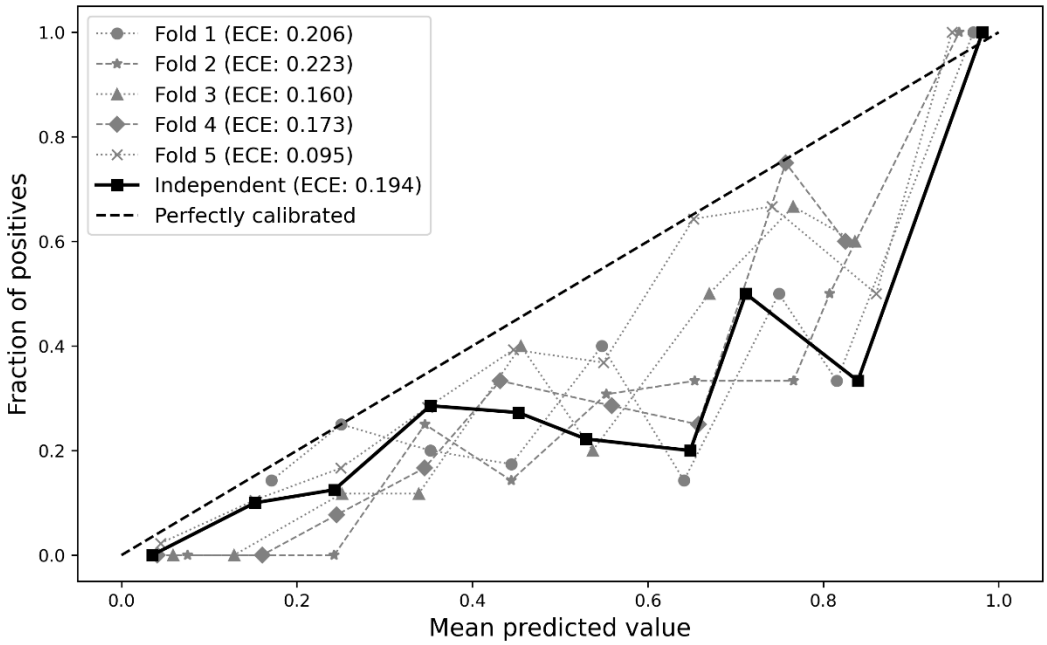


**Figure 8.** Calibration curves for the predictive model integrating both radiomics features and clinical factors, evaluated using cross-validation and independent test datasets, without calibration applied. Calibration curves for the five cross-validation folds (Fold 1-5) are shown with different markers (circle, star, triangle, diamond, and cross), while the solid black line with square markers represents the calibration curve for the independent test dataset.

### 3.6.5. Key features in model decision

Table 8 lists the radiomics features selected through the RFE process for the baseline models incorporating domain adaptation. The table presents information for both the model using only radiomics features and the model integrating both clinical factors and radiomics features. Across each of the five folds, only features selected at least twice were considered key contributors, with 14 and 18 features identified as key contributors in the decision-making process for both models. For the model using only radiomics features, the selected features were categorized into three primary groups: 2 shape-based features, 14 texture-based features, and 2 first-order statistical features. In the independent test dataset, 16 features were identified as key contributors, categorized into 4 shape-based features, 11 texture-based features, and 1 first-order statistical features. In the model integrating both clinical factors and radiomics features, the selected features were categorized into four primary groups: 4 shape-based features, 12 texture-based features, 1 first-order statistical features, and 2 clinical factors. In the independent test dataset, 16 features were identified as key contributors, categorized into 3 shape-based features, 9 texture-based features, 2 first-order statistical features, and 2 clinical factors.

**Table 8.** List of radiomics features (A to Y) and clinical factors (Z and α) selected at least twice across the five-fold cross-validation and independent test datasets. The symbols correspond to the feature names used in the frequency and coefficient analysis.

| Symbols | Feature names |
|---|---|
| A | Glrlm_RunLengthNonUniformity |
| B | Glrlm_RunEntropy |
| C | Shape_SurfaceVolumeRatio |
| D | Gldm_DependenceNonUniformityNormalized |
| E | Glszm_GrayLevelNonUniformity |
| F | Glszm_ZonePercentage |
| G | Glcm_ClusterTendency |
| H | Gldm_DependenceEntropy |
| I | Glszm_SizeZoneNonUniformityNormalized |
| J | Glcm_MaximumProbability |

| | |
|---|---|
| K | Glrlm_RunPercentage |
| L | Gldm_SmallDependenceHighGrayLevelEmphasis |
| M | Gldm_LargeDependenceEmphasis |
| N | Firstorder_Range |
| O | Firstorder_Minimum |
| P | Gldm_SmallDependeceEmphasis |
| Q | Shape_LeastAxisLength |
| R | Glcm_ClusterShade |
| S | Shape_Flatness |
| T | Shape_MinorAxisLength |
| U | Gldm_DependenceVariance |
| V | Shape_Sphericity |
| W | Shape_Maximum2DDiameterColumn |
| X | Glszm_SmallAreaHighGrayLevelEmphasis |
| Y | Glrlm_ShortRumLowGrayLevelEmphasis |
| Z | Tumor size |
| α | Molecular subtypes |

Figure 9 illustrates the importance of key features in the predictive model using only radiomics features by their selection frequency and contribution across the five-fold cross-validation and independent test datasets. The bar chart in the top row demonstrates that certain features (symbols A to C as described in Table 8) were consistently selected across all five folds. The coefficient heatmap in the second row represents that features A and B contributed positively to predicting the risk of LRR, whereas feature C contributed negatively. In the independent test dataset, features A, B, and C demonstrated similar contributions to those observed in the cross-validation, with features A and B having positive contributions and feature C having a negative contribution to predicting LRR risk.

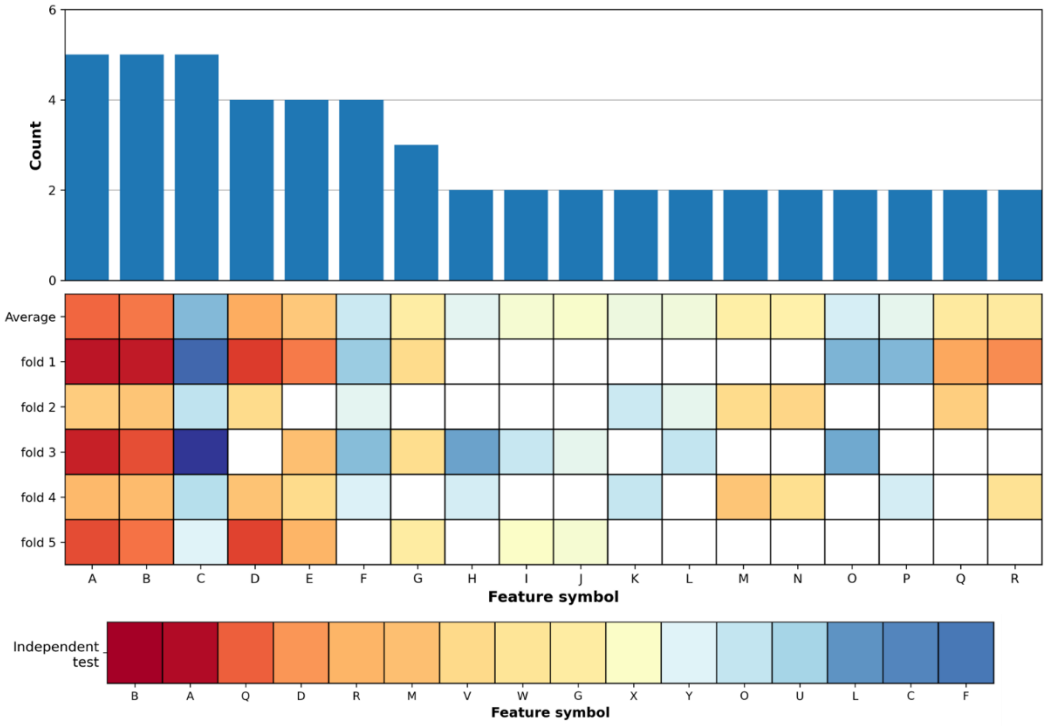


**Figure 9.** Feature selection frequency and coefficient analysis of key features in the predictive model using radiomics features only, evaluated using five-fold cross-validation and independent test datasets. The bar graph shows feature selection frequency, while the coefficient heatmap highlights feature importance, with red indicating positive and blue indicating negative contributions to predicting the risk of LRR. Feature symbols are detailed in Table 8.

Figure 10 illustrates the importance of key features in the predictive model that integrates both radiomics features and clinical factors, highlighting their selection frequency and contributions across the five-fold cross-validation and the independent test datasets. The bar chart in the top row highlights features A, B, and C (Table 8) consistently selected across all five folds. Features A and B contributed positively to predicting LRR, while feature C contributed negatively. In the bottom row, the coefficient heatmap for the independent test dataset indicates that features A, B, and C demonstrated similar contributions to those observed in the cross-validation, with features A and B contributing positively and feature C contributing negatively to predicting LRR risk. For the clinical factors, tumor size (Z) and molecular subtypes (α) were selected in both cross-validation and independent test datasets, with inconsistent selection across folds in cross-validation but consistent selection in the independent test dataset.

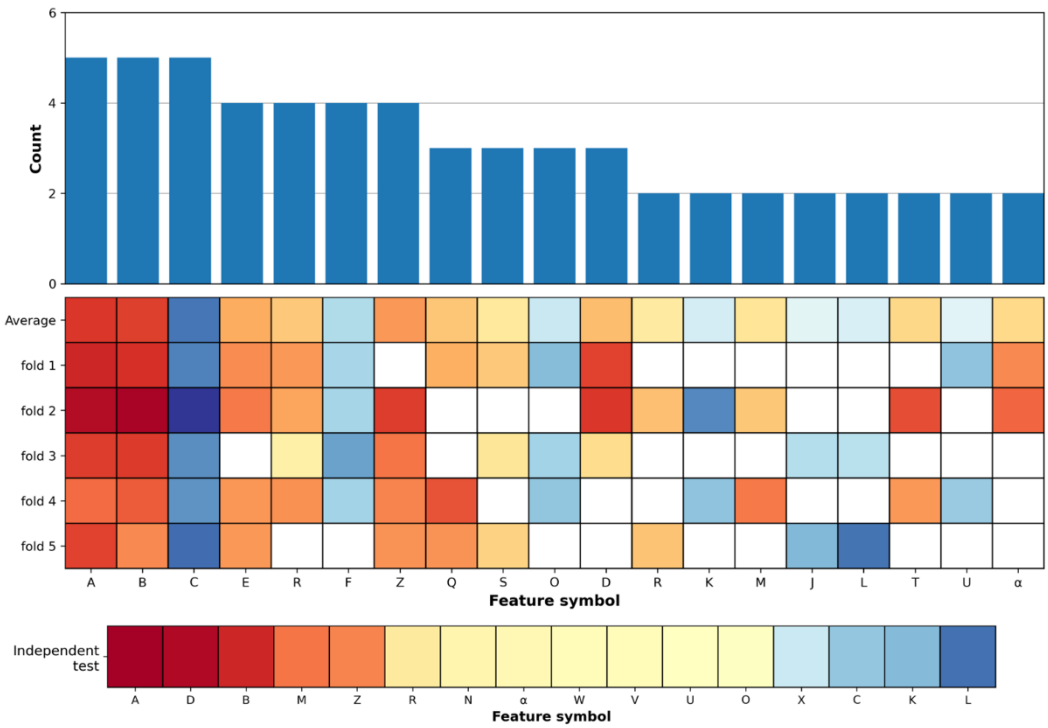


**Figure 10.** Feature selection frequency and coefficient analysis of key features in the predictive model integrating radiomics and clinical factors, evaluated using five-fold cross-validation and

independent test datasets. The bar graph shows feature selection frequency, while the coefficient heatmap highlights feature importance, with red indicating positive and blue indicating negative contributions to predicting the risk of LRR. Feature symbols are detailed in Table 8.

Table 9 presents the univariable analysis results for key features across the five-fold cross-validation and the independent test dataset, focusing on whether the odds ratio (OR) values are greater than or less than 1. Symbol A and B consistently showed OR values greater than 1, indicating an increased risk of LRR. Symbol C, in contrast, showed OR values less than 1 across all datasets, suggesting a risk reduction.

**Table 9.** Univariable analysis results for key features across cross-validation folds and the independent test dataset. P-values less than 0.05 are highlighted in bold.

| Features | OR (95% CI) | *P* value |
|---|---|---|
| Symbol A | | |
| Fold 1 | 1.002 (0.999-1.004) | 0.146 |
| Fold 2 | 1.001 (0.999-1.003) | 0.242 |
| Fold 3 | 1.002 (1.000-1.005) | **0.039** |
| Fold 4 | 1.002 (1.000-1.004) | 0.107 |
| Fold 5 | 1.003 (1.001-1.005) | **0.016** |
| Independent test dataset | 1.004 (1.001-1.006) | **0.002** |
| Symbol B | | |
| Fold 1 | 1.770 (0.865-3.623) | 0.118 |
| Fold 2 | 11.502 (3.234-40.909) | **0.001** |
| Fold 3 | 2.219 (1.068-4.609) | **0.033** |
| Fold 4 | 5.483 (1.771-16.980) | **0.003** |
| Fold 5 | 4.152 (1.552-11.108) | **0.005** |
| Independent test dataset | 5.583 (1.571-19.838) | **0.008** |
| Symbol C | | |
| Fold 1 | 0.323 (0.044-2.382) | 0.268 |
| Fold 2 | 0.001 (0.001-0.037) | **0.001** |
| Fold 3 | 0.060 (0.005-0.703) | **0.025** |
| Fold 4 | 0.003 (0.001-0.098) | **0.001** |
| Fold 5 | 0.027 (0.002-0.430) | **0.011** |
| Independent test dataset | 0.023 (0.001-0.443) | **0.013** |

***Abbreviations:*** OR, odd ratio; 95% CI, 95% confidence interval; Symbol A, Glrlm_RunLengthNonUniformity; Symbol B, Glrlm_RunEntropy; Symbol C, Shape_SurfaceVolumeRatio.

## 3.7. Discussion and conclusion

This study aimed to develop and validate an accurate predictive model for the risk of LRR in breast cancer patients. The baseline model, a foundational model developed in Chapter 2, was constructed using radiomics features extracted from T2-weighted fat-suppressed MRI sequences. However, it encountered challenges such as suboptimal performance and the inability to provide insights into key contributing features. In this chapter, enhancements to the model were made to optimize predictive performance, identify key contributing features, and ensure its robustness and generalizability across diverse datasets.

To comprehensively evaluate the factors contributing to the predictive performance of the model for LRR risk, several enhancement strategies were implemented and assessed against the baseline model. These enhancement strategies, including domain adaptation, feature selection techniques, clinical factor integration, and model calibration, were employed to optimize predictive performance and reliability. The best-performing model was developed by enhancing the baseline model through the incorporation of domain adaptation, the RFE process, and the integration of clinical factors, resulting in an average AUC of 0.757 for five-fold cross-validation and an AUC of 0.762 for the independent test dataset. Furthermore, the application of model calibration improved the reliability of predicted probabilities, as the AUC increased from 0.734 to 0.757 for cross-validation and from 0.698 to 0.762 for the independent test dataset, showing a closer alignment between predicted and actual risk. These results indicate that these enhancements contributed to a more robust and reliable model, demonstrating potential for clinical application.

In this study, several radiomics features were consistently identified as positive contributors to LRR risk prediction across both the five-fold cross-validation and independent test datasets, in models using only radiomics features as well as those integrating clinical factors. Specifically, Glrlm_RunLengthNonUniformity (Symbol A) and Glrlm_RunEntropy (Symbol B)[58] were consistently recognized as significant positive contributors to LRR risk in both the five-fold cross-validation and independent test datasets. These features, which capture the heterogeneity and complexity of tumor texture, may be crucial for understanding tumor aggressiveness and its likelihood of recurrence. Both features are derived from the Gray Level Run Length Matrix (GLRLM), a texture analysis technique that quantifies the occurrence of consecutive pixels with the

same intensity level in a specific direction.[59] Essentially, GLRLM provides information about the distribution of homogeneous runs of gray levels, which can describe the structural complexity and uniformity of a tumor. Symbol A reflects the variability in the length of homogeneous runs within the tumor, with higher values indicating greater heterogeneity. Similarly, Symbol B measures the randomness in the texture, where higher values are associated with increased complexity and irregularity within the tumor structure. These features represent the heterogeneous texture of the tumor, and previous studies have shown that such heterogeneity is often associated with poor treatment outcomes.[40,41,60]

In contrast, a feature representing the ratio between the surface area and volume of the tumor (Shape_SurfaceVolumeRatio, Symbol C)[61] was consistently identified as a negative contributor to LRR risk across both the five-fold cross-validation and independent test datasets, in models using only radiomics features as well as those integrating clinical factors. Lower values of this feature generally indicate that the tumor is more compact, while higher values suggest that the tumor is more irregular or elongated. The finding that lower Symbol C values may be linked to a higher risk of recurrence, possibly reflecting a tumor morphology that is more likely to recur due to being less responsive to surgery or radiotherapy.

In the integrated model incorporating clinical factors, tumor size (Symbol Z) emerged as a positive contributor to LRR risk across all folds, in addition to radiomics features, except for Fold 1. This suggests that larger tumor size may be associated with a higher risk of recurrence, reflecting its recognized significance as a prognostic factor in the risk of LRR. Similarly, molecular subtypes (Symbol α) also did not contribute consistently across all folds but showed positive contributions in Folds 1 to 2, indicating that more aggressive pathological features like IDC may correlate with an increased likelihood of recurrence in certain subsets of the dataset. These highlight the potential added value of clinical factors in complementing radiomics features for a more comprehensive risk prediction model.

To further support these findings, univariable analysis was conducted for key features across the cross-validation folds and the independent test dataset. For Symbol A, the OR consistently exceeded 1 across the folds, indicating an increased risk of LRR. The independent test dataset showed a

particularly significant OR of 1.004 (95% CI: 1.001-1.006, p = 0.002). This suggests that Symbol A is positively associated with the risk of LRR. For Symbol B, the OR varied across the folds, with a particularly high OR of 11.502 (95% CI: 3.234-40.909, p = 0.001) in Fold 2. The independent test dataset also revealed a significant OR of 5.583 (95% CI: 1.571-19.838, p = 0.008), supporting a strong positive association between Symbol B and LRR risk. For Symbol C, the OR consistently remained below 1 across the folds, suggesting a potential inverse relationship with LRR risk. The independent test dataset further supported this inverse association, with an OR of 0.023 (95% CI: 0.000–0.443, p = 0.013), indicating a significantly reduced likelihood of LRR.

While this study provides meaningful findings, several limitations should be acknowledged. First, TNBC is known to have the highest LRR risk among molecular subtypes.[38] In this study, no statistically significant difference in LRR risk was observed among subtypes, potentially due to the limited sample size and variability in treatment regimens. However, in real-world settings, the use of pembrolizumab based on KEYNOTE-522 has been shown to improve pathologic complete response rates in TNBC patients receiving neoadjuvant therapy, which may subsequently reduce LRR risk.[39] Importantly, this study did not include patients treated with pembrolizumab, highlighting the need for further studies to incorporate such patients into the modeling process. Second, while this study integrated clinical factors alongside radiomics features to enhance the model's predictive capability, there is potential for further exploration of other potential data sources. For instance, directly using MRI scans through convolutional neural networks (CNNs) could provide a more comprehensive assessment of tumor characteristics. CNNs are capable of capturing intricate spatial features from imaging data that may not be captured by manually defined radiomics features, thereby providing an additional layer of valuable information. Lastly, while Institution 2 was designated as the independent test set, data from Institutions 1, 3, and 4 were not utilized for independent testing because they were included in the training dataset to ensure sufficient sample size for model development. While this approach was necessary to address the limited number of LRR cases, it restricted the evaluation of the model's generalizability across multiple unseen institutions, potentially reducing the robustness of the findings. Further studies should include a larger registry with more LRR cases to strengthen the model’s training foundation and address limitations in sample diversity.

# 4. CONCLUSION

In this study, we have proposed a predictive model leveraging radiomics features extracted from T2-weighted MRI images with fat suppression to predict the risk of LRR in breast cancer patients. By incorporating multi-institutional patient registries, we improved the robustness of the model. The study introduced a foundational baseline model in Chapter 2 and subsequently enhanced its performance in Chapter 3 through strategies such as domain adaptation, different feature selection techniques, integration of clinical factors, and model calibration. These enhancements resulted in improved predictive performance, with the best-performing model achieving an average AUC of 0.757 in cross-validation and 0.762 for the independent test dataset, demonstrating potential for clinical application.

This study also highlighted specific texture features that reflect tumor heterogeneity and complexity as key contributors to the risk of LRR. Integrating clinical factors, such as tumor size and molecular subtypes, with radiomics features further enhanced the model's interpretability and predictive performance. These findings demonstrate the significance of texture features and clinical factors to achieve a more accurate and comprehensive understanding of risk factors, thereby contributing to improvements in the model's overall predictive capability.

However, the sample size in this study, particularly for high-risk subgroups like TNBC, was insufficient to comprehensively evaluate subtype-specific LRR risks. The absence of patients treated with modern therapies, such as pembrolizumab for TNBC, may limit the model's applicability to up-to-date treatment approaches.

In conclusion, this study leverages multi-institutional registries to establish a foundational baseline predictive model for assessing LRR risk in breast cancer patients. Several enhancements were subsequently investigated to optimize the model's performance, and key features contributing to LRR risk were identified. To further validate the proposed model’s predictive performance, prospective datasets should be analyzed in future studies.

## References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), pp.394-424.
2. Delaney, G., Jacob, S., Featherstone, C. and Barton, M., 2005. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *104*(6), pp.1129-1137.
3. Hall, E.J. and Giaccia, A.J., 2006. *Radiobiology for the Radiologist* (Vol. 6, p. 597).
4. Meyer, J. ed., 2011. *IMRT, IGRT, SBRT: advances in the treatment planning and delivery of radiotherapy*. Karger Medical and Scientific Publishers.
5. Ling, C.C., Yorke, E. and Fuks, Z., 2006. From IMRT to IGRT: frontierland or neverland?. *Radiotherapy and oncology*, *78*(2), pp.119-122.
6. Formenti, S.C. and Demaria, S., 2013. Combining radiotherapy and cancer immunotherapy: a paradigm shift. *JNCI: Journal of the National Cancer Institute*, *105*(4), pp.256-265.
7. Yu, S., Wang, Y., He, P., Shao, B., Liu, F., Xiang, Z., Yang, T., Zeng, Y., He, T., Ma, J. and Wang, X., 2022. Effective combinations of immunotherapy and radiotherapy for cancer treatment. *Frontiers in oncology*, *12*, p.809304.
8. Herskovic, A., Martz, K., Al-Sarraf, M., Leichman, L., Brindle, J., Vaitkevicius, V., Cooper, J., Byhardt, R., Davis, L. and Emami, B., 1992. Combined chemotherapy and radiotherapy compared with radiotherapy alone in patients with cancer of the esophagus. *New England Journal of Medicine*, *326*(24), pp.1593-1598.
9. Maduro, J.H., Pras, E., Willemse, P.H.B. and De Vries, E.G.E., 2003. Acute and long-term toxicity following radiotherapy alone or in combination with chemotherapy for locally advanced cervical cancer. *Cancer treatment reviews*, *29*(6), pp.471-488.
10. Kunkler, I.H., Williams, L.J., Jack, W.J., Cameron, D.A. and Dixon, J.M., 2015. Breast-conserving surgery with or without irradiation in women aged 65 years or older with early breast cancer (PRIME II): a randomised controlled trial. *The lancet oncology*, *16(3)*, pp.266-273.
11. Wang, J.Z., Huang, Z., Lo, S.S., Yuh, W.T. and Mayr, N.A., 2010. A generalized linear-

quadratic model for radiosurgery, stereotactic body radiation therapy, and high–dose rate brachytherapy. *Science translational medicine*, *2*(39), pp.39ra48-39ra48.

12. Yoo, S.K., Kim, T.H., Chun, J., Choi, B.S., Kim, H., Yang, S., Yoon, H.I. and Kim, J.S., 2022. Deep-learning-based automatic detection and segmentation of brain metastases with small volume for stereotactic ablative radiotherapy. *Cancers*, *14*(10), p.2555.
13. Yoo, S.K., Kim, K.H., Noh, J.M., Oh, J., Yang, G., Kim, J., Kim, N., Kim, H. and Yoon, H.I., 2024. Development of learning-based predictive models for radiation-induced atrial fibrillation in non-small cell lung cancer patients by integrating patient-specific clinical, dosimetry, and diagnostic information. *Radiotherapy and Oncology*, *201*, p.110566.
14. Huynh, E., Hosny, A., Guthier, C., Bitterman, D.S., Petit, S.F., Haas-Kogan, D.A., Kann, B., Aerts, H.J. and Mak, R.H., 2020. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, *17*(12), pp.771-781.
15. Janiesch, C., Zschech, P. and Heinrich, K., 2021. Machine learning and deep learning. *Electronic Markets*, *31*(3), pp.685-695.
16. Zhang, Y. and Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, *4*(1), p.25.
17. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), pp.206-215.
18. Field, M., Hardcastle, N., Jameson, M., Aherne, N. and Holloway, L., 2021. Machine learning applications in radiation oncology. *Physics and Imaging in Radiation oncology*, *19*, pp.13-24.
19. El Naqa, I., Karolak, A., Luo, Y., Folio, L., Tarhini, A.A., Rollison, D. and Parodi, K., 2023. Translation of AI into oncology clinical practice. *Oncogene*, *42*(42), pp.3089-3097.
20. Gillies, R.J., Kinahan, P.E. and Hricak, H., 2016. Radiomics: images are more than pictures, they are data. *Radiology*, *278*(2), pp.563-577.
21. Huang, Y., Liu, Z., He, L., Chen, X., Pan, D., Ma, Z., Liang, C., Tian, J. and Liang, C., 2016. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology*, *281*(3), pp.947-957.
22. Liu, X., Li, Y., Qian, Z., Sun, Z., Xu, K., Wang, K., Liu, S., Fan, X., Li, S., Zhang, Z. and

Jiang, T., 2018. A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. *NeuroImage: Clinical*, *20*, pp.1070-1077.

23. Koçak, B., Durmaz, E.Ş., Ateş, E. and Kılıçkesmez, Ö., 2019. Radiomics with artificial intelligence: a practical guide for beginners. *Diagnostic and interventional radiology*, *25*(6), p.485.
24. Hannoun-Levi, Jean-Michel, et al. "Partial breast irradiation as second conservative treatment for local breast cancer recurrence." *International Journal of Radiation Oncology* Biology* Physics* 60.5 (2004): 1385-1392.
25. Costeira, Beatriz, et al. "Long-term locoregional recurrence in patients treated for breast cancer." *Breast Cancer Research and Treatment* 202.3 (2023): 551-561.
26. Mukhtar, R.A., Chau, H., Woriax, H., Piltin, M., Ahrendt, G., Tchou, J., Yu, H., Ding, Q., Dugan, C.L., Sheade, J. and Crown, A., 2023. Breast conservation surgery and mastectomy have similar locoregional recurrence after neoadjuvant chemotherapy: results from 1462 patients on the prospective, randomized I-SPY2 trial. *Annals of surgery*, *278*(3), pp.320-327.
27. Kunkler, I.H., Williams, L.J., Jack, W.J., Cameron, D.A. and Dixon, J.M., 2023. Breast-conserving surgery with or without irradiation in early breast cancer. *New England Journal of Medicine*, *388*(7), pp.585-594.
28. Park, E.H., Jung, K.W., Park, N.J., Kang, M.J., Yun, E.H., Kim, H.J., Kim, J.E., Kong, H.J., Im, J.S. and Seo, H.G., 2024. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2021. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, 56(2), pp.357-371.
29. Agarwal, S., Pappas, L., Neumayer, L., Kokeny, K. and Agarwal, J., 2014. Effect of breast conservation therapy vs mastectomy on disease-specific survival for early-stage breast cancer. *JAMA surgery*, *149*(3), pp.267-274.
30. Early Breast Cancer Trialists' Collaborative Group, 2011. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials. *The Lancet*, *378*(9804), pp.1707-1716.
31. Elkhuizen, P.H., van de Vijver, M.J., Hermans, J.O., Zonderland, H.M., van de Velde, C.J. and Leer, J.W.H., 1998. Local recurrence after breast-conserving therapy for invasive breast

cancer: high incidence in young patients and association with poor survival. *International Journal of Radiation Oncology* Biology* Physics*, *40*(4), pp.859-867.

32. Arvold, N.D., Taghian, A.G., Niemierko, A., Abi Raad, R.F., Sreedhara, M., Nguyen, P.L., Bellon, J.R., Wong, J.S., Smith, B.L. and Harris, J.R., 2011. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. *Journal of clinical oncology*, *29*(29), pp.3885-3891.
33. Wapnir, I.L., Anderson, S.J., Mamounas, E.P., Geyer Jr, C.E., Jeong, J.H., Tan-Chiu, E., Fisher, B. and Wolmark, N., 2006. Prognosis after ipsilateral breast tumor recurrence and locoregional recurrences in five National Surgical Adjuvant Breast and Bowel Project node-positive adjuvant breast cancer trials. *Journal of Clinical Oncology*, *24*(13), pp.2028-2037.
34. Jacobson, J.A., Danforth, D.N., Cowan, K.H., d'Angelo, T., Steinberg, S.M., Pierce, L., Lippman, M.E., Lichter, A.S., Glatstein, E. and Okunieff, P., 1995. Ten-year results of a comparison of conservation with mastectomy in the treatment of stage I and II breast cancer. *New England Journal of Medicine*, *332*(14), pp.907-911.
35. Tseng, Y.D., Uno, H., Hughes, M.E., Niland, J.C., Wong, Y.N., Theriault, R., Blitzblau, R.C., Moy, B., Breslin, T., Edge, S.B. and Hassett, M.J., 2015. Biological subtype predicts risk of locoregional recurrence after mastectomy and impact of postmastectomy radiation in a large national database. *International Journal of Radiation Oncology* Biology* Physics*, *93*(3), pp.622-630.
36. Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A. and McGuire, W.L., 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *science*, *235*(4785), pp.177-182.
37. Slamon, D., Eiermann, W., Robert, N., Pienkowski, T., Martin, M., Press, M., Mackey, J., Glaspy, J., Chan, A., Pawlicki, M. and Pinter, T., 2011. Adjuvant trastuzumab in HER2-positive breast cancer. *New England journal of medicine*, *365*(14), pp.1273-1283.
38. Morrow, M., 2013. Personalizing extent of breast cancer surgery according to molecular subtypes. *The Breast*, *22*, pp.S106-S109.
39. Schmid, P., Cortes, J., Dent, R., Pusztai, L., McArthur, H., Kümmel, S., Bergh, J., Denkert, C., Park, Y.H., Hui, R. and Harbeck, N., 2021. VP7-2021: KEYNOTE-522: Phase III study of neoadjuvant pembrolizumab+ chemotherapy vs. placebo+ chemotherapy, followed by adjuvant

pembrolizumab vs. placebo for early-stage TNBC. *Annals of Oncology*, *32*(9), pp.1198-1200.

40. Kim, J.H., Ko, E.S., Lim, Y., Lee, K.S., Han, B.K., Ko, E.Y., Hahn, S.Y. and Nam, S.J., 2017. Breast cancer heterogeneity: MR imaging texture analysis and survival outcomes. *Radiology*, *282*(3), pp.665-675.
41. Park, H., Lim, Y., Ko, E.S., Cho, H.H., Lee, J.E., Han, B.K., Ko, E.Y., Choi, J.S. and Park, K.W., 2018. Radiomics signature on magnetic resonance imaging: association with disease-free survival in patients with invasive breast cancer. *Clinical Cancer Research*, *24*(19), pp.4705-4714.
42. Lee, Joongyo, et al. "Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer." *Oncology Letters* 26.4 (2023): 1-10.
43. Nitesh, V.C., 2002. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*, *16*(1), p.321.
44. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, pp.389-422.
45. Abdi, H. and Williams, L.J., 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), pp.433-459.
46. Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), pp.61-74.
47. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S. and Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, *77*(21), pp.e104-e107.
48. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.
49. Sun, B., Feng, J. and Saenko, K., 2017. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pp.153-171.
50. Street, W.N., Wolberg, W.H. and Mangasarian, O.L., 1993, July. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization* (Vol. 1905, pp. 861-870). SPIE.
51. Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. *Computers &*

*electrical engineering*, *40*(1), pp.16-28.

52. Ding, C. and Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, *3*(02), pp.185-205.
53. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), pp.267-288.
54. Jain, A.K., Duin, R.P.W. and Mao, J., 2000. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(1), pp.4-37.
55. Harris, S. and Harris, D., 2015. *Digital design and computer architecture*. Morgan Kaufmann.
56. Zadrozny, B. and Elkan, C., 2002, July. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699).
57. Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017, July. On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR.
58. Tang, X., 1998. Texture information in run-length matrices. *IEEE transactions on image processing*, *7*(11), pp.1602-1609.
59. Galloway, M.M., 1974. Texture analysis using grey level run lengths. *Nasa Sti/recon Technical Report N*, *75*, p.18555.
60. Dasgupta, A., Bhardwaj, D., DiCenzo, D., Fatima, K., Osapoetra, L.O., Quiaoit, K., Saifuddin, M., Brade, S., Trudeau, M., Gandhi, S. and Eisen, A., 2021. Radiomics in predicting recurrence for patients with locally advanced breast cancer using quantitative ultrasound. *Oncotarget*, *12*(25), p.2437.
61. Lorensen, W.E. and Cline, H.E., 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field* (pp. 347-353).

Abstract in Korean

# 유방보존치료 후 유방암 국소 재발을 예측하기 위한 다기관 레지스트리를 활용한 학습 기반 모델 개발

목적: 방사선 치료는 수술과 함께 유방보존치료를 가능하게 하는 필수적인 치료법이다. 그러나 일부 환자에서 국소 재발(LRR)이 발생하여 유방 보존이 실패할 수 있다. 본 연구는 다기관 레지스트리에서 추출한 라디오믹스 특징을 통합하여 유방암 환자의 LRR 위험을 예측하는 기계 학습 모델을 개발하고 검증하는 것을 목표로 하였다. 단일 자기공명영상(MRI) 시퀀스(T2 가중 지방 억제)를 활용하고 LRR 위험과 관련된 주요 특징들을 식별함으로써, 본 연구는 LRR 위험 예측 모델의 견고성과 임상 적용 가능성을 높여 개인 맞춤형 치료 계획 수립에 기여하고자 한다.

방법: 다기관 레지스트리를 기반으로 352명의 유방암 환자 데이터를 후향적으로 수집하고 분석하였다. 데이터셋은 T2 가중 지방 억제 MRI 스캔, 수동으로 윤곽을 그린 유방 종양, 및 진단 시 연령, 종양 크기, 병리학적 특성, 분자 아형 등의 임상적 요인으로 구성되었다. 종양 윤곽은 각 기관의 방사선종양학 전문의에 의해 수행되고 검증되었다. 클래스 불균형 문제를 해결하기 위해 오버샘플링 기법을 포함한 다양한 데이터 샘플링 방법을 탐색하고 평가하였으며, 최종적으로 균형 잡힌 모델 개발을 위해 LRR 환자와 LRR 발생하지 않은 환자를 각각 동일한 비율로 포함한 샘플을 무작위로 선택하여 모델 개발에 사용했다. 라디오믹스 특징으로는 수동으로 윤곽을 그린 관심 영역(ROIs)에서 추출된 형태 기반, 일차 통계, 텍스처 특징들이 포함되어 있고, 특징 추출 과정에서 MRI 스캔 정규화가 모델 성능에 미치는 영향이 평가되었다. 기계 학습 모델은 특징 선택 기법들과 주성분 분석(PCA)을 사용하고 로지스틱 회귀를 분류기로서 사용하여 개발하였다. 또한, 모델 성능을 향상시키기 위해 도메인 적응(domain adaptation) 기법을 적용하였으며, LRR 위험과 관련된 임상적 요인과 라디오믹스 특징을 통합한 모델을 개발하여 서로 다른 데이터 유형을 결합했을 때의 추가적인 예측 가치를 평가하였다. 모델의 성능은 5-fold 교차 검증 및 독립적인 테스트 데이터셋을 사용하여 평가되었으며, 확률 추정의 정확성을 향상시키기 위해 칼리브레이션을 적용하였다.

결과: MRI 스캔 정규화를 적용하고, 래퍼(wrapper) 방식의 재귀적 특징

제거(Recursive Feature Elimination, RFE)를 활용한 특징 선택을 수행하며, 라디오믹스 특징과 임상적 요인을 모두 입력으로 포함했을 때 최고의 성능이 달성되었다. 이러한 조건에서 모델은 교차 검증에서 평균 AUC 0.757 (95% 신뢰구간, 0.715-0.799)을, 독립적인 테스트 데이터셋에서 AUC 0.762를 달성하였다.

결론: 본 연구에서는 유방암 환자의 LRR 위험을 예측하기 위해 라디오믹스 특징과 LRR 위험과 관련된 임상적 요인을 통합한 예측 모델을 개발하였다. 연구 결과, 비침습적 바이오마커로서 라디오믹스는 임상적 요인과 결합될 때 개인 맞춤형 위험 평가를 향상시킬 가능성을 보여주었다. 본 모델의 예측력을 추가로 검증하기 위해 향후 전향적 데이터셋을 활용한 분석이 필요하다.

---

**핵심되는 말** : 국소재발, 유방암, 유방보존치료, 머신러닝, 라디오믹스

# PUBLICATION LIST

1. Yoo, S.K., Kim, K.H., Noh, J.M., Oh, J., Yang, G., Kim, J., Kim, N., Kim, H. and Yoon, H.I., 2024. Development of learning-based predictive models for radiation-induced atrial fibrillation in non-small cell lung cancer patients by integrating patient-specific clinical, dosimetry, and diagnostic information. *Radiotherapy and Oncology*, *201*, p.110566.
2. Choi, B.S., Yoo, S.K., Moon, J., Chung, S.Y., Oh, J., Baek, S., Kim, Y., Chang, J.S., Kim, H. and Kim, J.S., 2023. Acute coronary event (ACE) prediction following breast radiotherapy by features extracted from 3D CT, dose, and cardiac structures. *Medical physics*, *50*(10), pp.6409-6420.
3. Lee, J., Yoo, S.K., Kim, K., Lee, B.M., Park, V.Y., Kim, J.S. and Kim, Y.B., 2023. Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer. *Oncology Letters*, *26*(4), pp.1-10.
4. Yoo, S.K., Kim, H., Choi, B.S., Park, I. and Kim, J.S., 2022. Generation and evaluation of synthetic computed tomography (CT) from cone-beam CT (CBCT) by incorporating feature-driven loss into intensity-based loss functions in deep convolutional neural network. *Cancers*, *14*(18), p.4534.
5. Yoo, S.K., Kim, T.H., Chun, J., Choi, B.S., Kim, H., Yang, S., Yoon, H.I. and Kim, J.S., 2022. Deep-learning-based automatic detection and segmentation of brain metastases with small volume for stereotactic

ablative radiotherapy. *Cancers*, *14*(10), p.2555.