Applied Intelligence (2025) 55:901
https://doi.org/10.1007/s10489-025-06803-9

Explainable breast cancer prediction from 3-dimensional dynamic contrast-enhanced magnetic resonance imaging

Arslan Akbar[1] · Suya Han[1] · Naveed Urr Rehman[2] · Kanwal Ahmed[3] · Hassan Eshkiki[4] · Fabio Caraffini[4]

Accepted: 20 July 2025

Abstract

Deep learning models have been instrumental in extracting critical indicators for breast cancer diagnosis - the prevalent malignancy among women worldwide - from baseline magnetic resonance imaging. However, many existing models do not fully leverage the rich spatial information available in the 3D structure of medical imaging data, potentially overlooking important contextual details. This develops an explainable deep learning framework for classifying breast cancer that leverages the complete 3D and provides classification results alongside visual explanations of the decision-making process. The preprocessing pipeline is fed with 3D sequences containing 'tumour' and 'non-tumour' regions. It includes a 3D Adaptive Unsharp Mask (AUM) filter to reduce noise and augment image class, followed by normalisation and data augmentation. Classification is then achieved by training an augmented ResNet150 model. Three explainable artificial intelligence (XAI) techniques, including Shapley Additive Explanations, 3D Gradient-Weighted Class Activation Mapping, and Contextual Importance and Utility, are employed to provide improved interpretability. The model demonstrates state-of-the-art performance over the QIN-BREAST dataset, achieving testing accuracies of 98.861% for 'tumours' and 99.447% for 'non-tumours', as well as over the Duke Breast Cancer Dataset, where it achieves 99.104% for 'tumours' and 99.753% for 'non-tumours', while offering enhanced interpretability through XAI methods.

Keywords Breast cancer · Deep learning · DCE-MRI · Explainable AI · RESNET150

1 Introduction

Explainable Artificial Intelligence (XAI) significantly advances artificial intelligence (AI) engineering, enhancing process interpretability and adaptability for more reliable results in scientific and technological fields [1]. This is particularly important in the healthcare sector, where AI plays a major role in diagnosing various medical conditions [2], even in very early stages, which is the key for people diagnosed with life-threatening diseases [3].

Breast cancer represents the most commonly diagnosed malignancy in women, accounting for one-quarter of all cancer diagnoses and one-sixth of all cancer-related deaths. In this light, early diagnosis and intervention for breast cancer are essential [13], but challenging due to its subtle, hard-to-detect symptoms [14]. The literature highlights the complexity of the cellular structures of breast tumours, making it challenging to determine their origin or growth rate. In addition, the limitations of current diagnostic tools can lead to misdiagnoses or unclear results, potentially postponing the start of treatment. This is harmful to these diseases,

✉ Suya Han
iehansy@zzu.edu.cn

✉ Fabio Caraffini
fabio.caraffini@swansea.ac.uk

Arslan Akbar
engr.arslan@gs.zzu.edu.cn

Naveed Urr Rehman
naveed@gs.zzu.edu.cn

Kanwal Ahmed
Kanwal@henu.edu.cn

Hassan Eshkiki
h.g.eshkiki@swansea.ac.uk

1 School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China

2 School of Computing and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, Henan, China

3 School of Software, Henan University, Kaifeng 470001, Henan, China

4 Department of Computer Science, Swansea University, Swansea SA1 8EN, Wales, UK

Published online: 01 August 2025

where it has become evident that timely identification is one of the key factors in reducing mortality rates, lowering treatment costs, and improving prognosis and therapeutic outcomes [15].

In recent years, computer-aided detection (CAD) systems have gained traction in identifying cancerous cells. However, the effectiveness of CAD is highly dependent on the quality of soft tissue imaging. The introduction of baseline Dynamic Contrast-Enhanced (DCE) Magnetic Resonance Imaging (MRI) has improved the precision of breast tumour diagnosis due to its visual quality. However, image noise significantly limits prediction accuracy, requiring pre-processing techniques that remove noise while preserving image information. Furthermore, another barrier to early diagnosis and prognosis of breast cancer is related to costs, DCE-MRI being significantly more expensive than other screening methods and, therefore, being used primarily for high-risk patients. The risks of contrast allergy and other physical limitations further limit the use of magnetic resonance imaging [59].

Hence, maximising the extraction of information from the available MRI scan is essential for multiple reasons. The deep learning (DL) community has contributed to this goal by urging the development of automated feature extraction techniques [16] to increase the limited, manually crafted radiomic features currently available. It offers clear advantages for extracting features as it can detect 'hidden-to-human abnormal structures and tumour patterns at different extraction levels [17]. While 2D CNNs have been commonly used due to their efficiency and simpler data requirements [18], recent advances have demonstrated the strong performance of 3D deep learning models in Tumour analysis, particularly for brain and breast MRI [19]. This work builds on that foundation by introducing a customised 3D-aware processing pipeline, which captures critical spatial information across all three dimensions for improved analysis and classification [17].

This study presents an automated decision support system for healthcare professionals to make diagnoses from 3D DCE-MRI scans. Using DL models and three XAI methods, the resulting system not only predicts breast cancer but also provides explanations for such classification outcomes and additional graphical support to improve the interoperability of the DL model. Its contributions can be summarised as below:

- Development of a new customised 3D version of ResNet150 tailored for volumetric breast MRI data, incorporating depth-wise separable convolutions to reduce computational cost.
- Introduction of a new optimised preprocessing pipeline to enhance image quality while reducing noise.
- Provides interpretable overlays to highlight tumour and periTumoural regions, enabling clinicians to understand and validate AI decisions.

This article is structured as follows: Section 2 provides an overview of the latest research on breast cancer classification. Section 3 presents a detailed explanation of the proposed methodology. Section 4 describes the experimental setup and evaluation metrics used in the study. Section 5 discusses the results and provides analysis. Finally, Section 6 concludes the paper and highlights areas for future research.

2 Related works

This section reviews advancements in deep learning (DL) for medical imaging, with a focus on breast cancer segmentation and the integration of explainable artificial intelligence (XAI) in healthcare. The review is structured into three parts: general DL advancements in medical imaging, specific techniques for breast cancer segmentation, and XAI methods to enhance interpretability. A comparative analysis positions our proposed ResNet150X framework, which integrates an optimised 3D ResNet150 architecture with advanced XAI techniques for precise and interpretable breast cancer classification in 3D DCE-MRI data.

2.1 Deep learning in medical imaging

Convolutional Neural Networks (CNNs) have transformed digital image processing for healthcare [3]. CNNs excel in segmentation tasks, including tumour segmentation [20, 70], by extracting complex patterns from large datasets. CNN architectures such as U-Net have become popular in medical and biological fields for segmentation tasks [25, 70, 71]. Using these techniques, DL researchers are currently trying to automate breast cancer detection, segmentation, and classification [21, 28].

Most recent studies further extend DL capabilities by integrating blockchain technology and the Internet of Medical Things (IoMT) to models for breast cancer diagnostics [65], or sophisticated cascaded deep learning networks that leverage meaningful patterns in medical images [66], or hybrid approaches that exploit the joint use of artificial neural networks and multiple support vector machines [69]. For instance, a federated learning framework was introduced for breast tumor classification using magnified histopathological images, achieving 92.15% accuracy via IoMT [9]. However, its 2D focus and lack of explainability limit its applicability to 3D DCE-MRI.

However, DL models do not offer enough interpretability and transparent reasoning to be fully accepted within the healthcare domain. Efficient attempts using shallow models (see [68]) are plagued by similar considerations. In this context, XAI frameworks have recently gained recognition as a viable solution [22] seeking justification, transparency, information, and quantification of uncertainties [24] in healthcare DL models. Although not fully trusted [23], they are promising research directions for diagnostics.

2.2 Advances on automatic breast cancer segmentation

ResNet-based architectures and U-Net variants outperform conventional machine learning methods in understanding complex breast tumour structures [26]. The ResNet architecture [72] revolutionised DL by solving the vanishing gradient problem with residual learning, simplifying deep network training through residual optimisation. Its application in medical imaging, particularly breast tumour segmentation, has shown promising results [36].

Building on this foundation, ResNet-150 [73] enhances the performance and feature extraction capabilities of its predecessor. Its deeper architecture effectively captures intricate features that can make a difference in breast cancer segmentation due to improved boundary delineation in MRI images. Its residual connections reduce over-fitting and enhance generalisation on limited datasets. Compared to earlier models, ResNet-150 significantly improves segmentation accuracy, achieving higher Dice scores and IoU metrics over many medical segmentation tasks [35, 37, 38].

A 3D ResNet150 is available for improved spatio-temporal analysis in 3D data such as medical imaging [61]. This is the case with DCE-MRI. This variant offers robust feature extraction, versatility across tasks, and balanced complexity. Compared to Swin Transformers [74] and EfficientNet [75], it handles 3D data more efficiently, and it is more stable than Generative Adversarial Neural Networks and more generalisable than task-specific systems such as DeepLabV3+ [62]. A review of ResNet50 and Transformer-based models for breast cancer segmentation and diagnosis highlights advancements in image augmentation and multimodal analysis across histopathological and MRI datasets [8]. However, these models face interpretability challenges due to opaque decision-making processes and limited 3D generalisability, as they are optimized primarily for 2D imaging, reducing their effectiveness for volumetric data like DCE-MRI.

A framework, DeepMiCa, refines the detection of microcalcifications in mammograms through specialized convolutional architectures, delivering robust performance for early breast cancer identification [6]. However, its dependence on 2D mammography datasets and absence of explainable AI mechanisms constrain its generalisability to volumetric 3D DCE-MRI dataset. Similarly, Vision Transformer (ViT)-based approaches for breast cancer detection uses self-attention mechanisms to achieve superior feature extraction and diagnostic performance [7, 11]. Similarly, ViT models for mammographic breast cancer detection excel in pattern recognition through advanced self-attention mechanisms [7]. Nevertheless, their reliance on high-resolution images escalates computational requirements, and the absence of explainability tools limits practical deployment. In this light, the 3D ResNet150 is a state-of-the-art model for three-dimensional imaging.

2.3 Explainability methods in healthcare

The model in [29] enriches the prediction outcomes with a genetic, morphological, and clinical characteristics profile. First, it generates a heat map to visualise cancer cells and tumour-infiltrating lymphocytes (TiLS) in histological images. Then, it predicts molecular features from MRI data, including somatic mutations, protein expression, DNA methylation, and copy number variations. Its XAI functionalities provide explanations for classification choices, enabling a deeper understanding of the connections between the molecular and morphological characteristics of cancer [30]. Visualisation techniques are valuable and effective in healthcare applications. In this light, the authors of the case-based reasoning (CBR) framework in [27] decided to include visual explanations within the user interface.

The model in [31] classifies breast cancer tumours according to their shape through CNN architectures. Gradient-Weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME) are two XAI methods employed in this model to explain classification decisions.

An interesting framework for classifying invasive disease episodes (IDEs) in breast cancer patients [32] employs Shapley values [76], a highly used XAI method, to determine key factors influencing IDEs at five and ten years post-diagnosis. Shapley values analysis in this study reveals important factors that affect IDEs and model variables in disease progression. Table 1 presents a comparative summary of existing literature, highlighting model architectures, dataset characteristics, explainability techniques, and reported accuracy, to contextualize the contributions of the proposed ResNet150X framework.

Several studies recommend integrating XAI elements, like the model in [63] for breast cancer detection, the one in [64] for segmenting medical images to isolate tumours in mammograms, and the ensemble decision tree approach in [67] for extracting interpretable classification rules for

Table 1 Comparing identified approaches from the literature to our ResNet150X framework (last row). The boldfaced value indicate the best accuracy

Study	Model	Dataset	XAI method	Accuracy
Binder et al. [29]	Multi-modal CNN with genetic and morphological profiling	Histological & MRI data	Heatmaps, feature-based	86.49%
Hussain et al. [31]	CNN with shape-based tumour classification	Breast cancer MRI dataset	Grad-CAM, G LIME	85.72%
Massafra et al. [32]	XGBoost classifier for IDE prediction	Longitudinal clinical data	Shapley values	90.09%
Khater et al. [63]	Deep CNN for early detection	DCE-MRI Limited slices	Not specified	89.66%
Farrag et al. [64]	U-Net for tumour segmentation	Mammography images	Visual overlays	89.42%
Wang et al. [67]	Ensemble decision tree rules (Interpretable model)	Tabular breast cancer data	Rule-based explanations	91.70%
Lamy et al. [27]	CBR system with visual interface	MRI + clinical records	Visual UI	97.48%
Duwairi and Melhem [61]	ResNet150 variant for MRI classification	3D MRI	none	97.48%
	Improved 3D ResNet150 with optimised 3D-AUM	Full 3D DCE-MRI	SHAP, CIU, 3D Grad-CAM	**99.42**

breast cancer diagnosis. The objective of this research is to elaborate a robust XAI framework for precise breast cancer classification, leveraging the benefits of explainability to enhance its applicability in clinical practice.

3 Proposed approach

The ResNet150X is proposed as an efficient XAI framework for breast cancer classification.

In the proposed methodology, 3D DCE-MRI images obtained from the input datasets undergo initial standardisation and preprocessing steps designed to reduce noise while simultaneously enhancing image sharpness and contrast effectively. This enhancement is achieved through the application of an Adaptive Unsharp Mask (AUM) filter, specifically optimised for this task via a Bayesian Optimisation (BO) algorithm [79]. The AUM filter incorporates an adaptive scaling factor determined by the local variance within the images, which facilitates the sharpening of significant edges while concurrently suppressing noise in regions containing less diagnostic information. Standard data augmentation is used to obtain a good-sized dataset that includes a variety of geometrically transformed versions of the original images. Subsequently, the data is classified through a modified version of the ResNet-150 model [73] featuring custom layers to achieve higher performance in the processing of 3D DCE-MRI data.

To motivate the results, offer visual elucidation of the network's internal operations, and enhance transparency of its underlying mechanisms, this framework integrates three explainable AI methods, namely SHAP, CIU, and 3D Grad-CAM. The block diagram in Fig. 1 graphically shows the core modules and phases of the proposed explainable classification framework. Operationally, the articulated classification methodology is derived from the procedural steps in Algorithm 1.

Require: 3D DCE-MRI images from annotated datasets (Section 3.1)
Ensure: Classified breast cancer type (e.g., Tumour, non-Tumour)
Data Acquisition
Extract 3D DCE-MRI images from the annotated datasets
Split data into training, validation, and test sets
Pre-processing (Section 3.2)
Standardisation ▷ Scaling and resizing
Selection of region of interest ▷ Intensity-based threshold
Noise reduction ▷ AUM Filter
Filtering ▷ AUM Filter
Data augmentation ▷ e.g., rotation, flipping, intensity scaling
Classification (Section 3.3)
Use ResNet-150 as the base model
Replace the last layers ▷ 3D GAP, FC (binary)
Add batch normalisation and dropout layers
Training (Table 2)
Train the resulting model with a binary cross-entropy loss function and Adam optimiser
Explanations (Section 3.4)
Feature contributions to predictions ▷ SHAP
Analyse contextual feature importance ▷ CIU
Visualise tumour regions ▷ 3D Grad-CAM
return Classification outputs and explanations for clinical validation

Algorithm 1 The ResNet150X framework.

3.1 Input

All publicly accessible 3D DCE-MRI scans from the QIN-BREAST dataset [33] and the Duke University Breast

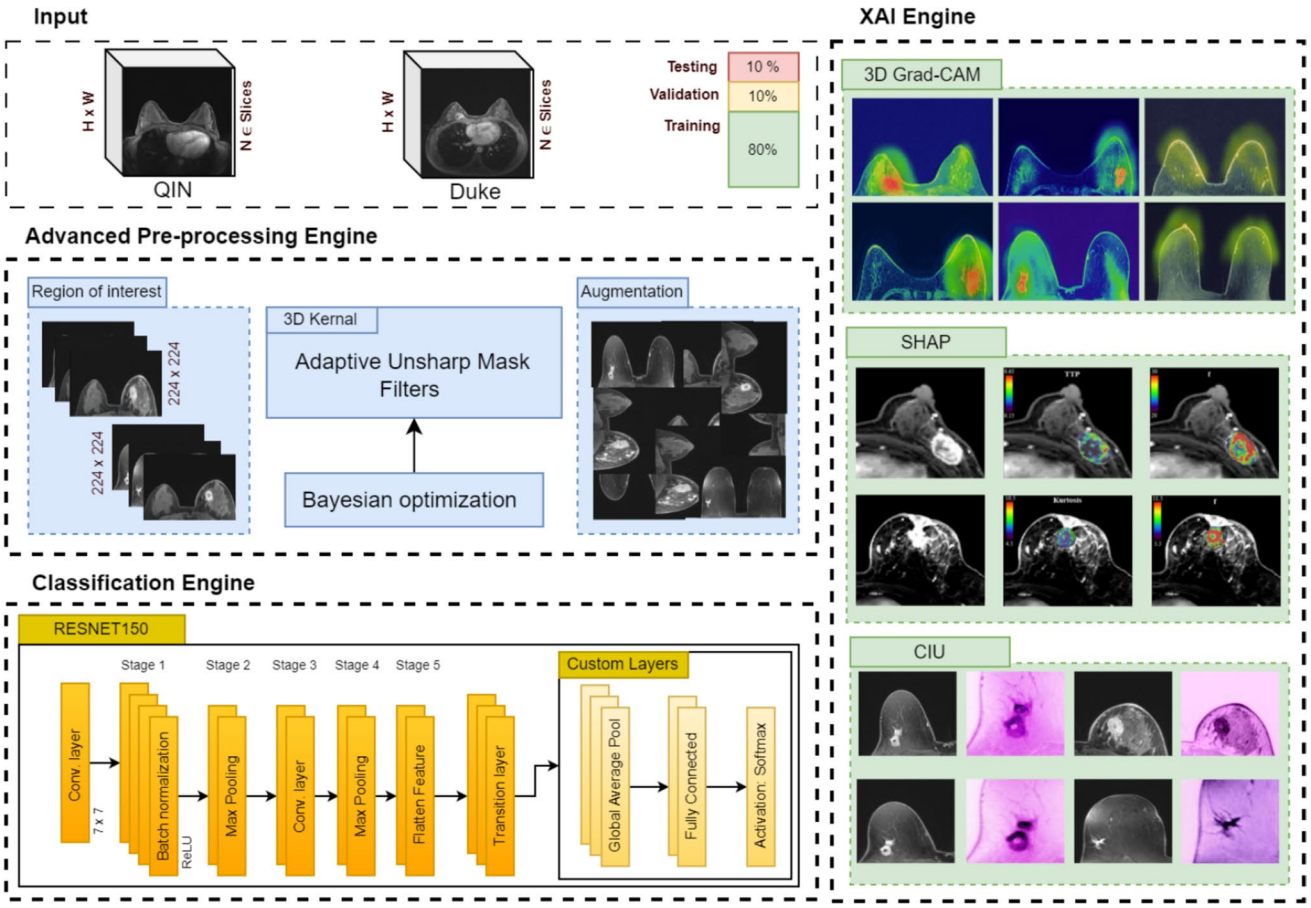

Fig. 1 Graphical representation of the proposed framework

Cancer Dataset [34] are used. The QIN-BREAST dataset, referred to as DS1 in the remainder of this article, comprises 76, 328 images derived from 672 series collected from 10 subjects throughout 20 investigations, where each scan was meticulously annotated with three-dimensional tumour bounding boxes by professional radiologists. The 3D DCE-MRI scans are labelled for the binary classification of breast cancer into tumour or non-tumour classes. This data set is part of an initiative aimed at standardising pharmacokinetic analysis and is widely used for machine learning applications in this field.

The Duke University Breast Cancer Dataset, referred to as DS2 in the remainder of this article, contains scans collected at Duke Medical Center between 2000 and 2014 from 922 patients through 5, 161 MRI series, totalling 773, 888 individual images. The dataset includes three to four post-contrast sequences for most cases, along with fat-saturated gradient echo T1-weighted pre-contrast sequences and non-fat-saturated T1-weighted sequences.

These two datasets are mainly concentrated on invasive breast cancer and provide comprehensive annotations supplied by researchers and radiologists, in addition to 3D bounding boxes that demarcate the tumour regions. They encompass a diverse collection of scans that illustrate various tumour morphologies, sizes, and locations, thereby enhancing the generalisability potential of the proposed framework, as demonstrated in Fig. 2.

3D data augmentation was applied at the series level after splitting the dataset to prevent data leakage and ensure only training data was expanded to enhance generalisation. Through the data augmentation process, the total number of images was significantly increased, with DS1 expanding by 4, 032 series and DS2 gaining an additional 30, 966 series.

Labels are automatically retained during augmentation, as transformations like flipping or scaling don't alter the ground truth. Each augmented sample inherits the original label without manual relabeling, following standard practice. In terms of class distribution, DS1 is distributed into 2, 000 tumour images and 2, 032 non-tumour images, maintaining a relatively balanced composition. DS2 presents a significantly larger proportion of tumour cases, i.e., 15, 000 images, while the non-tumour category comprises 15, 966 images.

The obtained image data set is partitioned into three distinct subsets: training, validation, and testing, following a split ratio of 80 : 10 : 10. The training set facilitates

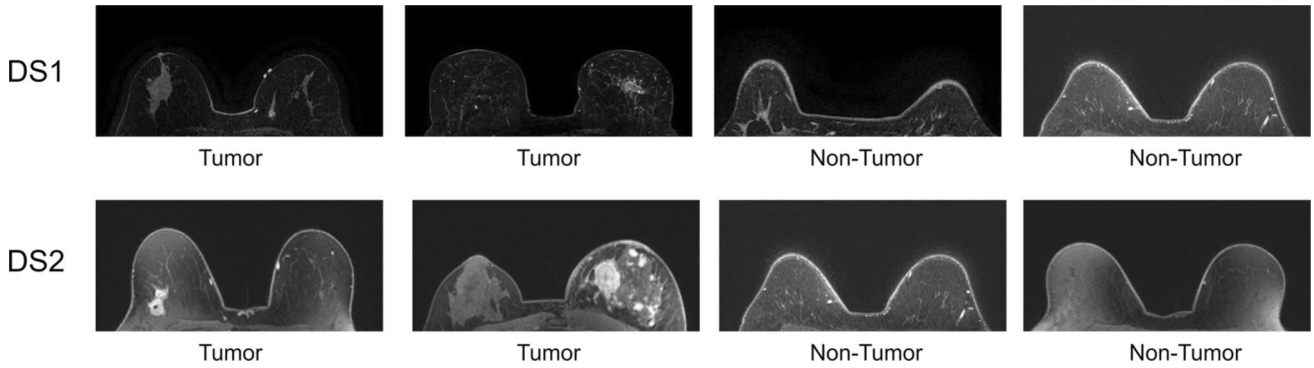

Fig. 2 Samples from DS1 and DS2

model development, while the validation set is employed for hyperparameter optimisation.

3.2 Advanced preprocessing engine

Ensuring meticulous data preparation is crucial to maintaining the integrity and effectiveness of the dataset for subsequent model training and validation.

3.2.1 Data standardisation

Original data from the source datasets undergo normalisation, which applies scaling and resizing operators to all images to enforce a common predefined size of 224×224 pixels (typical size accepted by ResNet150). This improves the model's ability to discriminate by minimising imaging data variability, allowing focus on subtle anatomical and pathological features. Afterward, they are converted into tensors and further normalised using the mean and standard deviation values for each channel [38].

3.2.2 Regions of Interest (ROI)

ROI are then selected by creating a breast mask from the 3D DCE-MRI using Otsu's threshold method [77], which finds the optimal intensity threshold by minimising intra-class variance. Morphological operations then clean the mask, retaining the largest connected breast component. The chest wall is excluded via anatomical cut-offs, and the mask defines the bounding box for image cropping.

3.2.3 Filtering

An AUM filter is designed to remove noise from breast health images while preserving critical structural features, such as edge details and intensity, in the 3D scans. Building on the traditional Unsharp Mask filter, the AUM filter enhances edges more sharply while reducing noise and artifacts through adaptive mechanisms that selectively enhance edges and preserve non-edge areas, avoiding standard noise amplification and halo effects, thereby improving image visibility and classification performance [60]. To adapt the filter to the 3D data type[1], a 3D kernel is integrated.

Selecting the most appropriate kernel is not straightforward. Larger kernel sizes are effective for noise reduction but may introduce blurring. Hence, an adaptive threshold approach is used to determine the level of sharpness applied to each voxel, informed by factors such as local intensity gradients and noise levels. By representing the 3D MRI image as $I(x, y, z)$, where (x, y, z) are the coordinates of a generic voxel in the 3D image, the filter is mathematically described as in (1).

$$\begin{aligned} B(x,y,z) &= \text{Low-pass filter}(I(x,y,z)),\\ U(x,y,z) &= I(x,y,z) - B(x,y,z),\\ \alpha(x,y,z) &= 1 + \beta \cdot \text{variance}(I(x,y,z)),\\ I_{\text{sharp}}(x,y,z) &= I(x,y,z) + \alpha(x,y,z) \cdot (I(x,y,z) - B(x,y,z)), \end{aligned} \tag{1}$$

where $B(x, y, z)$ represents the blurring mask and $U(x, y, z)$ denotes the unsharp mask. The adaptive scaling factor, $\alpha(x, y, z)$, is calculated based on local image characteristics to enhance the image adaptively. The enhanced 3D MRI image obtained after applying the adaptive unsharp mask is represented as $I_{\text{sharp}}(x, y, z)$. The original voxel intensity of the image is given by $I(x, y, z)$, and its blurred version is expressed as $B(x, y, z)$. Together, these components define the adaptive enhancement process for 3D MRI images.

To fine-tune the filter coefficients for optimal image enhancement, BO is applied, which models the objective function $f(\mathrm{x})$ using a *Gaussian Process*. BO iteratively selects the next parameter set x_{next} by maximizing the *Expected Improvement (EI)* as (2):

$$\text{EI}(\mathrm{x}) = (\mu(\mathrm{x}) - f^*)\Phi(z) + \sigma(\mathrm{x})\phi(z), \text{ with } z = \frac{\mu(\mathrm{x}) - f^*}{\sigma(\mathrm{x})}. \tag{2}$$

[1] AUM filters are meant for 2D images.

This approach balances exploration and exploitation to identify AUM settings that improve downstream classification performance efficiently. Moreover, the optimised AUM filter improves contrast between lesions and tissues, reduces noise while preserving boundaries, and improves the quality of the 3D DCE-MRI image for a more accurate diagnosis (examples in Fig. 3).

3.2.4 Data augmentation

Breast tumour analysis is challenging due to the complex and variable tumour morphology. To address this, data augmentation is employed to improve the robustness of the dataset and enhance model resilience, adaptability, abstraction, and generalisation in various clinical contexts [40].

Augmentation strategies include various transformations aimed at introducing various variations that replicate real-world tumour scenarios [39]. Vertical and horizontal flipping is also important to mimic different tumour locations within the breast. Scaling adjusts the model to detect different lesion sizes by cropping MRI scans to reflect clinical lesion diameter ranges. Horizontal and vertical translations improve tumour location accuracy, regardless of its location in breast tissue.

It should be stressed that data augmentation is not only a quantitative technique; it is a crucial part of making digital imaging models more accurate for medical use, which in turn helps with the detection and characterisation of breast tumours [41].

3.3 Classification engine

To address the volumetric nature of breast imaging data, the popular ResNet150 model is selected and replaced its standard 2D convolutional layers with 3D ones. Additionally, depth-wise separable convolutions are employed to reduce computational costs and parameter counts, ensuring efficiency without compromising performance. This involves separating the convolution into Depth-wise Convolution (channel-wise) and Point-wise Convolution (combining features), represented as (3) and (4).

$$Y(i,j,c) = \sum_{m=1}^{M}\sum_{n=1}^{N} X(i+m, j+n, c)\cdot W_c(m,n) \tag{3}$$

$$Z(i,j) = \sum_{c=1}^{C} Y(i,j,c)\cdot W_c \tag{4}$$

Here $Y(i, j, c)$) is the feature map after depth-wise convolution, ($X(i, j, c)$) is the input feature map, $(W_c(m, n))$ is the depth-wise convolution kernel, ($Z(i, j)$) is the feature map after point-wise convolution, (C) is the number of channels, and (M, N) is the dimension of the kernel. Depth-wise convolution processes each channel independently, while point-wise convolution aggregates the features, reducing computational cost and the number of parameters. In conclusion, ResNet150's original fully connected layer is replaced with a custom binary classification layer to distinguish between 'Tumour and 'Non-tumour' classes.

The architecture keeps ResNet150's deep residual structure with bottleneck designs, enhancing feature extraction while being computationally efficient. Each bottleneck

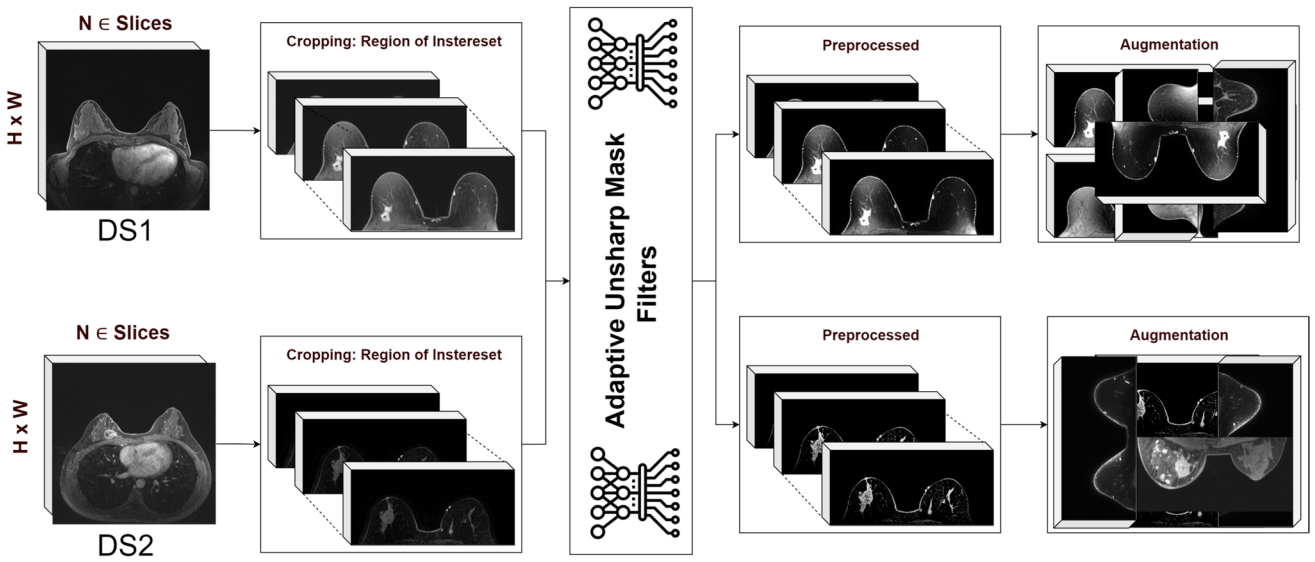

Fig. 3 Workflow for Image Preprocessing in the Proposed Framework

block follows a three-step process: 1×1 dimensionality reduction (5), 3×3 spatial feature extraction (6), and 1×1 dimensional restoration (7), balancing resources and performance. The residual block output is given by (8).

$$Y_1(i,j) = \sum_{k=1}^{K} X(i,j,k) \cdot W_1(k). \tag{5}$$

$$Y_2(i,j) = \sum_{m=1}^{M} \sum_{n=1}^{N} Y_1(i+m, j+n) \cdot W_2(m,n). \tag{6}$$

$$Y_3(i,j) = \sum_{k=1}^{K} Y_2(i,j,k) \cdot W_3(k). \tag{7}$$

$$F(X) = Y_3 + X. \tag{8}$$

The model uses Global Average Pooling (GAP) to condense feature maps before classification, reducing overfitting and preserving key information as shown in (9).

$$Y_{GAP}(c) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j,c), \tag{9}$$

where ($X(i, j, c)$) is the value of the input feature map at spatial location ((i, j)) and channel (c). (H, W) is the height and width of the feature map. The final classification layer outputs probabilities for 'Tumour' and 'Non-tumour' using the Softmax function is represented as (10),

$$P(y = c \mid x) = \frac{\exp(z_c)}{\sum_{i=1}^{C} \exp(z_i)}, \tag{10}$$

where (z_c) is the logit score for class $(c \in \{1, 2\})$.

3.4 XAI engine

Employing XAI methodologies facilitates the generation of interpretable outcomes and visually indicates critical regions within MRI scans that affect predictive results. This approach promotes collaborative decision-making processes between clinical practitioners and AI systems, increasing confidence and trust in deep learning models among patients and healthcare practitioners.

To improve transparency and aid clinical interpretation, we use three XAI techniques—3D Grad-CAM, SHAP, and CIU—tailored for volumetric MRI data. Each method highlights informative regions in the 3D input space, providing unique insights into the model's decision-making.

- **SHAP** (Shapley Additive Explanation) [45] provides insights into which features most impact a model's predictions. For breast tumour classification, SHAP assigns a value to each pixel of the image, indicating its contribution to the classification result. SHAP assigns importance to each voxel based on its marginal contribution to the model output. The Shapley value ϕ_i for a voxel i is given by (11),

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup i) - f(S) \right]. \tag{11}$$

Here, N is the set of all input voxels and S is a subset not containing the i^{th} voxels. Since exhaustive computation is infeasible in 3D, SHAP is approximated using KernelSHAP over voxel groups.

- **3D Grad-CAM** (Gradient-Waved Class Activation Mapping) [46] is a technique used to visualise the regions of a 3D image that are most important for a neural network to make the decision. It functions with pretrained 3D models and does not require any network modifications or retraining [43]. Highlighting the significant patterns in the volume of the image provides an interpretable way to understand the model's decision-making process. 3D Grad-CAM is extended from its 2D form by computing gradients of the class score y^c with respect to each feature map $A^k(x, y, z)$ in the 3D convolutional layer. The importance weights α_k^c are computed as shown in (12) and (13),

$$\alpha_k^c = \frac{1}{Z} \sum_x \sum_y \sum_z \frac{\partial y^c}{\partial A^k(x,y,z)}, \tag{12}$$

$$L_c^{\text{Grad-CAM}}(x,y,z) = \text{ReLU}\left(\sum_k \alpha_k^c A^k(x,y,z) \right), \tag{13}$$

where Z is the number of voxels in the feature map and ReLU ensures retention of positive influence. The resulting activation map is upsampled and overlaid on the MRI volume to visualise class-discriminative regions in 3D.

- **Context Importance and Utility** [44] focuses on utilising only relevant features to explain breast tumour classification based on Context Importance (CI) and Context Utility (CU). CI measures the extent to which the utility of the output changes when the input feature values are adjusted relative to the overall resultant image. CIU operates flexibly, even as the number of images increases, by incorporating an inverse operation.

In this process, all super-pixels are rendered transparent, enabling the effective identification of variations occurring within the breast tumours. This approach enhances the interpretability of the model by providing detailed insights into the significance of individual features and their impact on classification. Given a 3D MRI volume $x \in \mathbb{R}^{X \times Y \times Z}$, CIU computes two core metrics: CI and CU. These are defined as (14),

$$\mathrm{CI}(r) = \left| f(x) - f(x_{\setminus r}) \right| \text{ and } \mathrm{CU}(r) = f(x_r), \tag{14}$$

where $f(x)$ represents the model's prediction for the full volume, $x_{\setminus r}$ is the volume with region r masked (simulating exclusion), and x_r contains only region r, with all other voxels set to a neutral or transparent value. This formulation enables CIU to generate region-specific explanations by quantifying the contextual influence of voxel clusters on the prediction. Additionally, the method remains scalable and computationally efficient across high-dimensional 3D datasets by applying masking selectively over volumetric patches, making it particularly useful for interpretability in medical imaging applications.

4 Experimental and evaluation set-up

To replicate the training and fine-tuning of ResNet150X, the relevant information and parameter values are reported in Table 2.

Table 2 ResNet150X– Hyper-parameter settings

Component	Value	Description
Loss Function	Binary Cross-Entropy	Measures the difference between predicted and actual labels for binary classification.
Optimizer	Adam	Adjusts network parameters iteratively based on gradient updates from the loss function.
Learning Rate	0.005	Optimized learning rate for effective model training and convergence.
Batch Size	16	Number of samples per training batch, chosen based on performance metrics.
Learning Rate Decay	Cosine Annealing	Systematically reduces the learning rate over time to refine model training.
Epochs	10	Total number of iterations over the entire training dataset.
Dropout	N/A	The Proposed model relies on GAP and residual connections instead.
Termination criterion	N/A	The Proposed model uses learning rate decay, which helps convergence without needing to halt training early.

Hyperparameters for the ResNet150X model were based on prior research [48–50], empirical tuning and task-specific optimisation. The binary cross-entropy loss function was used for binary classification. The Adam optimiser, known for its adaptive learning rate, was chosen. A 0.005 learning rate was set to balance speed and stability. A batch size of 16 was settled on after tests to manage memory and stability. Cosine annealing was used for learning rate decay for gradual model refinement. The model was trained for 10 epochs, achieving sufficient convergence. Dropout and early termination were not used, as GAP and residual connections enhanced performance and convergence.

A comprehensive set of established performance metrics is used to assess the proposed model, including Accuracy, F2 Score, Area Under the Curve (AUC), F2 Score and Cohen Kappa scores [13]. These performance metrics, in particulr F1-Score and Cohen Kappa, are crucial to understanding the predictive capabilities of the model and its ability to distinguish between breast lesions accurately. The F2-score is a metric used to evaluate the performance of a binary classification model, emphasising recall more than precision. You can see the formula in (15),

$$F_\beta = \frac{(1+\beta^2) \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{(\beta^2 \cdot \mathrm{Precision}) + \mathrm{Recall}}, \tag{15}$$

where $\beta = 2$, $\mathrm{Precision} = \frac{TP}{TP+FP}$ and $\mathrm{Recall} = \frac{TP}{TP+FN}$, with TP, FP, and FN representing true positives, false positives, and false negatives, respectively.

Cohen's Kappa is another statistical metric used to measure the agreement between a classifier and ground truth when assigning categorical labels, accounting for the agreement that might occur by chance, as you can see in (16),

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{16}$$

where $P_o = \frac{TP+TN}{n}$ is the observed agreement and $P_e = \frac{(TP+FP)(TP+FN)+(TN+FP)(TN+FN)}{n^2}$ is the expected agreement.

Moreover, interpretability metrics, including SHAP values, CIU, and 3D Grad-CAM visualisations, provide qualitative insights by highlighting the areas of input data that most influence the model's predictions. These tools enhance the transparency of the decision-making process, aligning the model's predictions with clinical expectations. By incorporating these diverse metrics, the evaluation process ensures a robust assessment of the DL model's performance, reinforcing its reliability and effectiveness in detecting breast Tumours from MRI images within a clinical framework.

Table 3 ResNet150X vs baseline and state-of-the-art models on DS1. Boldfaced values indicate the best results

Ref.	Year	Technique	Accuracy	Precision	Recall	F1-Score
Comes et al. [48]	2024	3D BB sequence (3D customised CNN)	78.74%	78.20%	79.10%	78.65%
Iqbal and Sharif [49]	2024	Encoder-decoder architecture with feature fusion	94.45%	94.10%	94.80%	94.45%
Song et al. [50]	2023	CNN-EfficientNet and Transformer-PS-ViT	93.56%	93.00%	94.00%	93.50%
Iqbal and Sharif [51]	2023	Swin-Transformer + UNet	94.40%	93.80%	95.10%	94.45%
Muduli et al. [52]	2022	Customised CNN	93.14%	92.70%	93.60%	93.15%
Khamparia et al. [53]	2021	MVGG and ImageNet	94.32%	93.50%	94.90%	94.20%
Ahmed et al. [54]	2021	Customised VGG16	97.12%	96.70%	97.50%	97.10%
Shrivastava and Bharti [55]	2020	Stochastic Residual Gradient	91.41%	91.00%	91.80%	91.40%
Zhang et al. [56]	2020	Customised CNN	94.92%	94.60%	95.20%	94.90%
ResNet150X		Improved 3D ResNet150 with optimised 3D-AUM	**99.15%**	**99.42%**	**99.38%**	**99.40%**

Table 4 ResNet150X vs. baseline and state-of-the-art models for DS2. Boldfaced values indicate the best results

Ref.	Year	Technique	Accuracy	Precision	Recall	F1-Score
Comes et al. [48]	2024	3D BB sequence (3D customized CNN)	81.52%	80.90%	82.10%	81.49%
Iqbal and Sharif [49]	2024	Encoder-decoder architecture with feature fusion	95.13%	94.80	95.30%	95.05%
Song et al. [50]	2023	CNN-EfficientNet and Transformer-PS-ViT	92.46%	91.90%	93.00%	92.45%
Iqbal and Sharif [51]	2023	Swin-Transformer + UNet	92.16%	91.40%	92.90%	92.14%
Muduli et al. [52]	2022	Customised CNN	94.63%	94.10%	94.90%	94.50%
Khamparia et al. [53]	2021	MVGG and ImageNet	96.51%	96.20%	96.80%	96.50%
Ahmed et al. [54]	2021	Customised VGG16	98.18%	97.90%	98.40%	98.15%
Shrivastava and Bharti [55]	2020	Stochastic Residual Gradient	94.55%	94.10%	94.80%	94.45%
Zhang et al. [56]	2020	Customised CNN	93.65%	93.20%	94.10%	93.64%
ResNet150X		Improved 3D ResNet150 with optimised 3D-AUM	**99.42%**	**99.38%**	**99.53%**	**99.45%**

5 Results

The ResNet-150X model shows superior performance and greater interpretability, often missing in traditional methods, compared to baseline models such as those included in the Table 3 for comparison. For a clear and fair comparison, the methods listed in Tables 3 and 4 were reproduced on DS1 and DS2 using the same validation protocol, since their original studies did not report results on these datasets.

During the validation phase, the model showed remarkable accuracy, reaching a maximum of one hundred percent by the eighth epoch for the non-tumour class in DS2, as graphically reported in Fig. 4. The reliability of the model is further supported by the fact that its precision, recall, and F1 score were near 100% during that epoch. In the testing phase (Fig. 5), the model achieved an accuracy of 98.861% for tumour and 99.447% for non-tumour in DS1, and 99.104% for tumour and 99.753% for non-tumour in DS2. These outcomes underscore the model's exceptional reliability and high performance across both datasets (see performance in Fig. 6).

In data set DS1, the model exceeds expectations in accuracy and reliability (Table 5). Its near-perfect F2 score of 99.4% emphasises recall, ideal for medical diagnostics where minimising false negatives is crucial. A Cohen's Kappa of 97.5% shows near-perfect agreement between predictions and actual labels, accounting for chance and providing consistency even with unbalanced datasets.

Figure 7 presents the progression of AUC, Sensitivity, and Specificity over 10 epochs for DS1 and DS2. Both datasets show consistent improvement across all metrics, indicating effective model learning and convergence. For DS1,

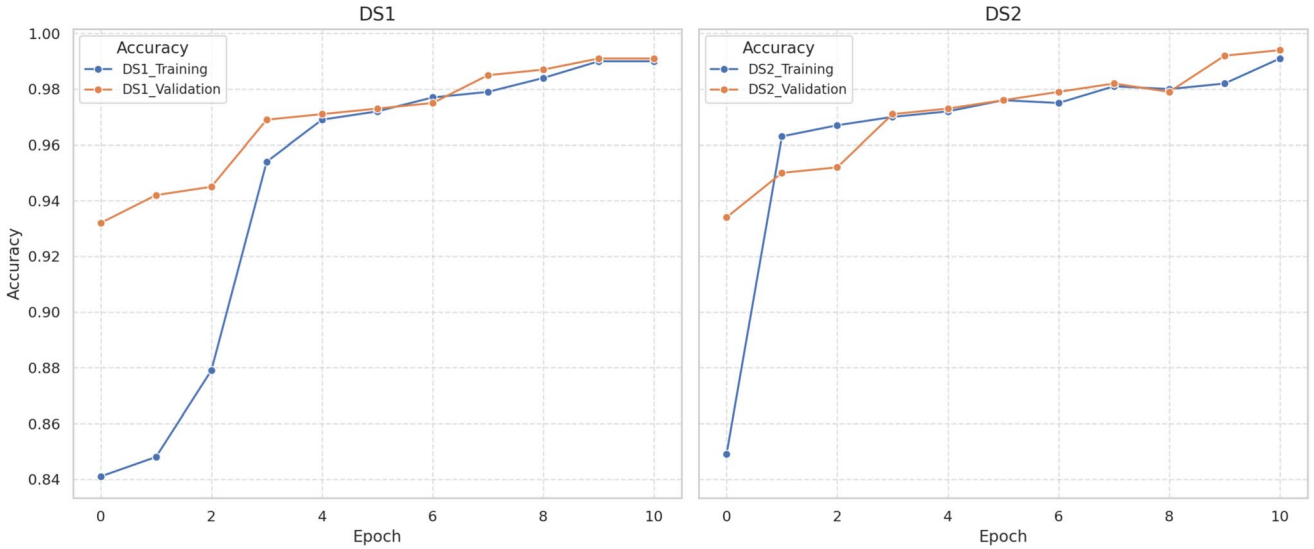

Fig. 4 Accuracy and loss trends for training and validation

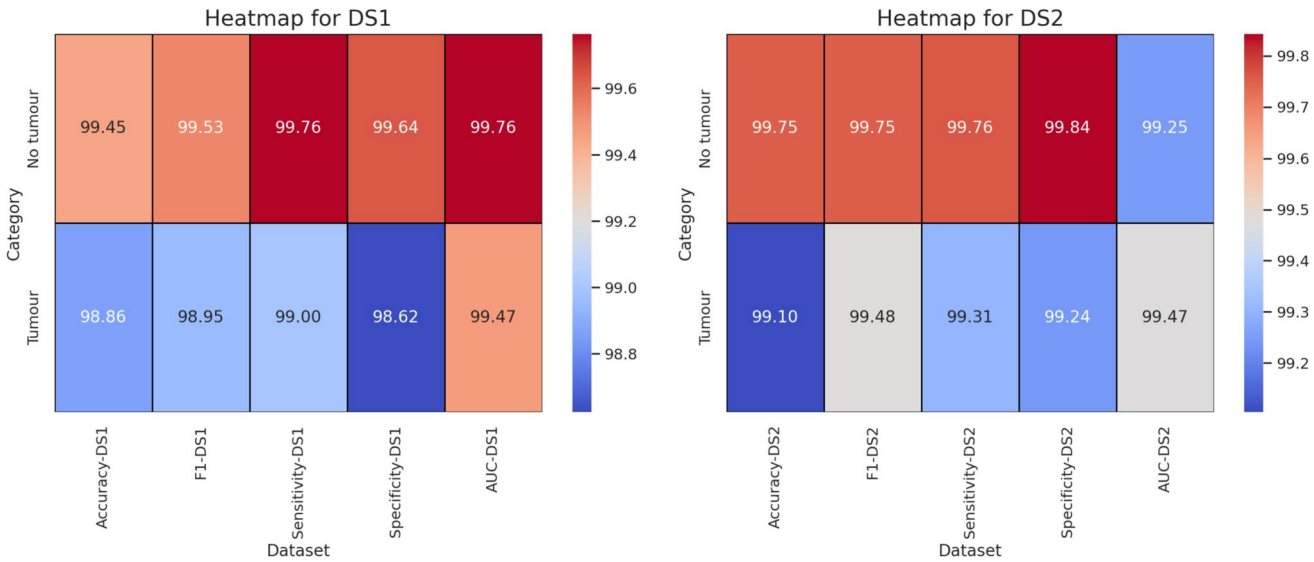

Fig. 5 Testing results (%) on DS1 and DS2

AUC increases from 97.0% to 99.3%, with sensitivity and specificity closely aligned, reflecting balanced and reliable classification. DS2 shows a similar upward trend, starting at 96.0% and reaching 99.0% AUC, confirming the model's robustness across varying data distributions. The non-linear growth patterns suggest typical model training behaviour, with rapid early improvement followed by convergence, and demonstrate strong diagnostic potential.

ResNet150X outperforms several state-of-the-art pre-trained models on the two datasets under investigation (Tables 6 and 7). With an exceptional accuracy (99.154% on DS1 and 99.428% on DS2), it outperforms its predecessor ResNet-150 and advanced models like MACUNet. It would be noted that MACUNet, leveraging multi-scale attention mechanisms, performed well (94.44% on DS1 and 96.177% on DS2), highlighting the benefits of attention for feature focus. EfficientNetB7 and InceptionResNet, which use compound scaling and inception modules with residual connections, respectively, also seem promising but were outperformed by MACUNet and ResNet-based models. U-Net [57] shows the lowest accuracy (77.770% on DS1 and 81.152% on DS2), reflecting the limitations of its vanilla encoder-decoder structure.

In a nutshell, the proposed model is very competitive and its high accuracy, combined with XAI integration for interpretability, underscores its suitability for clinical applications requiring precision and transparency, establishing it as a robust and trustworthy solution. Figure 8 visualises SHAP

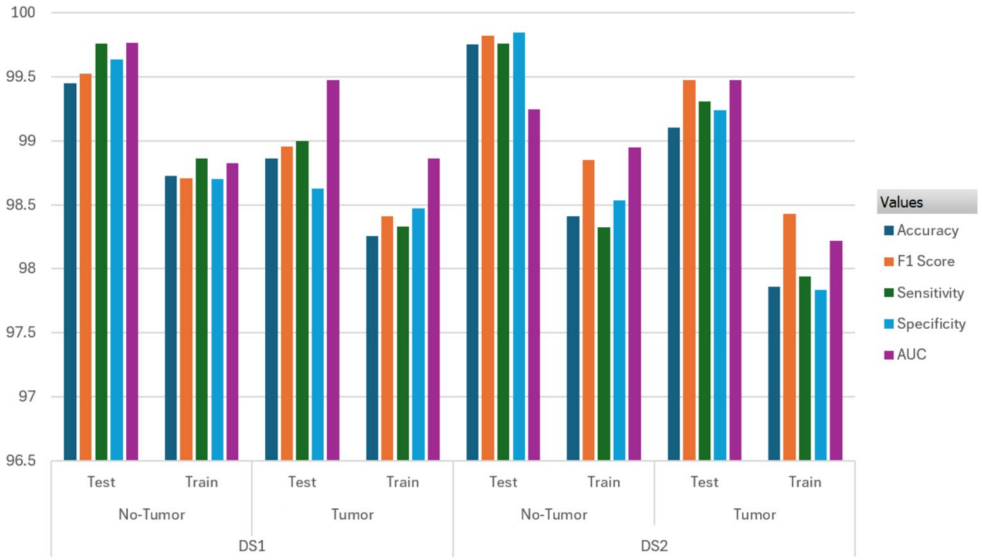

Fig. 6 Testing and training performance of the proposed ResNet150X over DS1 and DS2

Table 5 Reliability metrics

%Dataset	F2 Score	Cohen Kappa	Sensitivity	Specificity	AUC
DS1	0.99%	0.97%	99.38%	99.13%	99.62
DS2	0.97%	0.98%	99.53%	99.54%	99.36%

contributions by colouring relevant pixels according to the colour map scheme overlayed on MRI scans.

While the 'input' column shows the original image, the 'Raw-SHAP' column captures both positive and negative contributions across the tumour area, and the second 'Threshold-SHAP' column focuses on the most significant contributions. In other words, Raw-SHAP provides a comprehensive overview of the contributions made by different areas of the input image within the model. In contrast, Threshold-SHAP highlights the pivotal regions that are instrumental in influencing the decision-making process.

Although DS1 and DS2 comprise MRI scans in the same anatomical region, the SHAP visualisations of their respective model outputs reveal distinct differences in the spatial distribution and intensity of the metric. Within DS1, the two output images for each MRI scan exhibit unique intensity distributions, with one image demonstrating more prominent high-intensity regions (red) compared to the other. A similar pattern is identified in DS2, albeit with more diffuse alterations. These disparities arise from variations in the model's sensitivity to features within the tumour region. This visual representation of high- and low-intensity characteristics increases interpretability, allowing healthcare professionals better to understand the decision-making process of the proposed framework and cultivate trust in AI-driven methodologies. 3D-Grad-CAM also overlays heat

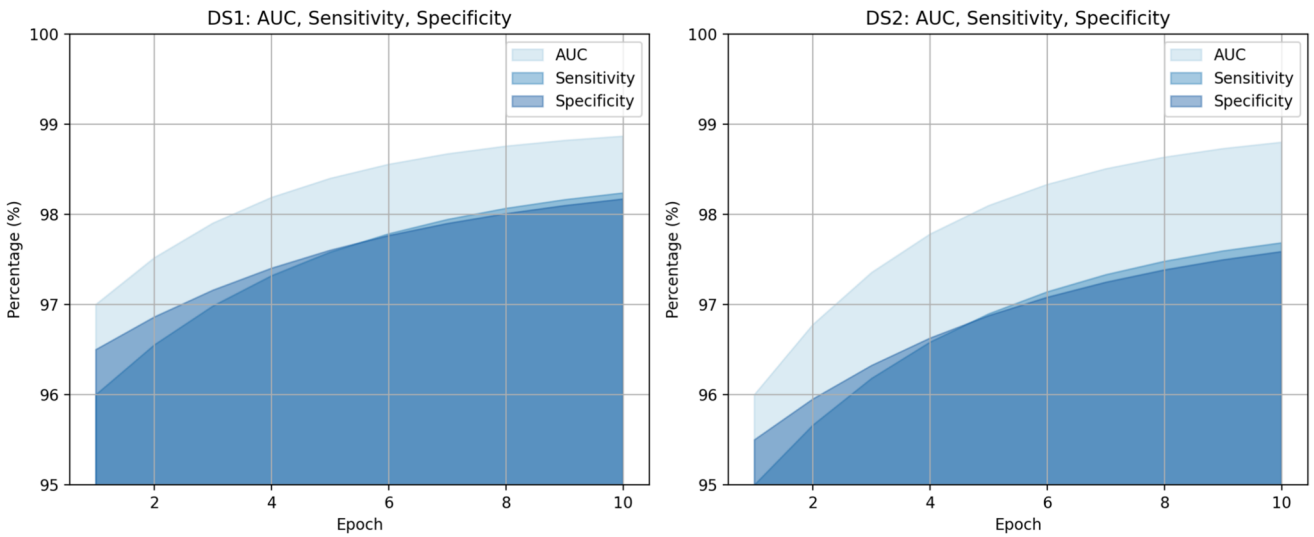

Fig. 7 Performance of the proposed ResNet150X over DS1 and DS2

Table 6 Comparison with pre-trained models using DS1. Boldfaced values indicate the best results

Model	Description	Accuracy	Precision	Recall	F1-Score
VGG16	Stacked CNN layers	84.33%	83.20%	84.90%	84.04%
EfficientNetB7	Scales depth, width, and resolution	86.95	86.50%	87.10%	86.80%
InceptionResNet	Combines Inception and residual connections	85.69%	84.80%	86.40%	85.59%
DenseNet	Dense connections to reuse features	82.18%	81.70%	83.20%	82.44%
MobileNet	Use depthwise separable convolutions	84.13%	84.60%	83.20%	83.89%
U-Net	Encoder-decoder with skip connections	77.77%	76.90%	78.40%	77.64%
MACUNet	U-Net with multi-scale attention mechanisms	94.44%	94.10%	94.80%	94.45%
ResNet150	Residual connections to fix vanishing gradients	97.45%	97.10%	97.80%	97.44%
ResNet150X	Improved 3D ResNet150 with optimised 3D-AUM	**99.15%**	**99.42%**	**99.38%**	**99.40%**

Table 7 Comparison with pre-trained models using DS2. Boldfaced values indicate the best results

Model	Description	Accuracy	Precision	Recall	F1-Score
VGG16	Stacked CNN layers	85.19%	84.50%	85.80%	85.14%
EfficientNetB7	Scales depth, width, and resolution	87.47%	87.00%	87.90%	87.44%
InceptionResNet	Combines Inception and residual connections	88.15%	87.60%	88.70%	88.15%
DenseNet	Dense connections to reuse features across layers	84.13%	83.70%	84.60	84.14%
MobileNet	Use depthwise separable convolutions	85.16%	85.50%	84.80%	85.15%
U-Net	Encoder-decoder with skip connections	81.15%	80.30	82.00%	81.14%
MACUNet	U-Net with multi-scale attention mechanisms	96.17%	95.90%	96.40%	96.15%
ResNet150	Residual connections to fix vanishing gradients	98.93%	98.70%	99.10%	98.90%
ResNet150X	Improved 3D ResNet150 with optimised 3D-AUM	**99.42%**	**99.38%**	**99.53%**	**99.45%**

maps onto MRI images to pinpoint areas that substantially impact the model's interpretations (Figs. 9 and 10).

To show how the focus of the model on critical regions of breast tumour evolves during training, the Epoch-Wise 3D Grad-CAM visualisations for DS1 and DS2 are reported in Figs. 11 and 12, respectively. Early epochs (columns 1-3) exhibit broader and less focused activations, indicating the model's initial exploration of general patterns in the data. As training advances (columns 4-6), activations become more focused on tumour regions, highlighting the model's ability to identify relevant features and ignore background noise. Over epochs, high-activation regions (red and yellow) align more with tumour areas, showing improved accuracy. In contrast, low-activation regions (green and blue) fade in irrelevant areas, indicating increased boundary identification confidence. Focus variability decreases in later epochs, indicating convergence and learning refinement, essential for reliable clinical deployment.

These heat maps for non-tumour cases emphasise the periTumoural area around the tumour as the critical zone for prediction, whereas heat maps for tumour cases identified either parts or the entirety of the intraTumoural area as saliency zones influencing the network's final decisions. Across multiple samples, the model consistently focuses on similar anatomical areas, reflecting a robust and reliable identification of key structures. Sharply concentrated overlays in some cases suggest confident feature detection, while more diffuse overlays indicate uncertainty or contributions from multiple regions.

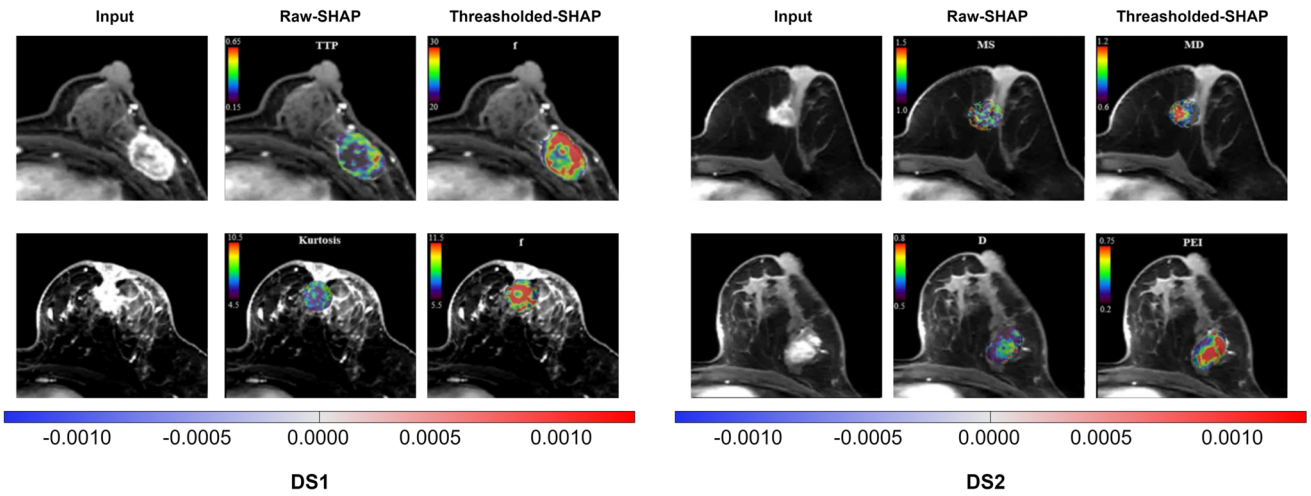

Fig. 8 Example of SHAP analysis and visualisations. The colour gradation spans from −0.0010, represented by blue, to 0.0010, represented by red, where blue denotes regions of lower intensity and red areas of higher intensity

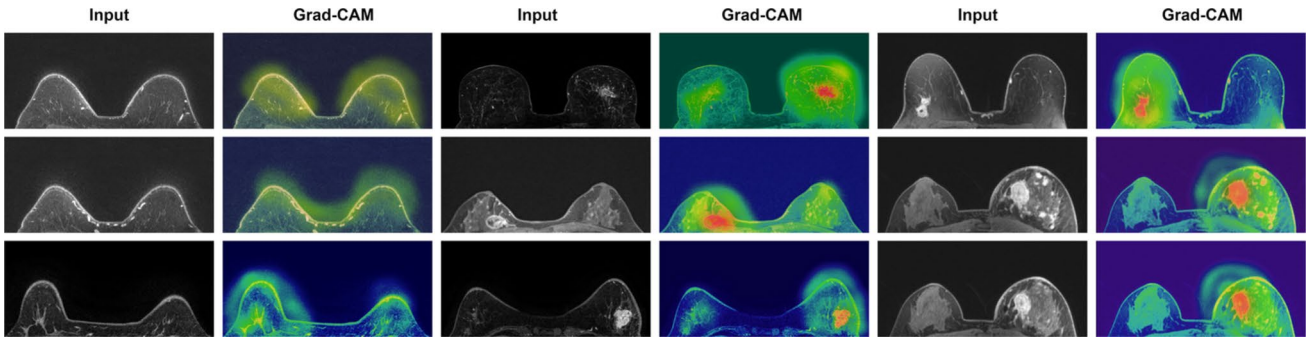

Fig. 9 Example of 3D Grad-CAM visualisations on DS1

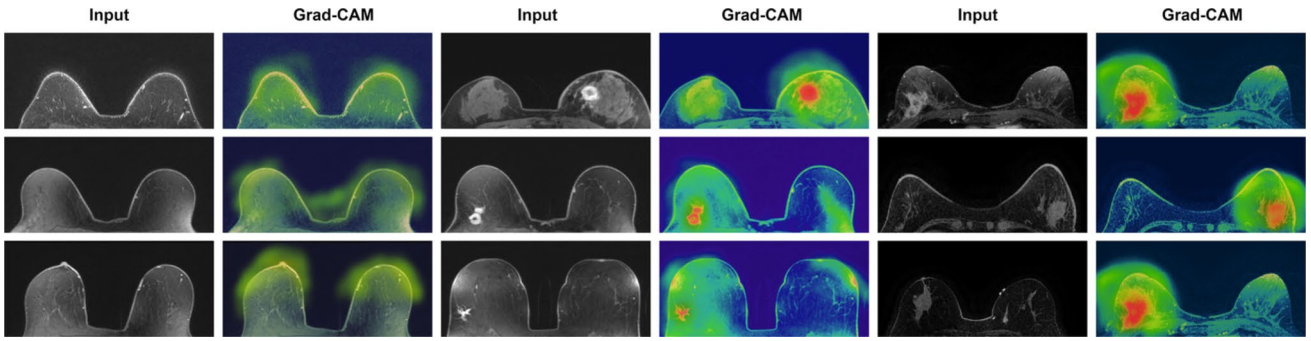

Fig. 10 Example of 3D Grad-CAM visualisations on DS2

The examples in Fig. 13 are annotated with the last method, namely CIU, with a threshold value of 0.01. In the CIU-based explanation module, each 3D MRI image is segmented into fifty superpixels-compact, homogeneous regions that group spatially and visually similar voxels. This segmentation helps maintain anatomical structure and local context while simplifying the interpretability analysis. The model then evaluates each superpixel by computing its Contextual Importance (CI) and Contextual Utility (CU).

CI quantifies how much the prediction outcome changes when the features within a superpixel are altered, thereby reflecting its sensitivity and impact. CU measures how positively or negatively a superpixel contributes to the classification decision, with values close to 1 indicating high utility (positive influence) and values near 0 denoting negligible or negative impact.

During inference, the CIU engine iteratively perturbs superpixels and observes the resulting changes in model predictions. Superpixels with the highest CI and CU scores

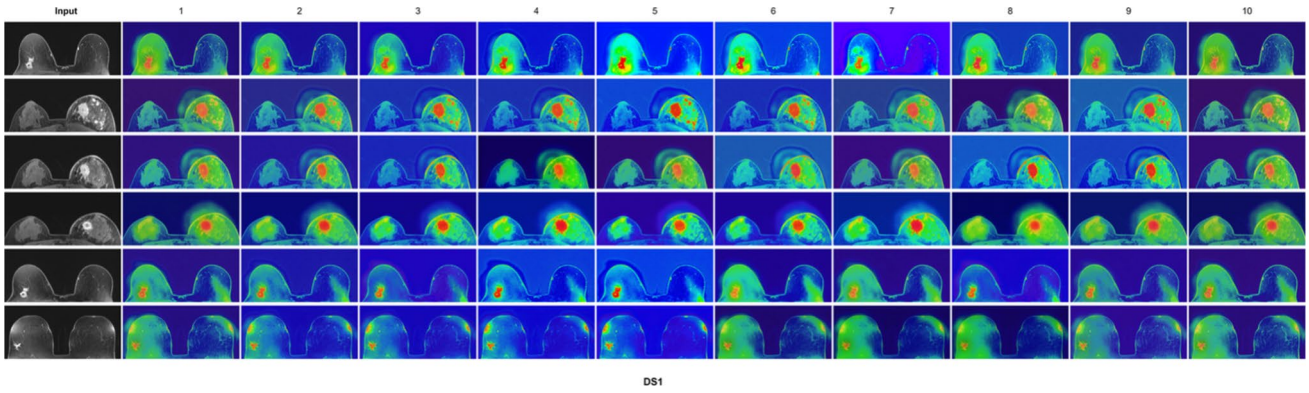

Fig. 11 Epoch-Wise 3D Grad-CAM Visualisations for DS1

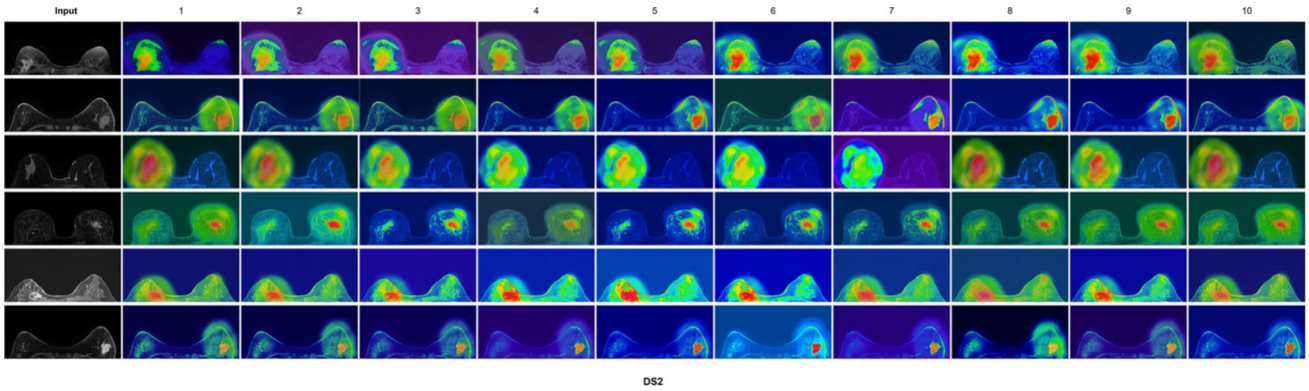

Fig. 12 Epoch-Wise 3D Grad-CAM Visualisations for DS2

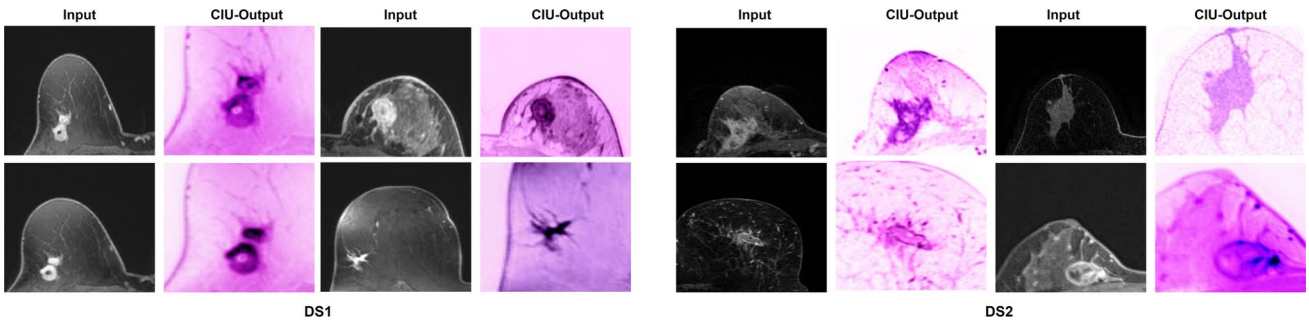

Fig. 13 Example of CIU Visualisations for DS1 and DS2

are considered the most influential in the decision-making process. These regions are visually highlighted on the CIU output map, allowing clinicians to identify critical anatomical structures, such as tumour boundaries or peri-tumour tissue, that directly influenced AI classification. This not only enhances model transparency but also fosters greater trust and clinical utility by aligning explanations with radiological expectations.

In the DS1 and DS2 datasets, the CIU overlays correspond to clinically relevant abnormalities [78], such as dense tissue or lesions, suggesting potential malignancies. The uniformity of visualisations between datasets indicates the robustness of the model in identifying significant regions, despite variations in imaging characteristics.

The utility of these visualisations in healthcare is significant, and their application should become more prevalent. Their function extends beyond the interpretation of models, offering substantial visual support to healthcare professionals to make informed diagnoses. In addition, systematic analysis of the voxels helps identify subtle abnormalities that can otherwise be missed in ambiguous cases. The CIU visualisations also play a crucial role in AI model training and validation. By assessing whether the highlighted regions align with the expected clinical characteristics, experts can determine whether the model emphasises the relevant areas

Table 8 Ablation study showing the effect of 3D AUM and customisation on ResNet150 across DS1 and DS2. Boldfaced values indicate the best results

Model Variant	Dataset	Accuracy	Precision	Recall	F1-Score
Baseline ResNet150	DS1	97.45%	97.10%	97.80%	97.44%
Baseline ResNet150 + 3D AUM		98.35%	98.10%	98.20%	98.15
Customized ResNet150		98.7%5	98.65%	98.85%	98.7%5
Full ResNet150X		**99.15%**	**99.42**	**99.38%**	**99.40%**
Baseline ResNet150	DS2	98.93%	98.70%	99.10%	98.90%
Baseline ResNet150 + 3D AUM		99.10%	98.95%	99.15%	99.05%
Customized ResNet150		99.28%	99.10%	99.40%	99.25%
Full ResNet150X		**99.42%**	**99.38%**	**99.53%**	**99.45%**

and implement the required adjustments. In addition, these overlays have the potential to enhance patient engagement by visually communicating diagnostic results, thus fostering greater confidence in AI-driven healthcare solutions.

5.1 Ablation study

The ablation study presented in Table 8 systematically evaluates the individual and combined contributions of the main components integrated into the proposed ResNet150X architecture: the 3D Attention-Upsampling Module (AUM) and the customised ResNet150 backbone. Results across both DS1 and DS2 datasets clearly show that each component brings measurable performance improvements over the baseline ResNet150 model. Specifically, adding the 3D AUM to the baseline improves accuracy from 97.45% to 98.35% on DS1 and from 98.93% to 99.10% on DS2. Similarly, incorporating the customized backbone raises performance to 98.75% (DS1) and 99.28% (DS2), with corresponding increases in precision, recall, and F1-score.

The full ResNet150X model, which combines both enhancements, achieves the highest scores across all evaluation metrics-99.15% accuracy on DS1 and 99.42% on DS2, with F1-scores reaching 99.40% and 99.45%, respectively. This demonstrates not only the complementary effectiveness of the two modules but also the robustness and generalisation capability of the proposed architecture. The consistent improvement in recall and F1-score is especially significant in clinical or critical applications, where the cost of false negatives can be high. Thus, the ablation results confirm that the full integration of 3D AUM and customisation in ResNet150X leads to a statistically and practically superior model, validating its role as a meaningful advancement over existing baselines.

5.2 Statistical significance study

The analysis in Table 9 shows that the proposed ResNet150X consistently and significantly outperforms a range of baseline and advanced models across both DS1 and DS2. All p-values are well below 0.05, indicating strong statistical significance. Models like InceptionResNet, EfficientNetB7, MobileNet, and 3D BB CNN show large negative t-statistics (e.g., -83.27 on DS1 and -57.51 on DS2 for InceptionResNet), reflecting a clear and consistent performance gap in favor of ResNet150X. These negative values arise because each model is statistically compared against ResNet150X, which consistently exhibits higher average scores. Even the base ResNet150 shows significant inferiority, underscoring the impact of the proposed enhancements.

ResNet150X achieves this improvement by integrating 3D Attention-Upsampling Modules (AUM), a customised ResNet150 backbone, and explainable AI (XAI) components. These features enhance both interpretability and representational capacity, helping the model capture deeper semantic information while maintaining training stability. Its consistent superiority across two datasets highlights

Table 9 Statistical comparison of top models against ResNet150X on DS1 and DS2

Model	DS1		DS2	
	T-Statistic	P-Value	T-Statistic	P-Value
InceptionResNet	-83.2663	0.000840	-57.5133	0.000008
3D BB CNN [48] (2024)	-86.0256	0.000948	-83.8527	0.000003
EfficientNetB7	-64.9866	0.000364	-74.2339	0.000003
MobileNet	-49.5198	0.000248	-112.7337	0.000001
Stochastic Gradient [55]	-67.8402	0.000712	-38.6475	0.000017
DenseNet	-45.3902	0.000441	-94.7456	0.000001
U-Net	-50.1372	0.000830	-60.6312	0.000008
Custom CNN [52]	-31.2398	0.000261	-33.4653	0.000033
SwinTrans+UNet [51]	-43.1773	0.000053	-27.3620	0.000092
ResNet150	-17.7481	0.000829	-7.0587	0.002380

Table 10 Performance of the proposed ResNet150x under varying class imbalance ratios for DS1 and DS2. Boldfaced values indicate the best results

Dataset	Class Ratio	Accuracy	Sensitivity	Specificity	F1	
			Tumour	non-Tumour	Tumour	non-Tumour
S1	1:1 (Balanced)	99.50%	99.00%	99.70%	99.40%	99.80%
	1:3 (Moderate Imb.)	99.65%	95.00%	99.85%	94.10%	99.90%
	1:9 (Severe Imb.)	**99.78%**	**88.00%**	**99.95%**	**86.50%**	**99.96%**
DS2	1:1 (Balanced)	99.60%	99.30%	99.78%	99.50%	99.83%
	1:3 (Moderate Imb.)	99.72%	96.00%	99.90%	95.20%	99.93%
	1:9 (Severe Imb.)	**99.85%**	**89.50%**	**99.97%**	**87.30%**	**99.98%**

strong generalisation ability, making it a reliable and effective solution for complex classification tasks such as those in medical or visual domains.

5.3 Data imbalance study

To investigate the effects of class imbalance, we conducted a controlled simulation using three different class distributions between Tumour and non-Tumour cases: balanced (1:1), moderately imbalanced (1:3), and severely imbalanced (1:9). Table 10 summarises the impact of these distributions on classification performance across both DS1 and DS2 datasets.

The results reveal that while overall accuracy marginally improves as the imbalance increases (e.g., DS1: 99.50% to 99.78%). A closer examination shows a sharp decline in Tumour sensitivity and F1-score—from 99.00% to 88.00% and 99.40% to 86.50% respectively in DS1—as the model becomes biased towards the overrepresented non-Tumour class. Meanwhile, non-Tumour specificity and F1-score continue to rise, indicating a skew in favour of the majority class. A similar trend is evident in DS2. These findings highlight the importance of using class-aware evaluation metrics and balancing techniques to mitigate the risks associated with real-world class imbalances, particularly in critical domains like medical diagnosis.

6 Conclusions and future work

The proposed framework, using 3D DCE-MRI images and three XAI techniques, achieved test accuracies of 98.861% for tumours and 99. 447% for non-tumours in the first dataset DS1 and 99. 104% for tumours and 99.753% for non-tumours in the second dataset DS2, proving its effectiveness in breast cancer diagnosis. The inclusion of XAI enhanced the framework's transparency by highlighting image regions affecting predictions, thus boosting credibility and understanding in healthcare. Despite obtaining an exceptional accuracy and high values in all the other metrics, there are areas where improvement can still be made. For example, the dataset may not reflect the variety of tumour types and imaging modalities encountered in real-world settings. Furthermore, the proposed framework's reliance on MRI as the sole diagnostic tool restricts its applicability in multi-modal diagnostic settings. Consequently, forthcoming research endeavours to enhance the model through the incorporation of additional images to secure a more varied dataset, the integration of multi-modal imaging data, and the employment of attention mechanisms that are presently underutilised for this task.

Author Contributions Conceptualisation A.A., H.E. and F.C.; methodology, A.A.; software, N.U.R.; validation, A.A. and K.A.; formal analysis, A.AM., K.JA.; funding acquisition, A.AR., M.A.; investigation, A.A. and K.A.; resources, S.H.; data curation, K.A.; writing—original draft preparation, A.A., H.E. and F.C.; writing - review & editing, A.AR., A.A., H.E. and F.C.; visualisation, N.U.R.; supervision, S.H.; all authors have read and agreed to the published version of the manuscript.

Data Availability All data in this article is publicly available and referenced.

Declarations

Competing interest The authors declare no conflicts of interest.

References

1. Srinivasu PN, Sandhya N, Jhaveri RH, Raut R (2022) From blackbox to explainable AI in healthcare: existing tools and case studies. Mob Inf Syst 2022(1):8167821
2. de Souza Jr LA, Mendel R, Strasser S, Ebigbo A, Probst A, Messmann H, Jp Papa, Palm C (2021) Convolutional Neural Networks

for the evaluation of cancer in Barrett's esophagus: explainable AI to lighten up the black-box. Comput Biol Med 135:104578
3. Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M, Rebholz-Schuhmann D, Decker S (2023) Explainable AI for bioinformatics: methods, tools and applications. Briefings Bioinf, bbad236–bbad236
4. Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J, Ouyang C, Gupta A, Gu Y, Deveci M (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 96:156–191
5. Alom MR, Farid FA, Rahaman MA, Rahman A, Debnath T, Miah ASM, Mansor S (2025) An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. Sci Rep 15(1):17531. https://doi.org/10.1038/s41598-025-67531-2
6. Gerbasi A, Clementi G, Corsi F et al (2023) DeepMiCa: automatic segmentation and classification of breast microcalcifications from mammograms. Comput Methods Programs Biomed 235:107483. https://doi.org/10.1016/j.cmpb.2023.107483
7. Demiroğlu U, Şenol B (2025) Evaluating Vision Transformer models for breast cancer detection in mammographic imaging. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi 14(1):287–313. https://doi.org/10.17798/bitlisfen.1583948
8. Zhang J, Wu J, Zhou XS et al (2023) Recent advancements in artificial intelligence for breast cancer: image augmentation, segmentation, diagnosis, and prognosis approaches. Semin Cancer Biol 96:11–25. https://doi.org/10.1016/j.semcancer.2023.08.004
9. Agbley BLY, Li JP, Haq AU et al (2023) Federated fusion of magnified histopathological images for breast tumor classification in the Internet of Medical Things. IEEE J Biomed Health Inf, pp 1–12. https://doi.org/10.1109/JBHI.2023.3256974
10. Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Deveci M (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 96:156–191. https://doi.org/10.1016/j.inffus.2023.03.008
11. Jahan I, Chowdhury MEH, Vranic S et al (2025) Deep learning and vision transformers-based framework for breast cancer and subtype identification. Neural Comput Appl 37:9311–9330. https://doi.org/10.1007/s00521-025-10984-2
12. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR (2022) Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Programs Biomed 226:107161
13. Mao YJ, Lim HJ, Ni M, Yan WH, Wong DWC, Cheung JCW (2022) Breast tumour classification using ultrasound elastography with machine learning: a systematic scoping review. Cancers 14(2):367
14. Chakraborty D, Ivan C, Amero P, Khan M, Rodriguez-Aguayo C, Başağaoğlu H, Lopez-Berestein G (2021) Explainable artificial intelligence reveals novel insight into Tumour microenvironment conditions linked with better prognosis in patients with breast cancer. Cancers 13(14):3450
15. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z (2020) Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc 27(7):1173–1185
16. Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, Jiang Q (2021) Improved machine learning-based predictive models for breast cancer diagnosis. Int J Environ Res Public Health 19(6):3211. https://doi.org/10.3390/ijerph19063211
17. Massafra R, Comes MC, Bove S, Didonna V, Gatta G, Giotta F, Fanizzi A, La Forgia D, Latorre A, Pastena MI, Pomarico D, Rinaldi L, Tamborra P, Zito A, Lorusso V, Paradiso AV (2022) Robustness evaluation of a deep learning model on sagittal and axial breast DCE-MRIs to predict pathological complete response to neoadjuvant chemotherapy. J Pers Med 12(6):953
18. Comes MC, Fanizzi A, Bove S, Didonna V, Diotaiuti S, La Forgia D, Latorre A, Martinelli E, Mencattini A, Nardone A, Paradiso AV, Ressa CM, Tamborra P, Lorusso V, Massafra R (2021) Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. Sci Rep 11(1):14123
19. Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, Van Sant EP, Wynn RT, Connolly E, Jambawalikar S (2019) Prior to initiation of chemotherapy, can we predict breast Tumour response? Deep learning convolutional neural networks approach using a breast MRI Tumour dataset. J Digit Imaging 32:693–701
20. Meena J, Hasija Y (2022) Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Comput Biol Med 146:105505
21. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G (2021) Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep 11(1):6968
22. Ladbury C, Zarinshenas R, Semwal H, Tam A, Vaidehi N, Rodin AS, Liu A, Glaser S, Salgia R, Amini A (2022) Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Transl Cancer Res 11(10):3853
23. Nazir S, Dickson DM, Akram MU (2023) Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med 156:106668
24. Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470
25. Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH (2021) nnU-Net for brain tumour segmentation. Brainlesion: glioma multiple sclerosis stroke and traumatic brain injuries, pp 118–132
26. Sunsuhi GS, Albin Jose S (2022) An Adaptive Eroded Deep Convolutional neural network for brain image segmentation and classification using Inception ResnetV2. Biomed Signal Process Control 78(103863):103863
27. Lamy JB, Sekar B, Guezennec G, Bouaud J, Séroussi B (2019) Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med 94:42–53
28. Rafiq A, Chursin A, Awad Alrefaei W, Rashed Alsenani T, Aldehim G, Abdel Samee N, Menzli LJ (2023) Detection and classification of histopathological breast images using a fusion of CNN frameworks. Diagnostics 13(10):1700
29. Binder A, Bockmayr M, Hägele M, Wienert S, Heim D, Hellweg K, Ishii M, Stenzinger A, Hocke A, Denkert C, Müller K-R, Klauschen F (2021) Morphological and molecular breast cancer profiling through explainable machine learning. Nat Mach Intell 3(4):355–366
30. Amoroso N, Pomarico D, Fanizzi A, Didonna V, Giotta F, La Forgia D, Latorre A, Monaco A, Pantaleo E, Petruzzellis N, Tamborra P, Zito A, Lorusso V, Bellotti R, Massafra R (2021) A roadmap towards breast cancer therapies supported by explainable artificial intelligence. Appl Sci 11(11):4881
31. Hussain SM, Buongiorno D, Altini N, Berloco F, Prencipe B, Moschetta M, Bevilacqua V, Brunetti A (2022) Shape-based breast lesion classification using digital tomosynthesis images: the role of explainable artificial intelligence. Appl Sci 12(12):6230
32. Massafra R, Fanizzi A, Amoroso N, Bove S, Comes MC, Pomarico D, Didonna V, Diotaiuti S, Galati L, Giotta F, La Forgia D, Latorre A, Lombardi A, Nardone A, Pastena MI, Ressa CM, Rinaldi L, Tamborra P, Zito A, Paradiso AV, Bellotti R, Lorusso V (2023) Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front Med 10:1116354

33. The Cancer Imaging Archive (TCIA) (2024) QIN-BREAST-DCE-MRI - The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net/collection/qin-breast-dce-mri/
34. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with Tumour locations (Duke-Breast-Cancer-MRI) - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. (n.d.). https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903
35. Kaur G, Sharma N, Gupta R (2023) Wheat leaf disease classification using EfficientNet B3 pre-trained architecture. In: 2023 international conference on Research Methodologies in Knowledge Management Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp 1–5
36. Rehman MU, Cho S, Kim JH, Chong KT (2020) BU-Net: Brain Tumour segmentation using modified U-Net architecture. Electronics (Basel) 9(12):2203
37. Aggarwal M, Tiwari AK, Sarathi MP, Bijalwan A (2023) An early detection and segmentation of Brain Tumour using Deep Neural Network. BMC Med Inform Decis Mak 23(1)
38. Kumar A, Nelson L, Singh S (2023) ResNet-50 transfer learning model for diabetic foot ulcer detection using thermal images. In: 2023 2nd International Conference on Futuristic Technologies (INCOFT)
39. Ullah F, Nadeem M, Abrar M (2024) Revolutionizing brain tumour segmentation in MRI with dynamic fusion of handcrafted features and global pathway-based deep learning. KSII Trans Internet Inf Syst 18(1)
40. Al Moteri M, Mahesh TR, Thakur A, Vinoth Kumar V, Khan SB, Alojail M (2024) Enhancing accessibility for improved diagnosis with modified EfficientNetV2-S and cyclic learning rate strategy in women with disabilities and breast cancer. Front Med 11:1373244
41. Sannasi Chakravarthy SR, Bharanidharan N, Vinoth Kumar V, Mahesh TR, Alqahtani MS, Guluwadi S (2024) Deep transfer learning with fuzzy ensemble approach for the early detection of breast cancer. BMC Med Imaging 24(1):82
42. Sharma D, Prabha C (2023) Security and privacy aspects of electronic health records: a review. In: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), pp 815–820
43. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48
44. Främling K (2022) Contextual importance and utility: a theoretical foundation. In: Australasian joint conference on artificial intelligence. Springer International Publishing, Cham, pp 117–128
45. Huang Y, Wang X, Cao Y, Li M, Li L, Chen H, Tang S, Lan X, Jiang F, Zhang J (2024) Multiparametric MRI model to predict molecular subtypes of breast cancer using Shapley additive explanations interpretability analysis. Diagn Interv Imaging 105(5):191–205
46. Raghavan K (2024) Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimed Tools Appl 83(19):57551–57578
47. Li M, Wang H, Qu N, Piao H, Zhu B (2024) Breast cancer screening and early diagnosis in China: a systematic review and meta-analysis on 10.72 million women. BMC Women's Health 24(1):97
48. Comes MC, Fanizzi A, Bove S, Didonna V, Diotiaiuti S, Fadda F, La Forgia D, Giotta F, Latorre A, Nardone A, Palmiotti G, Ressa CM, Rinaldi L, Rizzo A, Talienti T, Tamborra P, Zito A, Lorusso V, Massafra R (2024) Explainable 3D CNN based on baseline breast DCE-MRI to give an early prediction of pathological complete response to neoadjuvant chemotherapy. Comput Biol Med 172:108132
49. Iqbal A, Sharif M (2024) Memory-efficient transformer network with feature fusion for breast Tumour segmentation and classification task. Eng Appl Artif Intell 127:107292
50. Song P, Yang Z, Li J, Fan H (2023) DPCTN: dual path context-aware transformer network for medical image segmentation. Eng Appl Artif Intell 124:106634
51. Iqbal A, Sharif M (2023) BTS-ST: swin transformer network for segmentation and classification of multimodality breast cancer images. Knowl-Based Syst 267:110393
52. Muduli D, Dash R, Majhi B (2022) Automated diagnosis of breast cancer using multi-modal datasets: a deep convolution neural network based approach. Biomed Signal Process Control 71:102825
53. Khamparia A, Bharati S, Podder P, Gupta D, Khanna A, Phung TK, Thanh DN (2021) Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. Multidimension Syst Signal Process 32:747–765
54. Ahmed AS, Keshk AE, M Abo-Seida O, Sakr M (2022) Tumour detection and classification in breast mammography based on fine-tuned convolutional neural networks. IJCI Int J Comput Inf 9(1):74–84
55. Shrivastava N, Bharti J (2020) Breast Tumour detection and classification based on density. Multimed Tools Appl 79(35):26467–26487
56. Zhang C, Zhao J, Niu J, Li D (2020) New convolutional neural network model for screening and diagnosis of mammograms. PLoS ONE 15(8):e0237674
57. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, pp 234–241
58. Li R, Zheng S, Zhang C, Duan C, Su J, Wang L, Atkinson PM (2021) Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans Geosci Remote Sens 60:1–13
59. Morris EA (2002) Breast cancer imaging with MRI. Radiologic Clinics 40(3):443–466
60. Orhei C, Vasiu R (2023) An analysis of extended and dilated filters in sharpening algorithms. IEEE Access
61. Duwairi R, Melhem A (2023) A deep learning-based framework for automatic detection of drug resistance in tuberculosis patients. Egypt Inf J 24(1):139–148
62. Nasreen G, Haneef K, Tamoor M, Irshad A (2023) A comparative study of state-of-the-art skin image segmentation techniques with CNN. Multimed Tools Appl 82(7):10921–10942
63. Khater T, Hussain A, Bendardaf R, Talaat IM, Tawfik H, Ansari S, Mahmoud S (2023) An explainable artificial intelligence model for the classification of breast cancer. IEEE Access
64. Farrag A, Gad G, Fadlullah ZM, Fouda MM, Alsabaan M (2023) An explainable AI system for medical image segmentation with preserved local resolution: Mammogram Tumour segmentation. IEEE Access
65. Chaudhury S, Sau K (2023) A blockchain-enabled internet of medical things system for breast cancer detection in healthcare. Healthcare Analytics 4:100221
66. Asadi B, Memon Q (2023) Efficient breast cancer detection via cascade deep learning network. Int J Intell Netw 4:46–52
67. Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y (2020) An improved random forest-based rule extraction method for breast cancer diagnosis. Appl Soft Comput 86:105941
68. Alshayeji MH, Ellethy H, Gupta R (2022) Computer-aided detection of breast cancer on the Wisconsin dataset: an artificial neural networks approach. Biomed Signal Process Control 71:103141

69. Abdar M, Makarenkov V (2019) CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement 146:557–570
70. Enobun AE, Anakwenze UH, Taherkhani A, Anastassi Z, Caraffini F, Eshkiki H (2024) Segmenting breast ultrasound scans using a generative adversarial network embedding U-net. In: Xie X, Styles I, Powathil G, Ceccarelli M (eds) Artificial intelligence in healthcare. Springer Nature Switzerland, Cham, pp 149–159
71. Caraffini F, Eshkiki H, Mohammadpour M, Sullo N, George CH (2024) Towards improving single-cell segmentation in heterogeneous configurations of cardiomyocyte networks. In: Xie X, Styles I, Powathil G, Ceccarelli M (eds) Artificial intelligence in healthcare. Springer Nature Switzerland, Cham, pp 104–117
72. Xu W, Fu YL, Zhu D (2023) ResNet and its application to medical image processing: research progress and challenges. Comput Methods Programs Biomed 240:107660
73. Singh R, Sharma N (2024) Enhanced brain tumour segmentation in MRI scans using ResNet-150. In: 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). IEEE, pp 1619–1624
74. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
75. Elmejgari C, Nadir Y, Qbadou M (2024) Deep learning based techniques for breast cancer classification: a systematic review. In: 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). IEEE, pp 01–07
76. Nohara Y, Matsumoto K, Soejima H, Nakashima N (2022) Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed 214:106584
77. Serbet F, Kaya T (2025) New comparative approach to multilevel thresholding: chaotically initialized adaptive meta-heuristic optimization methods. Neural Comput Appl, 1–26
78. Peta J, Koppu S (2024) Explainable Soft Attentive EfficientNet for breast cancer classification in histopathological images. Biomed Signal Process Control 90:105828
79. Cole ER, Connolly MJ, Ghetiya M, Sendi ME, Kashlan A, Eggers TE, Gross RE (2024) SAFE-OPT: a Bayesian optimization algorithm for learning optimal deep brain stimulation parameters with safety constraints. J Neural Eng 21(4):046054