Computers in Biology and Medicine 188 (2025) 109721

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Multimodal deep learning fusion of ultrafast-DCE MRI and clinical information for breast lesion classification

Belinda Lokaj [a,b,*], Valentin Durand de Gevigney [a], Dahila-Amal Djema [d], Jamil Zaghir [b,c], Jean-Philippe Goldman [b,c], Mina Bjelogrlic [b,c], Hugues Turbé [b,c], Karen Kinkel [e], Christian Lovis [b,c], Jérôme Schmid [a]

[a] Geneva School of Health Sciences, HES-SO University of Applied Sciences and Arts Western Switzerland, Delémont, Switzerland
[b] Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
[c] Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland
[d] Hirslanden - Clinique des Grangettes, Geneva, Switzerland
[e] Réseau Hospitalier Neuchâtelois, Neuchâtel, Switzerland

ARTICLE INFO

*Keywords:*
Breast cancer
Ultrafast
Multimodal
Deep learning
Magnetic resonance imaging

ABSTRACT

*Background:* Breast cancer is the most common cancer worldwide, and magnetic resonance imaging (MRI) constitutes a very sensitive technique for invasive cancer detection. When reviewing breast MRI examination, clinical radiologists rely on multimodal information, composed of imaging data but also information not present in the images such as clinical information. Most machine learning (ML) approaches are not well suited for multimodal data. However, attention-based architectures, such as Transformers, are flexible and therefore good candidates for integrating multimodal data.

*Purpose:* The aim of this study was to develop and evaluate a novel multimodal deep learning (DL) model combining ultrafast dynamic contrast-enhanced (UF-DCE) MRI images, lesion characteristics and clinical information for breast lesion classification.

*Materials and methods:* From 2019 to 2023, UF-DCE breast images and radiology reports of 240 patients were retrospectively collected from a single clinical center and annotated. Imaging data were constituted of volumes of interest (VOI) extracted around segmented lesions. Non-imaging data were constituted of both clinical (categorical) and geometrical (scalar) data. Clinical data were extracted from annotated reports and were associated to their corresponding lesions. We compared the diagnostic performances of traditional ML methods for non-imaging data, an image model based on the DL architecture, and a novel Transformer-based architecture, the Multimodal Sieve Transformer with Vision Transformer encoder (MMST-V).

*Results:* The final dataset included 987 lesions (280 benign, 121 malignant lesions, and 586 benign lymph nodes) and 1081 reports. For classification with non-imaging data, scalar data had a greater influence on performances of lesion classification (Area under the receiver operating characteristic curve (AUROC) = 0.875 ± 0.042) than categorical data (AUROC = 0.680 ± 0.060). MMST-V achieved better performances (AUROC = 0.928 ± 0.027) than classification based on non-imaging data (AUROC = 0.900 ± 0.045), and imaging data only (AUROC = 0.863 ± 0.025).

*Conclusion:* The proposed MMST-V is an adaptative approach that can consider redundant information provided by multimodal information. It demonstrated better performances than unimodal methods. Results highlight that the combination of clinical patient data and detailed lesion information as additional clinical knowledge enhances the diagnostic performances of UF-DCE breast MRI.

* Corresponding author. Geneva School of Health Sciences, HES-SO University of Applied Sciences and Arts Western Switzerland, Delémont, Switzerland.
*E-mail address:* belinda.lokaj@hesge.ch (B. Lokaj).

https://doi.org/10.1016/j.compbiomed.2025.109721
Received 1 August 2024; Received in revised form 17 January 2025; Accepted 17 January 2025
Available online 19 February 2025

Abbreviations

| | |
|---|---|
| AB-MRI | Abbreviated Breast Magnetic Resonance Imaging |
| ACC | Accuracy |
| ACS | American Cancer Society |
| ADC | Apparent diffusion coefficient |
| AI | Artificial intelligence |
| ALN | Axillary lymph node |
| ANN | Artificial neural network |
| AUROC | Area under the receiver operating characteristic curve |
| BPE | Background parenchymal enhancement |
| CC | Craniocaudal |
| CCER | Geneva Cantonal Ethics Committee |
| CENTRA | Contrast ENhanced Timing Robust Angiography |
| CNN | Convolutional neural networks |
| CT | Computed tomography |
| DBT | Digital Breast Tomosynthesis |
| DCE | Dynamic contrast-enhanced |
| DCIS | Ductal carcinoma in situ |
| DeiT | Data-efficient image Transformer |
| DenseNet | Dense Convolutional Network |
| DiT | Dual-input Transformer |
| DL | Deep learning |
| IDC | Invasive ductal carcinoma |
| EER | Equal Error Rate |
| ILC | Invasive lobular carcinoma |
| IMLN | Intramammary lymph node |
| LSTM | Long short-term memory |
| MIP | Maximum Intensity Projection |
| ML | Machine learning |
| MLP | Multilayer perceptron |
| MLO | Mediolateral oblique |
| MMST-D | Multimodal Sieve Transformer – DenseNet-121 encoder |
| MMST-V | Multimodal Sieve Transformer – Vision Transformer encoder |
| MRI | Magnetic resonance imaging |
| MV-Swin-T | Multiview Swin Transformer |
| MVI | Microvascular invasion |
| MVT | Multi-view Vision Transformer |
| NLP | Natural language processing |
| ROC | Receiver operating characteristic |
| THRIVE | T1-weighted High Resolution Isotropic Volume Excitation |
| UMD | Unimodal DenseNet-121 |
| UMV | Unimodal Vision Transformer |
| UF | Ultrafast |
| ViT | Vision Transformer |
| VOI | Volume of interest |
| XAI | Explainable artificial intelligence |

## 1. Introduction

Breast magnetic resonance imaging (MRI) is the most sensitive imaging modality for breast cancer detection [1]. Compared to mammography, breast MRI allows complete examination of breast areas hard to access like the retroareolar, medial, and axillary regions [2]. Breast MRI provides dynamic contrast-enhanced (DCE)-MRI sequences – facilitating the examination of contrast media uptake kinetics. This approach contributes to the detection of more invasive and high-grade cancers compared to mammography [3–6]. Nevertheless, its use for screening is limited due to its high costs and lower equipment availability compared to other imaging methods such as ultrasound or mammography [7,8]. In order to reduce costs, new approaches aiming at reducing acquisition and interpretation time such as abbreviated (AB-MRI) and Ultrafast (UF)-DCE breast MRI have been increasingly investigated [1,2,9–11]. In particular, UF-DCE MRI allows examination of the whole breast in less than 2 min, while increasing specificity [2,12,13].

Artificial intelligence (AI) has gained greater attention and research focus in the field of breast imaging [14]. Among the diverse deep learning (DL) architectures, Convolutional neural networks (CNN) constitute the most dominant architecture applied to classification tasks in computer vision [15–17], including breast cancer detection by MRI [17]. The convolution operations allow extraction of high level features from images by compressing them to reduce their initial size with pooling methods for efficient classification [15]. Several studies on UF-DCE using AI have been conducted with varying patient sample sizes, ranging from 137 [10,18] to 488 patients [19]. All these studies demonstrated encouraging performances with area under the receiver operating characteristic curve (AUROC) values greater than 0.81. Among these, Dalmiş et al. [20] used UF-DCE images along with additional patient information, resulting in a significant improvement of performances.

More recently, Transformer network architectures, initially designed for natural language processing (NLP) tasks, have gained in importance. They rely on attention mechanisms that enable the capture of relationships between the different parts of the input sequence, thus to transform a sequence of words into another sequence of words, such as in machine translation for example [21]. Derived from this new success, Transformers have been investigated in many fields of imaging like detection or classification [22], and the Vision Transformer (ViT) is the first Transformer model that was successfully used for image classification [22–24]. The ViT consists in a standard Transformer encoder fed with a sequence of fixed-size image patches, with position embeddings [24].

In the medical imaging field, the number of ViT-based papers published in 2021 exceeded the cumulative number of CNN-based papers published in the years 2012–2015 [23]. And compared to CNN models, the ViT architecture uses a global dependence between images patches due to self-attention mechanisms that are not present in CNN models [22]. Transformers also offer the advantage of being flexible in designing a range of architectures, including hybrid models [25]. This adaptability may address the challenge of managing multivariable data [26], especially in the context of breast MRI multiparametric protocol, where there are limited studies on the evaluation of various sequence combinations or multimodality [27].

Majority of AI applications in healthcare predominantly rely on only one data modality, however clinicians for their diagnostic decision-making use a multitude of data sources, including multiple examinations, patient information and past history [28]. Although breast lesion classification primarily relies on image-based approaches, recent studies suggest that non-image information, such as patient clinical data, can further improve the classification [20,29]. Usually, radiological reports contain such indications and can provide additional information that are not contained in the images and thus may be useful for classification improvement.

The combination of information with DL models is often referred to as data fusion [30]. Early fusion merges modalities before model processing, while intermediate fusion encodes modalities separately for a final combined model. Late fusion processes each modality independently, merging predictions at the end. Transformers open new possibilities to deal with multimodal data fusion. A recent paper proposed a Transformer-based AI model that integrates both chest radiographs and clinical information to diagnose 25 pathologic conditions [31]. They found that integration of both imaging and non-imaging data in this multimodal model performed better than unimodal models.

Multimodal learning with Transformers faces challenges including the collection of large curated multimodal datasets that is much harder than for unimodal datasets, a lack of studies on the interpretability of such multimodal models, an increased model complexity with high computational demand, alignment and fusion of information from different modalities [32,33]. Several fusion techniques exist, such as merge attention, co-attention or cross-attention [32,33]. However, it was noted that exploring and measuring the interaction between modalities would be interesting to further improve multimodal learning [32].

The aim of this study was to evaluate and compare unimodal and multimodal classification by combining patient clinical details (e.g., age, menopausal status, BRCA gene mutation status), lesion attributes (e.g., size, volume, and position in the breast) and UF-DCE data for breast

lesion classification. We propose a Transformer-based architecture encompassing 3D imaging, scalar and categorical data, that are encoded - or tokenized - and then injected into a "Sieve" Transformer encoder. This sieve extracts both mutual and exclusive information from each encoded modality. These information are then fused by a last aggregator Transformer encoder for final classification. The main contributions of this paper are summarized as follows.

1. We developed a novel Multimodal Sieve Transformer (MMST) architecture to extract mutual and exclusive features among modalities for an efficient and flexible adaptive classification that proved superior to unimodal approaches.
2. We collected and annotated the largest dataset to-date with multimodal data including UF-DCE images for validation of our model.

## 2. Related works

### 2.1. Deep learning for breast DCE-MRI

According to the American Cancer Society (ACS), MRI screening is recommended for women with a lifetime breast cancer risk of 20–25 % or higher, determined by factors such as family history or genetic predisposition [34,35]. MRI is favored due to its high sensitivity in detecting breast tumors, which are rapidly visible on imaging because of angiogenesis, especially within the first 2 min following contrast media injection in DCE-MRI sequences [36]. Recent advancements in AI, particularly DL, have shown significant potential in enhancing breast cancer detection. AI models have the potential to assist in decision support, use as a triaging tool, and as a second reader in breast cancer diagnosis workflows [37]. A comprehensive review of DL applications in breast MRI between 2015 and 2022 analyzed 18 papers focused on breast cancer detection and screening. This review revealed that most studies utilized private datasets and primarily focused on DCE-MRI images, often using CNN architecture [17].

In the literature, only a limited number of papers on DL approaches explored the use of UF-DCE sequences. Jing et al. [19] utilized a ResNet-34 model to classify maximum intensity projection (MIP) UF-DCE images from both the left and right breasts of 837 examinations. Prior to classification, a 3D U-Net architecture was applied to segment the breast region, generating masks for more precise localization. Their findings indicated that the model allowed exclusion of normal breasts from analysis, potentially reducing the radiologist's workload by focusing attention on suspicious cases. Dalmiş et al. [20] employed a custom 3D DenseNet architecture with two dense blocks, each containing three convolutional layers. They trained separate CNNs on cubic bounding boxes derived from MIP UF-DCE images and T2-weighted (T2w) images of lesions. To enhance diagnostic performance, they integrated the CNN-generated likelihood values with apparent diffusion coefficient (ADC) values and patient-specific information, such as age and BRCA gene status, within a random forest classifier. The study demonstrated that combining all imaging and patient data resulted in superior performance compared to using individual data sources.

### 2.2. Transformer-based deep learning in medical imaging

Transformers, introduced by Vaswani et al. [21], were initially developed for NLP tasks but have now gained impact in computer vision, including medical imaging. The Transformer architecture consists of two main components: the encoder and the decoder, each composed of several layers. The encoder processes input data through multiple layers of self-attention and feed-forward networks. The decoder generates output sequences using similar layers but incorporates an additional cross-attention mechanism that focuses on the encoder's output, enhancing the model's ability to generate accurate predictions. Both components use positional encoding to maintain the order of input data [21,23,38].

In the context of medical imaging, ViT, introduced by Dosovitskiy et al. [24], was the first Transformer model that was successfully used for image classification [22–24]. It consists of a standard Transformer encoder fed with a sequence of fixed-size image patches, with position embeddings. Unlike CNNs, which focus on small, localized, regions of an image, Transformers capture global context and model long-range dependencies through their self-attention mechanism. This characteristic enables Transformers to effectively capture fine-grained patterns and spatial relationships across an entire image for tasks like segmentation or classification [23].

Matsoukas et al. [39] conducted a study comparing ViT and CNN for medical image classification. Their findings reveal that ViTs pre-trained on large datasets like ImageNet performed comparably to CNNs for medical image classification tasks. They also highlighted that, with small datasets, CNN tends to outperform ViT when models are trained from scratch. When self-supervised pre-training combined with fine-tuning is used, ViT slightly performs better over comparable CNNs as the number of training samples increases. The difference of superiority for ViT is expected to grow as more training data becomes available [39], therefore pretraining on large datasets is essential for ViT. Moreover, ViT lack the inductive biases of CNNs by design, but at the expense of being more computationally expensive [25,38].

In the literature, various configurations of Transformer architectures have been explored, primarily categorized into pure Transformer approaches and hybrid CNN-Transformer approaches. Some studies provided a comprehensive overview of pure Transformer models, such as ViT and its variants, while also discussing hybrid models that integrate CNNs with Transformers to leverage the strengths of both architectures [25,38,40]. A more in-depth exploration of multi-Transformer approaches, including detailed analyses of ViT architectures and CNN-Transformer hybrid variants, is presented in the detailed survey by Khan et al. [41] for computer vision.

Both approaches were used in medical imaging research. Cao et al. [42] used a Transformer-based model (MVI-TR) consisting of an encoder module, whose job is to extract features from region of interest of Computed Tomography (CT) 2D slice with the maximum tumor area, and a final classifier module for prediction of preoperative microvascular invasion (MVI) in hepatocellular carcinomas. Their model achieved superior performances compared to a contrastive learning model and a ResNet architecture. Fan et al. [43] developed a parallel bi-branch model (Trans-CNN Net) based on Transformer module and CNN module. The features were extracted from the two branches and fused in a feature fusion module for chest CT 2D image classification. Their approach performed better than the compared CNN (ResNet-152) and pure Transformer (Deit-B) architectures.

In the breast imaging field, several papers compared ViT and CNN models for ultrasound images classification and showed that ViT could achieve comparable performances, with even superior results when self-supervised pretrained ViT were used [44,45]. Lee et al. [46] also used a combination of 2D CNN architecture based on a ResNet-34 model to extract features of digital breast tomosynthesis (DBT) from individual slices that were fed into a TimeSformer architecture aiming to capture context from neighboring sections, before the prediction took place. The study demonstrated improved performances for breast cancer classification in DBT images compared to per-section baseline approach.

### 2.3. Transformers and multimodal data

In clinical practice, clinicians integrate multiple data sources for diagnostic decision-making process. As a result, there is growing interest in developing AI approaches that integrate multimodal data, such as combining various medical imaging modalities and/or patient information derived from textual data from radiology reports. Different fusion strategies exist for merging multimodal data, classified as early, intermediate and late fusion strategies by Huang et al. [30], the early fusion approach being the most common. Early fusion, or feature-level

fusion, involves combination of multiple data modalities into a single feature vector before input to the machine learning (ML) model, using techniques like concatenation or pooling. In joint or intermediate fusion, learned feature representations from the intermediate layers of neural networks are merged with features from other modalities, before input to a final model, thus this approach involves loss propagation back to the feature extracting model. And in late fusion, each modality is processed in a model independently and then predictions are merged at the end [30].

Several works exploited the use of early fusion. Chen et al. [47] implemented a multi-view Vision Transformer (MVT) architecture designed to capture information from both craniocaudal (CC) and mediolateral oblique (MLO) views of breasts. The MVT architecture consisted of a local transformer to individually process image information using patch and positional embeddings. These extracted features were subsequently concatenated and fed into a global transformer block, where inter-mammogram dependencies were learned before the case was classified via a final multilayer perceptron (MLP). Similarly, Sarker et al. [48] employed multiview mammogram data using a Multiview Swin Transformer (MV-Swin-T), achieving superior performance over a baseline Swin-T model that relied on a single-view mammogram. Tong et al. [22] used a dual-input Transformer (DiT) for predicting preoperative pathological complete response to neoadjuvant chemotherapy. Their model integrated ultrasound images of the lesion obtained both before and after chemotherapy treatment. The DiT architecture included four main modules: each image was encoded via a tokens-to-token patch embedding module, followed by shared positional and temporal embeddings that enhanced the encoding of patch vectors before they were input to a Transformer encoder. The combination of before-and-after image data led to improved performance in prediction. In case of late fusion, Hussain et al. [49] investigated multimodal data fusion by integrating textual radiology reports with mammogram images across four views for breast cancer classification. Here, a ViT extracted image features, while a long short-term memory (LSTM) model or an artificial neural network (ANN) extracted features from text data. These features were then fused in a final linear classification layer. They compared this approach to CNN-based feature extraction methods, with the highest performance achieved using a VGG19 + ANN combination, surpassing Transformer-based feature extraction architectures. The use of joint fusion was reported by Cai et al. [50] who demonstrated the superiority of ViT over CNN for feature extraction in a multimodal model for skin disease classification. Similarly, Khader et al. [31] used a Transformer encoder to extract features from chest X-ray images, which were then combined with clinical parameters in a final Transformer encoder. This approach highlighted the advantages of integrating imaging and non-imaging data for disease diagnosis.

Recent literature highlights the promising potential of Transformer architectures to enhance performance, particularly in multimodal applications where diverse data sources are integrated. Despite these advances, their application in 3D data and breast MRI remains relatively unexplored.

## 3. Materials and methods

This study is part of the Smart and Ultrafast Breast MRI (SUBREAM) project funded by the Swiss Cancer Research (KFS-5460–08-2021-R) and approved by the Geneva Cantonal Ethics Committee (CCER) (Project-ID: 2019-00716). Informed consent was obtained from each patient for the re-use of anonymized breast imaging reports and MRI examinations.

### 3.1. Study population

A total of 301 breast MRI examinations and 1081 breast radiology reports from 240 patients were collected and processed retrospectively between 2019 and 2023 at a clinical center (Hirslanden – Clinique des Grangettes, Geneva, Switzerland) without initial exclusion criteria. The reports included the radiological reports of the breast MRI examinations written in French, but also up to four previous breast imaging reports (MRI, ultrasound or mammography when available), as they often provide additional clinical information. Subsequent exclusion criteria were applied to this dataset to avoid error due to incomplete data such as missing slices or interrupted acquisition (n = 4), presence of metal clip artefacts reducing lesion visualization (n = 6), and presence of false nodule image artifact (n = 1). As a result, a total of 290 MRI scans were pre-annotated by B.L and verified by the breast radiologists D-A.D. and K.K. for final lesion detection and characterization. This classification was performed by the two expert radiologists, with more than 12 years' (D.A.D.) and 20 years' (K.K.) experience in breast MRI, using histopathology reports following biopsy or surgery, and one-year follow-up examination.

### 3.2. Breast imaging technique

Breast MRI examinations were performed with a 3T MRI scanner (Ingenia, Philips) using a 16-channel breast coil. Standard full breast MRI protocol was acquired including a bilateral axial UF-DCE MRI sequence (a research sequence not used for the diagnosis in the MRI reports); four-dimensional T1-weighted High Resolution Isotropic Volume Excitation MR sequence (4D-THRIVE) acquired just before intravenous gadolinium contrast injection within 1 min (TR/TE = 3.4/1.72 ms, slice thickness = 2.5 mm, matrix = 480x480 pixels, temporal resolution = 3.3sec, number of temporal phases = 14). This sequence is based on combination of fast imaging techniques, CENTRA (Contrast ENhanced Timing Robust Angiography) and keyhole, meaning that only center of k-space is sampled at each temporal step and peripheral k-space is copied from the reference scan, i.e. the first phase [51]. This allows excellent temporal resolution, particularly interesting for lesion enhancement investigation. Patients were scanned in prone position, and intravenous contrast injection was performed (contrast bolus at a rate of 2.5 mL/s followed by 20 mL NaCl flush with the same rate) immediately after the beginning of UF-DCE sequence acquisition.

### 3.3. Data preprocessing

Fig. 1 illustrates the data utilized in this study, comprising both *imaging* and *non-imaging data* rigorously collected and extracted.

*Imaging lesion data*: Segmentation of all lesions was performed with Philips IntelliSpace Portal 8.0 using semi-automatic 3D segmentation tool based on pixel intensity thresholding on the images of the last phase of the UF-DCE images presenting the maximum visible enhancement. This yielded a final dataset consisting of 987 segmented lesions divided into 3 lesion categories: 280 benign (B cat.), 121 malignant lesions (M cat.) and 586 benign lymph nodes (L cat.). Isotropic bounding boxes of 50 mm × 50 mm x 50 mm were generated from the 3D segmentations surrounding the center of the lesion and were used to extract 3D image volume of interest (VOI) from the last subtracted UF-DCE phase. It has been shown that bounding boxes including a small proportion of breast tissue around the lesion contribute to better accuracy compared to segmentation alone or to large bounding boxes [52]. The 50 mm size was therefore chosen as proposed by Dalmış et al. [20]. For lesions bigger than 50 mm (n = 4), multiple overlapping 50 $mm^3$ boxes were generated over the lesion, yielding multiple boxes for a single lesion.

*Non-imaging data*: Lesion characteristics, including lesion volume, center and size of the bounding box, and gravity center and diameters of the lesion, as well as elongation and flatness were calculated from the segmentations. Elongation and flatness definitions can be found in a study [53], where authors also investigated these shape descriptors in a ML model applied to diagnose bladder cancer in MRI. These descriptors were also used by Militello et al. [54] in a ML model combining shape and radiomics descriptors to characterize breast lesions in DCE-MRI. According to risk factors of breast cancer [55], and factors with

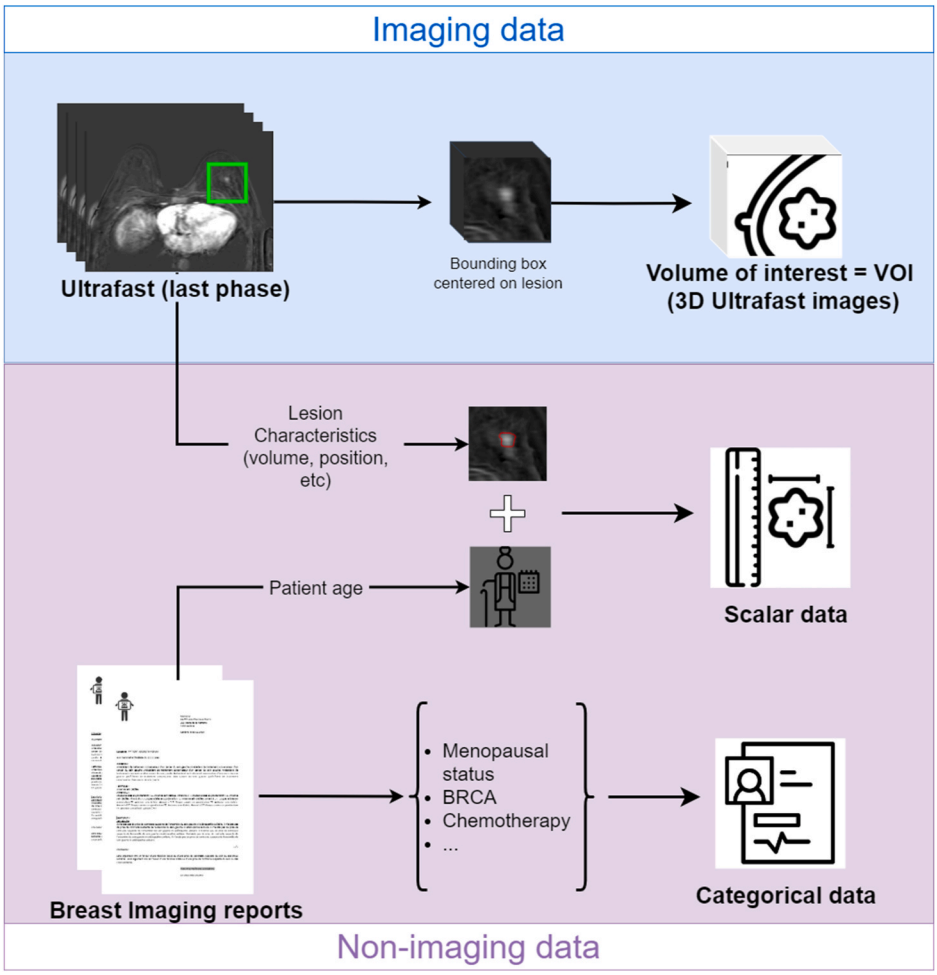


**Fig. 1.** Visual explanation of data used in this research.

potential influence on contrast enhancement of breast tissue, clinical information was annotated and extracted from the breast imaging reports using brat rapid annotation tool [56,57]. These clinical data information were contained in the "indications" part of the report, and were not correlated with the diagnosis, as they included menopausal status, contraception, personal and family history of breast cancer, BRCA mutation, and chemotherapy treatment status. For each clinical parameter, specific attributes were chosen, and when reports showed no corresponding information related to the clinical data it was categorized as "No info". Detailed attributes for each clinical parameter are provided in the Supplementary material S1. Finally for practical reasons, non-imaging data were categorized as two types: categorical data (extracted clinical parameters), and scalar data (lesion characteristics and patient age), see Fig. 1.

### 3.4. Machine learning and deep learning architectures

To discriminate lesions, different classification approaches were used. First, a traditional ML algorithm, relying on scalar and categorical data (non-imaging data), was employed. Then a DenseNet121 [58] DL model was trained and tested only on images acquired with UF-DCE only. Lastly, we developed a multimodal model combining imaging and non-imaging data. The proposed DL architectures were implemented using the PyTorch framework and the MONAI library [59]. ML classifier relied on Random Forests and was implemented with scikit-learn. The training and testing of the models were run on a computing unit consisting of 384 GB of CPU RAM and of an NVIDIA Tesla V100 SXM2 with 32 GB of GPU RAM.

We defined different classification scenarios involving the lesion categories B, L and M: two-class scenarios, benign lesions with lymph nodes versus malignant lesions (BL_M classification) and benign versus malignant lesions (B_M classification), as well as the three classes case, benign lesions versus lymph nodes and versus malignant lesions (B_L_M classification). Dataset was randomly split into train (60 %), validation (20 %), and test (20 %) sets. Balanced sampling and stratified five-fold cross-validation were performed respecting class prevalence, percentage of lesion type in the dataset (ensuring representative data), and patient-wise separation. Thus, all lesions from the same patient were kept in the same partition when building the folds for cross-validation. The same cross-validation folds were used for all experiments. Before being fed to models, images were processed as follows: 50 $mm^3$ VOIs were extracted from 3D images based on lesion mask bounding box centers. Then, the intensity of voxels of the resulting images was normalized between 0 and 1. Finally, images were resampled to 64x64x64 voxels volumes. Scalar data were also standardized (mean = 0; standard deviation = 1), and categorical data were converted to one-hot vectors.

#### 3.4.1. Non-imaging data model

*Random Forest classifier:* Among the various ML approaches that we tested, the Random Forest classifier [60] performed better using both scalar (SCA) and categorical (CAT) data (detailed parameters are provided in Supplementary material S2). Multiple experiments were performed with all SCA and/or all CAT features data as input, and also some individual SCA or CAT feature data.

### 3.4.2. Unimodal image models

For image-only classifications, we explored a CNN-based architecture and a Transformer-based one. In a previous work [61], we performed a comparative analysis between the two most commonly used models, ResNet-50 and DenseNet-121, utilizing two distinct input data types: subtracted images from the final phase of UF-DCE MRI and MIP images. The results of that study demonstrated that DenseNet-121 consistently outperformed ResNet-50, regardless of the data type input. We therefore chose DenseNet-121 for the CNN-based framework.

*Unimodal DenseNet-121 (UMD)*: The CNN-based model consisted in a DenseNet-121 composed with 121 densely connected convolutional layers [58], was trained and tested on VOIs. It is extensively used in medical imaging tasks due to its ability to effectively extract features from images while reducing the vanishing gradient problem [62]. Although DenseNet121 is commonly pretrained on 2D image datasets such as ImageNet, there is no equivalent for their 3D counterpart. Therefore, the model was trained from scratch on our 3D dataset. Training was made with the Adam optimizer [63], and used a learning

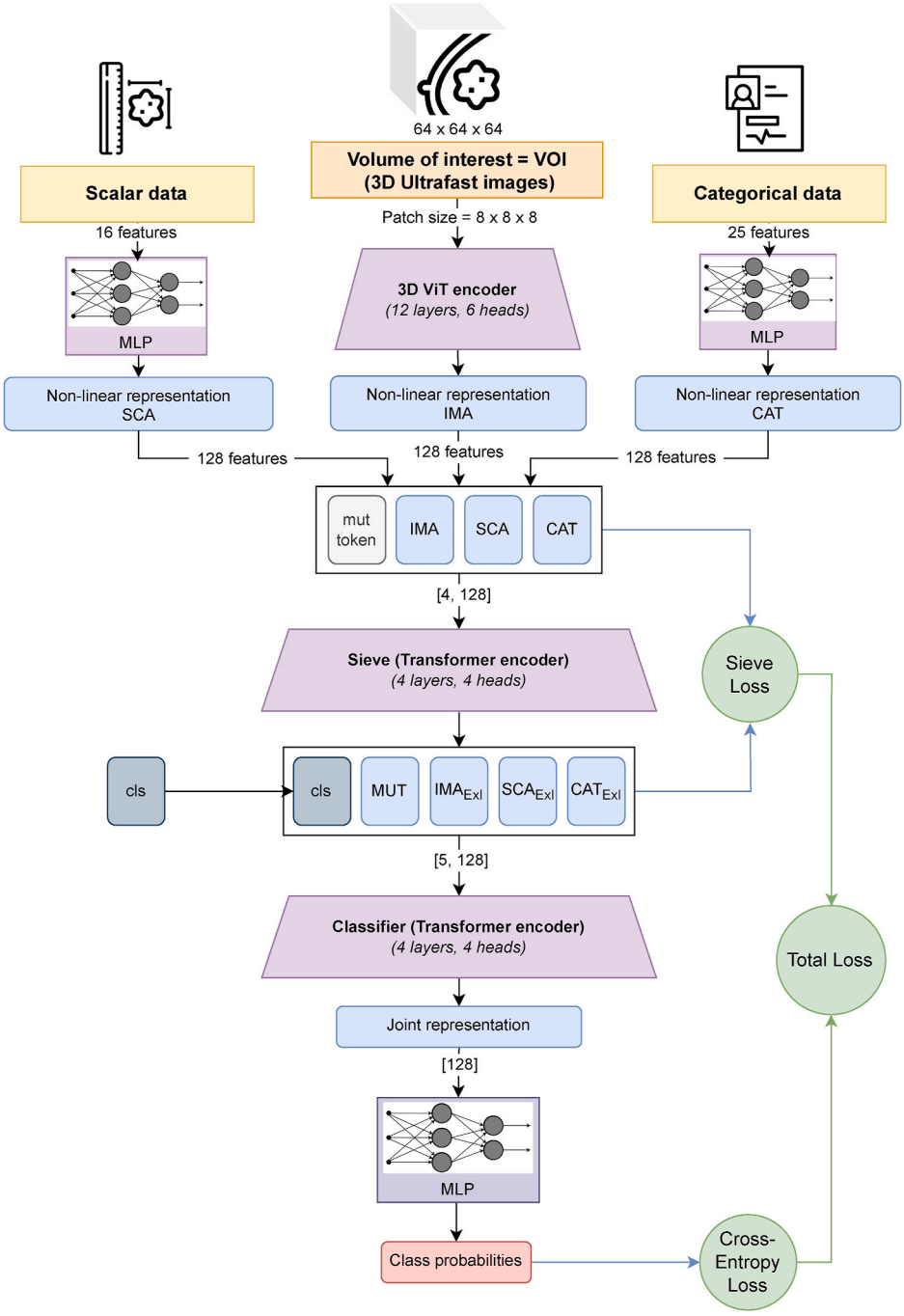


**Fig. 2.** MMST-V general architecture

Input Modalities: Use 2D scalar, 2D categorical, and VOI imaging data. First Stage: Process each modality separately using a 3D ViT encoder (or DenseNet-121 for MMST-D) for VOI images and an MLP for non-imaging data. Second Stage: Use a Sieve Transformer encoder to separate mutual and exclusive information from each modality, reducing feature redundancy. Third Stage: Fuse the representations using another Transformer encoder to create a joint representation. Output: Feed the joint representation to an MLP classifier to yield class probabilities.

MLP: Multi-Layer Perceptron, ViT: Vision Transformer, SCA: Scalar, CAT: Categorical, IMA: Images, MUT: Mutual, cls: class token, Exl: Exclusive.

rate of $10^{-4}$, 200 epochs with a batch size of 64. During the training, several data augmentations were applied, details are provided in Supplementary material S2.

*Unimodal Vision Transformer (UMV)*: The Transformer-based model consisted in a custom MONAI's 3D ViT [59] based on Dosovitskiy et al. work [24], that was not pretrained. We trained UMV only for the 3-class classification task (B_L_M), to compare with UMD.

#### 3.4.3. Multimodal Sieve Transformer models

*Multimodal Sieve Transformer with ViT (MMST-V):* The proposed DL architecture uses multiple modalities including non-imaging (scalar and categorical data) and imaging (VOIs) data in input and is based on a multi-stage Transformer encoder architecture. For the first stage, each modality is processed separately, VOI images by a small 3D ViT encoder [24] and non-imaging data by a MLP. In the second stage, representation features of all modalities are fed into a "Sieve" Transformer encoder trained to separate mutual and exclusive information from each modality, thus reducing redundancy of features. This involves training the Sieve Transformer encoder to maximize mutual information between the initial representation and a mutual token, while minimizing redundancy between the mutual token and the exclusive tokens, by calculating cross-correlation matrices between them [64,65]. This procedure results in one representation per modality carrying exclusive information along with an additional representation containing information shared between modalities. In the last stage, an additional Transformer encoder was used for fusing exclusive and mutual representations into a joint representation. This joint representation is then fed to a MLP classifier, yielding class probabilities.

The overall total loss function is the combination of sieve loss components, mutual $l_m$ and exclusive $l_e$ loss terms, with the traditional binary cross-entropy loss $l_{bce}$:

$$\mathscr{L}_{TOT} = \lambda_m l_m + \lambda_e l_e + \lambda_{bce} l_{bce} \quad (1)$$

where parameters λ are weighting factors of the different components of the loss function, whose computation details are given in the provided pseudocode in *Algorithm 1* - Supplementary material S3. The pipeline of the proposed MMST-V is illustrated in Fig. 2, and the main training hyperparameters for the model are presented in Supplementary material S3, along with the various data augmentations applied to the input data.

*Multimodal Sieve Transformer with DenseNet-121 (MMST-D):* For comparison, we also trained and tested an MMST-D model, which is based on MMST but uses a DenseNet-121 encoder to process the images (the same as used in UMD model). The MMST-D model was specifically trained for the 3-class classification task (B_L_M) for comparison with MMST-V.

### 3.5. Statistical analysis

Statistical analyses were conducted using Stata Statistical Software 16.1 (StataCorp LLC) and Torchmetrics [66]. ROC analysis was used for performance evaluation, as well as the metrics of AUROC, accuracy, sensitivity, and specificity. The threshold for sensitivity, specificity and accuracy was set at the Equal Error Rate (EER), meaning that both false positives and false negatives rates are equivalent. Thresholds at respectively 90 % and 95 % sensitivity were also calculated, as achieving these levels of sensitivity is clinically important and consistent with thresholds used in other studies [18,19]. The results are presented as mean ± standard deviation. A Wilcoxon-Mann-Whitney test was used to assess statistical differences between folds of models with a significance threshold $\alpha$ set at 0.05.

## 4. Results

### 4.1. Study population and breast lesion characteristics

In this study, from 240 patients we used 290 breast MRI examinations and 1081 radiology reports (MRI, ultrasound or mammography). The extraction of clinical information was made for each report, all clinical information was made for each patient using combination of all available radiology reports. The information combination followed several rules detailed in Supplementary material S1. Table 1 shows the population characteristics extracted from the radiology reports. Patient were divided according to the indication of breast MRI separating patient with known cancer and those without cancer at the time of MRI examination. Most patients did not have known cancer (78.2 %), compared to patient with known cancer (21.7 %). There was a high prevalence of missing information denoted as "*No info*", indicating instances where the information was not mentioned within the reports. Unknown BRCA mutation status (90.3 %) and chemotherapy (96.9 %) were among the most often criteria missing. No difference in age was observed (mean age = 54.8 ± 12.7, p = 0.7156) for patients with or without known cancer, and there was no significant difference of menopausal status, contraception, or in family history of breast cancer ($p > 0.05$). Almost half of women were on menopause (n = 145) and therefore did not take any contraception (n = 148), and one-third had family history of breast cancer (n = 104).

Our VOI lesion dataset exhibited predominant proportion of lymph nodes (59.4 %), followed by benign lesions (28.4 %) and malignant lesions (12.3 %). Further lesion characterization used the precise pathology results from biopsy or surgery, and Table 2 shows the main subtypes present in the lesion dataset. Lesion subtypes were usually similarly distributed in train, validation and test sets across the five folds (Supplementary material S5). Since there were multiple sub-lesion VOI generated for malignant lesions as they tend to be larger (>50 mm) and we carefully applied a patient-wise separation, ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC) presented a non-harmonious distribution between the folds as depicted in Fig. 3.

A significant difference of volume was found between benign and malignant lesions ($p < 0.05$), malignant lesions tending to have larger volume than benign lesions. IDC and DCIS were among the largest. There was also a significant difference in the position of the lesions depicted either with bounding box or gravity centers of the lesion along z axis and y axis. Lymph nodes used to be located towards axillary region (Fig. 4). Majority of lymph nodes were axillary lymph nodes in comparison to intramammary lymph nodes, thus explaining this difference. The main statistics of lesion characteristics by type of lesion are reported in Supplementary material S4. Certain lesion information exhibit redundancy, showing expected correlations among them. For instance, bounding box size, ellipsoid diameters, and lesion volume appeared to be correlated, displaying similar variations across different lesions. This is also true for position, bounding box center and lesion gravity center (Supplementary material S4).

### 4.2. Classification experiments

*Non-imaging data model:* Multiple combinations of scalar (SCA) and/or categorical (CAT) features as input were tested with Support Vector Machines (SVM), Random Forests and Logistic Regression classifiers. Initial findings revealed superior performances from Random Forests, that were therefore chosen for the experiments. Best performances were achieved when all data features were considered for all classification tasks (AUROC ranged in [0.891–0.903]).

**Table 1**
Study population characteristics and clinical information.

| | Total patient n=290 | No cancer n=227 (78.2%) | Cancer n=63 (21.7%) | p-value |
|---|---|---|---|---|
| **AGE (years)** | | | | |
| mean (± std) | 54.8 (±12.7) | 54.5 (±11.9) | 56.1 (±13) | *0.7156* |
| [min, max] | [22, 92] | [22, 81] | [36, 92] | |
| **MENOPAUSAL STATUS (n, %)** | | | | |
| Absent | 53 (18,3%) | 43 (18.9%) | 10 (15.8%) | |
| Present | **145 (50%)** | **119 (52.4%)** | **26 (41.2%)** | *0.191* |
| With substitute | 26 (8.9%) | 19 (8.4%) | 7 (11.1%) | |
| No info | 66 (22.75%) | 46 (20.2%) | 20 (31.7%) | |
| **CONTRACEPTION (n, %)** | | | | |
| With | 8 (2.7%) | 6 (2.6%) | 2 (3.2%) | |
| Without | **148 (51%)** | **121 (53.3%)** | **27 (42.8%)** | *0.133* |
| Other | 1 (0.3%) | . | 1 (1.6%) | |
| No info | 131 (45.2%) | 98 (43.2%) | 33 (52.4%) | |
| **FAMILY HISTORY OF BC (n, %)** | | | | |
| No | 76 (26.2%) | 56 (24.7 %) | 20 (31.7%) | |
| Yes | 104 (35.8%) | 86 (37.9) | 18 (28.5%) | *0.532* |
| No info | **110 (37.9%)** | **85 (37.4%)** | **25 (39.7%)** | |
| **PERSONAL HISTORY OF BC (n, %)** | | | | |
| No | 4 (1.4%) | 4 (1.7%) | - | |
| Yes | **186 (64.1%)** | **131 (57.7%)** | **55 (87.3%)** | *0.000* |
| Other | 19 (6.5%) | 18 (7.9%) | 1 (1.6%) | |
| No info | 81 (27.9%) | 74 (32.6%) | 7 (11.1%) | |
| **BRCA MUTATION (n, %)** | | | | |
| Negative | 9 (3.1%) | 9 (3.9%) | - | |
| Positive | 19 (6.5%) | 19 (8.4%) | - | *0.014* |
| No info | **262 (90.3%)** | **199 (87.6%)** | **63 (100%)** | |
| **CHEMOTHERAPY (n, %)** | | | | |
| No | **281 (96.9%)** | **225 (99.1%)** | **56 (88.8%)** | *0.000* |
| Yes | 9 (3.1%) | 2 (0.8%) | 7 (11.1%) | |

**BC** = Breast cancer.
**Bold style** = Category with higher proportion.

Further analysis indicated that Random Forests demonstrated enhanced performance with SCA data, specifically concerning the volume and position of lesions in comparison with categorical data. Lesion position had a greater influence for the classification tasks including lymph nodes. In contrast, when comparing only benign and malignant cases (less lesions to consider), the impact of volume was more substantial due to significant volume variations among the two different lesion types (benign and malignant). A comprehensive overview of the experimental results, including the impact of different data input types, is presented in Table 3, for the main categories with different classification scenarios. Performances of the other CAT features (such as menopausal status, family history of breast cancer or BRCA) taken alone were systematically lower than those shown in Table 3 (AUROC <0.590), except for patient history of breast cancer that was the best feature for CAT data (AUROC = [0.634, 0.648, 0.656]).

*Unimodal image models comparison*: For the 3-class classification task (B_L_M), a notable performance difference was observed between the UMD and UMV models, with UMD achieving an AUROC = 0.863 ± 0.025 compared to UMV's AUROC = 0.731 ± 0.056 (Table 4). The UMV model demonstrated the lowest performance across all experiments, with an AUROC below 0.800 and an average accuracy of 66.5 ± 9.9 %.

*Multimodal models comparison*: A comparison between the MMST-D and MMST-V models for the 3-class classification task (B_L_M) revealed a performance difference, with MMST-V outperforming MMST-D. Specifically, MMST-V achieved an AUROC of 0.928 ± 0.027, while MMST-D obtained an AUROC of 0.890 ± 0.030 (Table 4).

For all classification tasks (B_L_M, BL_M, and B_M), and considering only non-imaging, UMD and MMST-V, classification with solely non-imaging data (AUROC = [0.900, 0.891, 0.903]) achieved better performances than the UMD model using imaging data only (AUROC = [0.863, 0.863, 0.814]) for all classification scenarios. Performances of classification MMST-V performed systematically better than both non-imaging and UMD model for each classification task as depicted in Table 4 and Fig. 5. There were no overall significative differences between each classifier AUROC values across the five folds. The only significative difference was found between UMD and MMST-V AUROC values for B_L_M classification task (p-value = 0.0159).

Best performances were obtained with a 3-class classification and MMST-V (AUROC = 0.928, $ACC_{avg}$ = 88.2 %). Highest specificity was obtained with MMST-V at 90 % (Sp = 70.8 %) and 95 % (Sp = 47.0 %) sensitivity thresholds for both B_L_M and B_M classification tasks. For BL_M classification, the highest specificity was obtained with non-imaging data only (Sp = 36.7 %, 66.7 %) as depicted in Table 5.

## 5. Discussion

This study presents a comprehensive analysis of a large dataset comprising 987 breast lesions, including benign and malignant lesions, as well as lymph nodes commonly encountered in clinical settings. We completed this dataset with clinical information extracted from radiology reports, as they are essential for accurate diagnosis. Despite the inherent imbalance in the dataset, our approach enabled us to develop

**Table 2**
Lesion type.

| | |
|---|---|
| **B - BENIGN (28.4%)** | **280** |
| Fibrocystic breast changes | 175 |
| Fibroadenoma | 34 |
| Benign (type not specified) | 25 |
| Fat necrosis | 12 |
| Papilloma | 10 |
| Fibrosis | 8 |
| Post therapeutic scar remodeling | 5 |
| Galactophoric ducts inflammation | 3 |
| Nevus | 2 |
| Other benign lesions | 6 |
| **L - LYMPH NODES (59.4%)** | **586** |
| Benign lymph nodes | 586 |
| **M - MALIGNANT (12.3%)** | **121** |
| Invasive ductal carcinoma (IDC) | 60 |
| Cancer (no histopathology) | 21 |
| Ductal carcinoma in situ (DCIS) | 14 |
| Invasive lobular carcinoma | 11 |
| Lymphadenopathy (M-LN) | 10 |
| Atypical lobular hyperplasia | 2 |
| Atypical papilloma – B3 | 2 |
| Papillary carcinoma in situ | 1 |

and evaluate various classification tasks using different DL architectures.

### 5.1. *Influence of multimodal data on model performance*

The classification performance of unimodal imaging models was the lowest compared to non-imaging and multimodal (MMST) models. Although ViT have demonstrated high potential in previous studies [23,44], our unimodal ViT (UMV) model showed suboptimal performance relative to the unimodal DenseNet-121 (UMD) architecture. This lower performance could be attributed to several factors, including the relatively small size of our dataset and the lack of pretrained ViT models optimized for our specific medical imaging domain as highlighted in the literature [25,38,39].

We also explored addition of scalar data composed of lesion shape features, as proposed and investigated in the literature [54,67], and patient clinical information that provide a richer context to the image data [29,68,69]. The MMST-V emerged as the best performing model for three-class classification (B_L_M) when compared to non-imaging and unimodal approaches. When comparing MMST-D and MMST-V models, we observed that, contrary to the unimodal results, the MMST model with a ViT encoder outperformed MMST-D. Cai et al. [50] similarly reported that ViT used as a backbone generally performed better than CNNs for feature extraction in multimodal frameworks, showing an advantage over DenseNet encoders when pretrained. In our study, the multimodal approach and the Sieve encoder likely compensated for the lack of pretraining, contributing potentially to the improved performance of the MMST-V model.

The addition of lesion characteristics and clinical information notably enhanced classification performance, with SCA data, particularly lesion position and volume, showing significant influence. In our dataset, malignant lesions tended to exhibit larger volumes compared to benign lesions. This significant difference was also observed by Militello et al. in their balanced dataset composed of 57 benign and 54 malignant lesions [54] and also in the study of Abe et al. [70]. Lymph nodes were predominantly located in axillary regions, facilitating their discrimination based on solely position information or volume lesion. This observation can surely be attributed to the prevalence of axillary lymph nodes (ALN) versus intramammary lymph nodes (IMLN) in the breasts, that are positioned deeper and more externally in the breasts (Fig. 4). However, further granularity in categorizing lymph nodes, distinguishing between intramammary and axillary types, would provide better analysis for diagnosis, particularly regarding IMLN [71,72]. It should be noticed that some ALN could be positive (malignant) with cancer spreads towards lymph nodes. This was observed in our dataset and illustrated in Fig. 4, where ten red points denote axillary regions with such lymphadenopathy (M-LN) (see Table 2). Three classification tasks (B_L_M, BL_M, B_M) were chosen to evaluate and analyze clinical typical scenarios, as sometimes in clinical settings lymph nodes are easy to distinguish from breast lesions and sometimes they are not.

While the impact of lesion position was significant in classification

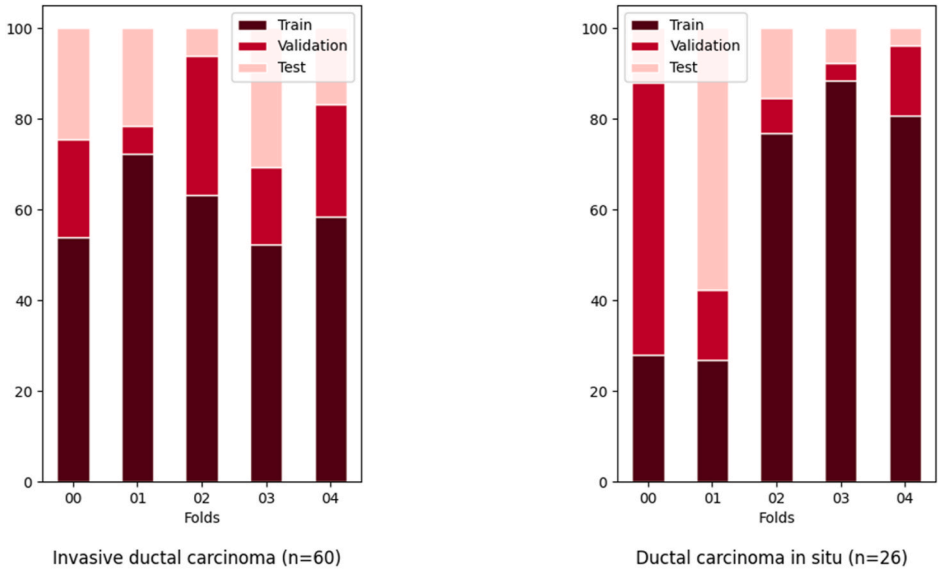


**Fig. 3.** IDC and DCIS distribution (train, validation, test sets) across the five folds.

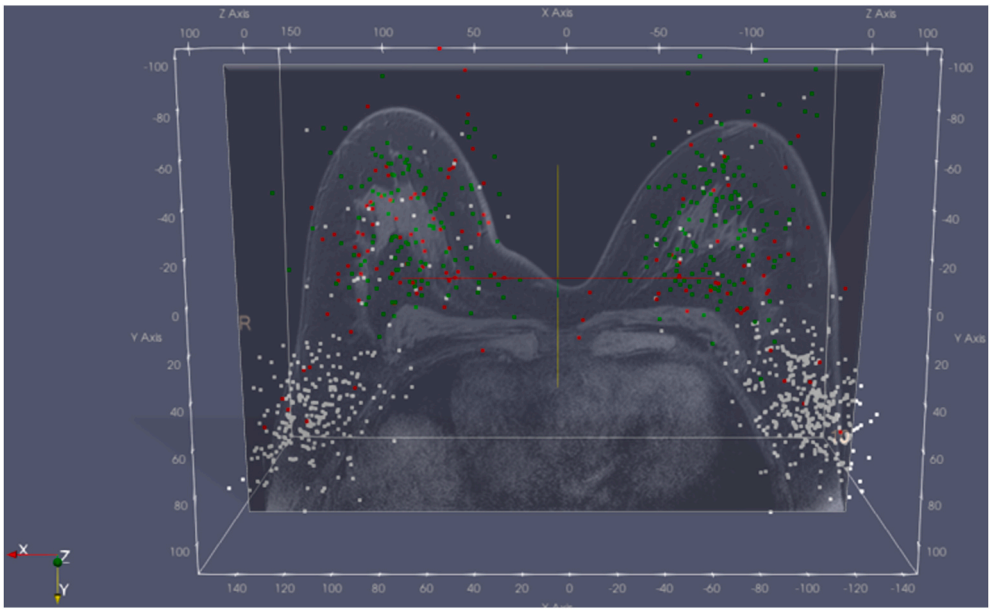


**Fig. 4.** Superposition of a T1w breast image with a standard breast morphology and the 3D positions of the 987 lesions (red dots: malignant lesions, green dots: benign lesions, white dots: lymph nodes). This visualization illustrates the high concentration of lymph node positions in the axillary regions. But it does not take into account differences in breasts and lesions size.

**Table 3**
Random forests 3-class and 2 class classification performances for different combination of scalar and categorical data (non-imaging data).

| Data | AUROC [%] | $ACC_{max}$ [%]* | $ACC_{avg}$ [%]* | Se [%]* | Sp [%]* |
|---|---|---|---|---|---|
| 3-class classification (B_L_M) | | | | | |
| **all CAT + SCA** | **90.0 (± 4.5)** | **92.0 (± 1.3)** | **84.3 (± 1.4)** | **83.1 (± 6.0)** | **83.4 (± 5.9)** |
| all SCA | 87.5 (± 4.2) | 91.9 (± 1.2) | 84.7 (± 1.3) | 77.3 (± 5.2) | 77.8 (± 5.4) |
| all CAT | 68.0 (± 6.0) | 87.2 (± 1.0) | 77.5 (± 3.2) | 63.7 (± 6.1) | 63.8 (± 6.7) |
| Patient age | 47.6 (± 8.9) | 51.1 (± 8.7) | 51.1 (± 8.7) | 42.9 (± 19.6) | 52.4 (± 11.2) |
| Lesion volume | 66.3 (± 8.5) | 87.0 (± 2.4) | 87.0 (± 2.4) | 38.1 (± 20.0) | 94.5 (± 4.2) |
| VOI center (xyz) | 70.1 (± 4.0) | 86.3 (± 1.5) | 77.7 (± 1.8) | 65.9 (± 2.9) | 65.6 (± 2.8) |
| 2-class classification (BL_M) | | | | | |
| **all CAT + SCA** | **89.1 (± 2.5)** | **91.6 (± 1.1)** | **83.4 (± 1.3)** | **80.6 (± 2.6)** | **81.1 (± 2.7)** |
| all SCA | 87.0 (± 2.0) | 91.4 (± 1.5) | 84.4 (± 0.8) | 79.5 (± 2.8) | 79.4 (± 3.7) |
| all CAT | 67.1 (± 5.4) | 87.1 (± 1.3) | 77.3 (± 2.4) | 62.7 (± 5.6) | 62.2 (± 3.8) |
| Patient age | 47.7 (± 8.6) | 48.5 (± 6.9) | 48.5 (± 6.9) | 46.8 (± 21.3) | 48.7 (± 9.7) |
| Lesion volume | 66.1 (± 6.8) | 87.2 (± 2.4) | 87.2 (± 2.4) | 36.9 (± 14.8) | 95.3 (± 3.3) |
| VOI center (xyz) | 71.3 (± 2.1) | 86.4 (± 1.0) | 78.2 (± 1.8) | 65.3 (± 5.6) | 64.0 (± 3.5) |
| 2-class classification (B_M) | | | | | |
| **all CAT + SCA** | **90.3 (± 6.0)** | **87.2 (± 3.9)** | **75.5 (± 4.0)** | **83.4 (± 5.5)** | **83.0 (± 5.0)** |
| all SCA | 85.9 (± 5.0) | 84.1 (± 4.2) | 73.5 (± 3.5) | 76.8 (± 4.7) | 76.9 (± 4.8) |
| all CAT | 71.4 (± 10.3) | 73.8 (± 5.3) | 62.6 (± 5.6) | 64.9 (± 10.3) | 64.9 (± 9.9) |
| Patient age | 45.0 (± 6.9) | 45.1 (± 7.2) | 45.1 (± 7.2) | 44.7 (± 13.9) | 45.3 (± 11.4) |
| Lesion volume | 62.1 (± 6.3) | 74.0 (± 3.9) | 74.0 (± 3.9) | 27.0 (± 13.4) | 97.2 (± 2.1) |
| VOI center (xyz) | 53.6 (± 7.6) | 67.7 (± 2.2) | 56.2 (± 1.9) | 53.6 (± 5.4) | 54.2 (± 5.4) |

AUROC = Area under ROC curve, $ACC_{avg}$ = Average accuracy, SCA = Scalar data, CAT = Categorical data.
*ACC, sensitivity and specificity were calculated at Equal Error Rate. Bold style = best performance.

involving lymph nodes lesions, performances diminished when distinguishing between benign and malignant cases only (B_M). As shown in Table 3, performance when using SCA data of VOI center (xyz) position in the breast goes from AUROC of 0.701 (B_L_M) and 0.713 (BL_M) to only 0.536 for B_M classification task.

Analysis of patient age revealed no significant difference between patients with and without cancer, and no changes across classification tasks. However, patients undergoing chemotherapy ($p < 0.05$) or with a personal history of breast cancer ($p < 0.05$) tended to have breast cancer in the dataset, introducing a slight bias towards cases with known breast

**Table 4**
Performances of the different classification tasks with the different models.

| | AUROC [%] | $ACC_{avg}$ [%]* | Se [%]* | Sp [%]* |
|---|---|---|---|---|
| 3-class classification (B_L_M) | | | | |
| Non-imaging | 90.0 (± 4.5) | 84.3 (± 1.4) | 83.1 (± 6.0) | 83.4 (± 5.9) |
| UMD | 86.3 (± 2.5) | 86.9 (± 1.6) | 76.1 (± 6.8) | 78.3 (± 4.1) |
| UMV | 73.1 (± 5.6) | 66.5 (± 9.9) | 67.9 (± 4.8) | 68.4 (±.4.9) |
| MMST-D | 89.0 (± 3.0) | 79.9 (± 3.5) | 80.2 (± 3.5) | 79.8 (± 3.5) |
| **MMST-V** | **92.8 (± 2.7)** | **88.2 (± 0.9)** | **86.7 (± 2.8)** | **86.3 (± 3.0)** |
| 2-class classification (BL_M) | | | | |
| Non-imaging | 89.1 (± 2.5) | 83.4 (± 1.3) | 80.6 (± 2.6) | 81.1 (± 2.7) |
| UMD | 86.3 (± 4.6) | 80.2 (± 2.8) | 77.7 (± 3.5) | 77.7 (± 3.5) |
| **MMST-V** | **90.5 (± 2.4)** | **86.4 (± 2.1)** | **82.7 (± 3.7)** | **82.8 (± 3.7)** |
| 2-class classification (B_M) | | | | |
| Non-imaging | 90.3 (± 6.0) | 75.5 (± 4.0) | 83.4 (± 5.5) | 83.0 (± 5.0) |
| UMD | 81.4 (± 10.0) | 68.7 (± 4.9) | 75.5 (± 9.4) | 75.6 (± 9.3) |
| **MMST-V** | **91.6 (± 2.9)** | **83.3 (± 4.6)** | **84.1 (± 5.3)** | **83.4 (± 5.4)** |

AUROC = Area under ROC curve, $ACC_{avg}$ = Average accuracy, Se = Sensitivity, Sp = Specificity.
*ACC, sensitivity and specificity were calculated at Equal Error Rate. Bold style = best performance.
UMD = Unimodal DenseNet-121, UMV = Unimodal ViT; MMST-D = Multimodal Sieve Transformer with DenseNet-121 backbone, MMST-V = Multimodal Sieve Transformer with ViT backbone.

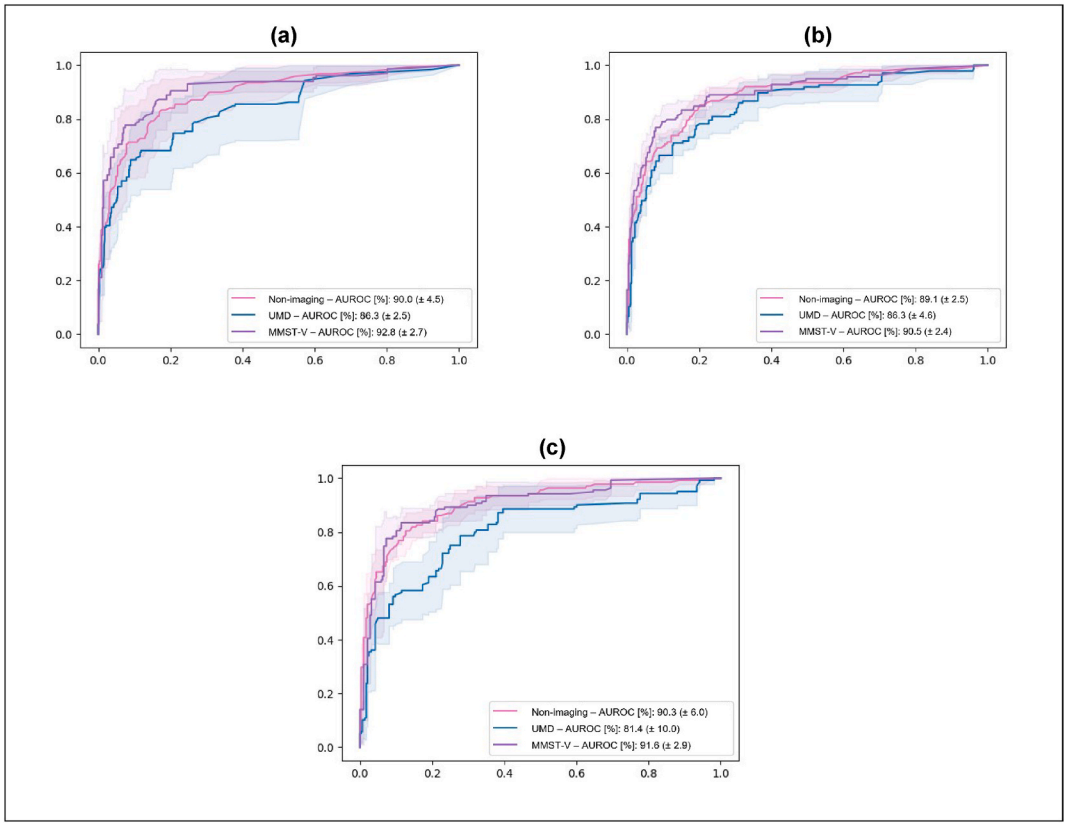


**Fig. 5.** Receiver operating characteristic (ROC) curves of Random Forest classifier, UMD and MMST-V for B_L_M classification **(a)**, BL_M classification **(b)**, and B_M classification **(c)**.

cancer at the time of examination (Table 1), meaning that maybe a difference should have been made between past history of breast cancer, and cancer known at the time of examination for staging MRI indication (assess extent of the disease). However, analysis of misclassified cases did not reveal any noticeable influence on the classification, as there were both false positives and true negatives with the same personal risk

**Table 5**
Specificity at different sensitivity thresholds.

| | B_L_M | BL_M | B_M |
|---|---|---|---|
| **Specificity at 95% Sensitivity [%]** | | | |
| Non-Imaging | 45.1 (± 23.3) | **36.7 (± 20.2)** | 40.1 (± 21.4) |
| UMD | 34.3 (± 16.7) | 32.8 (± 21.9) | 24.2 (± 22.7) |
| **MMST-V** | **47.0 (± 24.4)** | 34.6 (± 17.0) | **43.4 (± 18.5)** |
| **Specificity at 90% Sensitivity [%]** | | | |
| Non-Imaging | 65.4 (± 22.1) | **66.8 (± 13.6)** | 58.3 (± 27.8) |
| UMD | 53.1 (± 14.6) | 51.9 (± 22.2) | 44.6 (± 27.4) |
| **MMST-V** | **70.8 (± 14.4)** | 64.4 (± 14.0) | **70.1 (± 7.7)** |

Bold = best performance.

of breast cancer. Among the CAT data for non-imaging classification, patient history of breast cancer had greater importance than other clinical categorical data information. Mostly, this included the assessment of the extent of the disease, indicating that the patient has a known cancer at the time of the MRI examination. Similarly, the study of Holste et al. [73] found that the MRI indication, which included the assessment of the extent of the disease, was the most important feature.

Despite achieving an overall excellent classification performance, the MMST-V model exhibited some limitations. When evaluated at equal error rate (EER), 15 false negative cases were identified and much more false positive cases (n = 155). Missed cancers (n = 15) were mostly M-LN (n = 6), DCIS (n = 3) and IDC (n = 4) respectively. Two sensitivity thresholds of 95 % and 90 % were chosen. When sensitivity threshold was highest at 95 %, there were less false negatives, but a strong augmentation of false positive lesions (lower specificity of 47 %). Thus, while a 90 % sensitivity may not be clinically sufficient even more for screening perspectives, the model still maintains a high specificity (70.8 %) in comparison with literature (Table 6), implying a possible reduction of unnecessary biopsies. Reducing false negatives is essential in clinical practice to improve diagnostic accuracy and patient outcomes. However, setting sensitivity thresholds at 100 % is not ideal, as it leads to a significant increase in the false positive rate, which can impact patients with unnecessary follow-up procedures and heightened anxiety. Addressing residual false negatives will require a more comprehensive dataset, especially as cases of DCIS, IDC, and M-LN are currently underrepresented despite our data sampling techniques. Expanding these cases would enhance the generalizability of our findings across diverse patient presentations. Further insights could be gained by differentiating between patient groups, such as those undergoing routine screening and those receiving staging evaluations. For M-LN, which is typically associated with a known cancer diagnosis, this may improve detection of suspicious lymphadenopathy. Additionally, it is well-documented that MRI is less sensitive to DCIS, especially low-grade lesions, than mammography [74–76]. This is partially due to the generally lower contrast enhancement in low-grade DCIS, making these lesions more challenging to detect on MRI alone. Given the current diagnostic pathway, where mammography is standard in initial screenings, integrating mammographic findings with MRI data could refine lesion characterization and improve detection accuracy.

In order to reduce false positive lesions, one solution could be the addition of T2-weighted images. They have higher spatial resolution than UF-DCE images, thus allowing better morphological analysis of the lesion, and a better distinction between lesions that are easier to detect such as lymph nodes and some benign lesions. The addition of all phases of the UF-DCE images, and not only the last phase, could also provide useful information such as kinetic early enhancement parameters. Indeed, this could potentially allow correct classification of lesions with high background parenchymal enhancement (BPE) in the breast. In fact, it is known that the cycle period of the patient can induce high BPE that makes lesion visualization and detection more difficult [77,78]. An other possibility could be to propose in some cases an additional ultrasound on a case-by-case basis, similar to the approach commonly used in clinical practice for evaluating suspicious lesions (second-look ultrasound), which is a very accessible technique [79].

### 5.2. Comparison with previous works

In the literature, there are very few authors using AI and UF-DCE MRI sequence [12,80]. With a total of 987 breast lesions and patient information collected from 1081 breast radiology reports, as compared to the studies listed in Table 6, our dataset represents the largest single center dataset including UF-DCE MRI and clinical patient data (age, menopausal status, contraception, family and patient history of breast cancer, BRCA mutation carrier, and chemotherapy treatment).

Table 6 also highlights the performance of some of them addressing lesion classification. Among these works, the best performance was achieved with the analysis of 210 features (textural and dynamic) with a random forest classifier (AUROC = 0.8997), but still lower than our approach. Moreover, other studies with AI included traditional DCE-MRI imaging and non-imaging data. Lo Gullo et al. [68] combined radiomics features of all dynamic phases and clinical factors (menopausal status, age, lesion location) – achieving a sensitivity of 63.2 % (46.0–78.2) and a specificity of 91.4 % (82.3–96.8) for lesion classification. Holster et al. [73], combined MIP single-breast images and 18 clinical non-imaging features for breast image classification (not at lesion level), including clinical indication, mammographic breast density, age, and BPE. The best model was achieved with the combination of image and non-imaging data (AUROC = 0.903). Also, for prediction of pathologic complete response to neoadjuvant chemotherapy treatment, studies showed that combining DCE-MRI data and clinical data information improved the prediction performances [29,69], demonstrating the importance of integrating relevant clinical information for breast lesion classification.

Dalmiş et al. [20] is the only study that combines UF-DCE MRI imaging and non-imaging data (BRCA and age). However, their approach used MIP images as input data. In contrast, our study used only the last phase of UF-DCE at approximately 53 s, resulting in both cases in the use of 3D image data and thus the fourth temporal dimension inherent in UF-DCE sequence was not fully exploited. Future research should integrate the complete UF-DCE MRI sequence information including temporal as well as spatial information to fully exploit the potential of the sequence. UF-DCE is now recognized as a stand-alone clinical technique based on a recent meta-analysis from Ref. [81].

### 5.3. The added value of MMST-V model

There is no consensus on the best fusion strategy [82], however intermediate fusion (or joint) fusion, as defined by Huang et al. [83], and as used in our MMST-V model, provided in our case a better approach for lesion classification with multimodal data through interaction between features of multimodal data. To verify this, we tested in additional experiments only for 3-class classification the late fusion of UDM and Random Forest (non-imaging) predictions. Results revealed a slightly lower performance (AUROC = 0.912) compared to MMST-V. This finding is consistent with the study of Holste et al. [73], who tested Probability Fusion (equivalent to late fusion) and two types of joint fusion (type I = Learned Feature Fusion and type II Feature Fusion) for classification of MIP single-breast images from DCE-MRI with non-imaging features (age, clinical indication, breast density, etc.). They also demonstrated that Learned Feature Fusion (equivalent to the joint fusion we used) was the best approach achieving an AUROC of 0.903.

A review of Cui et al. [82] highlighted that the addition of more modalities for fusion may perform worse that with fewer modalities. In fact, the addition of information can introduce redundant information or

**Table 6**
Performance comparison of studies using AI and UF-DCE MRI.

| | Data | Approach | AUROC | Se [%] | Sp [%] |
|---|---|---|---|---|---|
| Platel et al. (2014) [10] | Patients (n=137)<br>Benign (n=71)<br>Malignant (n=83) | Combination of lesion kinetic and morphological features (VIBE + TWIST)<br>TWIST MRI sequence<br>SVM classifier | 0.87 | 90 | 52* |
| | | | | 95 | 39* |
| Milenkovic et al. (2017) [18] | Patients (n=137)<br>Benign (n=71)<br>Malignant (n=83) | 210 features analysis (35 textural features for each 6 dynamic features)<br>TWIST MRI sequence<br>Random Forest classifier | 0.8997 | 90 | 68.48 |
| | | | | 95 | 40.43 |
| Dalmiş et al. (2019) [20] | Breast lesions (n=576)<br>Benign (n=208)<br>Malignant (n=368) | Combination of MIP images, T2w images, ADC values and patient information (age and BRCA)<br>TWIST MRI sequence<br>Random Forest final classifier (images were first fed into a 3D CNN, then combined with patient information in a final Random Forest classifier) | 0.811 | 90 | 60* |
| | | | | 95 | 41* |
| Jing et al. (2022) [19] | Patients (n=488)<br>Normal breast (n=1501)<br>Abnormal breast (n=173) | MIP images of left and right breast with 3D segmentation mask<br>TWIST MRI sequence<br>ResNet-34 | 0.81 | 91 | 52 |
| | | | | 95 | 35 |
| **Present study** MMST-V B_M | Patients (n=240)<br>Benign (n=280)<br>Malignant (n=121)<br>Lymph nodes (n=586) | Combination of imaging and non-imaging (patient information and lesion characteristics) data<br>4D-THRIVE MRI Sequence | 0.916 | 90 | 70.1 |
| | | | | 95 | 43.4 |
| MMST-V BL_M | | | 0.905 | 90 | 64.4 |
| | | | | 95 | 34.6 |
| MMST-V B_L_M | | | 0.928 | 90 | 70.8 |
| | | | | 95 | 47.0 |

AUROC = Area under the ROC curve, Se = sensitivity, Sp = specificity.
*Specificity extrapolated and estimated at 90% and 95% sensitivity values based on ROC curves.

noise, negatively impacting model performance [82]. Therefore, we also tested a Multimodal Transformer without Sieve encoder (MMT), that achieved an AUROC of 0.909, not surpassing MMST-V (Table 7). These results underscore the importance of the Sieve encoder, and particularity of our MMST-V that addresses redundant information between modalities. In our dataset, some scalar features were known to be redundant as showed in the correlation matrix in Supplementary material S8, such as bounding box and lesion gravity center positions. We therefore tested the MMST-V with a hand-picked subset of 7 scalar features (age, volume, bounding box center position x, y, z, elongation and flatness). The MMST-V with reduced scalar features did not improve performance, highlighting that MMST-V addresses better the redundancy than a manual features preselection (Table 7).

### 5.4. Explainability

Model transparency constitutes a major barrier in the implementation of AI systems in clinical practice [26]. Therefore many explainability approaches were investigated in medical imaging [84,85], such as saliency maps – commonly used for explainable AI (XAI) [86,87]. In this study, saliency maps were generated from probability class at image level and non-imaging data level for best performing MMST-V model. Analysis of correct and non-correct predictions, and also distribution of the center of mass along the x, y, z axes of each saliency map intensities, revealed that model mostly focused on the lesion (center of VOI (32,32,32)), and sometimes its focus was outside the lesion (Supplementary material S9). The focus could be on another enhanced lesion present within the bounding box, or only on non-enhanced tissue. In the latter case, one possible explanation may be that the model, due to the multimodal context, may prioritize non-imaging features over imaging features. These observations are illustrated with some examples of saliency maps for various image lesions and their corresponding scalar saliency maps, that are provided in Supplementary material S6. The generation of scalar saliency maps

**Table 7**
Comparison of additional experiments and initial MMST for B_L_M classification.

| Experiments | AUROC [%] | $ACC_{avg}$ [%]* | Se [%]* | Sp [%]* |
|---|---|---|---|---|
| DenseNet121 + Random Forest (Late fusion) | 88.5 (± 4.4) | 79.6 (± 0.9) | 78.4 (± 4.7) | 78.3 (± 4.8) |
| MMT-V | 90.9 (± 3.7) | 86.0 (± 3.4) | 82.9 (± 4.8) | 83.1 (± 4.7) |
| MMST-V_7scalar | 88.8 (± 4.2) | 86.3 (± 1.4) | 81.6 (± 4.7) | 81.8 (± 4.8) |
| **MMST-V** | **92.8 (± 2.7)** | **88.2 (± 0.9)** | **86.7 (± 2.8)** | **86.3 (± 3.0)** |

AUROC = Area under ROC curve, $ACC_{avg}$ = Average accuracy, Se = Sensitivity, Sp = Specificity, MM = MMST-V without Sieve, MMST-V = Multimodal Sieve Transformer with ViT encoder, MMST-V_7scalar = MMST-V with reduced scalar features (only age, volume, bounding box position x,y,z, elongation, flatness).
*ACC, sensitivity and specificity were calculated at Equal Error Rate.
Bold style = best performance.

enabled us to evaluate the model attention across scalar data for each lesion. Indeed, the scalar saliency maps generated for the main types of lesions (benign, malignant and lymph nodes) showed that attention for lymph node was essentially attributed to position in the breast (mainly bounding box center and lesion gravity center). For malignant lesion, the importance of the size was also predominant (mainly lesion ellipsoid diameter and volume) (Supplementary material S7). Thus, these observations are in line with the same discriminatory tendencies of descriptive statistics presented in section 4.1.

In general, when the model failed to focus on a lesion, the focus was on enhanced lesions or vessels that were close to the targeted lesion within the bounding box, indicating possible influence of lesion surrounding on classification. This suggests that bounding box sizes should be adjusted to tightly encapsulate the lesion or consider capturing only the lesion VOI with a minimal margin of surrounding tissue, as many lesions have not the same dimensions and can be very small compared to the 50 $mm^3$ dimension of the VOI.

Analysis of all predictions also revealed incorrect prediction for two DCIS lesions localized close to each other in a patient with a marked diffuse glandular enhancement, reducing capability of correct prediction. Indeed, these same lesions were correctly predicted as cancer by the model in the follow-up images at another period of the menstrual cycle of the patient. Moreover, as shown in the saliency maps in Fig. 6 the focus of the MMST-V model was localized outside the lesion for the incorrect prediction, and on the lesion in the follow-up that was correctly predicted. It is also worth mentioning that clinicians also failed to diagnose the lesions initially due to the wrong cycle period of the patient, inducing a higher BPE and making lesion visualization and detection more difficult.

Finally, there are numerous cases where the approach failed, indicating that pixel attribution of saliency maps is potentially not reliable. There remains uncertainty when the model focuses on specific modalities (image, scalar or categorical data). In the context of a clinical setting, it is crucial to have a robust and reliable method of interpretability. As highlighted in literature, three other authors mentioned that saliency maps often lacks repeatability and reproducibility, and potentially sensitive to other elements that do not contribute to prediction [86,88,89]. It is important therefore to further investigate and assess carefully the sensitivity and robustness of the used explainability methods [86]. The analysis of the center of mass of saliency maps, while providing a quick and useful overview of the results, may not be optimal as it does not account for the lesion volume within the VOI. For a more comprehensive evaluation, further analysis should be conducted, including comparisons with other XAI methods such as occlusion maps and attention maps.

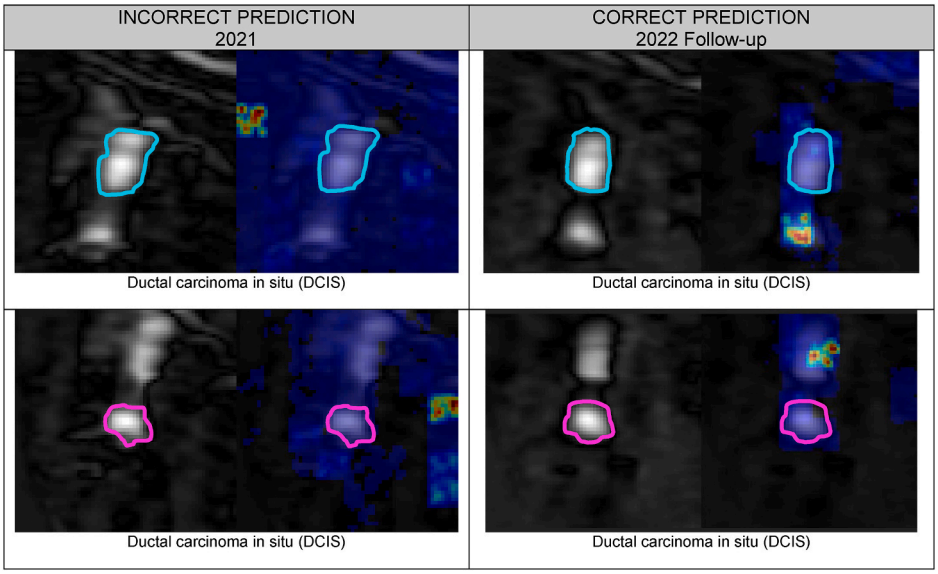


**Fig. 6.** Example of saliency maps for predictions of two DCIS lesions from the same patient but at different examination dates (MMST-V, B_L_M classification, the lesion is in the center of the image. The first DCIS outlined in light blue, and second DCIS outlined in light pink).

### 5.5. Limitations and perspectives

Our study has some limitations. First, the imaging data were imbalanced regarding the main lesion types, characterized by a higher proportion of lymph nodes compared to benign and malignant lesions. However, there are very few studies that included lymph nodes in their dataset, and we chose to keep them as their clinical prevalence is high and it may be difficult in some cases to differentiate them, particularly IMLNs [71,72]. Also in this study, it appeared that the differentiation between ALN and IMLN would have provided a more accurate analysis of IMLN diagnosis. Second, non-imaging data, more precisely the categorical data, were not always complete among patients and presented imbalance. There was a lot of missing data because not all clinical information were reported in the unstructured breast imaging reports. However, missing information were attributed to the category of "no information" to reflect clinical conditions. Merging some categories for each class in the categorical data could be a solution that can be explored to deal with the imbalance and missing data. Also, this limitation highlights the importance and effort required to collect and annotate multimodal data. But unfortunately, incomplete multimodal data are often found in real-world clinical practice, as such data are more difficult to obtain in general, and constitute an important limitation for multimodal frameworks [82]. Further investigations are needed to address the problem of incomplete data and their influence on MMST-V performances. Furthermore, no external validation was made and thus the generalizability of our model could not be tested, due to the lack of similar multimodal data not collected yet. External validation may help the investigation on data variability and incomplete data. Our models were not pretrained due to the unavailability of large multimodal data set and particularly for 3D imaging data [90]. Multimodal models inherently contribute to complex networks with an increased number of trainable parameters, leading to potential risk of overfitting, especially when dataset size is limited [91]. Thus, even if our UF-image dataset size is comparable to those literature (Table 7), a larger dataset would likely be more beneficial, enhancing the confidence in our results. This is particularly important for determining whether changes in number of features significantly impact the model performance. The MMST-V model had the highest number of trainable parameters (26.5M) and required the most training time (Supplementary material S10), reflecting its complexity and computational demands, due to Transformer-based approach. However, there is increasing interest in enhancing computational and memory efficiency of Transformers. Numerous techniques have been developed to make Transformers faster and more lightweight, as reviewed comprehensively by Fournier et al. [92]. Exploration and application of those methods may help improve the efficiency of multimodal framework approaches. Also, effective and efficient unimodal feature extraction is a crucial step prior to the fusion process, particularly when dealing with heterogeneous multimodal data (e.g., textual, imaging data) as noted by Cui et al. [82]. Therefore, the choice of encoder type (e.g., ViT, CNN) for feature extraction is important and must be tested, as it can significantly impact fusion quality and overall model performance.

With multimodal architectures, explainability efforts aim not only to clarify each modality's individual contribution but also to illustrate the interactions between different data types. Achieving explainability is complex, as it requires assessing the interpretability of each modality as well as their combined impact. In our study, we attempted to separate and analyze each modality's explainability independently; however, the explainability of the combined multimodal interactions remains an area for future investigation.

This study highlights the potential clinical applications of UF-DCE aided by AI, where we purposely reduced the number of acquisition sequences required for breast MRI examination, or optimized the sequences used for breast cancer diagnosis. By incorporating a T2-weighted sequences and the temporal dimension of the UF-DCE, we can potentially increase specificity, while keeping short examination time and patient comfort. Additionally, we note that the segmentation step is crucial as it provides valuable information about shape and position descriptors of lesions, which was useful and essential for lesion classification and is already reported in the BI-RADS system [93]. Thus, it is important to develop in future work an automatic segmentation method to avoid laborious manual processing, such as proposed in the literature with UF-DCE [94].

## 6. Conclusion

Our Transformer-based model trained on combined clinical imaging and non-imaging data showed superior classification performances compared to previous works and to models solely trained on unimodal data for classification of breast lesion with UF-DCE MRI. Despite the persistence of false negatives and false positives, our approach leverages diverse data sources and automatically addresses the issue of redundancy inherent of using information from multiple sources, thus fostering future multimodal studies that are needed to enforce our findings and ensure the robustness of our approach.

## CRediT authorship contribution statement

**Belinda Lokaj:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Valentin Durand de Gevigney:** Writing – review & editing, Validation, Supervision, Software, Resources, Methodology. **Dahila-Amal Djema:** Resources. **Jamil Zaghir:** Writing – review & editing. **Jean-Philippe Goldman:** Writing – review & editing. **Mina Bjelogrlic:** Writing – review & editing. **Hugues Turbé:** Writing – review & editing. **Karen Kinkel:** Writing – review & editing, Resources. **Christian Lovis:** Writing – review & editing, Supervision. **Jérôme Schmid:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis.

## Ethics statement

This study is part of the Smart and Ultrafast Breast MRI (SUBREAM) project funded by the Swiss Cancer Research (KFS-5460–08-2021-R) and approved by the Geneva Cantonal Ethics Committee (CCER) (Project-ID: 2019-00716). Informed consent was obtained from each patient for the re-use of anonymized breast imaging reports and MRI examinations.

## 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.

## Acknowledgment

This study is part of the Smart and Ultrafast Breast MRI (SUBREAM) project funded by the Swiss Cancer Research (KFS-5460–08-2021-R) and approved by the Geneva Cantonal Ethics Committee (CCER) (Project-ID: 2019-00716).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109721.

## References

[1] C.K. Kuhl, S. Schrading, K. Strobel, et al., Abbreviated breast magnetic resonance imaging (MRI): first postcontrast subtracted images and maximum-intensity

projection—a novel approach to breast cancer screening with MRI, J. Clin. Oncol. 32 (2014) 2304–2310, https://doi.org/10.1200/JCO.2013.52.5386.
[2] Y. Gao, S.L. Heller, Abbreviated and ultrafast breast MRI in clinical practice, Radiographics 40 (2020) 1507–1527, https://doi.org/10.1148/rg.2020200006.
[3] C.K. Kuhl, Abbreviated magnetic resonance imaging (MRI) for breast cancer screening: rationale, concept, and transfer to clinical practice, Annu. Rev. Med. 70 (2019) 501–519, https://doi.org/10.1146/annurev-med-121417-100403.
[4] C.K. Kuhl, H.H. Schild, Dynamic image interpretation of MRI of the breast, J. Magn. Reson. Imag. 12 (2000) 965–974, https://doi.org/10.1002/1522-2586(200012)12:6<965::AID-JMRI23>3.0.CO;2-1.
[5] D. Sheth, H. Abe, Abbreviated MRI and accelerated MRI for screening and diagnosis of breast cancer, Top. Magn. Reson. Imag. 26 (2017) 183–189, https://doi.org/10.1097/RMR.0000000000000140.
[6] J. Gillman, H.K. Toth, L. Moy, The role of dynamic contrast-enhanced screening breast MRI in populations at increased risk for breast cancer, Womens Health 10 (2014) 609–622, https://doi.org/10.2217/WHE.14.61.
[7] K. Kinkel, Protocole abrégé en IRM mammaire : erreur ou certitude, Imag. Femme 27 (2017) 149–151, https://doi.org/10.1016/j.femme.2017.03.004.
[8] F. Tollens, P.A.T. Baltzer, M. Dietzel, et al., Cost-effectiveness of digital breast tomosynthesis vs. Abbreviated breast MRI for screening women with intermediate risk of breast cancer—how low-cost must MRI Be? Cancers 13 (2021) 1241, https://doi.org/10.3390/cancers13061241.
[9] C.K. Kuhl, The changing world of breast cancer: a radiologist's perspective, Invest. Radiol. 50 (2015) 615–628, https://doi.org/10.1097/RLI.0000000000000166.
[10] B. Platel, R. Mus, T. Welte, et al., Automated characterization of breast lesions imaged with an ultrafast DCE-MR protocol, IEEE Trans. Med. Imag. 33 (2014) 225–232, https://doi.org/10.1109/TMI.2013.2281984.
[11] R.M. Mann, R.D. Mus, J. van Zelst, et al., A novel approach to contrast-enhanced breast magnetic resonance imaging for screening: high-resolution ultrafast dynamic imaging, Invest. Radiol. 49 (2014) 579–585, https://doi.org/10.1097/RLI.0000000000000057.
[12] M. Kataoka, M. Honda, A. Ohashi, et al., Ultrafast dynamic contrast-enhanced MRI of the breast: how is it used? Magn. Reson. Med. Sci. 21 (2022) 83–94, https://doi.org/10.2463/mrms.rev.2021-0157.
[13] J.C.M. van Zelst, S. Vreemann, H.-J. Witt, et al., Multireader study on the diagnostic accuracy of ultrafast breast magnetic resonance imaging for breast cancer screening, Invest. Radiol. 53 (2018) 579, https://doi.org/10.1097/RLI.0000000000000494.
[14] M. Codari, S. Schiaffino, F. Sardanelli, R.M. Trimboli, Artificial intelligence for breast MRI in 2008–2018: a systematic mapping review, Am. J. Roentgenol. 212 (2019) 280–292, https://doi.org/10.2214/AJR.18.20389.
[15] G. Chartrand, P.M. Cheng, E. Vorontsov, et al., Deep learning: a primer for radiologists, Radiographics 37 (2017) 2113–2131, https://doi.org/10.1148/rg.2017170077.
[16] T. Pang, J. Wong, W. Ng, C. Chan, Deep learning radiomics in breast cancer with different modalities: overview and future, Expert Syst. Appl. 158 (2020), https://doi.org/10.1016/j.eswa.2020.113501.
[17] R. Adam, K. Dell'Aquila, L. Hodges, et al., Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review, Breast Cancer Res. 25 (2023) 87, https://doi.org/10.1186/s13058-023-01687-4.
[18] J. Milenković, M.U. Dalmış, J. Žgajnar, B. Platel, Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast DCE-MRI protocol, Med. Phys. 44 (2017) 4652–4664, https://doi.org/10.1002/mp.12408.
[19] X. Jing, M. Wielema, L.J. Cornelissen, et al., Using deep learning to safely exclude lesions with only ultrafast breast MRI to shorten acquisition and reading time, Eur. Radiol. 32 (2022) 8706–8715, https://doi.org/10.1007/s00330-022-08863-8.
[20] M.U. Dalmiş, A. Gubern-Mérida, S. Vreemann, et al., Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI, Invest. Radiol. 54 (2019) 325–332, https://doi.org/10.1097/RLI.0000000000000544.
[21] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008, https://doi.org/10.48550/arXiv.1706.03762.
[22] T. Tong, D. Li, J. Gu, et al., Dual-input transformer: an end-to-end model for preoperative assessment of pathological complete response to neoadjuvant chemotherapy in breast cancer ultrasonography, IEEE J Biomed Health Inform 27 (2023) 251–262, https://doi.org/10.1109/JBHI.2022.3216031.
[23] F. Shamshad, S. Khan, S.W. Zamir, et al., Transformers in medical imaging: a survey, Med. Image Anal. 88 (2023) 102802, https://doi.org/10.1016/j.media.2023.102802.
[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: transformers for image recognition at scale. Proc IEEECVF Conf Comput Vis Pattern Recognit, 2021, https://doi.org/10.48550/arXiv.2010.11929.
[25] J. Li, J. Chen, Y. Tang, et al., Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives, Med. Image Anal. 85 (2023) 102762, https://doi.org/10.1016/j.media.2023.102762.
[26] B. Lokaj, M.-T. Pugliese, K. Kinkel, et al., Barriers and facilitators of artificial intelligence conception and implementation for breast imaging diagnosis in clinical practice: a scoping review, Eur. Radiol. (2023), https://doi.org/10.1007/s00330-023-10181-6.
[27] A. Meyer-Bäse, L. Morra, U. Meyer-Bäse, K. Pinker, Current status and future perspectives of artificial intelligence in magnetic resonance breast imaging, Contrast Media Mol. Imaging 2020 (2020) 1–18, https://doi.org/10.1155/2020/6805710.
[28] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nat. Med. 28 (2022) 1773–1784, https://doi.org/10.1038/s41591-022-01981-2.
[29] N. Khan, R. Adam, P. Huang, et al., Deep learning prediction of pathologic complete response in breast cancer using MRI and other clinical data: a systematic review, Tomography 8 (2022) 2784–2795, https://doi.org/10.3390/tomography8060232.
[30] S.-C. Huang, A. Pareek, S. Seyyedi, et al., Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, Npj Digit Med 3 (2020) 1–9, https://doi.org/10.1038/s41746-020-00341-z.
[31] F. Khader, G. Müller-Franzes, T. Wang, et al., Multimodal deep learning for integrating chest radiographs and clinical parameters: a case for transformers, Radiology 309 (2023) e230806, https://doi.org/10.1148/radiol.230806.
[32] P. Xu, X. Zhu, D.A. Clifton, Multimodal learning with transformers: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 12113–12132, https://doi.org/10.1109/TPAMI.2023.3275156.
[33] X. Wang, G. Chen, G. Qian, et al., Large-scale multi-modal pre-trained models: a comprehensive survey, Mach Intell Res 20 (2023) 447–482, https://doi.org/10.1007/s11633-022-1410-8.
[34] D. Saslow, C. Boetes, W. Burke, et al., American cancer society guidelines for breast screening with MRI as an adjunct to mammography, CA A Cancer J. Clin. 57 (2007) 75–89, https://doi.org/10.3322/canjclin.57.2.75.
[35] R.M. Mann, C.K. Kuhl, L. Moy, Contrast-enhanced MRI for breast cancer screening, J. Magn. Reson. Imag. 50 (2019) 377–390, https://doi.org/10.1002/jmri.26654.
[36] V. Molleran, M.C. Mahoney, The BI-rads breast magnetic resonance imaging lexicon, Magn. Reson. Imag. Clin. N. Am. 18 (2010) 171–185, https://doi.org/10.1016/j.mric.2010.02.001.
[37] O. Díaz, A. Rodríguez-Ruíz, I. Sechopoulos, Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects, Eur. J. Radiol. 175 (2024) 111457, https://doi.org/10.1016/j.ejrad.2024.111457.
[38] K. Han, Y. Wang, H. Chen, et al., A survey on visual transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 87–110, https://doi.org/10.1109/TPAMI.2022.3152247.
[39] C. Matsoukas, J.F. Haslum, M. Söderberg, K. Smith, Is it time to replace CNNs with transformers for medical images?. https://doi.org/10.48550/arXiv.2108.09038, 2021.
[40] Z. Liu, Q. Lv, Z. Yang, et al., Recent progress in transformer-based medical image analysis, Comput. Biol. Med. 164 (2023) 107268, https://doi.org/10.1016/j.compbiomed.2023.107268.
[41] A. Khan, Z. Rauf, A. Sohail, et al., A survey of the Vision Transformers and their CNN-Transformer based Variants, Artif. Intell. Rev. 56 (2023) 2917–2970, https://doi.org/10.1007/s10462-023-10595-0.
[42] L. Cao, Q. Wang, J. Hong, et al., MVI-TR: a transformer-based deep learning model with contrast-enhanced CT for preoperative prediction of microvascular invasion in hepatocellular carcinoma, Cancers 15 (2023) 1538, https://doi.org/10.3390/cancers15051538.
[43] X. Fan, X. Feng, Y. Dong, H. Hou, COVID-19 CT image recognition algorithm based on transformer and CNN, Displays 72 (2022) 102150, https://doi.org/10.1016/j.displa.2022.102150.
[44] B. Gheflati, H. Rivaz, Vision transformers for classification of breast ultrasound images, 44th Annu Int Conf IEEE Eng Med Biol Soc EMBC (2022) 480–483, https://doi.org/10.1109/EMBC48229.2022.9871809.
[45] G. Zhou, B. Mosadegh, Distilling knowledge from an ensemble of vision transformers for improved classification of breast ultrasound, Acad. Radiol. 31 (2024) 104–120, https://doi.org/10.1016/j.acra.2023.08.006.
[46] W. Lee, H. Lee, H. Lee, et al., Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images, Radiol Artif Intell 5 (2023) e220159, https://doi.org/10.1148/ryai.220159.
[47] X. Chen, K. Zhang, N. Abdoli, et al., Transformers improve breast cancer diagnosis from unregistered multi-view mammograms, Diagnostics 12 (2022) 1549, https://doi.org/10.3390/diagnostics12071549.
[48] S. Sarker, P. Sarker, G. Bebis, A. Tavakkoli, MV-Swin-T: mammogram classification with multi-view SWIN transformer, Proc IEEE Int Symp Biomed Imaging (2024) 2024, https://doi.org/10.1109/isbi56570.2024.10635578.
[49] S. Hussain, M. Ali, U. Naseem, et al., Performance evaluation of deep learning and transformer models using multimodal data for breast cancer classification, in: Cancer Prevention, Detection, and Intervention, 2024, pp. 59–69.
[50] G. Cai, Y. Zhu, Y. Wu, et al., A multimodal transformer to fuse images and metadata for skin disease classification, Vis. Comput. 39 (2023) 2781–2793, https://doi.org/10.1007/s00371-022-02492-4.
[51] K.A. Mende, J.M. Froehlich, C. von Weymarn, et al., Time-resolved, high-resolution contrast-enhanced MR angiography of dialysis shunts using the CENTRA keyhole technique with parallel imaging, J. Magn. Reson. Imag. 25 (2007) 832–840, https://doi.org/10.1002/jmri.20879.
[52] J. Zhou, Y. Zhang, K.-T. Chang, et al., Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue, J Magn Reson Imaging JMRI 51 (2020) 798–809, https://doi.org/10.1002/jmri.26981.
[53] Özdemir H, Azamat S, Sam Özdemir M Can Only the Shape Feature in Radiomics Help Machine Learning Show That Bladder Cancer Has Invaded Muscles? Cureus 15:e45488. https://doi.org/10.7759/cureus.45488.
[54] C. Militello, L. Rundo, M. Dimarco, et al., 3D DCE-MRI radiomic analysis for malignant lesion prediction in breast cancer patients, Acad. Radiol. 29 (2022) 830–840, https://doi.org/10.1016/j.acra.2021.08.024.
[55] S. Łukasiewicz, M. Czeczelewski, A. Forma, et al., Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment

strategies—an updated review, Cancers 13 (2021) 4287, https://doi.org/10.3390/cancers13174287.
[56] P. Stenetorp, S. Pyysalo, G. Topić, et al., Brat: a web-based tool for NLP-assisted text annotation, in: F. Segond (Ed.), Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.
[57] J. Zaghir, B. Lokaj, K. Kinkel, et al., Efficient clinical information extraction from breast radiology reports in French, Stud. Health Technol. Inf. 316 (2024) 1780–1784, https://doi.org/10.3233/SHTI240776.
[58] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 2261–2269.
[59] M.J. Cardoso, W. Li, R. Brown, et al., MONAI: an open-source framework for deep learning in healthcare, arXiv (2022), https://doi.org/10.48550/arXiv.2211.02701.
[60] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[61] B. Lokaj, A.-D. Djema, K. Kinkel, et al., 1402-3 Combining ultrafast MRI sequence with artificial intelligence (AI) for breast cancer detection, ECR 2023 Book Abstr Insight into Imaging 14 (Suppl 4) (2023) 217, https://doi.org/10.1186/s13244-023-01522-6.
[62] T. Zhou, X. Ye, H. Lu, et al., Dense convolutional network and its application in medical image analysis, BioMed Res. Int. 2022 (2022) 1–22, https://doi.org/10.1155/2022/2384830.
[63] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv (2017), https://doi.org/10.48550/arXiv.1412.6980.
[64] J. Zbontar, L. Jing, I. Misra, et al., Barlow twins: self-supervised learning via redundancy reduction, Proc 38th Int Conf Mach Learn. (2021), https://doi.org/10.48550/ARXIV.2103.03230.
[65] A. Bardes, J. Ponce, Y. LeCun, VICReg: variance-invariance-covariance regularization for self-supervised learning, ICLR 2022 - Int Conf Learn Represent (2021). https://arxiv.org/abs/2105.04906.
[66] N. Detlefsen, J. Borovec, J. Schock, et al., TorchMetrics - measuring reproducibility in PyTorch, J. Open Source Softw. 7 (2022) 4101, https://doi.org/10.21105/joss.04101.
[67] F. Bianconi, I. Palumbo, M.L. Fravolini, et al., Form factors as potential imaging biomarkers to differentiate benign vs. Malignant lung lesions on CT scans, Sensors 22 (2022) 5044, https://doi.org/10.3390/s22135044.
[68] R. Lo Gullo, I. Daimiel, Saccarelli C. Rossi, et al., Improved characterization of sub-centimeter enhancing breast masses on MRI with radiomics and machine learning in BRCA mutation carriers, Eur. Radiol. 30 (2020) 6721–6731, https://doi.org/10.1007/s00330-020-06991-7.
[69] S. Joo, E.S. Ko, S. Kwon, et al., Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer, Sci. Rep. 11 (2021) 18800, https://doi.org/10.1038/s41598-021-98408-8.
[70] H. Abe, N. Mori, K. Tsuchiya, et al., Kinetic analysis of benign and malignant breast lesions with ultrafast dynamic contrast-enhanced MRI: comparison with standard kinetic assessment, AJR Am. J. Roentgenol. 207 (2016) 1159–1166, https://doi.org/10.2214/AJR.15.15957.
[71] G. Gunes, P. Crivellaro, D. Muradali, Management of MRI-detected benign internal mammary lymph nodes, Indian J. Radiol. Imag. 32 (2022) 197–204, https://doi.org/10.1055/s-0042-1750180.
[72] A.G. Bitencourt, E.V. Ferreira, D.C. Bastos, et al., Intramammary lymph nodes: normal and abnormal multimodality imaging features, Br. J. Radiol. 92 (2019) 20190517, https://doi.org/10.1259/bjr.20190517.
[73] G. Holste, S.C. Partridge, H. Rahbar, et al., End-to-End learning of fused image and non-image features for improved breast cancer classification from MRI, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3287–3296.
[74] I.-M.A. Obdeijn, C.E. Loo, A.J. Rijnsburger, et al., Assessment of false-negative cases of breast MR imaging in women with a familial or genetic predisposition, Breast Cancer Res. Treat. 119 (2010) 399–407, https://doi.org/10.1007/s10549-009-0607-7.
[75] H.I. Greenwood, C.K. Maldonado Rodas, R.I. Freimanis, et al., Magnetic resonance imaging insights from active surveillance of women with ductal carcinoma in situ, Npj Breast Cancer 10 (2024) 1–9, https://doi.org/10.1038/s41523-024-00677-9.
[76] C.D. Lehman, Magnetic resonance imaging in the evaluation of ductal carcinoma in situ, JNCI Monogr 2010 (2010) 150–151, https://doi.org/10.1093/jncimonographs/lgq030.
[77] S. Vreemann, M.U. Dalmis, P. Bult, et al., Amount of fibroglandular tissue FGT and background parenchymal enhancement BPE in relation to breast cancer risk and false positives in a breast MRI screening program: a retrospective cohort study, Eur. Radiol. 29 (2019) 4678–4690, https://doi.org/10.1007/s00330-019-06020-2.
[78] M. Honda, M. Kataoka, M. Iima, et al., Background parenchymal enhancement and its effect on lesion detectability in ultrafast dynamic contrast-enhanced MRI, Eur. J. Radiol. 129 (2020) 108984, https://doi.org/10.1016/j.ejrad.2020.108984.
[79] A. Bumberger, P. Clauser, M. Kolta, et al., Can we predict lesion detection rates in second-look ultrasound of MRI-detected breast lesions? A systematic analysis, Eur. J. Radiol. 113 (2019) 96–100, https://doi.org/10.1016/j.ejrad.2019.02.008.
[80] Kataoka M, Honda M, Sagawa H, et al Ultrafast Dynamic Contrast-Enhanced MRI of the Breast: From Theory to Practice. J. Magn. Reson. Imag. n/a: https://doi.org/10.1002/jmri.29082.
[81] Y. Amitai, V.A.R. Freitas, O. Golan, et al., The diagnostic performance of ultrafast MRI to differentiate benign from malignant breast lesions: a systematic review and meta-analysis, Eur. Radiol. (2024), https://doi.org/10.1007/s00330-024-10690-y.
[82] C. Cui, H. Yang, Y. Wang, et al., Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review, Prog. Biomed. Eng. 5 (2023) 022001, https://doi.org/10.1088/2516-1091/acc2fe.
[83] S.-C. Huang, A. Pareek, S. Seyyedi, et al., Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, Npj Digit Med 3 (2020) 136, https://doi.org/10.1038/s41746-020-00341-z.
[84] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, J Imaging 6 (2020) 52, https://doi.org/10.3390/jimaging6060052.
[85] M.A. Gulum, C.M. Trombley, M. Kantardzic, A review of explainable deep learning cancer detection models in medical imaging, Appl. Sci. 11 (2021) 4573, https://doi.org/10.3390/app11104573.
[86] J. Zhang, H. Chao, G. Dasegowda, et al., Revisiting the trustworthiness of saliency methods in radiology AI, Radiol Artif Intell 6 (2024) e220221, https://doi.org/10.1148/ryai.220221.
[87] M. Champendal, H. Müller, J.O. Prior, CS dos Reis, A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging, Eur. J. Radiol. 169 (2023), https://doi.org/10.1016/j.ejrad.2023.111159.
[88] A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, Proc. AAAI Conf. Artif. Intell. 33 (2019) 3681–3688, https://doi.org/10.1609/aaai.v33i01.33013681.
[89] P.-J. Kindermans, S. Hooker, J. Adebayo, et al., The (Un)reliability of saliency methods, in: W. Samek, G. Montavon, A. Vedaldi, et al. (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 267–280.
[90] X. Pei, K. Zuo, Y. Li, Z. Pang, A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions, Int. J. Comput. Intell. Syst. 16 (2023) 44, https://doi.org/10.1007/s44196-023-00225-6.
[91] C. Cui, H. Yang, Y. Wang, et al., Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review, Prog. Biomed. Eng. 5 (2023) 022001, https://doi.org/10.1088/2516-1091/acc2fe.
[92] Q. Fournier, G.M. Caron, D. Aloise, A practical survey on faster and lighter transformers, ACM Comput. Surv. 55 (2023) 1–40, https://doi.org/10.1145/3586074.
[93] E. Morris, C. Comstock, C. Lee, ACR BI-RADS® magnetic resonance imaging, in: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, American College of Radiology, Reston, VA, 2013.
[94] F. Ayatollahi, S.B. Shokouhi, R.M. Mann, J. Teuwen, Automatic breast lesion detection in ultrafast DCE-MRI using deep learning, Med. Phys. 48 (2021) 5897–5907, https://doi.org/10.1002/mp.15156.