Breast Cancer: Targets and Therapy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Uncertainty-Aware Machine Learning for Predicting Axillary Lymph Node Metastasis Using Breast MRI

Sevgi Ünal[1], Remzi Gürfidan[2]

[1]Department of Radiology, Izmir Katip Celebi University Ataturk Training and Research Hospital, Izmir, Türkiye; [2]Database, Network Design and Management, Isparta University of Applied Science, Isparta Vocational School of Information Technologies, Isparta, Türkiye

Correspondence: Sevgi Ünal, Email sevgiunal84@gmail.com

Purpose: Axillary lymph node metastasis (ALNM) is a significant prognostic factor in breast cancer and has an impact on staging, treatment and survival. The objective of this study is to create a machine learning model that will be able to predict axillary lymph node metastasis (ALNM) in a preoperative setting using breast MRI-derived and clinicopathological variables, while also achieving probability calibration and uncertainty-aware prediction for more reliable risk estimates.

Patients and Methods: For this retrospective single-centre study, 204 patients who underwent contrast-enhanced breast MRI from 2021 to 2024 were selected. The dataset comprised of 23 independent variables, respectively, representing demographic, clinical, radiological, histopathological, molecular characteristics along with a binary target variable indicating ALNM status. The data was split into 60% called training set, 20% calibration set, and 20% test set. Candidate models were evaluated based on ROC-AUC on the training subset that was used for screening. Subsequently, calibration was performed in a held-out calibration set, following which class-conditional conformal prediction was applied to quantify predictive uncertainty.

Results: Out of all the models we evaluated, the Conformal-Calibrated Interpretable Risk Model (CCIRM) was found to be the best model, achieving a test accuracy of 0.9268, a weighted F1 score of 0.9270 and AUC of 0.937. With a precision of 0.9545 and a recall of 0.9130 in the ALNM-positive class, the model was potent. In addition to discrimination strength, the framework produces calibrated risk estimates and uncertainty-aware prediction sets that enable transparent interpretation of model outputs in clinically borderline cases.

Conclusion: The proposed approach includes the model selection based on the machine learning and probability calibration, as well as conformal prediction to achieve reliability, beyond label prediction, and uncertainty-aware risk estimation for preoperative ALNM assessment. The results from this analysis suggest that CCIRM is a potential methodological framework for trustworthy clinical decision support, although the prospective and multicentre external validation should be done before it becomes applicable to the clinical setting.

Keywords: axillary lymph node, breast MRI, interpretable machine learning, conformal prediction, breast cancer

Introduction

Breast cancer is one of the leading causes of cancer-related death among women worldwide.[1] Axillary lymph node metastasis (ALNM) is a major prognostic factor affecting staging, treatment planning, and overall survival in breast cancer.[1,2] The five-year survival rate decreases from 98.6% to 84.4% in patients with ALNM.[2] Although sentinel lymph node biopsy (SLNB) and axillary lymph node dissection (ALND) remain the gold standard methods for assessing axillary lymph node (ALN) status, both are invasive procedures associated with potential complications.[1,3] Therefore, current guidelines such as NCCN, ESMO, and ASCO recommend imaging-based evaluation of ALN status before treatment planning.[4] In this context, accurate non-invasive preoperative prediction of ALNM may help reduce unnecessary invasive axillary procedures.

https://doi.org/10.2147/BCTT.S599211
Received: 16 February 2026
Accepted: 17 April 2026
Published: 1 May 2026

Breast MRI provides comprehensive morphological and functional information, including tumour size, shape, margin characteristics, contrast enhancement patterns, tissue heterogeneity, and vascularity.[4,5] Recent advances in artificial intelligence have led to a growing number of radiomics and machine learning studies aiming to predict ALNM using features derived from breast MRI.[6,7] Most of these studies have focused on dynamic contrast-enhanced MRI (DCE-MRI) and diffusion-weighted imaging (DWI), which reflect tumour angiogenesis and cellularity, respectively.[7,8] Diffusion-related parameters such as the apparent diffusion coefficient have also been associated with pathological features linked to nodal spread.[8,9] In addition, MRI-based ALNM assessment is not limited to DCE-MRI and DWI, as findings such as peritumoral oedema, peripectoral oedema, and adjacent vascular features may also contribute to risk evaluation.[4]

With all these development, there are still some methodological limitations in the existing ALNM prediction literature. Many previous MRI-based radiomics and machine learning studies have predominantly been targeted toward discrimination performance, mostly through AUC-based comparison. Most of them provided limited information on probability calibration, uncertainty quantification and external validation.[5,6,10–16] A model may therefore show outstanding classification ability even if the risk estimates it produces do not reliably match the actual probability of nodal involvement. This limitation is important clinically because an overconfident and miscalibrated prediction may mislead preoperative axillary assessment and may lead to inappropriate reassurance or unnecessary invasive axillary procedures.[3,4] Furthermore, a number of previous investigations were carried out in small or selected cohorts and lacked sufficiently broad validation strategies to support generalisability.[6,10–13,16]

The literature on ALNM prediction is not limited to image-only radiomics approaches. Previous studies have demonstrated that the inclusion of imaging findings with clinicopathological, histopathological, or molecular variables may improve predictive modelling by capturing complementary dimensions of tumour biology and disease behaviour.[5,10,13] Because tumours could not only exhibit pathological aggressiveness and belong to a molecular subtype that are risk factors for nodal metastases, such multimodal predictive frameworks may have more clinical applicability. In this view, calibration is crucial since the clinical interpretation may not only rely on the predicted class label but also on the reliability of their risk estimates. Furthermore, conformal prediction may offer an extra safeguard by signalling when the model output is uncertain, thus assisting in avoiding overconfident interpretation in borderline cases.[3,4,6,10]

The models of clinical identification will need to be adapted to the clinical decision-threshold context in which they are used. In breast cancer treatment, preoperative assessment of nodal risk may help in closely scrutinizing axillary findings and may influence a less liberal approach to the decision-making process for SLNB or ALND.[3,4] For a clinically meaningful ALNM prediction model, it is important to not only develop a model that provides good discrimination but also offers calibrated, uncertainty-aware predictions for transparent decision support in context.

In this instance, the predictive model has clinical value not only when it correctly distinguishes patients with and without nodal metastasis but also when the model's estimated probabilities are trustworthily made, and uncertainty is made explicit. A highly discriminative model but poorly calibrated may yield risk estimates that are difficult to interpret in practice. This is especially true if preoperative decisions penalise false negatives and false positives. Also, in borderline cases, a forced single-label prediction may obscure an uncertainty in the model. As such, strategies that collaboratively tackle discrimination, calibration, and predictive uncertainty could offer a more clinically responsible framework for decision support in breast imaging.

To address this gap, the researchers developed a framework called Conformal-Calibrated Interpretable Risk Model (CCIRM). This framework is a participatory uncertainty aware machine learning for preoperative prediction of ALNM. In addition, it is based on breast MRI-derived and clinical and pathological features. CCIRM was designed to yield clinically more reliable and cautiously interpretable risk estimation by integrating model selection, probability calibration and class-conditional conformal prediction in a single evaluation pipeline, quite unlike standard predictive models which largely focus on classification performance. Thus, the present study's contribution is not only a measure of predictive performance but also a methodological emphasis on reliability, transparency, and uncertainty-aware clinical decision support.

The main contributions of this study are summarised below:

- Development of a multimodal machine learning framework for preoperative ALNM prediction using breast MRI-derived and clinicopathological variables,
- Explicit integration of probability calibration to improve the reliability of estimated risk scores,
- Application of class-conditional (Mondrian) conformal prediction to generate uncertainty-aware prediction sets,
- Evaluation of model performance from a trustworthy clinical AI perspective, extending beyond discrimination metrics alone.

Materials and Methods

This section contains information about the data set used in the study, which includes details about the patients' visual characteristics.

Patients

This study was conducted in accordance with the Helsinki Declaration, as updated in 2013. This retrospective study was approved by the Health Research Ethics Committee of İzmir Katip Çelebi University (09.10.2025/0601), and the requirement for individual informed consent for retrospective analysis was waived. A comprehensive retrospective screening of clinical, pathological, and radiological databases was conducted to predict ALNM with ML in breast cancer patients diagnosed at İzmir Atatürk Training and Research Hospital, Katip Çelebi University, between June 2021 and August 2024. This was a retrospective single-centre modelling study designed for internal development and evaluation of an uncertainty-aware clinical prediction framework.

The inclusion criteria for the study are as follows:

1. Female patients diagnosed with breast cancer based on histopathological findings,
2. Patients who underwent bilateral breast MRI prior to surgery,
3. Patients who underwent axillary lymph node biopsy, ALND, or SLNB for LNM assessment.

The exclusion criteria are as follows:

1. Patients with a history of breast surgery or chemotherapy prior to MRI examination,
2. Patients with incomplete or low-quality MRI images,
3. Patients whose histopathological results are unavailable.

The clinical-pathological data were obtained directly from the medical record system and pathology reports. The clinical-pathological information collected included the following variables: age, side of the breast with the lesion, multifocality, histopathological grade, histopathological type, presence of accompanying ductal carcinoma in situ (DCIS), ER status, PR status, HER-2 status, Ki67 expression level, and ALN involvement. According to the immunohistochemical evaluation guidelines for breast cancer published by the College of American Pathologists (CAP), ER and PR positivity were defined as the presence of nuclear staining in $\geq$1% of tumour cells, while ER and PR negativity was defined as <1% nuclear staining in the presence of a positive internal control. A threshold value of 14% was used to distinguish Ki67 expression levels.

HER2 negativity (HER2−) is defined as cases with an IHC score of 0 or IHC 1+ and no HER2 gene amplification detected by in situ hybridisation (ISH); HER2 positivity (HER2+) is defined as cases with an IHC score of 3+ or IHC 2+ and HER2 gene amplification detected by ISH. All cases were divided into four molecular subtypes based on ER, PR, HER2, and Ki-67 expression status:

1. Luminal A: ER/PR positive, HER2 negative, low Ki-67
2. Luminal B: ER/PR positive and high Ki-67 or HER2 positive
3. HER2+: ER and PR negative, HER2 positive
4. Triple negative: ER, PR and HER2 negative

The study period and MRI acquisition period were identical. All breast MRI examinations included in this study were performed between June 2021 and August 2024. The reference standard for ALNM status was based on histopathological evaluation obtained from sentinel lymph node biopsy (SLNB), axillary lymph node dissection (ALND), or biopsy specimens.

Image Acquisition

All examinations were performed on a Siemens MAGNETOM Lumina (Germany) system with a 3.0 T magnetic field strength, using an 18-channel bilateral breast coil. The patient was placed in the prone position and underwent morphological evaluation using axial T2-weighted fat-suppressed TSE/TIRM (approximately TR/TE 5000–6000/60–80 ms, slice thickness 3–4 mm, interval ≤0.5 mm, FOV 300–360 mm) for morphological evaluation. Diffusion-weighted imaging (DWI, EPI) was then performed with b-values of 0 and 800–1000 s/mm^2 (additional high b 1500 s/mm^2 when necessary), and ADC maps were automatically generated (slice thickness 3–4 mm, FOV 300–360 mm, fat-suppressed SPAIR). A pre-contrast axial 3D T1-weighted fat-suppressed GRE (VIBE/Dixon-VIBE) sequence was used for reference (approximately TR/TE 4–5/1.3–2.0 ms, flip angle 9–12°, isotropic voxel 0.8–1.2 mm, GRAPPA acceleration 2–3). Dynamic contrast-enhanced imaging was performed using the standard 3D T1 fat-suppressed VIBE/Dixon-VIBE protocol: 1 pre-contrast + at least 5 post-contrast phases, with a planned target temporal resolution of 60–90 seconds per phase (TR/TE 4–5/1.3–2.0 ms, flip angle 9–12°, isotropic 0.8–1.2 mm). Gadolinium chelate was administered intravenously at a dose of 0.1 mmol/kg, injected at a rate of 1.5–2.0 mL/s and flushed with 20 mL saline; the first post-contrast phase was initiated ~15–20 seconds after bolus. Pixel-based subtraction from pre-contrast volumes to post-contrast volumes was applied in all post-processing steps; MPR (axial/sagittal/coronal) and MIP reconstructions were created. Additionally, Dixon-based fat suppression and automatic shimming were preferred to reduce field homogeneity issues and increase fat suppression consistency; breath-hold/relaxed breathing protocols and short phase durations were used to limit motion artefacts.

Intratumoral ADC, peritumoral ADC, ADC ratio, peritumoral T2, peripectoral T2, DWI rim sign, background parenchymal enhancement (BPE), type of enhancement, adjacent vessel sign (adjacent vessel sign), tumor size, multi-focal/multicentric status, and Breast Imaging Reporting and Data System (BI-RADS) categorization.

Breast MRI image analyses were performed by a radiologist with over 5 years of breast MRI experience. For tumor size measurement, the largest diameter was assessed using an electronic digital caliper with DCE-MRG. BPE (minimal/mild or moderate/marked), tumor mass enhancement pattern (mass, non-mass), and BI-RADS category (4–5) were analyzed according to the second edition of the BI-RADS MRI lexicon. When evaluating the adjacent vessel sign in subtraction images, the presence of vessels entering the contrasted lesion or touching the lesion margin was considered a positive adjacent vessel finding. ADC mapping was performed on DWI images based on the lesion spread shown in DCE-MRI. ROIs were placed on the ADC maps as large oval or round ROIs adjacent to the tumor contour and the largest cross-sectional area of the tumor in the breast parenchyma. The peritumoral–tumoral ADC ratio was calculated as peritumoral maximum ADC/tumoral ADC.

Dataset

For this research work, a clinical data set of 204 patients undergoing breast magnetic resonance imaging (MRI) studies has been employed to make predictions about the involvement of the axillary lymph nodes. The data set has 23 independent features including demographic, clinical, radiology, histopathology, and molecular attributes, in addition to one target feature specifying the status of axillary lymph nodes. The involvement of axillary lymph nodes has been treated as a binary target feature, expressed as present (1) or absent (2), following histopathologic analysis. Before model development, several data preprocessing measures have also been employed to manage missing values and to maintain numerical scales in features. The data set has been considered apt for supervised binary classification problems and has been employed to model several machine learning predictors to determine axillary lymph node metastasis. The target feature classes are provided in Table 1.

The dataset for "Breast MRI-Based Lymph Node Prediction" collects information on diverse variables pertaining to the biologic, histopathologic, and molecular aspects of patients' breasts through their MRI analyses. There are 204

Table 1 Class Distribution of the Target Variable

Class Label	Axillary Lymph Node Status	Number of Patients	Percentage (%)
1	Present	116	56.9%
2	Absent	88	43.1%
Total	—	204	100%

patients included, with the dataset consisting of 23 distinctive variables. These variables include demographic information, MRI observations, diffusion imaging metrics, histopathologic types, as well as biomarkers. These variables distinguish between the presence (coded as "1") or absence (coded as "2") of involvement in the axillary lymph node. Visual examples from the same patient from whom the data set was obtained is shown in Figure 1. Figure 1 illustrates multimodal MRI findings of a centrally located malignant mass in the right breast. In Figure 1a, the yellow arrow indicates peritumoral edema observed on T2-weighted imaging, while the blue arrow highlights edema in the prepectoral region on the same sequence. In Figure 1b, the green arrow demonstrates the malignant mass lesion on contrast-enhanced T1-weighted (DCE) images. In Figure 1c, diffusion-weighted imaging (DWI) reveals diffusion restriction of the malignant lesion, indicated by the orange arrow. Additionally, axillary lymph node metastasis (ALNM) is shown in Figure 1c with a purple arrow in the right axilla. In Figure 1d, the red arrow marks the presence of the adjacent vessel sign on maximum intensity projection (MIP) images. Finally, Figure 1e further emphasizes the lesion characteristics on processed imaging, supporting the malignant nature of the mass.

This dataset will aid in the development of models aiming to predict "axillary lymph node metastasis in patients with breast cancer." Further information on the dataset's properties will be found in Table 2.

The predictor set comprises 23 independent variables representing a structured combination of demographic, radiological, diffusion-related, histopathological, and molecular characteristics. Variables were encoded using clinically meaningful binary or categorical schemes (eg, presence/absence indicators, BI-RADS categories, tumour size thresholds), while continuous measurements such as age and ADC-derived parameters were retained in their numeric form. Histopathological type was represented using one-vs-rest binary indicators (IDC, ILC, and other subtypes) to preserve interpretability and reduce ambiguity in multiclass histology information. A detailed description of each feature, including data type, encoding strategy, and clinical interpretation, is provided in Table 3.

Patient Characteristics

The study population consisted of a total of 204 female patients aged between 36 and 82 years (mean age 48.9 years). The MRI quantitative values and clinical-pathological characteristics of ALNM+ and ALNM- patients included in the training and test sets are presented in Table 4.

Table 4 is here

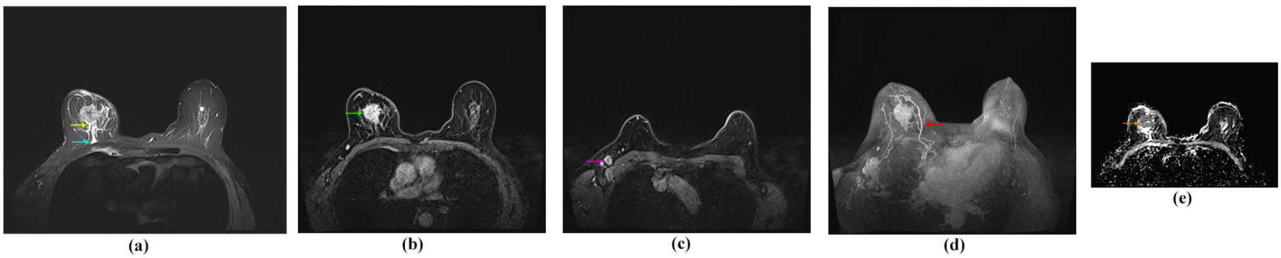

Figure 1 Multimodal MRI findings of a centrally located malignant mass in the right breast. (**a**) T2-weighted image showing peritumoral edema (yellow arrow) and prepectoral edema (blue arrow). (**b**) Contrast-enhanced T1-weighted dynamic contrast-enhanced image demonstrating the malignant mass lesion (green arrow). (**c**) Diffusion-weighted image showing axillary lymph node metastasis in the right axilla (purple arrow). (**d**) Maximum intensity projection image demonstrating the adjacent vessel sign (red arrow). (**e**) Processed image further emphasizing the lesion characteristics of the malignant mass (Orange arrow), supporting its malignant nature.

https://doi.org/10.2147/BCTT.S599211

Table 2 General Characteristics and Feature Categories of the Dataset Used in This Study

Attribute	**Description**
Dataset Name	Breast MRI–Based Lymph Node Prediction Dataset
Data Source	Clinical breast magnetic resonance imaging (MRI) examinations
Total Number of Patients	204
Number of Features	23 independent variables + 1 target variable
Target Variable	Axillary lymph node status
Target Variable Definition	Axillary lymph node involvement (present = 1, absent = 2)
Learning Type	Supervised learning
Problem Type	Binary classification
Data Modality	Clinical, radiological, histopathological, and molecular features
Missing Data Handling	Missing values were addressed during the preprocessing stage
Labelling Method	Based on histopathological and clinical evaluation
Intended Use	Prediction of axillary lymph node metastasis using machine learning methods
Feature Category	**Description**
Demographic Features	Patient age
Clinical Characteristics	Tumour laterality, tumour size, BI-RADS score, multifocal/multicentric status
MRI Findings	Background parenchymal enhancement (BPE), enhancement pattern, peritumoral/prepectoral edema, diffusion-weighted imaging (DWI) rim sign
Diffusion Parameters	Intratumoral ADC, peritumoral ADC, ADC ratio
Histopathological Features	Invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), and other histological subtypes
Molecular Biomarkers	Estrogen receptor (ER), progesterone receptor (PR), Ki-67 index, molecular subtype
Pathological Findings	Presence of DCIS, tumour grade
Clinical Characteristics	Tumour laterality, tumour size, BI-RADS score, multifocal/multicentric status
MRI Findings	Background parenchymal enhancement (BPE), enhancement pattern, peritumoral/prepectoral edema, diffusion-weighted imaging (DWI) rim sign
Diffusion Parameters	Intratumoral ADC, peritumoral ADC, ADC ratio

Table 3 Description of the Independent Variables Used in the Study

Feature Name	**Type**	**Encoded Values/Range**	**Description**
Age	N	Continuous (years)	Age of the patient at the time of MRI examination
Tumour Laterality	C	Right = 1, Left = 2	Laterality of the breast tumour
BI-RADS Score	C	BI-RADS 4 = 1, BI-RADS 5 = 2	Radiological assessment category indicating malignancy likelihood
Background Parenchymal Enhancement (BPE)	C	Minimal/Mild = 1, Moderate/Marked = 2	Degree of background parenchymal enhancement on breast MRI
Enhancement Pattern	C	Mass = 1, non-mass = 2	Type of contrast enhancement observed on MRI
Adjacent Vessel Sign	C	Present = 1, Absent = 2	Presence of adjacent vessel sign on MRI
Peritumoral Edema	C	Present = 1, Absent = 2	Presence of edema surrounding the tumour
Prepectoral Edema	C	Present = 1, Absent = 2	Presence of edema in the prepectoral region
DWI Rim Sign	C	Present = 1, Absent = 2	Presence of rim sign on diffusion-weighted imaging
Intratumoral ADC	C	Continuous	Apparent diffusion coefficient (ADC) measured within the tumour
Peritumoral ADC	N	Continuous	ADC value measured in the peritumoral region
ADC Ratio (Peritumoral/Intratumoral)	N	Continuous	Ratio of peritumoral ADC to intratumoral ADC values

(*Continued*)

Table 3 (Continued).

Feature Name	Type	Encoded Values/Range	Description
DCIS Accompaniment	C	Present = 1, Absent = 2	Presence of ductal carcinoma in situ accompanying the invasive tumour
Tumour Size	C	< 20 mm = 1, ≥ 20 mm = 2	Maximum tumour diameter based on imaging
Multifocal/Multicentric Disease	C	Present = 1, Absent = 2	Presence of multifocal or multicentric tumour lesions
Tumour Grade	C	Low grade (1–2) = 1, High grade (3) = 2	Histopathological tumour grade
Estrogen Receptor (ER) Status	C	Positive = 1, Negative = 2	Estrogen receptor expression status
Progesterone Receptor (PR) Status	C	Positive = 1, Negative = 2	Progesterone receptor expression status
Ki-67 Index	C	< 20% = 1, ≥ 20% = 2	Cellular proliferation index
Molecular Subtype	C	Luminal A = 1, Luminal B = 2, HER2+ = 3, Triple-negative = 4	Molecular classification of breast cancer
Histopathological Type – IDC	B	Yes = 1, No = 0	Invasive ductal carcinoma indicator
Histopathological Type – ILC	B	Yes = 1, No = 0	Invasive lobular carcinoma indicator
Histopathological Type – Other	B	Yes = 1, No = 0	Other histopathological subtypes

Abbreviations: N, Numerical; C, Categorical; B, Binary.

Table 4 Description of the Patient Characteristics Used in the Study

Features	ALNM+(116)	ALNM-(88)
Age (years)	53,3 (36–82)	54,7(34–82)
Direction	-	-
Right	66(%59,4)	45(%40,9)
Left	50(%53,7)	43(%46,2)
Molecular subtype	-	-
Luminal A	80(%54,7)	66((%45,2)
Luminal B	9(%50)	9(%50)
HER-2 overexpression	18(%81,8)	4(%18,1)
Triple-negative	9(%50)	9(%50)
ER status	-	-
Positive	101(%55,8)	80(%44,1)
Negative	15(%65,2)	8(%34,7)
PR	-	-
Positive	98(%56,6)	75(%43,3)
Negative	18(%58)	13(%41,9)
Histologic grade	-	-
High grade	28(%80)	7(%20)
Low grade	88(%52)	81(%47,9)
Ki-67	-	-
High expression	49(%74,2)	17(%25,7)
Low expression	67(%48,5)	71(%51,4)
Multifocal/multicentric	-	-
Present	49(%73,1)	18(%26,8)
Absent	67(%48,9)	70(%51)
Tumor size	-	-
20mm>	32(%30,7)	72(%69,3)
20mm<	84(%84)	16(%16)

(*Continued*)

Table 4 (Continued).

Features	ALNM+(116)	ALNM-(88)
DCIS	-	-
Present	87(%62,5)	52(%37,4)
Absent	29(%44,6)	36(%55,3)
Histopathological type	-	-
IDC	67(%57.8)	45(%51.1)
ILC	22(%59,4)	15(%40,5)
Other	6(%30)	14(%70)
Combined	21(%60)	14(%40)
DWI	-	-
Tumoral ADC (10−6mm2/s)	823,6	766,3
Peritumoral maximal ADC (10−6mm2/s)	1538,6	1341,02
Peritumoral-tumoral ADC ratio	3,57	1,84
DWI rim sign	-	-
Present	81(%76,4)	25(%23,5)
Absent	35(%35,7)	63(%64,2)
Prepectoral edema	-	-
Present	24(%96)	1(%4)
Absent	92(%51,3)	87(%48,6)
Peritumoral edema		-
Present	63(%76,8)	19(%23,1)
Absent	53(%43,4)	69(%56,5)
Adj vessel sign	-	-
Present	60(%77,9)	17(%22)
Absent	56(%44)	71(%55,9)
Type of enhancement	-	-
Mass	74(%54,8)	61(%45,1)
Nonmas	42(%54,5)	35(%45,4)
BPE	-	-
Minimal/mild	46(%44,6)	57(%55,3)
Moderate /marked	70(%69,3)	31(%30,6)
BI-RADS	-	-
4	15(%31,9)	32(%68)
5	101(%64,3)	56(%35,6)

Abbreviations: ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; DCIS, ductal carcinoma in situ; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma.

Proposed Model CCIRM (Calibrated & Conformal Clinical Risk Model)

This study proposes a hybrid decision support model that offers reliable probability estimates (calibration) and uncertainty awareness (conformal prediction) with high discriminatory power (AUC/F1) for predicting axillary lymph node status from tabular clinical-pathological variables. The proposed method simultaneously addresses two needs that are increasingly prominent in clinical risk modelling in the literature. The first is not only to predict the correct class label, but also to ensure that the predicted probabilities reflect the true event probabilities (calibration). The second is to reduce overconfidence in clinical decision-making by clearly flagging situations where the model is "uncertain" (conformal uncertainty sets).

The modelling pipeline excludes fields such as identity/name from the raw dataset that do not directly contribute to clinical prediction. As the "histopathological type" variable contains multi-categorical content, it is converted to a binary representation (one-hot/binary indicator) for IDC/ILC/Other before being input into the model. Additionally, for software compatibility and reproducibility, all attribute names are standardised (removal of special characters, conversion of

spaces to underscores, etc). Non-numeric fields are forced to numeric type, and missing values are handled with median imputation to ensure consistency across models. These steps reduce the risk of data leakage while enabling different model families to operate in a comparable manner on the same data representation.

To minimize data leakage, preprocessing operations were incorporated into a unified modelling workflow and applied separately within the relevant training process. Missing-value imputation was performed using median-based imputation, and feature scaling was applied only for model families requiring scale-sensitive input, such as linear and kernel-based classifiers. Accordingly, preprocessing parameters were derived from the corresponding training data only and were not estimated using calibration or test data. No preprocessing parameter was estimated from the calibration or test subsets. All imputations, scaling operations, and model-specific transformations were fitted using the relevant training data only and then applied to the held-out subsets, thereby preserving the independence of calibration and final test evaluation.

In the proposed approach, the data is divided into three parts: train, calibration, and test. The train part is used for model selection and basic model training. The calibration part is reserved for probability calibration and (if applicable) learning conformal thresholds. The test part is used solely for the final performance report. The aim of this design is to prevent the hyperparameter search/learning process from artificially improving test performance and to perform the calibration step on a separate hold-out set independent of the distribution on which the model was trained. Thus, the test set assumes a more reliable role as "unseen data" for both classification performance and the evaluation of calibration/conformal properties.

A single model family is not optimal in all cases for clinical tabular data. Therefore, the method defines a candidate pool consisting of model families with different inductive biases. The candidate pool includes Logistic Regression, SVC, Random Forest/Extra Trees, and HistGradientBoosting. Candidates are screened only on TRAIN using RandomisedSearchCV in a Stratified K-fold scheme, and ROC-AUC is used as the target metric. This choice is appropriate for measuring discrimination power independent of the decision threshold, particularly in the context of clinical problems where class imbalance may occur. The selected best model is transferred to the calibration phase by model selection-induced optimism bias is reduced. Probability calibration was then applied only to the selected model using the independent calibration subset, so that calibration quality could be assessed without contaminating model selection or final test evaluation.

Hyperparameter optimisation was performed using RandomizedSearchCV under stratified k-fold cross-validation. The number of cross-validation folds, the number of randomized search iterations, and the predefined search ranges for each candidate algorithm were specified in advance to ensure reproducibility, and the corresponding model-specific settings are summarised in the relevant section.

In the present framework, the nonconformity score was defined as one minus the calibrated probability assigned to the candidate class. Class-specific thresholds were then estimated from the calibration set using the empirical quantile associated with the predefined significance level. A class label was included in the prediction set if its nonconformity score did not exceed the corresponding class-conditional threshold. This class-based thresholding strategy was preferred to avoid the limitations of a single global threshold in the presence of class-specific calibration differences.

The reliability of model probabilities is critical in clinical decision-making processes. Therefore, the probabilities generated by the base model are calibrated on the calibration set using CalibratedClassifierCV (cv="prefit"). Both the Sigmoid (Platt scaling) and Isotonic regression methods are tested. The calibration method is selected based on the LogLoss (negative log-likelihood) on the calibration set. LogLoss is suitable for measuring calibration quality because it penalises not only correct/incorrect decisions but also how "accurate" the probabilities are. The calibration method with the lowest LogLoss is selected to obtain the final "calibrated model". This step was considered essential because clinically meaningful risk interpretation depends not only on the predicted class, but also on how closely predicted probabilities correspond to observed event frequencies. In this respect, calibration improves the interpretability of estimated risk scores and reduces the likelihood of overconfident probability outputs being used in a misleading manner.

In clinical practice, identifying "uncertain" samples is valuable in scenarios where the costs of false negatives/false positives are high. To this end, the method applies Mondrian (class-conditional) conformal prediction on calibrated probabilities.

The nonconformity score is calculated using Equation 1.

$$s = 1 - p(\text{true class}) \tag{1}$$

A separate threshold is calculated for each class using Equation 2.

$$q_c = Q_{1-\alpha}(s \mid y = c) \tag{2}$$

This approach produces a fairer and more balanced "certain/uncertain" distribution by setting class-based thresholds instead of a single global threshold, particularly when class imbalances exist or classes exhibit different calibration behaviours. Results are reported using metrics such as coverage (the proportion of the true class within the set) and uncertain rate (|S|>1). Thus, the model clarifies uncertainty situations rather than making a "single decision" in a clinical scenario. The CCIRM process architecture is shown in Figure 2.

The classic "train one model – report accuracy" approach produces three fundamental problems in clinical applications. These problems and the motivating factors behind the development of this model are listed below.

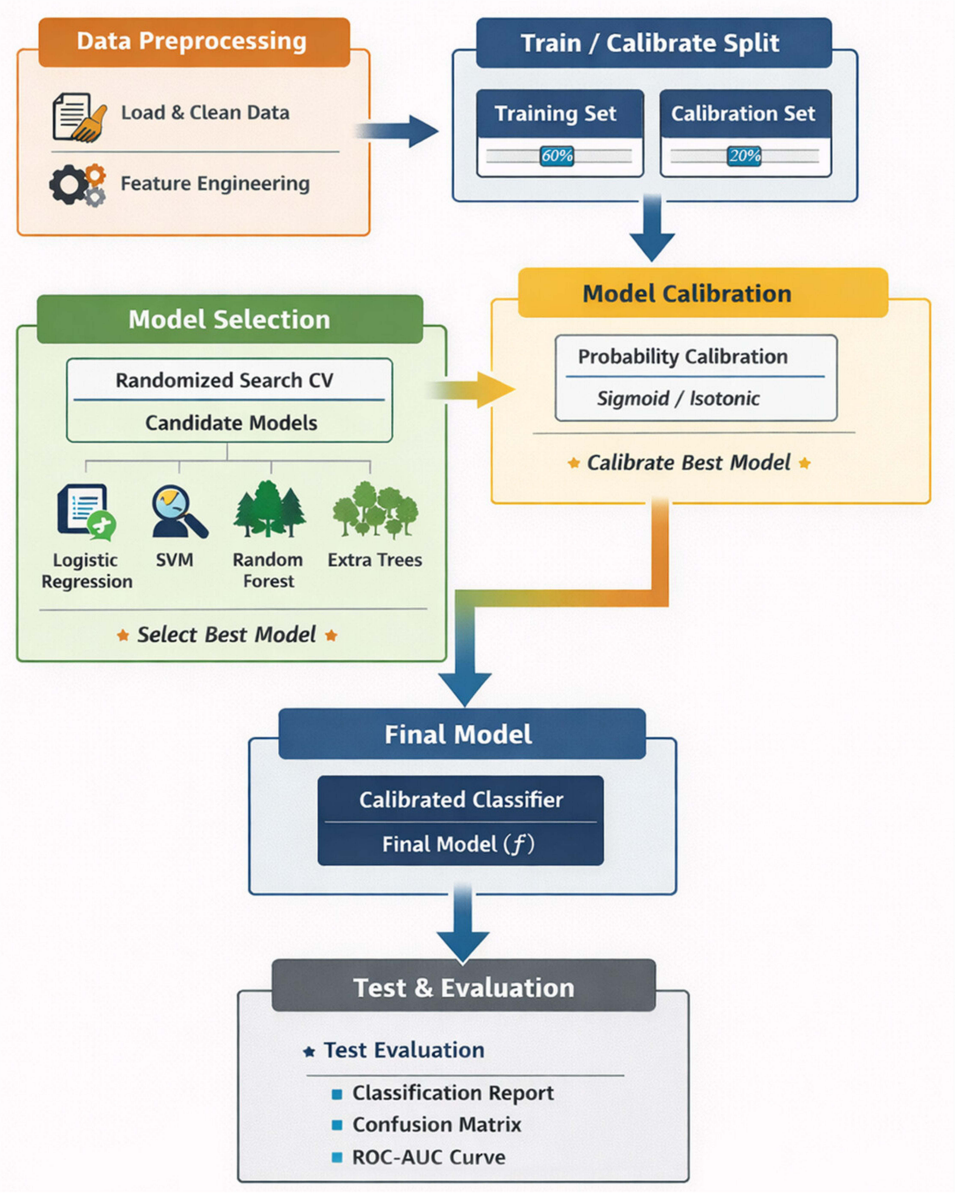

Figure 2 Workflow of the CCIRM classification framework, including data preprocessing, train/calibration split, model selection, probability calibration, final model construction, and test evaluation.

- Model selection bias and generalisability risk: A fixed algorithm is not optimal for every dataset. Train-only selection from a candidate model family allows for selecting the inductive bias best suited to the data representation.
- Uncalibrated probabilities: Although ensemble trees such as ExtraTrees/RandomForest can provide high accuracy, the probabilities they produce are often not "well calibrated". This is a critical weakness when making probabilistic interpretations in clinical decisions, such as stating a "70% risk". The proposed approach corrects the probabilities using a calibration set and increases reliability by selecting with LogLoss.
- Invisibility of uncertainty: The standard ExtraTrees model produces a single label for each example and does not naturally flag cases where it is "uncertain." Conformal prediction separates "certain/uncertain" examples with a defined coverage guarantee target. This is particularly valuable in clinical decision support systems from a safety and ethical perspective.

The scientific contribution of this approach lies in integrating the dimensions of prediction reliability and uncertainty awareness into the model, going beyond the goal of "high accuracy with a single metric". Current clinical artificial intelligence literature emphasises the necessity of calibration and uncertainty quantification, particularly in risk score generation, from the perspective of clinical safety. By combining the triad of separation (AUC/F1) + calibration (LogLoss/reliability) + uncertainty sets (coverage/efficiency) into a single end-to-end pipeline, it offers a safer, more reliability-oriented, and more clinically cautious framework for decision support. This transforms the method from merely a "high-performance classifier" into a risk-focused, reliable, and clinically applicable family of models. CCIRM pseudo code is shown in Table 5.

Table 5 CCIRM Pseudo Code

Algorithm 1.	*CCIRM: Calibrated & Conformal Clinical Risk Model (Train-Only Search + Hold-out Calibration + Mondrian Conformal)*
Input:	Tabular dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ Target keyword for label column (e.g, "axillar") Random seed r Split ratios: p_{test}, p_{calib} Candidate model set $\mathcal{M}$ with hyperparameter spaces $\{\Omega_m\}$ Search budget T(number of randomized trials), CV folds K Calibration methods $\mathcal{C} = \{\text{sigmoid}, \text{isotonic}\}$ Conformal significance level α(optional) Positive label y^+(for ROC-AUC)
Output:	Calibrated predictive model $\hat{f}$(probabilistic) Test metrics: classification report, confusion matrix, ROC-AUC, LogLoss Mondrian conformal prediction sets $\Gamma_\alpha(x)$+ coverage/uncertainty
Procedure:	
Load and preprocess dataset	Read Excel file into dataframe df. Drop identifier/name columns (e.g, ID, "name"). Locate target label column y using the keyword. If "histopathologic type" exists encode into three binary indicators $x^{\text{hist}} \rightarrow (\text{hist\_IDC}, \text{hist\_ILC}, \text{hist\_OTHER})$ and remove original histopathology column.Sanitize feature names (remove special characters; enforce uniqueness). Convert non-numeric columns to numeric (coerce invalid to NaN). Drop suspicious features containing predefined keywords. Define $X \leftarrow \text{df} \setminus \{y\}$.
Split data into train/calib./ test	Stratified split: $(X_{\text{traincal}}, y_{\text{traincal}}), (X_{\text{test}}, y_{\text{test}}) \leftarrow \text{Split}(X, y, p_{\text{test}}, r)$ Stratified split: $(X_{\text{train}}, y_{\text{train}}), (X_{\text{calib}}, y_{\text{calib}}) \leftarrow \text{Split}(X_{\text{traincal}}, y_{\text{traincal}}, p_{\text{calib}}, r)$
Train-only model selection via randomized search	Initialize best score $\text{s} \leftarrow -\infty$; best model$\text{m} \leftarrow \emptyset$.For each candidate model $m \in \mathcal{M}$: Define preprocessing pipeline P_m(median imputer; scaling for linear/kernel models). b) Run Randomized Search CV on TRAIN only: $\hat{\theta}_m \leftarrow \arg\max_{\theta \sim \Omega_m}^{T} \text{CVScore}(\text{ROC} - \text{AUC}, P_m(\theta), X_{\text{train}}, y_{\text{train}}, K)$ c) If obtained CV score s_m>s: update $\text{s} \leftarrow s_m, \text{s} \leftarrow P_m(\hat{\theta}_m)$

(*Continued*)

https://doi.org/10.2147/BCTT.S599211

Table 5 (Continued).

Fit selected base model on TRAIN	f←Fit(m, X_{train}, y_{train})
Hold-out probability calibration on CALIB (choose best by LogLoss)	For each calibration method $c \in \mathcal{C}$: a) Fit calibrated model using prefit base:$f_c \leftarrow \text{Calibrate}(f, c, X_{calib}, y_{calib})$ b) Compute calibration LogLoss on CALIB: $\ell_c \leftarrow \text{LogLoss}(y_{\text{calib}}, f_c(X_{\text{calib}}))$ Select best calibration: $\hat{f} \leftarrow \arg\min_{c \in \mathcal{C}} \ell_c$
Evaluate on TEST	Predict class labels: $\hat{y}_{\text{test}} \leftarrow \hat{f}(X_{\text{test}})$ Predict probabilities: $\hat{p}_{\text{test}} \leftarrow \hat{f}_{\text{proba}}(X_{\text{test}})$ Compute: classification report, confusion matrix (fixed class order), ROC-AUC (using y^+), and LogLoss.
Mondrian conformal prediction sets	Obtain calibrated probabilities on CALIB:$\hat{p}_{\text{calib}} \leftarrow \hat{f}_{\text{proba}}(X_{\text{calib}})$ For each class k: compute class-conditional nonconformity scores $s_i = 1 - \hat{p}_{\text{calib}}(y_i \mid x_i), \quad i \in \{i : y_i = k\}$ and quantile threshold: $q_k \leftarrow Q_{1-\alpha}(\{s_i\}_{y_i=k})$ For each test instance x, define prediction set:$\Gamma\alpha(\text{x}) = \{k : \hat{p}_{test}(k\|x)) \geq \ 1 - q_k\}$ If $\Gamma_\alpha(x) = \emptyset$, set $\Gamma_\alpha(x) = \{\text{argmax}_k \hat{p}_{\text{test}}(k \mid x)\}$. Report conformal coverage and uncertainty rate:$\text{Coverage} = \frac{1}{\|D_{\text{test}}\|} \sum^1 [y \in \Gamma_\alpha(x)], \text{Uncertainty} = \frac{1}{\|D_{\text{test}}\|} \sum^1 [\| \Gamma_\alpha(x) \| > 1]$

Statistical Evaluation of Discrimination, Calibration, and Uncertainty

Model performance was assessed from three complementary perspectives: discrimination, calibration, and predictive uncertainty. Discrimination performance on the independent test subset was evaluated using ROC-AUC, accuracy, precision, recall, and F1 score. Calibration quality was assessed to determine whether predicted probabilities were consistent with observed outcomes, and where possible this evaluation may be supported by metrics such as log loss, Brier score, and calibration plots. In addition, the uncertainty-aware component of the framework was assessed using conformal prediction outputs, including empirical coverage, class-conditional coverage, uncertainty rate, and prediction set size. This evaluation strategy was adopted to ensure that the proposed framework was examined not only as a classifier, but also as a clinically interpretable and reliability-oriented decision-support model.

Multiple machine learning algorithms were implemented and evaluated using predefined hyperparameter configurations to ensure a fair and reproducible comparison. The proposed CCIRM framework integrates probability calibration (sigmoid and isotonic) with conformal prediction, while conventional classifiers such as SVC, Random Forest, Extra Trees, CatBoost, XGBoost, and Decision Tree were configured using commonly accepted parameter settings from the literature. For all models, a consistent data partitioning strategy and fixed random seed were applied to eliminate variability arising from data splits. The complete set of algorithms and their corresponding hyperparameter values are summarized in Table 6.

The motivation to use RandomizedSearchCV over the exhaustive grid search approach was based on the high dimensionality of the search space, as well as the fact that the data sample size was not large. An exhaustive grid search algorithm searches over the whole parameter space, which becomes costly in terms of computation, especially in the case of a high-dimensional search space. It has been observed that RandomizedSearchCV compares favorably to grid search with a lot fewer function evaluations, especially if a significant number of the involved hyperparameters are not driving the overall problem, meaning the objective function. Moreover, as the experimental design here strictly enforced that the models be selected on the basis of the train data alone to avoid the information-leakage problem, the randomized search method provides a reasonable trade-off between the computational cost and the optimization of the hyperparameters. This observation has been ascertained by the existing observation that randomized search outperforms grid search in the case of a typical clinical application of a machine-learning algorithm.

ROC-AUC was set as the primary optimization objective because of its threshold-independent way of evaluating the discriminatory capability of the model. This characteristic is of prime importance in a clinical setting, where the model is intended to aid decision-making. Accuracy, on the other hand, being a threshold-dependent evaluation criterion, is not ideal for this purpose. Accuracy can be biased if the cost of false positives versus false negatives differs in a particular clinical setup, which in the case of axillary lymph node prediction, it did. Therefore, it was easy to generate models based

Table 6 Machine Learning Algorithm Hyperparameter Values

Algorithms	Hyperparameter Values
CCIRM	"Calib_Methods": ["Sigmoid", "Isotonic"], "Alpha": 0.10, "Random_State": 42, Calibration_Bins": 10, "Test_Size": 0.20, "Calib_Size": 0.20
SVC	"Test_Size": 0.2, "Random_State": 42, "Kernel": "Rbf", "C": 1.0, "Gamma": "Scale", "Degree": 3, "Probability": True
Random Forest	"Test_Size": 0.2, "Random_State": 42, "N_Estimators": 500, "Max_Depth": None, "Min_Samples_Split": 2, "Min_Samples_Leaf": 1, "Max_Features": "Sqrt", "Class_Weight": None, "N_Jobs": −1
Extra Trees	"Test_Size": 0.2, "Random_State": 42, "N_Estimators": 500, "Max_Depth": None, "N_Jobs": −1, "Tree_Step": 50
CatBoost	"Test_Size": 0.2, "Random_State": 42, "Iterations": 500, "Learning_Rate": 0.05, "Depth": 6, "L2_Leaf_Reg": 3.0, "Verbose": 50
XGBoost	"Test_Size": 0.2, "Random_State": 42, ", 'Learning_Rate': 0.05, 'Max_Depth': 6, Subsample": 0.9, "Colsample_Bytree": 0.9, "Reg_Lambda": 1.0, "N_Jobs": −1, "Verbose": 50
Decision Tree	"Test_Size": 0.2, "Random_State": 42, "Max_Depth": None, "Min_Samples_Split": 2, "Min_Samples_Leaf": 1, "Class_Weight": None, "Depth_Sweep_Max": 25

on the use of ROC-AUC as the primary optimization criterion, whereas accuracy, precision, recall, F1, etc. were measured on the independent test data.

To ensure a stringent and leak-proof evaluation pipeline, the entire dataset was split into three distinct subsets—training, calibration, and testing. The fitting of the model and tuning of the hyperparameters were done exclusively with the training set, whereas the calibration set had a different role for the probability calibration and the conformal threshold estimation. Only the final test set was used for performance evaluation in an unbiased manner. The chosen splitting ratios (around 60% training, 20% calibration, and 20% testing) were meant to give a fair chance to the three conflicting requirements: a large enough data set for the model to learn robustly, a small enough data set for the probability calibration to be reliable, and an independent test set large enough to provide stable performance estimates.

Results

The results obtained from the Breast MRI-Based Lymph Node Prediction Dataset demonstrate that the evaluated machine learning models achieved varying levels of predictive performance on the independent test subset. Among these models, the proposed Conformal-Calibrated Interpretable Risk Model (CCIRM) achieved the highest overall classification performance, with a test accuracy of 0.9268 and a weighted F1 score of 0.9270. CCIRM also maintained a favourable balance between precision and recall across both classes, with a precision of 0.9545 and a recall of 0.9130 for the ALNM-positive class.

Compared with conventional classifiers, CCIRM showed the strongest overall test performance. The SVC model with RBF kernel ranked second, with a test accuracy of 0.9024, whereas Random Forest and Extra Trees achieved test accuracies of 0.8780 and 0.8537, respectively. CatBoost and XGBoost produced similar performance levels, both yielding a test accuracy of 0.8293, while the standalone Decision Tree model showed the weakest overall performance, with a test accuracy of 0.7073. Taken together, these findings indicate that while several modern classifiers provided reasonable discrimination, the integrated calibration and conformal framework of CCIRM was associated with the most favourable overall performance profile in the present test cohort.

The standalone Decision Tree model was the least performing one of all the tested algorithms with a test accuracy of 0.7073 and a macro F1-score of 0.706, thus pointing out the incapacity of shallow non-ensemble models to represent the complex and nonlinear relations that are characteristic of multimodal clinical and radiological data. All in all, the findings suggest that although the cutting-edge ensemble and kernel-based techniques offer quite a good predictive ability, the incorporation of probability calibration and conformal prediction within the CCIRM framework results in excellent discrimination and stability. This improvement is particularly evident in the positive class recall and overall accuracy, underscoring the clinical and methodological value of the proposed approach is shown in Table 7.

Table 7 Machine Learning Algorithm Performance Metric Values

Algs.	Params.	Evaluation Metrics				
CCIRM	**Info**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.9545	0.9130	0.9333	23	0.9268
	Negative	0.8947	0.9444	0.9189	18	
	Accuracy	–	–	0.9268	41	
	Macro Avg.	0.9246	0.9287	0.9261	41	
	Weighted Avg.	0.9287	0.9268	0.9270	41	
SVC	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.952	0.870	0.909	23	0.9024
	Negative	0.850	0.944	0.895	18	
	Acc.	-	-	0.902	41	
	Macro Avg.	0.901	0.907	0.902	41	
	Weighted Avg	0.907	0.902	0.903	41	
Random Forest	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.909	0.870	0.889	23	0.878
	Negative	0.842	0.889	0.865	18	
	Acc.	-	-	0.878	41	
	Macro Avg.	0.876	0.879	0.877	41	
	Weighted Avg	0.880	0.878	0.878	41	
Extra Trees	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.905	0.826	0.864	23	0.8537
	Negative	0.800	0.889	0.842	18	
	Acc.	-	-	0.854	41	
	Macro Avg.	0.852	0.857	0.853	41	
	Weighted Avg	0.859	0.854	0.854	41	
CatBoost	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.900	0.783	0.837	23	0.8293
	Negative	0.762	0.889	0.821	18	
	Acc.	-	-	0.829	41	
	Macro Avg.	0.831	0.836	0.829	41	
	Weighted Avg	0.839	0.829	0.829	41	
XGBoost	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.864	0.826	0.844	23	0.8293
	Negative	0.789	0.833	0.811	18	
	Acc.	-	-	0.829	41	
	Macro Avg.	0.827	0.830	0.828	41	
	Weighted Avg	0.831	0.829	0.830	41	
Decision Tree	**Info.**	**Precision**	**Recall**	**F1-Score**	**Support**	**Test Acc.**
	Positive	0.762	0.696	0.727	23	0.7073
	Negative	0.650	0.722	0.684	18	
	Acc.	-	-	0.707	41	
	Macro Avg.	0.706	0.709	0.706	41	
	Weighted Avg	0.713	0.707	0.708	41	

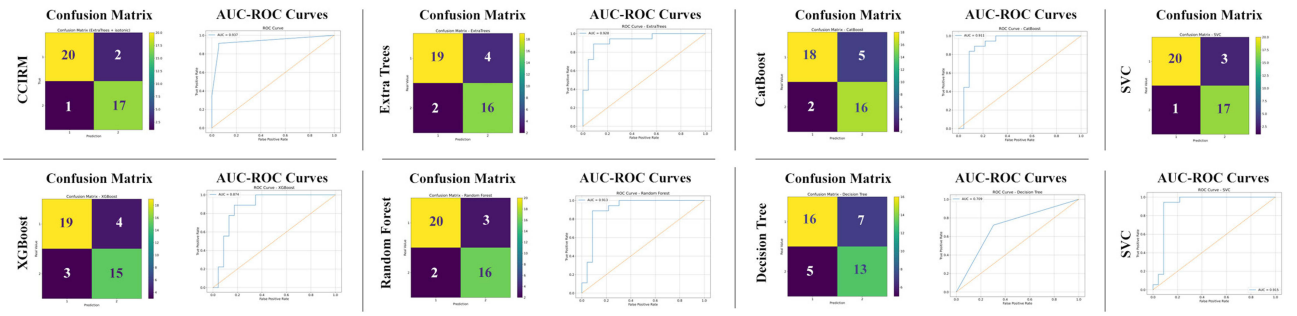

Figure 3 Confusion matrices and ROC curves of the machine learning models evaluated for CCIRM classification, including Extra Trees with isotonic calibration, Extra Trees, CatBoost, SVC, XGBoost, Random Forest, and Decision Tree.

Calibration and Uncertainty-Aware Performance of CCIRM

The framework not only enhances discrimination performance. The proposed framework was also intended to improve reliability and transparency through better probability calibration and conformal prediction. Essentially, this means that CCIRM is not limited to generating class labels but rather designed to provide more clinically interpretable risk estimates, while also signifying when model predictions should be considered uncertain. In the context of the pre-operative ALNM assessment, it is particularly important to be aware of the presence of uncertainty because an overconfident prediction in borderline cases may be difficult to justify. Thus, the methodological value of CCIRM is not only reflected in its competitive classification performance but it also emphasizes calibrated risk estimation and explicit uncertainty processing. A confusion matrix visualizes the model's performance according to classes. It does so by showing the model's true positive, true negative, false positive, and false negative predictions. Also, it helps us to understand the prediction error distribution for each class in detail. Similarly in receiver operating characteristic (ROC) curves the area under the curve (AUC) is the measure of how well a model discriminates between classes measured over a range of thresholds; it is a threshold- independent measure of classification performance. The confusion matrices and ROC–AUC curves obtained from all evaluated machine learning models are illustrated in Figure 3 of this study.

The confusion matrix and ROC curves provide different perspectives on performance of the model. The confusion matrices gave a class-specific ability to assess the true positive, true negative, false positive, and false negative predictions made by a model. In contrast, the ROC analysis quantified the threshold-independent discrimination ability of each model. Out of all evaluated algorithms, CCIRM generated relatively few false-positive and false-negative predictions. Clinically significant conclusions can be drawn from missed metastatic staging and overcalling of nodal involvement in clinical treatment planning. The findings of the confusion matrix corroborated that the ROC-AUC value of CCIRM was the highest at 0.937 among all methods indicating that CCIRM performed the best in overall class-separation performance across different decision thresholds.

Kernel-based SVC gets a pretty good score of 0.915 and a confusion matrix similar to Random Forest; still, the error pattern shows a bit falser negative than CCIRM. That makes SVC less perfect for cases were missing a true positive can hurt and better for cases where a true positive can be easier to miss even if we see it in the image. Decision Tree has the worst score with a score of 0.709 and a confused confusion matrix, showing that one tree alone cannot get the complex nonlinear signals from multi modal breast MRI data.

In summary, the combined analysis of confusion matrices and ROC curves shows the methodological benefits of the CCIRM method. By combining probability calibration with uncertainty-aware modelling, CCIRM not only enhances overall discrimination but also creates a more clinically relevant pattern of errors compared to standard machine learning tools. These results, shown in Table 7, support the idea that CCIRM is better for both prediction and clinical decision support.

Discussion

In this retrospective study, we developed an uncertainty-aware machine learning framework for preoperative prediction of axillary lymph node metastasis by integrating breast MRI-derived and clinicopathological variables. The contribution

https://doi.org/10.2147/BCTT.S599211

of the present work is methodological as well as predictive, because the proposed CCIRM framework was designed not only to achieve strong discrimination performance, but also to provide calibrated probability estimates and explicit uncertainty-aware outputs. In contrast to conventional classifiers that return only point predictions, CCIRM enables risk estimates to be interpreted with greater caution and transparency, which may be particularly valuable in clinically borderline cases.

The accurate assessment of ALNM in the preoperative period is of critical importance for prognosis and treatment decisions in breast cancer. Numerous studies have been conducted on this subject. In a study by Zhang et al, DCE MRI, DWI, and the combined DCE MRI+DWI model were compared. The AUC values for DCE MRI + DWI and combined models were determined to be 0.793, 0.774, and 0.864, respectively, and a significant difference was found between the predictive efficiencies of the combined models.[11] Although combined models have complementary advantages, the data coverage is insufficient for ALNM prediction.

In another study featuring T2-weighted, fat-saturated T1-weighted, and DCE MRI sequences, the AUC value was 0.781.[12]

In yet another study, the AUC value was reported as 0.82 in a model developed using genetic and clinicopathological data in addition to radiomic analysis of DCE MRI images.[13]

Studies have been conducted using radiomic analysis to evaluate the peritumoral region based on MRI images.[14,15,17] Liu et al found that in DCE MRI images, the AUC of the combined model was 0.867 in peritumoral, intratumoral, and combined radiomic studies, which was higher than the other two models, and emphasised the importance of the peritumoral region in ALNM prediction.[14] This study evaluated a single sequence image and did not evaluate clinical-pathological data and data from DCE MRI and T2-weighted sequences, which are important markers for ALNM in MRI examinations. In another study emphasising the importance of the peritumoral region, a fusion nomogram model incorporating radiomic signatures and habitat signatures, which evaluated the 4 mm peritumoral region and clinical-pathological data, was found to perform best in the training set (AUC: 0.977).[17] However, the most significant limitation of this study is that it only involves radiomic analysis of DCE MRI images. T2-weighted and DWI images were not evaluated. However, the importance of the peritumoral region has also been demonstrated in the literature; Peritumoral oedema detected on T2-weighted MRI is associated with invasive tumour characteristics and poor prognosis, while DWI/ADC measurements (particularly the peritumour-tumour ADC ratio and peritumoral ADC values) provide prognostic information related to invasion and metastasis.[17–19]

Although many previous ALNM prediction studies have reported promising discrimination performance, most comparisons in the literature have been based primarily on AUC and related classification metrics. By contrast, calibration quality and predictive uncertainty have been addressed less consistently. For this reason, the present study should be interpreted not simply as another performance-oriented classifier comparison, but as a methodological effort to incorporate reliability and uncertainty awareness into ALNM prediction. Accordingly, comparison with earlier studies should consider differences not only in discrimination metrics, but also in whether model outputs are calibrated and whether uncertainty is explicitly represented.

In this study, seven different models were developed using quantitative data obtained from MRI sequences and clinical-pathological data, with the aim of non-invasively predicting ALNM in patients with invasive breast cancer at an early stage. Among the developed models, CCIRM demonstrated robust performance and achieved the highest area under the curve (AUC = 0.937) among all tested algorithms. Furthermore, this model provided a test accuracy of 0.9268 and a weighted F1 score of 0.9270. It showed a high balance between precision and sensitivity, achieving precision and sensitivity values of 0.9545 and 0.9130, respectively, for the positive class ALNM. Its most notable difference from the other developed models is its high sensitivity. Importantly, the added value of CCIRM should be understood not only in terms of improved discrimination, but also in terms of its ability to provide calibrated risk estimates and uncertainty-aware outputs within a clinically interpretable evaluation framework.

Beyond its discrimination performance, the proposed framework may offer additional methodological and clinical value by explicitly incorporating predictive uncertainty into preoperative ALNM assessment. In breast cancer care, both false-negative and false-positive predictions may influence surgical planning and treatment decisions.[3,4] Therefore, a model that can identify not only predicted class labels but also cases with greater uncertainty may be more useful

than a purely deterministic system. In this respect, the conformal component of CCIRM may help distinguish between higher-confidence predictions and cases that require more cautious interpretation, thereby supporting radiological assessment rather than replacing it.[5,6,10] Such uncertainty-aware outputs may be particularly relevant in routine workflow, where borderline cases may benefit from closer image review, multidisciplinary discussion, or additional clinicopathological correlation.[4,8,18,19]

Our findings should also be interpreted in the context of previous radiomics and deep learning studies on ALNM prediction. Although many prior studies have reported promising AUC values using DCE-MRI, multiparametric MRI, radiomics, and deep learning approaches,[5–7,11–15,17,18] calibration quality and predictive uncertainty have been less consistently incorporated into these models.[5,6,10] Accordingly, the contribution of the present study lies not only in predictive performance, but also in its emphasis on reliability and uncertainty-aware clinical decision support. Nevertheless, because the present study was retrospective, single-centre, and based on a relatively small test cohort, the current results should be regarded as promising but preliminary until confirmed by prospective, multicentre, and externally validated studies.[5,6,10] Although the present framework was designed to support uncertainty-aware interpretation, the current manuscript provides stronger evidence for discrimination performance than for fully quantified conformal efficiency. Therefore, the uncertainty-aware contribution should be viewed as an important methodological direction of the framework, while further quantitative conformal reporting will be necessary in future studies to establish the magnitude of this advantage more precisely.

Although the performance observed in the present cohort was strong, direct comparison with previously published studies should be interpreted cautiously because of differences in study design, patient population, MRI protocols, feature sets, and validation strategies. This is because our study is based on predicting ALNM by evaluating all MRI sequences of the mass findings without being limited to a few sequences and training the model with clinical-pathological data. Furthermore, our model has a low rate of false positive results, which is an important parameter in clinical practice from a radiological perspective.

Overall, the proposed CCIRM framework appears to offer a promising approach for uncertainty-aware preoperative ALNM prediction by integrating discrimination, calibration, and conformal estimation within a single modelling pipeline. However, its potential clinical value should be interpreted cautiously, because the present findings were obtained from a retrospective single-centre cohort, detailed conformal performance was not comprehensively quantified, and the model was not externally validated. Accordingly, the current results should be regarded as encouraging but preliminary evidence of methodological feasibility rather than definitive proof of routine clinical applicability.

Limitations

Several limitations of this study. A single radiologist performed MRI image evaluation and feature extraction in the first place and hence, the measurements might be subjected to operator dependence and inter-observer variability. All scans were acquired in a single centre using a single institutional workflow. This may limit the generalisability of the proposed pipeline to other scanners, acquisition protocols and patient populations. The acceptable cohort size for an initial modelling study was certainly achieved. However, the independent test subset remained relatively small in size and this, in turn, limits the statistical stability. It also limits the precision of the reported performance estimates. Another significant limitation is the lack of external validation. The performance of the framework on internal testing was strong, however, when applied to populations with different prevalence profiles, clinical characteristics, or imaging practice, its discrimination, calibration, and uncertainty properties may differ. As a result, the present findings should be interpreted as internally encouraging but not yet firm evidence of clinical generalisability. As well as that, although the current work emphasizes calibration and uncertainty-aware modeling, further analyses, including confidence interval estimation, decision-curve analysis, clinical threshold assessment, and formal interpretability evaluation, would be required to strengthen the clinical relevance and explanatory transparency of the framework. Future studies should therefore include prospective pilot evaluation, multicentre external validation, multi-vendor imaging data, and complementary interpretability analyses to determine whether the proposed approach can maintain its reliability and usefulness under real-world clinical conditions.

Conclusions

In this study, we developed a hybrid clinical machine learning framework that combines model selection, probability calibration, and conformal prediction for preoperative estimation of axillary lymph node metastasis in breast cancer. The proposed CCIRM model demonstrated strong discrimination performance while also being designed to provide more reliable probability estimates and uncertainty-aware outputs. In this respect, the main contribution of the study lies not only in predictive performance, but also in its emphasis on calibration, reliability, and clinically cautious risk estimation.

The conformal prediction component may be particularly valuable in radiology-oriented decision support because it enables identification of cases with greater predictive uncertainty rather than forcing a deterministic output in every instance. Nevertheless, the present findings should be interpreted as preliminary from a clinical implementation perspective, since external validation and formal clinical utility analyses were not performed. Accordingly, prospective multicentre validation and additional clinical utility assessment will be necessary before broader applicability can be established.

Data Sharing Statement

Reasonable requests may be made with the corresponding author's permission.

Ethical Statement

This study was conducted with the approval of the Scientific Research and Publication Ethics Board of İzmir Katip Çelebi University of Health Research Ethics Committee, under the official correspondence dated 09/10/2025 (Decision No: 0601). All collected data were anonymized, and no personal identifiable information was used during the analysis. Data processing was performed solely using hidden patient identifiers. The research was carried out in full compliance with the principles of the Declaration of Helsinki. With the official correspondence dated 09/10/2025 and numbered 0601, ethics certificate was obtained from İzmir Katip Çelebi University of Health Research Ethics Committee.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Dong F, Li J, Wang J, Yang X. Diagnostic performance of DCE-MRI radiomics in predicting axillary lymph node metastasis in breast cancer patients: a meta-analysis. *PLoS One*. 2024;19:e0314653. doi:10.1371/journal.pone.0314653
2. Wang Q, Lin Y, Ding C, et al. Multi-modality radiomics model predicts axillary lymph node metastasis of breast cancer using MRI and mammography. *Eur Radiol*. 2024;34(9):6121–6131. doi:10.1007/s00330-024-10638-2
3. Aydın H. Meme kanseri evrelemesinde aksiller radyoloji: mammografi, ultrasonografi, manyetik rezonans görüntüleme ve lenfanjiografik yöntemler. *Türk Radyoloji Dergisi*. 2017;36:69–75. doi:10.5152/turkjradiol.2017.747
4. Shi W, Su Y, Zhang R, et al. Prediction of axillary lymph node metastasis using a magnetic resonance imaging radiomics model of invasive breast cancer primary tumor. *Cancer Imaging*. 2024;24(1):122. doi:10.1186/s40644-024-00771-y
5. Guo F, Sun S, Deng X, et al. Predicting axillary lymph node metastasis in breast cancer using a multimodal radiomics and deep learning model. *Front Immunol*. 2024;15:1482020. doi:10.3389/fimmu.2024.1482020
6. Lee CF, Lin J, Huang Y-L, et al. Deep learning-based breast MRI for predicting axillary lymph node metastasis: a systematic review and meta-analysis. *Cancer Imaging*. 2025;25(1):44. doi:10.1186/s40644-025-00863-3
7. Zhang X, Liu M, Ren W, et al. Predicting of axillary lymph node metastasis in invasive breast cancer using multiparametric MRI dataset based on CNN model. *Front Oncol*. 2022;12:1069733. doi:10.3389/fonc.2022.1069733
8. Ilıca T, Oren C, Ercıktı N, et al. A targeted high-resolution axillary MRI to detect axillary lymph node metastasis with ADC values. *Med J Mugla Sitki Kocman Univ*. 2017;4:17–23.

9. Aydede YS. *Diagnostic Performance of the Deep Learning Method Trained Using MRI AND F-18 FDG-PET/BT Images in the Evaluation of Axillary Lymph Node Metastasis in Breast Cancer Patients*. Sağlik Bilimleri Üniversitesi Adana Tip Fakültesi Genel Cerrahi Anabilim. DALI, Adana; 2024.
10. Zhao X, Wang M, Wei Y, et al. Overview of multimodal radiomics and deep learning in the prediction of axillary lymph node status in breast cancer. *Acad Radiol*. 2025;32:6623–6641. doi:10.1016/j.acra.2025.07.017
11. Zhang D, Shen M, Zhang L, He X, Huang X. Establishment of an interpretable MRI radiomics-based machine learning model capable of predicting axillary lymph node metastasis in invasive breast cancer. *Sci Rep*. 2025;15(1):26030. doi:10.1038/s41598-025-10818-0
12. Cheng Y, Wu M, Tang W, et al. Prediction of axillary lymph node metastasis in breast cancer based on MRI: a novel domain adaptative radiomics pipeline for multicenter studies. *Med Phys*. 2025;52:e70122. doi:10.1002/mp.70122
13. Lai J, Chen Z, Liu J, et al. A radiogenomic multimodal and whole-transcriptome sequencing for preoperative prediction of axillary lymph node metastasis and drug therapeutic response in breast cancer: a retrospective, machine learning and international multicohort study. *Int J Surg*. 2024;110:2162–2177. doi:10.1097/JS9.0000000000001082
14. Liu Y, Li X, Zhu L, et al. Preoperative prediction of axillary lymph node metastasis in breast cancer based on intratumoral and peritumoral DCE-MRI radiomics nomogram. *Contrast Media Mol Imaging*. 2022;2022:6729473. doi:10.1155/2022/6729473
15. Ding J, Chen S, Serrano Sosa M, et al. Optimizing the peritumoral region size in radiomics analysis for sentinel lymph node status prediction in breast cancer. *Acad Radiol*. 2022;29:S223–S228. doi:10.1016/j.acra.2020.10.015
16. Ermiş İ. *Genetic Factors Affecting Axillary Lymph Node Metastasis in Early Stage Breast Cancer*. In: T.C Gazi Üniversitesi Tıp Fakültesi Genel Cerrahi Anabilim Dalı. Ankara; 2019.
17. Park NJY, Jeong JY, Park JY, et al. Peritumoral edema in breast cancer at preoperative MRI: an interpretative study with histopathological review toward understanding tumor microenvironment. *Sci Rep*. 2021;11(1):12992. doi:10.1038/s41598-021-92283-z
18. Zhao S, Li Y, Ning N, et al. Association of peritumoral region features assessed on breast MRI and prognosis of breast cancer: a systematic review and meta-analysis. *Eur Radiol*. 2024;34(9):6108–6120. doi:10.1007/s00330-024-10612-y
19. Moradi B, Gity M, Banihashemian M, et al. Correlation of peri-tumoral edema determined in T2 weighted imaging with apparent diffusion coefficient of peritumoral area in patients with breast carcinoma. *Breast Imaging Iran J Radiol*. 2020;17:97978.