Computer Methods and Programs in Biomedicine 265 (2025) 108765

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Enhancing radiomics features via a large language model for classifying benign and malignant breast tumors in mammography

Sinyoung Ra [a], Jonghun Kim [b], Inye Na [b], Eun Sook Ko [c], Hyunjin Park [b,*]

[a] Department of Artificial Intelligence, Sungkyunkwan University, Suwon, Republic of Korea
[b] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[c] Samsung Medical Center, Department of Radiology, School of Medicine, Sungkyunkwan University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Breast cancer
Enhanced radiomics
Mammogram
Large language model
Extensible Learning

ABSTRACT

Background and Objectives: Radiomics is widely used to assist in clinical decision-making, disease diagnosis, and treatment planning for various target organs, including the breast. Recent advances in large language models (LLMs) have helped enhance radiomics analysis.

Materials and Methods: Herein, we sought to improve radiomics analysis by incorporating LLM-learned clinical knowledge, to classify benign and malignant tumors in breast mammography. We extracted radiomics features from the mammograms based on the region of interest and retained the features related to the target task. Using prompt engineering, we devised an input sequence that reflected the selected features and the target task. The input sequence was fed to the chosen LLM (LLaMA variant), which was fine-tuned using low-rank adaptation to enhance radiomics features. This was then evaluated on two mammogram datasets (VinDr-Mammo and INbreast) against conventional baselines.

Results: The enhanced radiomics-based method performed better than baselines using conventional radiomics features tested on two mammogram datasets, achieving accuracies of 0.671 for the VinDr-Mammo dataset and 0.839 for the INbreast dataset. Conventional radiomics models require retraining from scratch for an unseen dataset using a new set of features. In contrast, the model developed in this study effectively reused the common features between the training and unseen datasets by explicitly linking feature names with feature values, leading to extensible learning across datasets. Our method performed better than the baseline method in this retraining setting using an unseen dataset.

Conclusions: Our method, one of the first to incorporate LLM into radiomics, has the potential to improve radiomics analysis.

1. Introduction

Breast cancer is a highly prevalent and devastating disease affecting women [1]. Mammography is widely used to detect abnormalities in breast tissues. Efficient methods for analyzing mammographs to classify benign and malignant tumors are critical for therapeutic choices. Radiomics extracts hundreds of quantifiable features from routine medical imaging data and has been used to decode complex tumor phenotypes, enabling personalized medicine [2–4]. This approach has enhanced the decision-making process in cancer diagnosis, underscoring its transformative potential in oncological care, including breast cancer [5].

Radiomics features are categorized into intensity, shape, and texture, which are processed using machine learning methods [6]. Recent studies have highlighted the potential of radiomics features to distinguish malignant lesions on mammography [5,7]. Notably, one study [8] used fractal-based features to differentiate between malignant and benign lesions. Recently, another study [9] demonstrated that combining quantitative radiomics features with machine learning algorithms enhanced diagnostic performance using mammographic images. It surpassed the accuracy of experienced radiologists. Using radiomics features to analyze breast lesions has also proven to be effective with diverse magnetic resonance imaging modalities besides mammograms [10–14]. Despite these active radiomics studies, radiomics features are

* Corresponding author at: Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, 16419, 2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do, Republic of Korea, , Republic of Korea.
E-mail address: hyunjinp@skku.edu (H. Park).

https://doi.org/10.1016/j.cmpb.2025.108765
Received 6 January 2025; Received in revised form 27 March 2025; Accepted 3 April 2025
Available online 3 April 2025

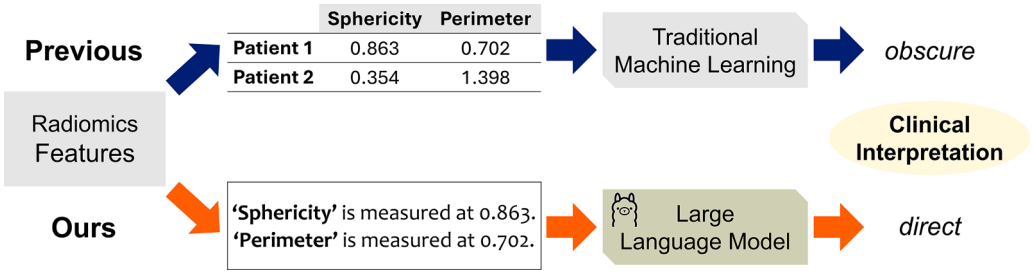

Fig. 1. Comparison of our method with previous radiomics analysis: our proposed method leverages not only feature value but also the feature name compared with traditional radiomics analysis.

often treated merely as numerical values without considering their underlying meaning; for example, feature A, is treated as 0.3 rather than the entropy texture feature treated as 0.3. Clinical interpretation of the names of the radiomics features is often discussed post hoc, but is not directly included in machine learning models. Thus, the current approach may be mathematically correct, but does not leverage the connection between the name of the feature with its value, thus obscuring the clinical interpretation.

Large language models (LLMs) have undergone significant evolution, marking a new era in artificial intelligence capabilities. These models, which excel at processing and synthesizing vast amounts of text, code, and various data forms, have achieved unprecedented progress in recent years. LLMs are engineered by training on massive text datasets such as medical texts, enabling them to identify intricate patterns and relationships. This sophisticated understanding can generate new data that align with learned patterns. Recently, considerable research has focused on the diverse applications of LLMs in the medical field. For instance, chatbot studies, such as Med-PaLM [15], ChatDoctor [16], PMC-LLaMA [17], and LLaVA-Med [18], have sought to develop LLMs specifically tailored for the medical domain. In addition, LLMs are now increasingly used in multiple tasks. Lee et al. [19] combined LLMs with a generative model to generate medical reports from medical images and to answer complex questions. Oh et al. [20] successfully implemented multimodal target volume contouring using an LLM. These advancements underscore the potential of LLMs in revolutionizing various aspects of medicine. However, studies specifically exploring the application of LLMs in relation to radiomics features remain scarce.

LLMs are widely recognized for their exceptional capabilities in processing and generating natural language, excelling in tasks that involve understanding and generating text. Traditionally, their strength has been observed primarily in textual comprehension, which has led to concerns about their ability to handle numerical data effectively. However, recent studies have provided evidence that LLMs not only treat numbers as strings of tokens but also capture meaningful quantitative representations. Zhu et al. (2025) [21] demonstrated that LLMs encode numerical values in a nearly linear manner relative to their true magnitudes. Along the same lines, Hanna et al. (2023) [22] showed that models such as GPT-2 can perform relational numerical operations (e.g., determining "greater-than"), indicating an internal sensitivity to numerical ordering. McCoy et al. (2024a) [23] revealed that numerical representations are shaped by an autoregressive training process, which leads to an inherent quantitative structure within the model. Moreover, Marjieh et al. (2025) [24] explored how LLMs form representational spaces that entangle both string-like and numerical properties, supporting the idea that these models acquire an understanding of numbers beyond simple token prediction.

We propose a novel approach that uses an LLM to consider the meaning of features in addition to their values (Fig. 1). LLMs may link the names of features with learned clinical knowledge and literature, enhancing radiomics features. Thus, our method is mathematically correct and leverages the meanings of the features. We propose enhancing radiomics features with LLM and demonstrate their efficacy in distinguishing between benign and malignant tumors in breast mammography.

Our main contributions can be summarized as follows:

- **Enhancing radiomics features with LLM**: We propose a novel method for designing an` input prompt tailored to radiomics features and fine-tuning the LLM for our downstream task. Our study is one of the first attempts to leverage LLM to enhance radiomics features.
- **Classifying breast masses with improved performance**: Our enhanced radiomics features were better at classifying benign and malignant breast masses, as validated on two independent datasets, compared with other baselines using unenhanced radiomics features.
- **Facilitating extensible learning**: Radiomics models require retraining from scratch in an unseen dataset, whereas our method can effectively reuse the common features between the training and unseen datasets. This is because of the explicit link between the feature name and feature value, which allows extensible learning across datasets.

2. Materials and methods

2.1. Datasets

This study was approved by the Institutional Review Board of Sungkyunkwan University. We used a Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset to train our mass segmentation model [25]. Our research leveraged the VinDr-Mammo training dataset to train an LLM to classify benign and malignant masses using radiomics features [26]. To assess the performance of the fine-tuned LLM model, we employed the VinDr-Mammo and INbreast datasets as internal and external testing sets, respectively [27]. Institutional Review Board approvals were obtained for the respective studies, with written informed consent.

CBIS-DDSM, originating from the United States, is a carefully selected subset derived from the DDSM dataset, comprising 10,239 images curated by a mammographer with specialized training. Segmentation masks are present for the lesions in the dataset.

VinDr-Mammo, which originated in Vietnam, provides digital mammography images with detailed assessments at both the breast and lesion levels. The dataset comprises 5000 mammography examinations, each accompanied by four standard views, for a total of 20,000 images. Each breast in the dataset has been assessed using the Breast Imaging Reporting and Data System (BI-RADS) classification and breast density [28]. Additionally, for malignant findings, the dataset provides information on the category of abnormalities (such as mass, calcification, asymmetry, distortion, and other associated features), lesion locations (represented by bounding boxes), and corresponding BI-RADS classifications.

INbreast, which originated in Portugal, contains 410 mammograms derived from 115 unique cases. These mammograms are categorized by specialists into various abnormalities, including asymmetry, calcification, clustering of Merkel cell carcinoma, mass, distortion, spiculated

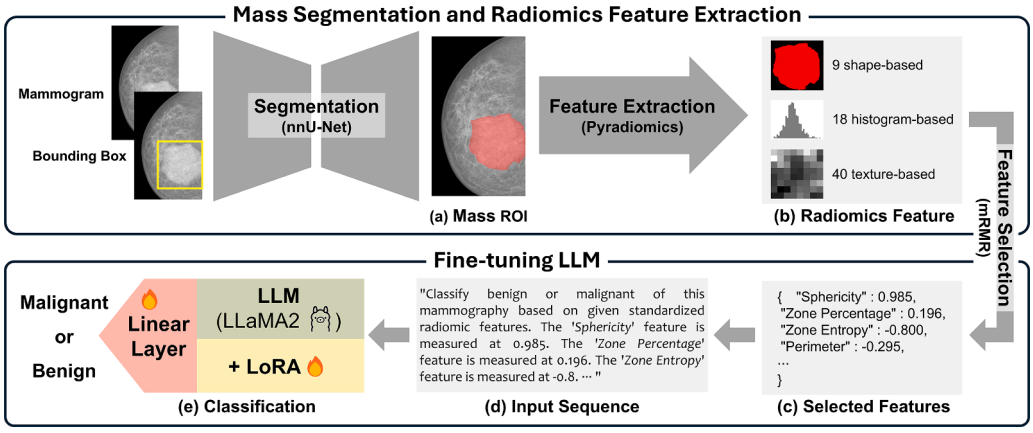

Fig. 2. Overview of the method: (a) Placement of mass ROIs; (b) Radiomics feature extraction through Pyradiomics: a total of 67 radiomics features were extracted per breast mass; (c) Feature selection by mRMR (minimum redundancy maximum relevance) algorithm: we selected eight features to help construct the input sequence; (d) Input sequence construction and formatting; (e) Fine-tuning the LLM for the classification task.

region, and pectoral muscle. In 90 cases, both mediolateral oblique (MLO) and craniocaudal (CC) views are available for each breast, whereas 25 cases have two views of only one breast. Additionally, BI-RADS classifications ranging from 1 to 6 are provided for each case, and biopsies had been performed in 56 cases.

2.2. Preprocessing

All mammograms used in this study underwent minimum–maximum normalization and only images containing masses were selected for analysis. For classification based on BI-RADS grading, categories 1 and 2 were classified as normal, category 3 as benign, and categories 4–6 as malignant. Subsequently, only the benign and malignant cases were included in the experimental dataset for binary classification. The VinDr-Mammo dataset was employed for training and comprised 785 samples, with an additional 197 samples allocated for validation and 235 for internal testing. Furthermore, the INbreast dataset served as an external testing dataset, comprising 111 samples.

An overview of the proposed method is shown in Fig. 2. First, we placed mass regions of interest (ROIs) on the mammogram using a pre-trained segmentation network (details are provided later). Radiomics features were then extracted from the ROI and mammogram. This was followed by a feature selection procedure to retain only the features related to the target task of classification. Using prompt engineering, we devised an input sequence that reflected the selected features and the target task. The input sequence was fed to the selected LLM, which was fine-tuned using low-rank adaptation (LoRA).

2.3. Developing a mass segmentation model

The VinDr-Mammo dataset lacks contours for masses and only provides bounding boxes; therefore, obtaining contours for masses is necessary to extract radiomics features. In collaboration with board-certified radiologists, we enhanced the delineation of mass boundaries in the VinDr-Mammo dataset using a semi-automatic process. This segmentation was achieved by incorporating both images and their corresponding bounding boxes as inputs to the nn-UNet model [29–31], previously trained using ROI annotations from the CBIS-DDSM dataset. The inclusion of bounding boxes alongside the images provides a more precise context for the model, facilitating enhanced identification and segmentation of the masses by allowing the model to focus on relevant areas of interest. This collaborative effort enabled us to accurately segment masses in mammography images by combining the expertise of radiologists with the capabilities of neural network models.

The CBIS-DDSM dataset only contained film-based images, because of which, accurately extracting radiomics features was inherently challenging and restricted its usage solely to developing the segmentation model. The nature of film images complicates the precise extraction of radiomics features, which is critical for comprehensive analysis and classification in radiomics studies.

2.4. Radiomics feature extraction

Radiomics features were extracted from contoured masses in the mammograms using PyRadiomics, a Python package tailored for radiomics analysis [32]. This process involves quantifying various shape, intensity and texture features to capture the diverse aspects of tumor morphology and heterogeneity. A comprehensive suite of 67 radiomics features was extracted, which included 9 shape-based features, 18 first-order statistical features, and 40 texture-based features [33]. Among the texture-based features, 24 were derived from the gray-level co-occurrence matrix and 16 from the gray-level size zone matrix. We excluded wavelet-based features from the typical radiomics feature set because they are challenging to interpret and typically offer only marginal performance gains [34–37]. This selective approach, which focuses on first-order statistics, shape, gray-level co-occurrence matrix, and gray-level size zone matrix, has been widely adopted in the field [36–38]. All radiomics features were subjected to z-score standardization using the mean and standard deviation derived from the training set.

2.5. Preparation of input sequence for training the LLM

Model training did not differentiate between the left/right and CC/MLO mammographic views. Instead, all the image variations were incorporated into a single unified model. This approach leveraged a comprehensive dataset and facilitated enhanced learning outcomes through exposure to a broad range of cases during the training process. By not segregating the dataset based on specific view types, the model could benefit from the increased diversity and quantity of training samples. This could potentially improve the generalization and performance across various imaging scenarios.

The initial instruction of the input sequence to the LLM was to classify masses in mammograms as benign or malignant using the given radiomics features. Subsequently, the names and values of these features were articulated in sentences conforming to a predefined format. Owing to the limitations of the length of the input sequence, we could not feed many radiomics features into the LLM. Thus, we selected eight radiomics features using the minimum redundancy maximum relevance (mRMR) algorithm [6,39]. We specifically chose mRMR because it effectively minimizes redundancy among highly correlated radiomics features while maximizing their relevance to the classification task. In addition,

Input Sequence: Classify benign or malignant of this mammography based on given standardized radiomic features. The 'Sphericity' feature is measured at 0.985 . The 'Zone Percentage' feature is measured at 0.196 . The 'Small Area High Gray Level Emphasis' feature is measured at 0.379 . The 'Zone Entropy' feature is measured at -0.8 . The 'Maximum' feature is measured at 1.235 . The 'Perimeter' feature is measured at-0.295 . The 'Maximum Diameter' feature is measured at-0.226 . The 'Major Axis Length' feature is measured at-0.171 .

Label: 1(Malignant)

Fig. 3. Example of input sequence and label for the large language model.

mRMR allows us to fix the number of selected features, which is a critical advantage, given the input-length limitations of LLMs. This fixed feature set ensures that the LLM receives consistent and concise inputs, making it particularly suitable for our application. This approach is supported by previous studies [6,37,39,40].

This feature selection process was undertaken to use the data efficiently within the imposed constraints to ensure a focused and effective analysis of the classification task. The text in the box (Fig. 3) illustrates a typical input sequence for a malignant mass. This sequence was generated using ChatGPT to format the radiomics feature names and their corresponding values into a coherent, standardized template, which was then provided as an input to the LLaMA2-based classifier.

2.6. Fine-tuning LLM with LoRA

LLMs learn from trillions of samples and thus contain knowledge of radiomics features. However, such knowledge must be adapted to a specific downstream task to improve the performance. Therefore, we employed the LoRA fine-tuning of an existing LLM to adapt the existing learned knowledge of radiomics to our classification task [41,42]. This approach allowed us to retain the pre-existing knowledge of LLM while adapting its capabilities to our specific requirements, ensuring that the valuable radiomics context was not lost during the adaptation process. LoRA fine-tuning strategically modifies only a fraction of the model's parameters, thus maintaining the integrity of the original structure and knowledge base of the model. It does this while achieving significant performance improvements in the targeted classification tasks. This method underscores our commitment to leverage and augment the pre-trained model's strengths without compromising its existing knowledge assets. Specifically, we used the same training–test split (785 training, 197 validation, and 235 test samples) for both LoRA fine-tuning and radiomics analysis.

2.7. Comparison with other methods

We conducted a comprehensive comparison to assess the classification performance of our proposed method relative to traditional machine-learning algorithms using enhanced radiomics features. The compared methods were random forest (RF), gradient boosting (GB), support vector machine (SVM), naïve Bayes (NB), logistic regression (LR), and multilayer perceptron (MLP) [37,43,44]. Each model was implemented using the scikit-learn library. The RF model used 100 estimators, while the GB model employed a learning rate of 0.1 and a maximum depth of 3. The SVM model uses a radial basis function kernel, and the NB model is based on a Gaussian distribution. The LR model applies L2 regularization with the Laplace solver, and the MLP model comprises a deep neural network with a hidden layer size of 1024 neurons and a ReLU activation function. In addition to these models, we included XGBoost [45], an effective gradient-boosting algorithm known to achieve state-of-the-art performance in various tabular data tasks. TransTab [46], a recent tabular deep-learning model designed to utilize both feature names and values was also added. Both models are

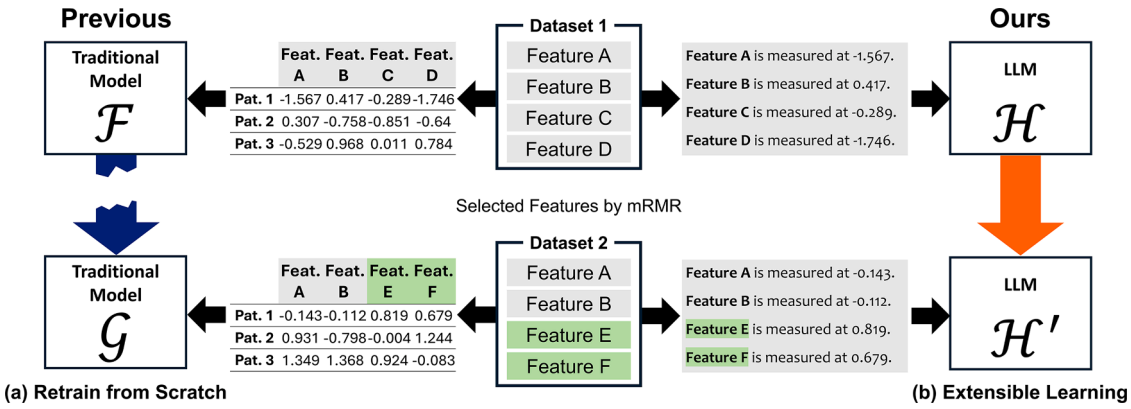

Fig. 4. Extensible learning aspects of our method: (a) Existing radiomics models require retraining from scratch for an unseen dataset where selected features are altered. (b) Our proposed method explicitly incorporates feature names, enabling continual learning for an unseen dataset by leveraging the direct link between feature names and values. The common features between datasets are shown in green. From an input perspective, traditional models rely on numerical values without incorporating feature names. Consequently, the same features must be input in the same order, especially in external datasets. Conversely, our proposed model explicitly includes feature names alongside the numerical values. This allows for robustness and continual learning, even if some features change in external datasets, thereby eliminating the need to retrain the model from scratch.

Table 1
Performance comparison in the internal and external test set (bold case is the best performance in a given column).

Dataset	VinDr-Mammo (internal test)				INbreast (external test)			
Model	Acc	F1	AUC	Sens	Acc	F1	AUC	Sens
RF	0.631	0.624	0.631	0.622	0.768	0.822	0.748	0.811
XGBoost	0.604	0.618	0.605	0.649	0.786	0.838	0.761	**0.838**
GB	0.622	0.615	0.622	0.613	0.768	0.827	0.735	**0.838**
SVM	0.653	0.645	0.653	0.640	0.821	0.861	0.814	**0.838**
NB	0.627	0.500	0.623	0.378	0.804	0.836	0.826	0.757
LR	0.653	0.667	0.654	**0.703**	0.768	0.817	0.760	0.784
MLP	0.653	0.661	0.654	0.685	0.732	0.776	0.746	0.703
TransTab	0.649	0.652	**0.727**	0.667	0.750	0.800	**0.872**	0.757
Ours	**0.671**	**0.673**	0.700	0.685	**0.839**	**0.873**	0.862	**0.838**

considered strong baselines for structured data analysis and serve as competitive comparisons with the proposed approach.

While deep learning models, such as convolutional neural networks (CNNs) and vision transformers (ViTs), have demonstrated impressive performance in image analysis tasks [47–51], our research adopted a fundamentally different approach by focusing on enhancing radiomics features through the semantic understanding capabilities of LLMs. Unlike CNNs and ViTs, which excel at pixel-level feature extraction and end-to-end image analysis, our method integrates clinical knowledge to improve the interpretability and extensibility of preexisting radiomics features. Therefore, direct comparisons with these image-based deep--learning models are inappropriate and were not considered.

2.8. Extensible learning in external dataset with different radiomics features

Conventionally, a radiomics model may suffer from performance degradation when applied to an unseen external dataset with different characteristics, requiring the selection of a new set of radiomics features specific to the external dataset (Fig. 4) [52]. Although this new set of features differs from existing ones, a degree of overlap is expected. Under these circumstances, traditional radiomics models must be retrained from scratch, whereas our model can be effectively trained using the explicit specification of the feature name. A common feature can be placed in an order different from the existing set of features but can still be easily leveraged by the retraining process because of the explicit link with the feature name. We tested this by comparing existing radiomics models trained from scratch with our model retrained on the external dataset. Our retraining approach is a form of extensible learning that efficiently adapts to changes in the dataset characteristics.

Although traditional models can perform transfer learning when the exact feature order is preserved, they generally require retraining, if external datasets introduce a new or slightly altered set of radiomics features. By contrast, our method explicitly includes feature names in the input sequence, allowing the model to identify and leverage overlapping features between different datasets. This mechanism supports extensible learning by linking semantically meaningful feature names to their corresponding values. Furthermore, our method encodes some degree of semantic knowledge about radiomics features and can relate them to specific disease contexts, such as breast cancer, using free-form task descriptions in the prompt. This enables the model to align features across datasets based not only on name similarity but also on clinical relevance.

2.9. Implementation details

We chose the LLaMA2–7B model as the baseline LLM using the default settings [17,20]. A binary cross-entropy loss function was employed to quantify the discrepancy between the predicted outcomes and actual labels, thereby optimizing the learning process. The LoRA configuration used a rank value of 16 to target all layers, including “Q” and “V” attention, consistent with the original LoRA study [53,54]. The network parameters were optimized using the AdamW optimizer with a learning rate of 0.0001. This process was extended to over 500 training epochs to ensure comprehensive model training and optimal convergence. In addition, a cosine-learning rate scheduler was implemented to dynamically adjust the learning rate throughout training, enabling a more refined and effective optimization process. Experimental setups were conducted within PyTorch in Python using CUDA 12.0, on an NVIDIA Quadro RTX 8000 48GB. The model that demonstrated the best performance in the validation set was selected as the final model for further evaluation. Subsequent assessments were conducted on both internal and external test datasets to rigorously evaluate the generalization capabilities of the model and its performance across diverse data scenarios.

3. Results

We compared our results with those of other methods to demonstrate the effect of integrating LLMs into radiomics, highlighting their meaningful contribution to enhancing breast lesion classification.

3.1. Selected features from the training

Using the mRMR algorithm, we selected eight radiomics features from the VinDr-Mammo training set. These features were consistently applied in the external testing set without any alterations and included sphericity, perimeter, maximum diameter, and major axis length features in the shape category; a histogram feature, maximum; and small area high gray level emphasis, zone percentage, and zone entropy features in the texture category.

3.2. Performance of the proposed method

In the internal test set derived from VinDr-Mammo, our method gave the most superior performance compared with the comparative methods across various evaluation metrics. Specifically, our method demonstrates the highest accuracy (ACC), F1-score (F1), and area under the receiver operating characteristic curve (AUC) metrics. Regarding sensitivity (Sens), our method achieved the second-highest performance among the evaluated methods. Table 1 presents the results of the study. To evaluate the proposed method on an external dataset, we applied it to the INbreast dataset, which was not used during the initial training. To mitigate the distribution shift between the datasets, we performed retraining (i.e., adaptation) using half of the INbreast data starting from the model originally trained on the VinDr-Mammo dataset. The remaining half of the INbreast data were used for testing. Our method exhibited better performance across all evaluation metrics than the comparative methods. Specifically, our method achieved the highest accuracy, F1-score, AUC, and Sens, underscoring its robustness and effectiveness in external validation (Table 1).

Table 2
Performance comparison on the external test set (INbreast) with new radiomics features (bold case is the best performance in a given column).

Model	Acc	F1	AUC	Sens
RF	0.768	0.822	0.748	0.811
XGBoost	**0.821**	0.865	0.801	0.865
GB	0.750	0.806	0.734	0.784
SVM	0.768	0.835	0.709	0.892
NB	0.768	0.827	0.735	0.838
LR	0.768	0.822	0.748	0.811
MLP	0.714	0.758	0.733	0.676
TransTab (transfer learning)	0.679	0.743	0.818	0.703
TransTab (from scratch)	0.732	0.810	0.808	0.865
Ours (extensible learning)	**0.821**	**0.868**	**0.853**	0.892
Ours (from scratch)	0.786	0.857	0.721	**0.973**

3.3. Additional validation with different radiomics features in extensible learning

Owing to the data shift in the external dataset, we performed a new round of mRMR-based feature selection using half of the INbreast dataset. This process yielded a new set of eight radiomics features specific to the external dataset. Among them, four features—*sphericity, perimeter (shape), maximum (histogram)*, and *small area high gray level emphasis (texture)*—overlapped with those selected from the VinDr-Mammo dataset, whereas the other four features —*elongation (shape), maximum probability, joint energy, high gray level zone emphasis (texture)*—were newly selected, reflecting distinct data characteristics. Half of the INbreast dataset was used to train various models from scratch, and the remaining half was used for testing. By explicitly specifying the feature names, our method can readily accommodate the addition of new features without disrupting the learning process, thereby facilitating continuous adaptation and refinement. Table 2 shows the comparisons of our methods in an extensible learning fashion with other methods that were retrained from scratch using half of the external test set as the training set. Our method performs better than the other methods, demonstrating its ability for extensible learning. Specifically, we tested a tabular deep learning model (i.e., TransTab) [46] in the same setting, where feature names could be specified, but lacked contextual disease information (i.e., breast cancer classification). The results of TransTab were worse than ours, possibly owing to missing context information related to disease classification.

3.4. Ablation study

We then conducted ablation studies to quantify the incremental effects of the various components of our method.

LoRA rank: In our LoRA ablation studies, we experimented with rank values of 4, 8, 16, and 32 (Table 3). This analysis highlights the importance of the hyperparameter selection. A rank value of 16 consistently produced the best results for both the internal and external datasets.

Various configurations of model components: We compared various configurations of model components to evaluate their impact on performance (Table 4). The "Head Only" configuration involved freezing the LLM model and training only the final classifier layer. Using LoRA for training proved to be more efficient than "Head Only" configuration. We also compared the configuration of feature selection using the mRMR method with a widely used set of radiomics features, including the pixel surface, sphericity, perimeter, mean, variance, entropy, small-area high-gray-level emphasis, and zone entropy. The results showed that combining LoRA with the mRMR feature selection achieved the best overall performance, with significant improvements in accuracy, F1 score, and sensitivity.

Table 4
Ablation on various components (bold case is the best performance in a given column).

Feature selection	LoRA	Head only	Acc	F1	AUC	Sens
		✓	0.630	0.584	**0.710**	0.540
	✓		0.630	0.613	0.674	0.611
✓		✓	0.671	0.629	0.699	0.595
✓	✓		**0.671**	**0.673**	0.700	**0.685**

Effect of censoring feature names: To examine the impact of semantic information contained in the radiomics feature names, we conducted an ablation study in which the original feature names (e.g., "Sphericity" and "Zone Percentage") were replaced with generic labels (e.g., "Feature A" and "Feature B"). The results showed a significant decrease in the performance when the semantic content of the feature names was removed, indicating that these names provided valuable information for the model (Table 5). Notably, this performance drop was observed not only in our LLM-based approach but also in Transtab [46], where feature names were explicitly specified. These findings confirm that a robust link exists between feature names and their numerical values and that this connection is crucial for effective feature representation.

4. Discussion

This study contributes to radiomics analysis through the innovative integration of LLMs with radiomics features. By contextually incorporating the names and characteristics of each radiomics feature rather than focusing solely on their numerical values, our method opens the possibility of enhancing the interpretative capabilities of radiomics analysis. This method performed better at classifying benign and malignant breast lesions than traditional machine learning techniques. Furthermore, our integration of LLMs facilitated extensible learning, enabling the dynamic application of various radiomics features across different datasets. This flexibility is essential for handling the variability in features across various medical imaging contexts, thereby improving the robustness and adaptability of the method. These advancements pave the way for the development of a comprehensive method that can be generalized across a wide array of medical imaging scenarios, thus enhancing diagnostic accuracy and adaptability in clinical settings. Collectively, these contributions represent a significant advancement in the application of machine learning in medical diagnostics, showcasing the transformative potential of LLMs in medical image analysis.

Importantly, the integration of explainable artificial intelligence (XAI) aspects into our framework further enhanced its clinical value. By

Table 3
Ablation on the rank of LoRA (bold case is the best performance in a given column).

Dataset	VinDr-Mammo (internal test)				INbreast (external test)			
LoRA rank	Acc	F1	AUC	Sens	Acc	F1	AUC	Sens
4	0.604	0.574	0.670	0.541	0.821	0.861	0.847	**0.838**
8	0.662	0.661	0.692	0.667	0.821	0.857	0.866	0.811
16	**0.671**	**0.673**	**0.700**	**0.685**	**0.839**	**0.873**	0.862	**0.838**
32	0.649	0.633	0.684	0.613	0.804	0.836	**0.879**	0.757

Table 5
Ablation on censoring feature name (bold case is the best performance in a given column).

Dataset	VinDr-Mammo (internal test)				INbreast (external test)			
Model	Acc	F1	AUC	Sens	Acc	F1	AUC	Sens
TransTab (Censored)	0.520	0.542	0.554	0.577	0.661	0.796	0.482	1.000
TransTab	0.649	0.652	**0.727**	0.667	0.750	0.800	**0.872**	0.757
Ours (Censored)	0.647	0.570	0.695	0.487	0.804	0.841	0.849	0.784
Ours	**0.671**	**0.673**	0.700	**0.685**	**0.839**	**0.873**	0.862	**0.838**

providing clear, interpretable explanations that link identified radiomics features to known pathological characteristics, our results conform to ethical and legal standards [55–57] enhancing stakeholder trust and acceptance [57]. We identified eight radiomics features that are important for classifying benign and malignant cases. Among these, two shape features—sphericity and perimeter—are highly interpretable and confirmable in the existing literature [56]. Spherical tumors are likely to be low-risk and thus related to benign masses [58], whereas larger tumors are likely to be high-risk and related to malignant masses [59]. Texture features, such as zone entropy, are associated with tumor heterogeneity and thus could be linked with malignant masses [60]. These XAI aspects of our study not only improve the classification performance but also make the model's decision-making process more interpretable and trustworthy, paving the way for future advancements in radiomics analysis.

Our study had some limitations. First, the approach was tested only in the context of breast mammography in a classification task. Further studies are required to confirm the validity of our approach for diverse imaging modalities and tasks. Second, we selected a particular LLM to demonstrate the proposed approach. Additional experiments using other established LLMs are required to determine the validity of the proposed approach. Third, our approach may provide a better understanding of radiomics features. Recent LLMs can provide explanations of the given data; therefore, our enhanced radiomics features using LLM can be further developed to provide even better interpretability of radiomics features.

Ethical statement

This study was approved by the Institutional Review Board of Sungkyunkwan University.

CRediT authorship contribution statement

Sinyoung Ra: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Jonghun Kim:** Writing – original draft, Supervision. **Inye Na:** Investigation. **Eun Sook Ko:** Formal analysis. **Hyunjin Park:** Writing – original draft, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Research Foundation (RS-2024–00408040), AI Graduate School Support Program (Sungkyunkwan University) (RS-2019-II190421), ICT Creative Consilience Program RS-2020-II201821), and the Artificial Intelligence Innovation Hub Program (RS-2021-II212068).

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2011) 69–90.
[2] H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006.
[3] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, Radiomics: the process and the challenges, Magn. Reson. ImAging 30 (2012) 1234–1248.
[4] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. Van Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker, Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer (1965) 48 (2012) 441–446.
[5] A. Conti, A. Duggento, I. Indovina, M. Guerrisi, N. Toschi, Radiomics in breast cancer classification and prediction. Seminars in Cancer Biology, Elsevier, 2021, pp. 238–250.
[6] W. Zhang, Y. Guo, Q. Jin, Radiomics and its feature selection: a review, Symmetry. (Basel) 15 (2023) 1834.
[7] F. Pesapane, P. De Marco, A. Rapino, E. Lombardo, L. Nicosia, P. Tantrige, A. Rotili, A.C. Bozzini, S. Penco, V. Dominelli, How radiomics can improve breast cancer diagnosis and treatment, J. Clin. Med. 12 (2023) 1372.
[8] R.M. Rangayyan, T.M. Nguyen, Fractal analysis of contours of breast masses in mammograms, J. Digit. Imaging 20 (2007) 223–237.
[9] N. Mao, P. Yin, Q. Wang, M. Liu, J. Dong, X. Zhang, H. Xie, N. Hong, Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study, Journal of the American College of Radiology 16 (2019) 485–491.
[10] H.M. Whitney, N.S. Taylor, K. Drukker, A.V. Edwards, J. Papaioannou, D. Schacht, M.L. Giger, Additive benefit of radiomics over size alone in the distinction between benign lesions and luminal a cancers on a large clinical breast MRI dataset, Acad. Radiol. 26 (2019) 202–209.
[11] J. Zhou, Y. Zhang, K.T. Chang, K.E. Lee, O. Wang, J. Li, Y. Lin, Z. Pan, P. Chang, D. Chow, Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue, Journal of Magnetic Resonance Imaging 51 (2020) 798–809.
[12] V.S. Parekh, M.A. Jacobs, Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI, NPJ. Breast. Cancer 3 (2017) 43.
[13] S. Bickelhaupt, D. Paech, P. Kickingereder, F. Steudle, W. Lederer, H. Daniel, M. Götz, N. Gählert, D. Tichy, M. Wiesenfarth, Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography, J. magnetic resonance imaging 46 (2017) 604–616.
[14] J. Kim, H. Park, Radiomics-guided multimodal self-attention network for predicting pathological complete response In breast MRI, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 2024, pp. 1–5.
[15] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, Large language models encode clinical knowledge, Nature 620 (2023) 172–180.
[16] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, Cureus. (2023) 15.
[17] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, Y. Wang, PMC-LLaMA: toward building open-source language models for medicine, Journal of the American Medical Informatics Association (2024) ocae045.
[18] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: training a large language-and-vision assistant for biomedicine in one day, Adv. Neural Inf. Process. Syst. 36 (2024).
[19] S. Lee, W.J. Kim, J. Chang, J.C. Ye, LLM-CXR: instruction-finetuned LLM for CXR image understanding and generation, The Twelfth International Conference on Learning Representations (2023).
[20] Y. Oh, S. Park, H.K. Byun, Y. Cho, I.J. Lee, J.S. Kim, J.C. Ye, LLM-driven multimodal target volume contouring in radiation oncology, Nat. Commun. 15 (2024) 9186.
[21] F. Zhu, D. Dai, Z. Sui, Language models encode the value of numbers linearly, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 693–709.
[22] M. Hanna, O. Liu, A. Variengien, How does GPT-2 compute greater-than?: interpreting mathematical abilities in a pre-trained language model, Adv. Neural Inf. Process. Syst. 36 (2023) 76033–76060.

[23] R.T. McCoy, S. Yao, D. Friedman, M.D. Hardy, T.L. Griffiths, Embers of autoregression show how large language models are shaped by the problem they are trained to solve, Proceedings of the National Academy of Sciences 121 (2024) e2322420121.
[24] R. Marjieh, V. Veselovsky, T.L. Griffiths, I. Sucholutsky, arXiv preprint, 2025.
[25] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (2017) 1–9.
[26] H.T. Nguyen, H.Q. Nguyen, H.H. Pham, K. Lam, L.T. Le, M. Dao, V. Vu, VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography, Sci. Data 10 (2023) 277.
[27] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (2012) 236–248.
[28] C. D'Orsi, L. Bassett, S. Feig, Breast Imaging Reporting and Data System (BI-RADS), Breast imaging Atlas, 4th edn., American College of Radiology, Reston, 2018.
[29] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2021) 203–211.
[30] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Springer, Munich, Germany, 2015, pp. 234–241. October 5-9, 2015proceedings, part III 18.
[31] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, P.F. Jaeger, nnu-net revisited: a call for rigorous validation in 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 488–498.
[32] J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (2017) e104–e107.
[33] M.E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, G. Cook, Introduction to radiomics, Journal of Nuclear Medicine 61 (2020) 488–495.
[34] Z. Kong, C. Jiang, R. Zhu, S. Feng, Y. Wang, J. Li, W. Chen, P. Liu, D. Zhao, W. Ma, 18F-FDG-PET-based radiomics features to distinguish primary central nervous system lymphoma from glioblastoma, NeuroImage: Clinical 23 (2019) 101912.
[35] A. Demircioğlu, Reproducibility and interpretability in radiomics: a critical assessment, diagnostic and interventional radiology (Ankara, Turkey), (2024).
[36] I. Na, J. Kim, E.S. Ko, H. Park, RadiomicsFill-Mammo: synthetic mammogram mass manipulation with radiomics features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 723–733.
[37] H.-h. Cho, S.-h. Lee, J. Kim, H. Park, Classification of the glioma grading using radiomics analysis, PeerJ. 6 (2018) e5982.
[38] K. Sun, Z. Jiao, H. Zhu, W. Chai, X. Yan, C. Fu, J.-Z. Cheng, F. Yan, D. Shen, Radiomics-based machine learning analysis and characterization of breast lesions with multiparametric diffusion-weighted MR, J. Transl. Med. 19 (2021) 1–10.
[39] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEe Trans. Pattern. Anal. Mach. Intell. 27 (2005) 1226–1238.
[40] L. Xu, P. Yang, W. Liang, W. Liu, W. Wang, C. Luo, J. Wang, Z. Peng, L. Xing, M. Huang, A radiomics approach based on support vector machine using MR images for preoperative lymph node status evaluation in intrahepatic cholangiocarcinoma, Theranostics. 9 (2019) 5374.
[41] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.
[42] Y. Mao, Y. Ge, Y. Fan, W. Xu, Y. Mi, Z. Hu, Y. Gao, A survey on lora of large language models, Front. Comput. Sci. 19 (2025) 197605.
[43] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J. Aerts, Machine learning methods for quantitative radiomic biomarkers, Sci. Rep. 5 (2015) 1–11.
[44] A.A.K. Abdel Razek, A. Alksas, M. Shehata, A. AbdelKhalek, K. Abdel Baky, A. El-Baz, E. Helmy, Clinical applications of artificial intelligence and radiomics in neuro-oncology imaging, Insights. ImAging 12 (2021) 1–17.
[45] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[46] Z. Wang, J. Sun, Transtab: learning transferable tabular transformers across tables, Adv. Neural Inf. Process. Syst. 35 (2022) 2902–2915.
[47] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324.
[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[50] M.L. Abimouloud, K. Bensid, M. Elleuch, O. Aiadi, M. Kherallah, Mammography breast cancer classification using vision transformers, in: International Conference on Intelligent Systems Design and Applications, Springer, 2023, pp. 452–461.
[51] W.M. Salama, M.H. Aly, Deep learning in mammography images segmentation and classification: automated CNN approach, Alexandria Eng. J. 60 (2021) 4701–4709.
[52] M. Wennmann, L.T. Rotkopf, F. Bauer, T. Hielscher, J. Kächele, E.K. Mai, N. Weinhold, M.S. Raab, H. Goldschmidt, T.F. Weber, Reproducible radiomics features from Multi-MRI-scanner test–Retest-study: influence on performance and generalizability of models, J. Magnet. Resonance Imag. (2024).
[53] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, Adv. Neural Inf. Process. Syst. (2024) 36.
[54] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, G. Neubig, Towards a unified view of parameter-efficient transfer learning, Int. Confer. Learn. Represent. (2022).
[55] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a "right to explanation, AI. Mag. 38 (2017) 50–57.
[56] C. Militello, F. Prinzi, G. Sollami, L. Rundo, L. La Grutta, S. Vitabile, CT radiomic features and clinical biomarkers for predicting coronary artery disease, Cognit. Comput. 15 (2023) 238–253.
[57] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, Queue 16 (2018) 31–57.
[58] H. Li, X. Meng, T. Wang, Y. Tang, Y. Yin, Breast masses in mammography classification with local contour features, Biomed. Eng. Online 16 (2017) 1–12.
[59] E.J. Lee, Y.-W. Chang, Prospective analysis of breast masses using the combined score for quantitative ultrasonography parameters, Sci. Rep. 12 (2022) 16205.
[60] F. Davnall, C.S. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G.J. Cook, V. Goh, Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights. ImAg. 3 (2012) 573–589.