Neural Computing and Applications (2025) 37:3759–3776
https://doi.org/10.1007/s00521-024-10790-2

ORIGINAL ARTICLE

# Efficient breast cancer detection using neural networks and explainable artificial intelligence

Tamilarasi Kathirvel Murugan[1] · Pritikaa Karthikeyan[1] · Pavithra Sekar[1]

Received: 16 May 2024 / Accepted: 7 November 2024 / Published online: 16 December 2024


**Abstract**
The growing dependence on deep learning models for medical diagnosis underscores the critical need for robust interpretability and transparency to instill trust and ensure responsible usage. This study investigates the efficacy of various explainable artificial intelligence (XAI) techniques in comprehending deep learning models utilized for breast cancer classification from down sampled histopathology images. A comparative assessment of multiple convolutional neural network (CNN) architectures, encompassing standard CNNs, ResNet, VGG-16, and VGG-19, on down sampled images was conducted. The primary goal is to pinpoint the model exhibiting the highest accuracy and subsequently employ three prominent XAI methods—LIME, SHAP, and Saliency Maps—to get insights into the top-performing model. This study identifies VGG-19 as the best-performing model with an accuracy of 92.59% and demonstrates that among various XAI techniques, LIME provides the most accurate and clinically relevant explanations for breast cancer classification from down sampled histopathology images. These findings, validated by medical professionals, enhance the interpretability and reliability of deep learning models in clinical settings, promoting their responsible integration into healthcare practices. This validation was further corroborated through consultation with medical professionals, including doctors specializing in breast cancer diagnosis. This research endeavors to deepen the understanding of the model's rationale and instill confidence in its outputs. The outcomes of this study hold significant promise in elevating the interpretability and reliability of deep learning models tailored for breast cancer diagnosis, thus facilitating their responsible integration into clinical settings.

**Keywords** Breast cancer · Explainable artificial intelligence · Convolutional neural networks (CNN)

## 1 Introduction

In the past few decades, substantial advancements have been made in the battle against breast cancer; however, it continues to pose a formidable challenge in healthcare worldwide. With millions of individuals impacted each year, the imperative to improve diagnostic methods and treatment approaches is undeniable. Breast cancer presents a formidable global health obstacle, placing a considerable strain on healthcare infrastructures and posing a significant risk to the health of women globally. Early detection remains crucial in minimizing the effects of this widespread illness, enhancing treatment effectiveness, and ultimately preserving lives.

For a variety of reasons, diagnosing breast cancer in clinical practice is still extremely difficult. Since early-stage breast cancer frequently has no symptoms, it can be challenging to identify using conventional techniques like mammography. Factors including breast tissue density, imaging variability, and the possibility of human error and weariness among radiologists and pathologists can all be limitations of these techniques. For example, mammography, although a common screening method, is prone to false positives in younger populations and has a sensitivity of about 77% for women with thick breasts. Even with technological advances, current AI models frequently lack the interpretability and openness required to be fully accepted in therapeutic settings. With an estimated 2.3 million new cases worldwide each year, breast cancer is the most frequent malignancy among women, underscoring the

✉ Tamilarasi Kathirvel Murugan
tamilarasi.k@vit.ac.in

[1] School of Computer Science Engineering, Vellore Institute of Technology, Chennai, Tamilnadu, India

significance of handling these issues. Since the 5-year survival rate for cases of localized breast cancer can approach 99%, it is imperative that cases are detected early and accurately. For cases diagnosed at a later stage, the survival rate lowers to 27%. Clinical instances often highlight the serious repercussions of diagnostic errors or delays, emphasizing the need for more trustworthy and comprehensible diagnostic instruments. By assessing and improving the interpretability of deep learning models used for breast cancer classification, this research seeks to close these gaps in knowledge in order to improve diagnostic precision and successfully incorporate these models into clinical practice.

The convergence of advanced technologies, particularly in Artificial Intelligence (AI) and its Explainable AI (XAI) subset, has emerged as a transformative force in revolutionizing breast cancer detection methodologies. Traditionally, the primary modalities for breast cancer screening and diagnosis have included mammography, ultrasound, and histopathological analysis. While these techniques have been foundational in identifying potential malignancies, they are not without limitations. Mammography, despite its widespread use, may yield false positives or negatives, and the interpretation of results is subject to the expertise of the radiologist. Histopathological examination, though highly accurate, is invasive and time-consuming. The advent of AI, and more specifically XAI, presents an opportunity to augment and refine these diagnostic processes. Explainable AI has become increasingly vital in addressing the complexity of sophisticated machine learning models, aiming to ensure transparent and understandable decision-making, especially in clinical settings. In the domain of breast cancer detection, where the stakes are high, the interpretability of AI algorithms is paramount for building trust among healthcare professionals and engendering confidence in patients. Pretrained CNN networks like VGG-16, VGG-19 and ResNet-50 offer several benefits. They offer pre-learned features that have been retrieved from huge datasets, allowing for quicker convergence and requiring less intensive training. Utilizing pre-trained CNNs enhances performance, making them indispensable in practical applications. To improve model performance, hyper parameter tuning is a crucial subsequent machine learning step. It entails optimizing the settings of parameters like batch size, learning rate and regularization strength.

Explainable Artificial Intelligence (XAI) methods serve as indispensable tools in unraveling the complexities inherent in black-box machine learning algorithms. These sophisticated techniques offer pathways to decipher the intricate decision-making processes underlying these algorithms, providing post hoc explanations that shed light on the rationale behind a model's predictions. By emphasizing the characteristics influencing the decision-making process, XAI not only enhances the accuracy of breast cancer diagnosis but also enables a deeper comprehension of the traits and patterns indicative of malignancies. This heightened understanding fosters greater trust among healthcare professionals and patients alike, underscoring the importance of transparency and interpretability in clinical settings.

Against the backdrop of this transformative potential, the comprehensive exploration delves into the extensive literature and recent advancements in breast cancer detection with XAI. It is imperative to recognize the critical imperative of model transparency and interpretability, necessitating the augmentation of inquiry with state-of-the-art subsequent explanation approaches. By harnessing methodologies such LIME, SHAP and saliency mapping, the proposed system endeavors to bring the underlying decision-making rationale of the chosen model. Through elucidating the discriminative features driving classification outcomes, these interpretability methods provide a nuanced understanding of model behavior, thereby facilitating informed clinical decision-making and enhancing trust in computational pathology systems.

In essence, the concerted efforts in advancing the frontier of breast cancer detection with XAI not only promise heightened diagnostic accuracy but also pave the way for a future where healthcare professionals and patients are empowered with actionable intelligence, ultimately leading to improved patient outcomes and a more effective response to the global challenge of breast cancer. Besides AI and XAI, novel technologies like digital pathology and telemedicine are paramount in increasing the accessibility to the diagnosis process, and also in the efficiency in handling breast cancer findings. Digital pathology causes histopathological films digitization, thus remote invasions occurring, the collaboration among the experts, and the incorporation of AI algorithms for the automated analysis. Besides, telemedicine platforms carry out the outreach of the specialist medical services to the poor populated groups, permitting early screenings, consultations, and follow ups, consequently eliminating the differences in the detection and treatment processes of the cancer of the breast [21].

Furthermore, the adoption of multimodal data fusion, which involves imaging, genomic, and clinically oriented data, could help to improve the reliability and personalization of algorithms employed in breast cancer diagnosis [22]. These methods make use of different types of data and information to gain more comprehensive information about tumor heterogeneity, therapeutic responses, and patient outcomes, thus, allowing for personalized and more effective treatment options. As the landscape of breast cancer diagnosis continues to keep on evolving, the

security of data privacy and ethics are of particular importance now than ever before. The data governance framework, informed consent and a stricter adherence to regulatory standards are essential to protect patient privacy and ethical practices extended by AI in healthcare solutions. Through this interdisciplinary amalgamation of computational pathology, deep learning, and interpretability methodologies, the research aims to make significant strides in advancing the frontier of breast cancer diagnosis. By offering novel insights into the application of deep learning in histopathological analysis and elucidating the rationale behind model predictions, the proposed system aspires to empower healthcare practitioners with actionable intelligence, ultimately contributing to improved patient outcomes and the fight against breast cancer on a global scale.

In delving into the intricate interplay among Artificial Intelligence (AI), Explainable AI (XAI), and conventional diagnostic methods, a systematic examination seeks to illuminate the transformative potential of these groundbreaking technologies in reshaping the landscape of breast cancer detection. Embarking on this comprehensive journey, the investigation not only navigates the technical intricacies of these methodologies but also delves into their profound socio-ethical implications.

The aim is to forge a holistic understanding of how XAI can seamlessly integrate into clinical practice, thereby fostering a paradigm shift toward improved diagnostic precision and informed decision-making. Within this multifaceted exploration, the endeavor is to unravel the underlying decision-making processes of AI and XAI models, thereby shedding light on the mechanisms driving diagnostic accuracy. By elucidating the decision rationale of these models, XAI emerges not only as a tool for enhancing diagnostic precision but also as a vital aid for clinicians in comprehending and validating results. This transparency serves as a cornerstone, ensuring that the integration of AI in breast cancer detection remains firmly grounded in the principles of medical ethics.

Moreover, as this transformative landscape is navigated, the critical importance of fostering a collaborative and informed decision-making process between healthcare providers and intelligent algorithms is recognized. By instilling trust and confidence through transparency, the path is paved for a future where AI augments clinical expertise, ultimately leading to improved patient outcomes and a more effective response to the global challenge of breast cancer.

## 2 Literature review

The literature survey encompasses a diverse array of innovative approaches in utilizing artificial intelligence (AI) and machine learning for breast cancer diagnosis. BI-RADS-NET-V2 model for ultrasound images, achieving high accuracy and providing clinicians with both semantic and quantitative explanations was introduced, thereby enhancing trust in the diagnostic results [1]. An investigation of AI comprehensive explanation tools emphasizes the importance of interpretability in AI predictions, showcasing SHAP as a superior technique over LIME for providing trustworthy explanations in breast cancer ultrasound images [2]. Capsule networks emerge as a potent tool in breast cancer classification. Their ability to capture orientational and spatial knowledge, overcoming limitations associated with traditional convolutional neural networks (CNNs) [3]. One of the study conducted introduces CapsNetMMD, a deep learning technique employed for recognizing breast cancer-related genes through the integration of multi-omics data, demonstrating superior performance compared to other machine learning techniques [4]. The research by Peta and Koppu emphasizes the significance of leveraging machine learning and techniques, such as CNNs and capsule networks, for effective breast cancer prediction, with emphasis on comparing their performance and exploring avenues for future enhancements [5]. Tackling the security concerns associated with AI-based breast cancer diagnosis, ElGamal image encryption method that was broadened is a federated learning framework, enhancing overall security and performance through convolutional capsule twin attention tuna optimal network model [6]. In the domain of transfer learning, utilized ResNet, EfficientNet, and DenseNet versions to classify histopathology images of invasive ductal carcinoma, with Resnet101 demonstrating superior performance [7].

A novel approach for breast metastases detection, employing capsule networks with Baye's routing for classification, followed by patch-level segmentation, providing a comprehensive solution for identifying tumors in breast cancer images was introduced. [8]. Few-shot learning was explored for histopathology image classification, addressing challenges related to generalization by employing Prototypical networks and MAML on different datasets [9]. An AI screening model for breast cancer utilizing mammography results has been introduced, incorporating TabNet, node transition probability, and graph neural networks for improved accuracy and correlation of features [10]. An explainable AI method was used for breast cancer spread prediction, employing a CatBoost classifier with LIME to enhance trustworthiness and interpretability, addressing

issues of complexity and bias in results was introduced [11].

A comprehensible machine learning evaluation of long-term mental well-being outcomes following breast cancer detection, utilizing k-means clustering and an XGBoost model with SHAP values that are used for identifying crucial features influencing patient's mental health states was conducted [12]. Linear projections were employed, Radviz visualization techniques, and Decision Tree induction algorithms for building models to differentiate between malignant and benign breast cancers, emphasizing the interpretability of CART as a white-box method [13]. Explainable framework with machine learning for classifying multiple medical datasets, utilizing SHAP to analyze complex models and combining basic ML models for disease detection, demonstrating high accuracy and potential for early identification was introduced [14]. An EfficientNet-based Machine Learning Algorithm for Breast Cancer identification using Mammograms, showcasing the effectiveness of smart computer methods for accurate diagnosis was developed [15]. An AI-driven simulation for breast cancer treatment paths, employing Markov Decision Process, was developed to integrate physician guidelines. The model is trained to recommend sequential treatments, highlighting the importance of adapting to varying training performance [16]. A comparative study was conducted on the prediction of breast cancer utilizing optimized algorithms, introducing a divide and conquer kernel with SVM for accurate diagnosis, highlighting the superiority of a hybrid Radial Basis Function Neural Network in terms of performance [17].

A SVM-ANN optimized algorithm was developed for classifying breast cancer data into benign and malignant categories, demonstrating notable accuracy and holding potential for early detection. [18]. An evaluation of deep learning models was carried out on the BreakHis dataset, incorporating both up-sampling and down-sampling techniques to tackle the issue of imbalanced datasets. The study underscores the significance of balancing classes to enhance model performance. [19]. Several efforts were directed toward identifying breast cancer threat using XGBoost, highlighting the significance of considering risk factors such as family history and physical inactivity, with Random Forest demonstrating superior performance [20]. Collectively, these studies contribute to the evolving landscape of AI applications in breast cancer research, addressing issues of trust, interpretability, and effective diagnosis. A present study introduces a new method to detect breast cancer with the use of heat images taken by infrared camera, where an autoencoder that is driven by Gaussian pyramid for noise removal and quality enhancement. By combining deep learning with the denoising autoencoder using ensemble classifier DenseNet-201, this model has outperformed all previous models in terms of accuracy in detecting breast cancer. Also, an attention-guided Grad-CAM enhances transparency by highlighting salient regions in imagery, fostering understanding and collaboration between AI-driven diagnostics and clinical expertise [26].

Other strategies will advocate for an intelligible machine learning-based method of diagnosing lung cancer using a mixture of models such as Decision Trees, Logistic Regression, Random Forest and Naive Bayes classifier. The study has managed to achieve a high degree of success in early identification since it is able to establish Logistic Regression and Random Forest classifiers with an accuracy level of 97% on 'Lung Cancer Detection' dataset obtained from Kaggle. Additionally, the integration of SHAP and LIME XAI models enhances interpretability, facilitating insights into the predictive mechanisms of the employed models [27]. One approach proposed a deep learning-based Convolutional Neural Network (CNN) utilizing InceptionResNetV2 and InceptionV3 transfer learning models for the classification of colon cancer types: adenocarcinoma and benign cancer. Utilizing histopathological image datasets, the study aims to enhance cancer diagnosis accuracy using artificial intelligence techniques [28]. Integration of deep learning models into the Internet of Medical Things (IoMT) for accurate breast cancer screening is necessary. However, a critical limitation lies in the lack of interpretability of these models, hindering their adoption by caregivers.

To address this, a study proposes an end-to-end explainable AI framework for analyzing breast cancer prediction models using mammography data, facilitating better understanding and evaluation by healthcare professionals [29]. A study focuses on the development of AI-based models for early detection of esophageal malignancy, an important area of cancer research due to its high mortality rate. However, the opacity of complex AI models poses a challenge, necessitating the incorporation of Explainable AI (XAI) techniques for transparency and trustworthiness. By utilizing LIME, the study enhances the interpretability of deep learning models, achieving a notable accuracy of 88.75% with DenseNet-201 on actual endoscopic images [30].

## 3 Proposed methodology

With popular architectures like VGG-16, VGG-19, and ResNet, that offer an interpretable deep learning approach for breast cancer identification using Convolutional Neural Networks (CNNs) and transfer learning techniques, the proposed system starts by optimizing the previously trained VGG-16, VGG-19, and ResNet models using a sizable

dataset of histological pictures of breast cancer. The next step is data preparation, which involves scaling, augmentation, and normalization of the images. After that feature extraction is performed to obtain the most pertinent features from the pictures, which helps distinguish between healthy and unhealthy tissues. Then the retrieved features are used to train a cost-sensitive CNN classifier as shown below in Fig. 1. This classifier is trained to detect the presence of breast cancer while considering the imbalanced nature of medical datasets, where the occurrence of malignant cases is often significantly lower than benign cases. The approach employs techniques such as class weighting or undersampling to mitigate the impact of class imbalance.

Moving to the online stage, the trained CNN models are utilized for real-time breast cancer detection in new cases. Furthermore, Local Interpretable Model-agnostic Explanations (LIME), SHAP (SHapley Additive exPlanations) and salient maps are leveraged to explain the predictions made by the models. LIME generates local surrogate models around individual predictions, providing insight into how different regions of the input image contribute to the final decision. This facilitates better understanding and trust in the model's predictions, particularly in critical medical applications like breast cancer diagnosis. SHAP provides insights into the importance of features in breast histopathology datasets by quantifying the contribution of each feature to model predictions, aiding in the interpretation and understanding of the underlying mechanisms driving predictions. Salient maps for breast histopathology highlight regions of interest within tissue images, aiding in the identification and interpretation of key features associated with diagnostic characteristics (Figs. 2 and 3).

### 3.1 Dataset

This study utilizes the "Breast Histopathology Images" dataset, which is available through Kaggle, a renowned online platform for data science and machine learning enthusiasts. This dataset is an essential tool for improving image analysis in medicine, especially when it comes to diagnosing breast cancer. The dataset is made up of a sizable number of histological pictures taken from samples of breast tissue. These pictures show two different groups: the IDC + group, which has irregular cell structures, high cell density, and aberrant tissue patterns; and the no-cancer group, which has regular, healthy cell arrangements and normal tissue architecture as shown below in Fig. 4.

For medical practitioners, these images are essential because they provide comprehensive insights into the structure and makeup of breast tissues. The careful structuring of this dataset is one thing to observe, with images neatly categorized into IDC + and no-cancer groups. This labeling approach provides researchers and practitioners with a structured framework for training and evaluating machine learning algorithms, facilitating efforts to achieve increased accuracy in breast cancer detection and classification. By ensuring that images are neatly categorized into these groups, researchers and practitioners can streamline their efforts in developing and refining machine learning algorithms for breast cancer detection and classification. This structured approach facilitates the implementation of robust evaluation methodologies, enabling the assessment of algorithm performance and the identification of areas for improvement.

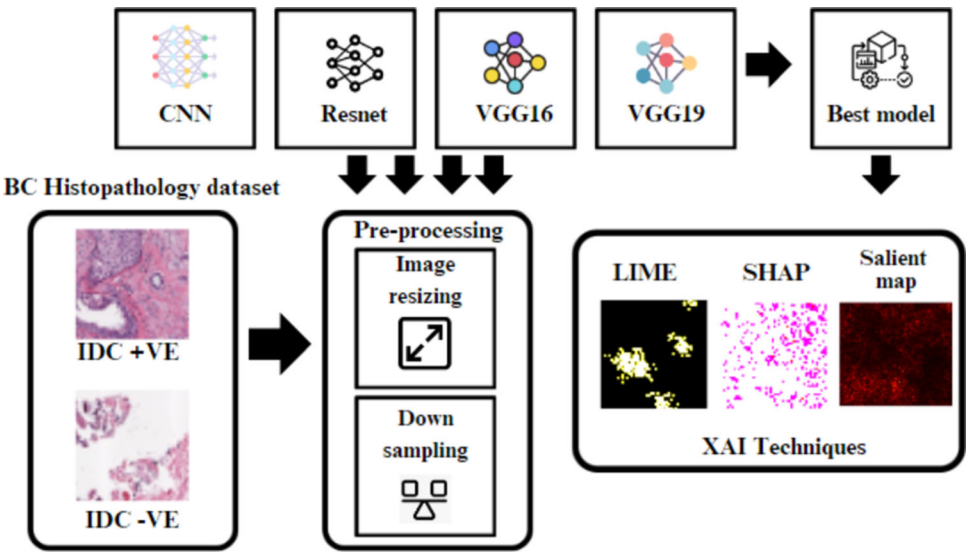


**Fig. 1** Architecture of the proposed modeling framework for breast pathological classification

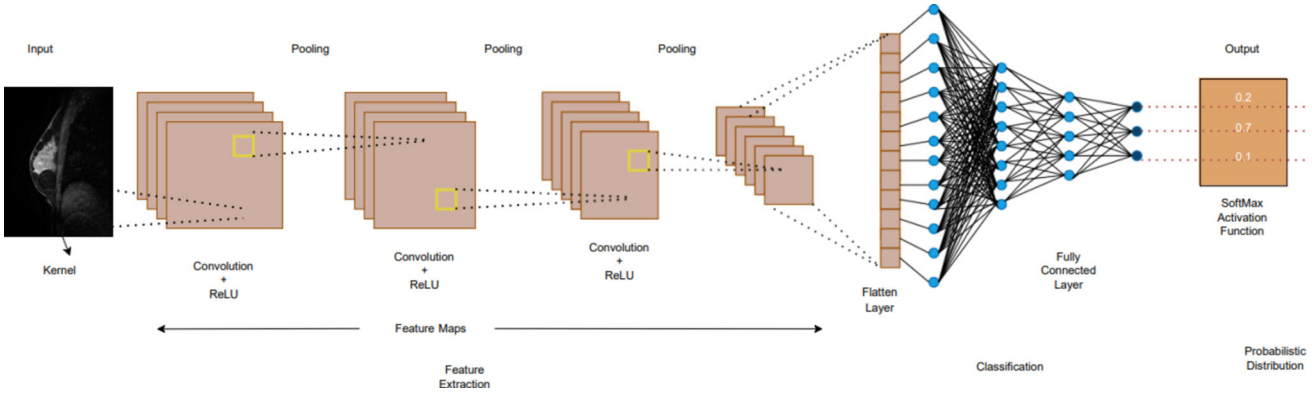


**Fig. 2** Convolutional neural network architecture diagram

### 3.2 Experimental setup

The experimental configuration makes use of the free edition of Google Colaboratory, a well-known cloud-based platform that provides free access to GPU and TPU resources required for efficient machine learning model training. This decision granted us access to powerful computational resources without incurring extra expenses, thus facilitating the training of complex models without being constrained by hardware limitations. To ensure the reproducibility of our experiments, we provide the following setup details. The experiments were executed in Google Colab, utilizing a virtual environment with access to GPUs. The software environment included Python 3.8 and libraries such as TensorFlow 2.8.0, Keras 2.8.0, and scikit-learn 1.0.2, along with other essential packages. For reproducibility, we set the random seed to 42 using NumPy and TensorFlow to ensure consistent results. The dataset used was sourced from Kaggle, and preprocessing steps such as normalization and augmentation were uniformly applied. To implement the popular models, valued for their versatility and reliability in building machine learning solutions existing frameworks were leveraged. These frameworks provide a comprehensive array of tools and functionalities suitable for various types of machine learning models, ranging from neural networks to traditional statistical models. Throughout the implementation process, Python was utilized as the primary programming language, esteemed for its simplicity, readability, and extensive libraries, offering an optimal environment for developing the machine learning solutions.

### 3.3 Preprocessing

Two commonly employed techniques in the preprocessing phase for breast cancer detection are undersampling and image scaling. The class imbalance that frequently exists in medical datasets—where benign cases may outweigh malignant cases—is addressed by undersampling. The class distribution before undersampling is shown in Figs. 5 and 6. below. Undersampling guarantees a balanced distribution between the classes, preventing bias toward the dominant class and facilitating effective learning from both classes. It does this by randomly choosing a subset of instances from the majority class (benign cases).

Image resizing, on the other hand, uniformizes input image dimensions to a set size appropriate for the input layer of the CNN model. Medical images can change in size as a result of different acquisition modes or equipment settings, which makes standardization essential. Resizing the images allows for quick processing and feature extraction by ensuring consistency and compatibility with the CNN model. Transfer learning for breast cancer diagnosis performs better and is more effective when undersampling and scaling images work together to enhance the preprocessing pipeline.

### 3.4 Convolutional neural network

The CNN architecture functions by sequentially processing breast histopathological images to detect signs of cancer. Initially, as shown above in Fig. 2, convolutional layers extract features from the input images, capturing patterns indicative of malignant or benign tissue characteristics. Batch normalization enhances training stability by normalizing layer activations, while max pooling reduces spatial dimensions, making the network robust to variations. Dropout layers mitigate overfitting by randomly deactivating neurons during training. Following feature extraction, dense layers learn complex relationships in the data, aiding in classification. Finally, the output layer predicts the likelihood of cancer presence based on learned features. This architecture aims to accurately classify breast tissue as either malignant or benign, facilitating early identification and treatment of breast cancer. The following equations are fundamental to the Convolutional

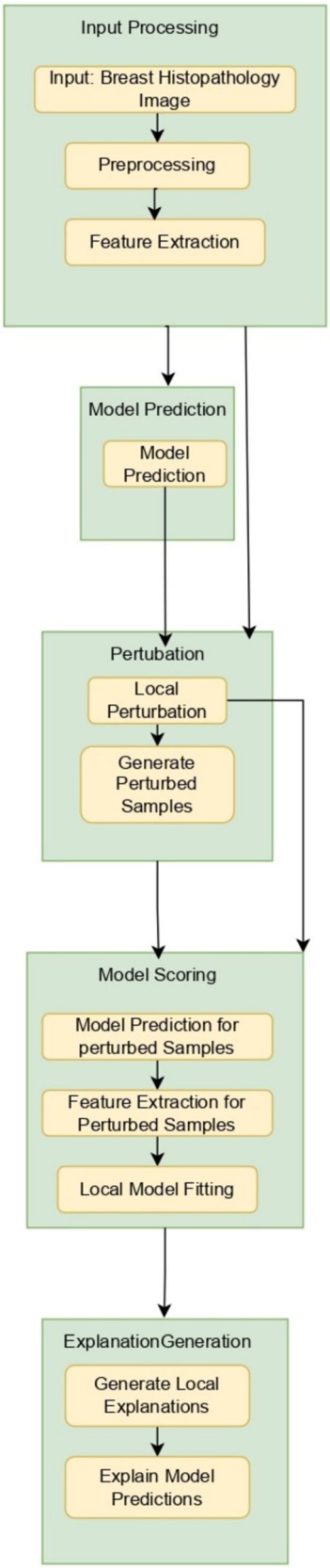


**Fig. 3** LIME interconnected framework for diagnosis

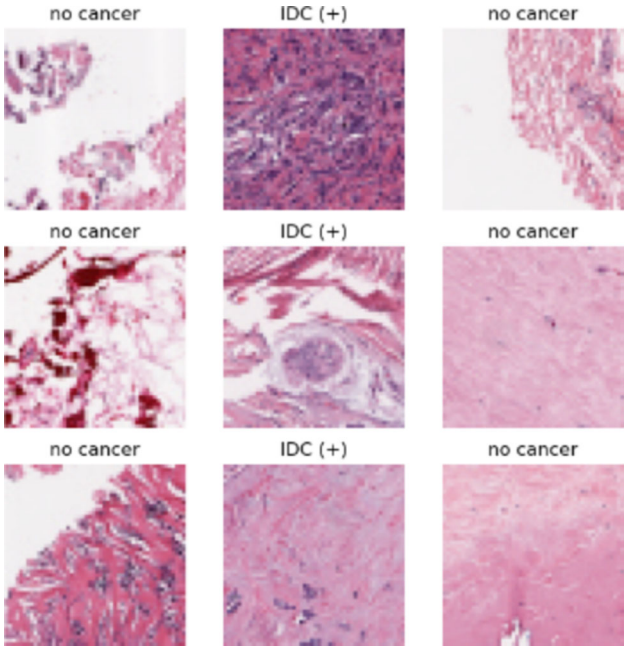


**Fig. 4** Dataset of breast histopathology

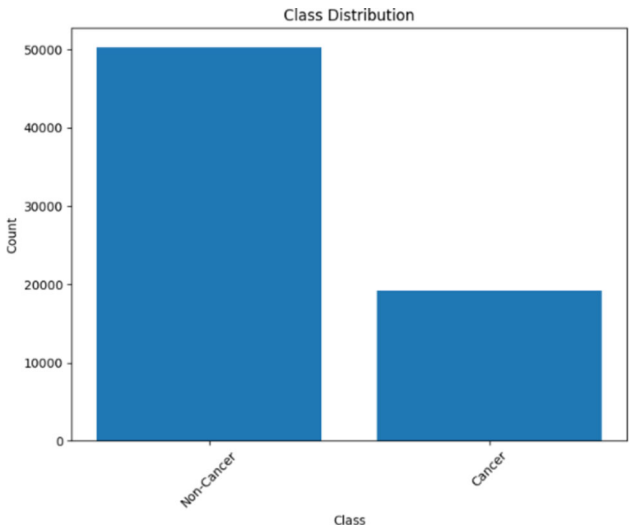


**Fig. 5** Class distribution bar plot for cancerous and non-cancerous samples

Neural Network (CNN) architecture and have been utilized in this study for various computational operations and model components:

### 3.4.1 Convolution operation

$$Output(x,y) = \sum_{i=0}^{f_h}\sum_{j=0}^{f_w}\sum_{k=0}^{C_{in}} (Input(x+l, y+m, n) \times Filter(l,m,n)) + b \quad (1)$$

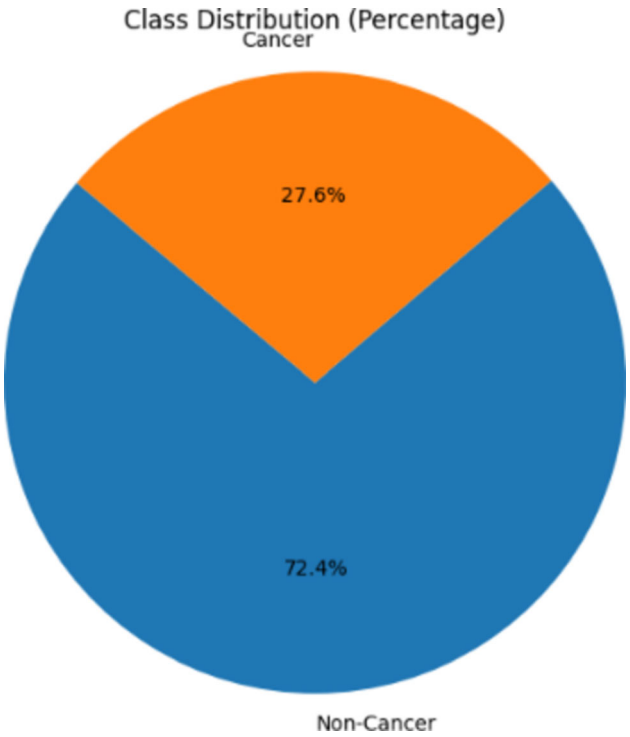


**Fig. 6** Class distribution percentage for cancerous and non-cancerous samples

where Input(x,y,k) is the input value at position (x,y) of the nth channel.

The weight of the convolutional filter at position (l, m) in the nth channel is denoted as Filter(l, m, n).

b is the bias term. $f_h$ and $f_w$ are the dimensions representing the height and width of the filter.

$C_{in}$ is the indicative total number of input channels.

#### 3.4.2 Activation function (ReLU)

$$ReLU(x) = max(0, x) \tag{2}$$

#### 3.4.3 Pooling operation (max pooling)

$$Output(x, y, k) = max\big(Input\big(x \cdot s_x + i, y \cdot s_y + j, k\big)\big)$$

$s_x$ and $s_y$ are the stride along the x and y axes, respectively.

#### 3.4.4 Fully connected layer

$$Output = Activation\left(\sum_{i=1}^{n} (Weight_i \times Input_i) + Bias\right) \tag{4}$$

$Weight_i$ is the weight assigned to the ith input in the fully linked layer. $Input_i$ is the $i^{th}$ input in the fully connected layer. Bias is the bias term in the fully connected layer.

#### 3.4.5 Softmax function

$$\mathrm{P(y = jx)} = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{5}$$

$z_j$ is the input to the softmax function for class j. P(y = j|x) is the likelihood of class j given input x.

### 3.5 Transfer learning techniques

Using pre-trained CNNs that have been trained on enormously large image datasets is the basis for transfer learning with models such as ResNet, VGG-16 and VGG-19.. The trained convolutional layers of the pre-trained models are frozen and they are first used as feature extractors. After that, they are refined using the breast cancer histopathology dataset, with their weights significantly altered to better suit the particular goal of detecting breast cancer. The models can acquire task-specific characteristics necessary for differentiating between benign and malignant breast tissue through this process of fine-tuning. To categorize the breast cancer photos, more layers—usually completely linked layers—are put on top of the pre-trained layers after fine-tuning. Through the use of optimization and backpropagation techniques, the model gains the ability to map the extracted features to the appropriate classes during training. The best-performing model is chosen for future investigation or deployment in real-world scenarios based on these evaluation metrics. Transfer learning using pre-trained models is a useful method for breast cancer detection tasks since it has benefits including improved generalization, faster convergence, and the capacity to learn from small amounts of data.

### 3.6 Model evaluation

A confusion matrix provides a comprehensive overview of the classification outcomes, encompassing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Such visualization facilitates a detailed examination of the model's ability to distinguish between different classes. By analyzing the distribution of predictions across the matrix, valuable insights into accuracy, sensitivity, specificity, and overall performance are gleaned. Consequently, the model achieving the highest accuracy was selected for further analysis and deployment, highlighting its superior predictive capabilities. This

meticulous evaluation process ensures the identification of a robust and dependable model for breast cancer detection.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total number of Predictions}} \times 100\% \quad (6)$$

$$\text{Precision} = \frac{\textit{True positives}}{\textit{True positives} + \textit{False positives}} \quad (7)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (8)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

### 3.7 ResNet-50

The Residual Network (ResNet), introduced by He et al., emerged as the most effective and reliable neural network during the annual ILSVRC 2015 competition. ResNet-50, a specific variant of ResNet, showcased notable advantages such as faster convergence and improved classification accuracy. It achieved this by utilizing identity shortcuts, where the output values mimic the input values, enabling training on 224 × 224-pixel color images. The ResNet architecture is constructed by stacking various different residual units, and its configurations can vary based on the number of total residual units and layers employed. The ResNet-50 architecture used in this study comprised 49 convolutional layers along with a fully connected layer. While ResNet includes fully connected layers and convolutional pooling, similar to other pretrained networks like VGG, it is important to note that ResNet is significantly deeper. In fact, ResNet is deeper than VGG-16 by 8 times hence results in a larger number of features that can be learned and, consequently a higher potential for improved classification accuracy.

### 3.8 VGG-19

VGG-19 is an architecture introduced in 2014 as an extension of VGG-16 by the Visual Geometry Group (VGG) at the University of Oxford. It was developed to attain higher accuracy in various classification tasks. The key feature of VGG-19 is its simplicity and uniformity. Throughout the network, 3 × 3 filters are used in the convolutional layers, enabling localized feature representation. The max pooling layers downsample the spatial dimensions, enabling the capture of increasingly higher-level and abstract features. With 16 convolutional layers and 3 fully connected layers, VGG-19 has a total of 19 layers. Its deeper layout allows It can learn more complicated patterns in input data, but this raises processing and memory requirements. VGG-19 has been widely used and has shown impressive performance on benchmark datasets like ImageNet. In 2014, it was named the winner of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Its success can be attributed to the layered convolutional structure, which enables detailed understanding of intricate information. While VGG-19 is primarily utilized for image classification tasks, its architecture has also been applied to other computer vision problems, including object recognition.

### 3.9 VGG-16

VGG-16, which is also referred to as Visual Geometry Group-16, was developed by Karen Simonyan and Andrew Zisserman for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014, where it achieved the highest performance among all competing networks. The configuration of the convolutional layers is specifically designed to excel in image classification tasks. In VGG-16, convolution operations employ 3 × 3 filters. These choices significantly influence the resulting output image at each convolution layer. A dropout layer with a rectified linear unit (ReLU) activation function with a dropout ratio of about 0.5 follow each convolutional layer to address overfitting. ReLU introduces nonlinearity to the model, while dropout helps prevent excessive reliance on specific features. During the convolution operation, the filter scans the input image, performing mathematical operations and producing an output image. This process involves shifting the filter by a certain number of pixels determined by the stride, resulting in the output image The VGG-16 model's architecture and design choices have played a crucial role in its exceptional performance in image classification tasks, establishing it as a highly influential deep neural network. Max pooling is another important component of VGG-16, contributing to breaking down and resizing input images. This process enhances the extraction of important features while optimizing memory usage. The early layers of VGG-16 primarily learn basic image; while, the deeper layers are allocated to features such as edges to capture more intricate and complex image features.

### 3.10 LIME

LIME (Local Interpretable Model-agnostic Explanations) offers a methodical approach for interpreting breast cancer histopathology images. Specifically tailored for the intricate nuances of histopathological analysis, LIME operates by first selecting a data instance, typically corresponding to an individual histopathology image, from the breast histopathology dataset. It then generates perturbations or variations around this selected instance, introducing subtle changes or noise while preserving clinically relevant features. These perturbed images are put into the machine

learning model to see the related predictions, allowing LIME to study how the model performs locally near the selected data instance[23]. Leveraging this information, LIME constructs an interpretable model, often a linear regression model that approximates the behavior of the original black-box model in this local region. This interpretable model captures the relationship between the input features (histopathology image pixels) and the model's predictions. LIME determines the contribution of each pixel to the model's prediction for the selected data instance by assigning relevance scores to input features based on the interpretable model's coefficients.

Through visualizations of these importance scores, clinicians can discern the specific histological features driving the model's decision-making process for individual histopathology images.

This localized interpretability empowers clinicians with actionable insights into the factors influencing the model's predictions, guiding validation and enhancing understanding of the diagnostic process in breast cancer histopathology analysis. LIME requires model-specific information to evaluate a model's local fidelity. Local fidelity quantifies how well a model captures the features surrounding a specific prediction. In every suggested model f:Rd → R, where f(x) represents the likelihood of class x, the locality surrounding x was defined by utilizing πx(z) as a measure of proximity between instances.

$$\xi(\mathrm{x}) = argmax_{g\in \mathrm{G}} \mathrm{L}(\mathrm{f},\ \mathrm{g}, \pi_x) + \Omega(\mathrm{g}) \tag{10}$$

where the degree to which g misrepresented f in the locality denoted by πx was measured using the fidelity function $\mathrm{L}(\mathrm{f},\ \mathrm{g}, \pi_x)$.

In order to maximize the number of interpretations, Ω(g) was minimized, and the fidelity function was also reduced. The workflow of LIME is illustrated in Fig. 3. above.

### 3.11 SHAP

SHAP (SHapley Additive exPlanations) stands as a powerful tool in the realm of breast cancer histopathology analysis, offering a robust method for interpreting machine learning models [24]. Its comprehensive approach provides invaluable insights into the myriad factors influencing the model's predictions, guiding clinicians toward a deeper understanding of the underlying features that signify breast cancer. By dissecting the contribution of individual features within the histopathology images, SHAP unravels the intricate web of histopathological characteristics guiding the model's decision-making process. This interpretability is paramount for clinicians, allowing them to validate the model's decisions and glean actionable insights into the diagnostic process.

Through intuitive visualizations such as summary plots or force plots, SHAP sheds light on the specific regions within the histopathology images that wield the greatest significance for the model's prediction. By pinpointing these key morphological patterns associated with malignancy, SHAP facilitates the identification of crucial diagnostic markers. Leveraging SHAP explanations empowers researchers to derive clinical insights that transcend traditional diagnostic methodologies. These insights inform not only diagnosis but also prognosis and treatment planning for breast cancer patients, offering a holistic approach to patient care. In essence, SHAP serves as a cornerstone in the quest for enhanced diagnostic accuracy and clinical understanding in breast cancer histopathology analysis, bridging the gap between machine learning algorithms and real-world clinical applications.

### 3.12 Saliency map

Salient maps, a visualization technique commonly employed in deep learning model interpretation, offer valuable insights into the decision-making process of models applied to breast histopathology image analysis. These maps work by highlighting regions within the histopathology images that significantly influence the model's prediction, aiding in the identification of morphological features crucial for classification, such as malignant versus benign tissue. The process typically involves computing gradients of the model's output concerning the pixels of the input image, effectively quantifying the contribution of each pixel to the final prediction [25]. These gradients are then normalized to emphasize regions of higher importance relative to others. By overlaying the normalized salient map onto the original image, researchers can visually discern which areas the model prioritizes when making its decision. This interpretative tool facilitates the identification of histological patterns indicative of cancerous growth, such as irregular cell structures or dense nuclei. Validation of salient maps involves comparison with expert annotations or ground truth labels to ensure alignment with known pathological features associated with breast cancer. Iterative refinement of both the model and salient map generation process further enhances their utility in clinical settings, fostering trust, transparency, and improved decision-making. The saliency map M for a given class c is obtained by averaging the gradients across the feature maps:

$$M_c^l = \frac{1}{z}\sum_i \sum_j \frac{\delta Y_e}{\delta A_{ij}^l} \tag{11}$$

where $M_c^l$ is the saliency map for class c at the l-th convolutional layer.

$Y_c$ is the model's output score for class. $A_{ij}^{l}$ represents the activation of the i-th row and j-th column of the feature map of the l-th convolutional layer.

Z is a normalization term to scale the gradients appropriately.

## 4 Experiment and result

The evaluation of various pre-trained convolutional neural network (CNN) models on a given dataset was conducted with meticulous consideration of diverse hyperparameters and performance metrics. Diving into the intricacies of model selection and training configurations, the examined models encompassed a comprehensive models including CNN, VGG-16, ResNet, and VGG-19. A key method for guaranteeing sufficient representation of various subgroups within a dataset is stratified sampling. To apply this method to a breast histopathology dataset, one must first define the population and its subgroups. All histopathological pictures make up the population, and class labels—such as IDC + (malignant) and no-cancer (benign)—are used to designate the subgroups (strata). After that, the dataset is split up into these strata, making sure that every stratum only includes pictures from a single class. Partitioning the dataset into training, validation, and test sets yields the sample size for each stratum. To ensure representative samples, photos from each stratum are then chosen for the corresponding sets using random sampling. This method guarantees that each class is proportionately represented in the test, validation, and training sets. This is crucial for imbalanced datasets, such as those used in medical research. By guaranteeing sufficient representation of minority classes, stratified sampling lowers bias and increases the model's generalizability while also boosting the model's robustness and dependability.

Facilitating a robust assessment framework, the dataset underwent a division into distinct training and testing subsets, having a ratio of 70% for training and 30% for testing, ensuring allocation of data for model training and validation. For the optimization, the Adam optimizer, known for its adaptive learning rate methodology, was uniformly employed across all models. Emphasizing the significance of batchwise data processing, varying batch sizes ranging from 30 to 40 samples were employed, with an optimal balance between computational efficiency and gradient precision during the training.

In the training paradigm, a critical aspect was the iteration count, encapsulated by the number of epochs. This training metric spanned a range from 35 epochs to a more extensive 500 epochs, reflecting the diverse complexities inherent in the considered models. Notably, ResNet emerged as the frontrunner, achieving an overall accuracy of 95.40% on the test dataset, following a rigorous training spanning 500 epochs with a batch size of 40. This extensive training period highlights the model's ability to learn intricate patterns and generalize well to hidden data. Meanwhile, the CNN model showcased an impressive accuracy of 94.78% after only 35 epochs, underscoring its efficiency in capturing important features within the dataset as shown in Figs. 7 and 8. This rapid convergence shows the effectiveness of the model architecture in extracting relevant information from the input data.

Similarly, VGG-16, although requiring an extensive training period of 500 epochs like ResNet, demonstrated commendable performance with an accuracy of 94.07%. This underscores the robustness of its architectural design in tackling complex image classification tasks and its ability to achieve competitive performance even with prolonged training periods.

In a slightly contrasting change, VGG-19 exhibited a respectable accuracy of 92.59% after 500 epochs, despite utilizing a smaller batch size of 30. This delineates the subtle trade-offs between batch size and convergence dynamics, with VGG-19 showcasing a balance between training efficiency and performance accuracy. Overall, the experimental results underscore the diverse capabilities and characteristics of the considered models, providing valuable insights into their training behaviors and performance outcomes (Tables 1 and 2).

A thorough investigation of pre-trained Convolutional Neural Network (CNN) models illuminates the intricate interplay between their architecture, training configuration, and dataset characteristics (Figs. 9, 10, 11, 12, 13, 14, 15, 16 and 17).

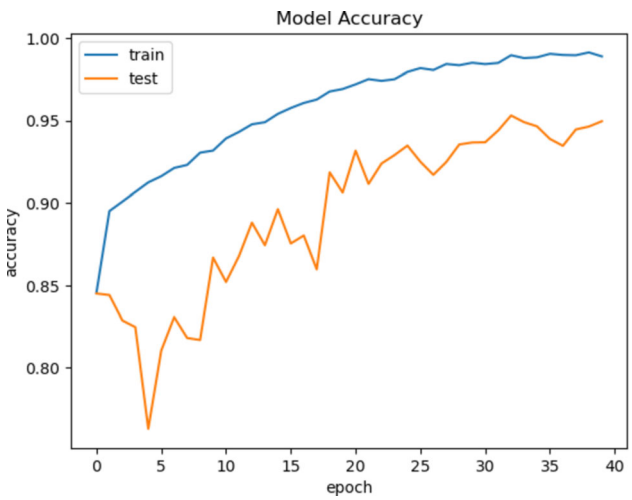


**Fig. 7** Precision Evaluation of the CNN Model for breast cancer diagnosis

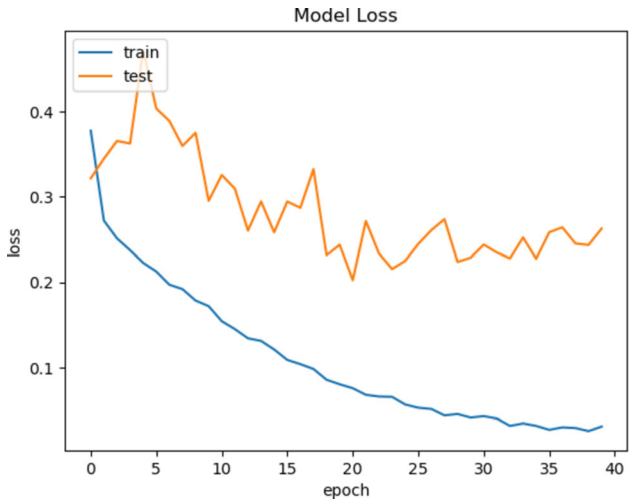


**Fig. 8** Model loss for the CNN model for breast cancer diagnosis

When comparing the performance of the CNN, VGG-16, ResNet-50, and VGG-19 models across various metrics, it is clear that each model has distinct strengths. ResNet-50 emerges as the most well-rounded model, with the highest accuracy at 95.40% and the best recall at 0.8356, making it ideal for applications where identifying true positives is crucial. It also maintains strong specificity and precision, indicating it balances overall performance well. On the other hand, VGG-16 excels in specificity (0.9889) and precision (0.9146), making it the best choice when avoiding false positives is critical, although it has a lower recall, meaning it might miss more true positives compared to ResNet-50. CNN offers a solid performance across all metrics with an accuracy of 94.78%, but it does not lead in any specific area, making it a good all-around option, especially for balanced datasets. VGG-19, while having the lowest accuracy at 92.59%, still shows strong specificity (0.9771), which could be advantageous in scenarios where minimizing false positives is a priority.

Furthermore, it emphasizes the ongoing efforts to refine and optimize performance within the domain of artificial intelligence and machine learning. Following meticulous evaluation of multiple CNN architectures, including ResNet, VGG-16, VGG-19, and others, on breast histopathology images, ResNet emerged as the superior

**Table 2** Comparative performance analysis of CNN vs VGG-16 vs ResNet-50 vs VGG-19

| Pretrained model | Specificity | Precision | Recall | Accuracy |
|---|---|---|---|---|
| CNN | 0.9680 | 0.8204 | 0.8008 | 94.78 |
| VGG-16 | 0.9889 | 0.9146 | 0.7610 | 94.07 |
| ResNet-50 | 0.9854 | 0.9065 | 0.8356 | 95.40 |
| VGG-19 | 0.9771 | 0.8557 | 0.7948 | 92.59 |

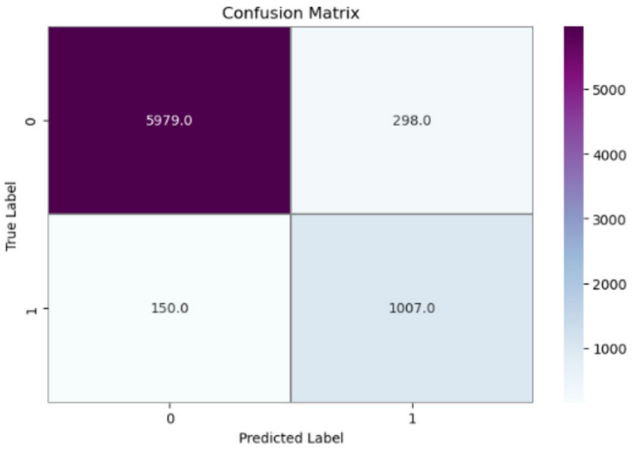


**Fig. 9** Confusion matrix of the CNN model on breast histopathology images

model. The prominence of ResNet underscores its resilience and suitability for the given task, reaffirming its potential to advance diagnostic capabilities in the medical domain.

To get a glimpse into the method of making decisions of the ResNet model, the LIME technique was utilized. LIME produces localized interpretations for individual predictions by altering the input data and observing how it impacts the model's output as shown below in Fig. 18. By doing so, it helps us understand which features or regions of the input images contribute the most to the model's predictions.

The table presents a comparative analysis of the performance metrics for different pre-trained convolutional

**Table 1** Comparative performance analysis of CNN vs VGG-16 vs ResNet-50 vs VGG-19

| Pretrained model | Train-test split ratio | Optimizer | Initial learning rate | Batch size | Number of epochs | Accuracy |
|---|---|---|---|---|---|---|
| CNN | 0.70:0.30 | ADAM | 0.0001 | 35 | 40 | 94.78 |
| VGG-16 | 0.70:0.30 | ADAM | 0.0001 | 500 | 40 | 94.07 |
| ResNet-50 | 0.70:0.30 | ADAM | 0.0001 | 500 | 40 | 95.40 |
| VGG-19 | 0.70:0.30 | ADAM | 0.0001 | 500 | 30 | 92.59 |

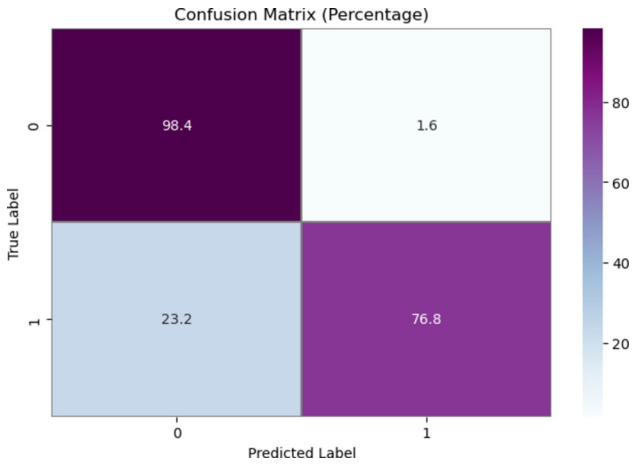


**Fig. 10** Confusion matrix percentage of the CNN model on breast histopathology images

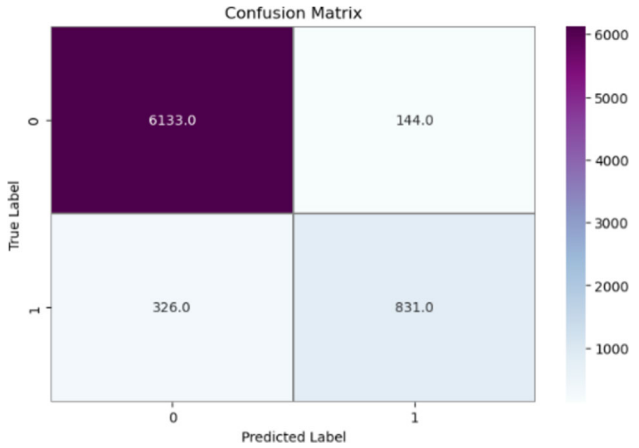


**Fig. 11** Confusion matrix of the VGG-16 model on breast histopathology images

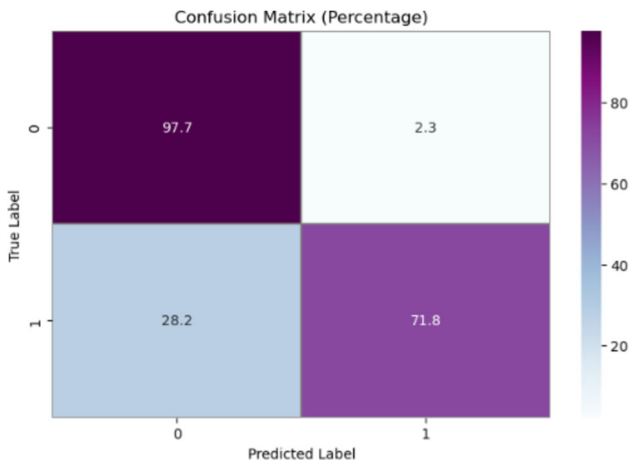


**Fig. 12** Confusion matrix percentage of the VGG-16 model on breast histopathology images

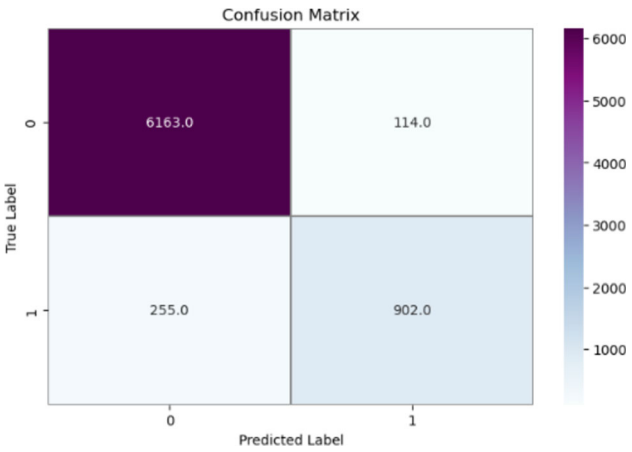


**Fig. 13** Confusion matrix of the ResNet-50 model on breast histopathology images

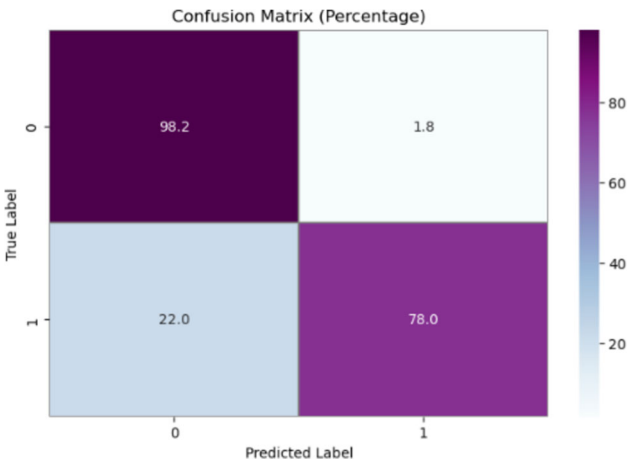


**Fig. 14** Confusion matrix percentage of the ResNet-50 model on breast histopathology images

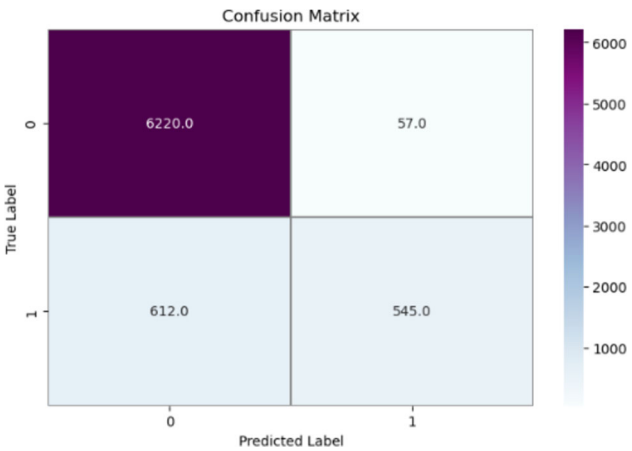


**Fig. 15** Confusion matrix of the VGG-19 model on breast histopathology images

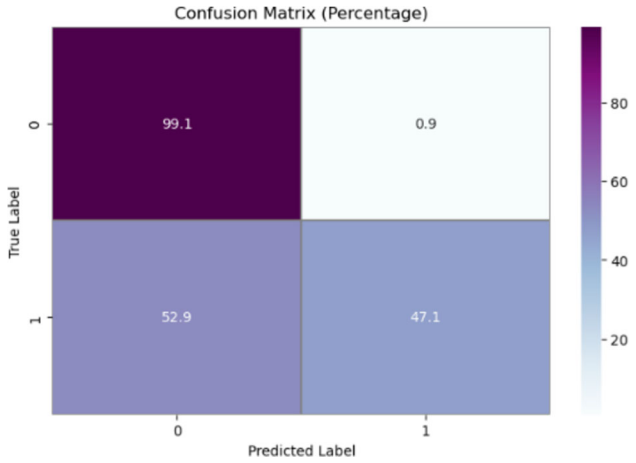


**Fig. 16** Confusion matrix percentage of the VGG-19 model on breast histopathology images

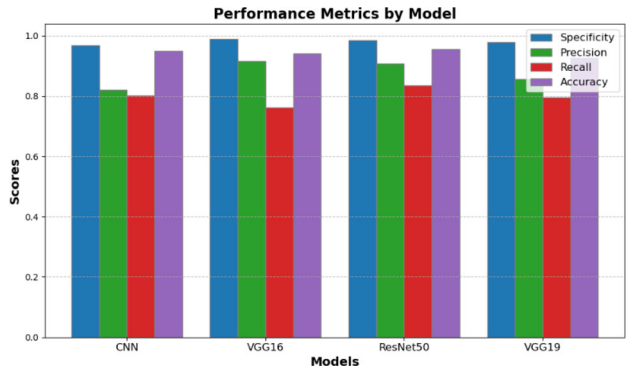


**Fig. 17** Comparative analysis of classification accuracy: visualizing performance metrics with accuracy graphs

neural network (CNN) models—CNN, VGG-16, ResNet-50, and VGG-19—applied to a specific task.

These metrics include specificity, precision, recall, and accuracy, each offering insights into the models' effectiveness in distinguishing between cancerous and non-cancerous samples. Specificity measures the model's ability to accurately identify non-cancerous cases; while, precision evaluates its precision in identifying cancerous cases. Recall assesses the model's capability to capture all positive cases accurately. Finally, accuracy provides an overall indication of the models' correctness in their predictions. Among the models, ResNet-50 demonstrates the highest accuracy of 95.40%, indicating its superior performance in correctly classifying both cancerous and non-cancerous samples. Additionally, it exhibits high specificity and recall, suggesting fewer false positives and negatives. VGG-16 and VGG-19 also display competitive performance but with slightly lower accuracy compared to ResNet-50. This comprehensive evaluation aids in understanding the strengths and weaknesses of each model, facilitating informed decision-making in selecting the most suitable model for the given task.

The original image patch is extracted from a histopathological slide stained with H&E (Hematoxylin and Eosin), aimed at highlighting tissue structure and cell morphology. The image appears pixelated due to digital zooming, a common practice in digital pathology for detailed examination. For the Positive LIME Explanation, upon applying LIME to the ResNet model, this panel showcases regions highlighted in yellow. These regions positively contribute to the model's prediction, likely indicating features associated with cancerous tissue. The model identifies these areas as significant for distinguishing between different classes of breast histopathology images. In the Complete LIME Explanation, the panel presents a comprehensive view of features identified by the LIME algorithm as relevant for the model's prediction. Along with the positive contributions highlighted in yellow, it includes all features considered important by the model, regardless of their impact on the prediction. The yellow outlines may represent boundaries of features that the model deems crucial for accurate classification.

In addition to the bright red areas, the saliency map also showcases regions in black and yellow, each signifying distinct characteristics in the model's process of making decisions. While the red regions denote significant influences on the model's output, the black areas suggest regions of lesser importance, possibly representing background or irrelevant features within the breast tissue as shown above in Fig. 19. On the other hand, the yellow regions highlight additional features that add toward the model's predictions, albeit with less influence compared to the red regions.

By incorporating these distinct color-coded regions, the saliency map offers a comprehensive and visually informative representation of the features and their varying degrees of relevance in guiding the model's classifications of breast histopathology images. Expanding the size of the saliency map enhances its clarity and facilitates a more detailed examination of the features contributing to the model's process of making decisions.

In the SHAP legend for breast histopathology analysis, the color spectrum ranges from blue to white to red, each representing different levels of influence on the model's predictions as shown above in Fig. 20.

Areas shaded in blue indicate features could have an adverse effect on the model's results. In the context of breast histopathology, blue regions may correspond to histological elements or patterns that are associated with benign tissue characteristics. These features are less likely to contribute to the model's classification of malignant or abnormal tissue.

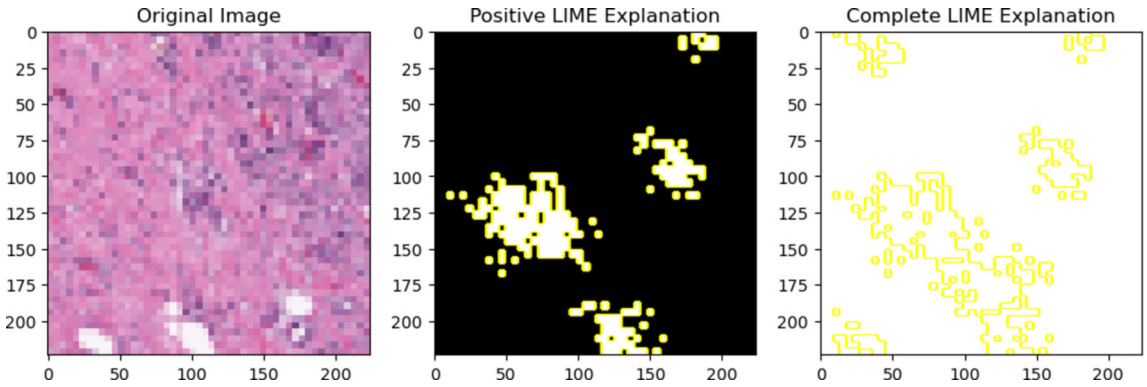


**Fig. 18** LIME explanation for breast histopathology dataset

Neutral or white regions suggest features that have minimal influence on the model's predictions. These areas may represent background elements or noise within the histopathological images that are neither strongly indicative of benign nor malignant tissue characteristics. The model's decisions are less affected by these neutral features.

Regions shaded in red signify features that exert a positive influence on the model's output. In the context of breast histopathology analysis, red areas likely highlight histological structures or patterns associated with malignant or abnormal tissue. These features are considered crucial by the model in distinguishing between benign and malignant tissue types or identifying specific histopathological patterns indicative of disease.

By interpreting the SHAP legend in this way, researchers can acquire knowledge into the varying significance of different histological features in the process of making decisions by the model for breast histopathology classification tasks.

In comparing deep learning models with traditional machine learning methods for your project, several key factors emerge. Deep learning models like ResNet-50 and VGG-16 often outperform classical algorithms such as SVM and Random Forest in handling high-dimensional data like breast histopathology images, thanks to their ability to automatically extract features. However, traditional methods, which rely on manual feature engineering, are typically more interpretable since the features are easier to understand. While deep learning models are often viewed as "black boxes," interpretability techniques like saliency mapping, LIME, and SHAP are improving the understanding of their decision-making processes. On the downside, deep learning models are computationally intensive and require longer training times; whereas, traditional algorithms are faster and less resource-demanding, making them suitable for smaller datasets or limited computational resources.

To assess the reliability of the model's predictions, the confidence intervals were calculated for the classification accuracy across different architectures. VGG-16 showed a 95% confidence interval of (80.5%, 88.3%); while, VGG-19 had an interval of (82.1%, 90.0%). In contrast, ResNet101 exhibited the highest performance with a confidence interval of (86.7%, 93.2%), demonstrating its statistical robustness and superiority among the evaluated models. By incorporating methods such as LIME, SHAP, and saliency maps, we further validated the model's predictions, reinforcing our findings regarding ResNet101 as the optimal architecture for breast cancer histopathology analysis.

LIME, Saliency Maps, and SHAP are three widely used methods for explaining machine learning models, each with its own strengths and weaknesses. LIME is model-agnostic, offering interpretable, local explanations by approximating the model around a specific prediction, but it can be computationally expensive and may not fully capture the model's true behavior. Saliency Maps, on the other hand, provide fast, visually intuitive explanations by highlighting influential regions in an image. However, they may suffer from gradient saturation and focus on irrelevant details, making interpretation difficult for non-experts. SHAP, grounded in cooperative game theory, fairly attributes each feature's contribution to the model, offering both local and global insights. Despite its strong theoretical foundation, SHAP can be computationally slow and complex to interpret, especially for large datasets or models. Each method is valuable depending on the use case, model type, and available computational resources.

## 5 Conclusion

After very careful analysis of several pre-trained models (ResNet-50, VGG-16 and VGG-19) on breast histopathology images, invaluable insights have been obtained about the subtle interplay amid model architecture as well as

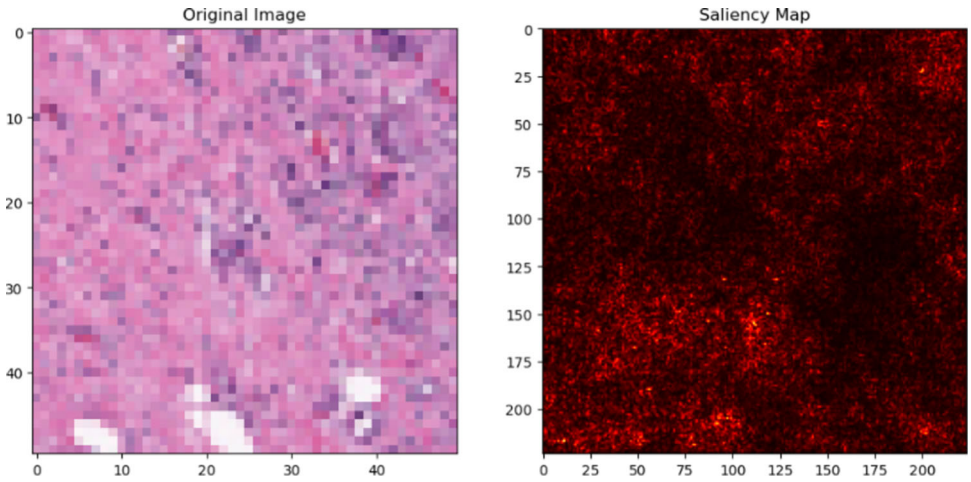


**Fig. 19** Saliency map explanation for breast histopathology dataset

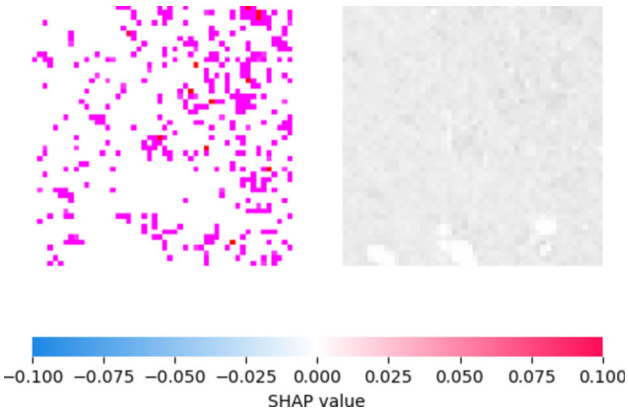


**Fig. 20** SHAP explanation for breast histopathology dataset

training parameters and dataset attributes. Thorough investigation of different hyperparameters and performance metrics showed that ResNet-50 was the best performer with excellent accuracy and consistency in image classification tasks. The evaluation went beyond mere pursuit of accuracy to become an extensive study into the complexities within deep learning, explaining what each architectural variant has for its strong points and limitations.

Moreover, this research expanded into interpretability that employed advanced techniques including saliency mapping LIME and SHAP to expose intricate decision-making processes of ResNet model. These interpretive methods were instrumental in identifying areas involved in classifying accurate breast histopathology images. With the ever-increasing need for innovation and exploration, challenging the boundaries of knowledge is paramount while leveraging AI's transformative power to reshape healthcare for our posterity. In conclusion, striving for perfection in AI. Pursuing, interpretability appears as a key issue in the understanding of deep learning models by penetrating their "black box" and thereby finding out why they actually make those decisions. These methods are useful in explaining how input features relate to model predictions for example saliency mapping, LIME, and SHAP. Their use can offer more insight into how these systems work thus leading to better optimization strategies aimed at improving their efficiency and reliability. Moreover, interpretability is not only critical for model performance but also fosters trust and transparency in AI-driven solutions that are crucial for their acceptance and application in real-life contexts. Easing the complexity associated with decision-making process by AI algorithms instills confidence among stakeholders such as healthcare providers, patients or even regulatory entities. This facilitates the responsible integration of AI technologies into clinical practice which may revolutionize healthcare delivery and improve patient outcomes.

In navigating the intricate landscape of AI and machine learning in healthcare, experts assert that upholding principles of ethics, fairness, and accountability is imperative. They emphasize that the conscientious development and deployment of AI-driven solutions necessitate careful consideration of potential biases, unintended consequences, and ethical implications. Prioritizing these considerations, they argue, can mitigate risks and ensure that AI technologies serve the best interests of individuals and communities. In conclusion, experts suggest that the intersection of AI, interpretability, and healthcare presents a fertile ground for innovation and advancement. By deepening comprehension of AI algorithms, enhancing their interpretability, and upholding ethical standards, they believe that the full potential of these technologies can be harnessed to address pressing challenges in healthcare and enhance the well-being of countless individuals worldwide. As this transformative journey unfolds, they advocate for a steadfast commitment to excellence, integrity, and the pursuit of a healthier, more equitable future for all.

## 6 Limitations and future scope

The strides made in understanding CNN models for classifying histopathology images have been significant, but several limitations and areas for future work remain. Firstly, the issue of dataset bias is crucial; the effectiveness of CNN models heavily relies on the quality and diversity of training data. Access to larger, more representative datasets encompassing various tissue types, imaging modalities, and patient demographics could enhance generalization and robustness. Secondly, while interpretability techniques like saliency mapping, LIME, and SHAP provide valuable insights, there is a need for novel approaches that capture complex interactions among features. Clinical validation is essential, requiring collaboration with expert pathologists to review model predictions and correlate them with patient outcomes. Ethical considerations surrounding privacy and data security are paramount, ensuring responsible development and deployment of AI tools. A promising future enhancement could be the development of hybrid XAI models that combine visual interpretability with quantitative metrics. This approach would provide a multi-dimensional understanding of feature significance, enhancing both the transparency and reliability of CNN models in histopathology. Additionally, opportunities for integrating XAI in oncology extend beyond histopathology; for instance, applying XAI in radiology could improve model interpretability in tumor detection. Furthermore, research into how XAI techniques can be applied to ensemble models is vital for unlocking their full potential. By addressing these limitations and pursuing innovative directions in XAI, we can advance AI-driven medical image analysis, ultimately enhancing diagnostic accuracy, clinical decision-making, and patient outcomes in healthcare.

## Declarations

**Conflict of interest** The authors declare that there are no conflicts of interest in this paper.

## References


1. Zhang B, Vakanski A, Xian M (2023) BI-RADS-NET-V2: a composite multi-task neural network for computer-aided diagnosis of breast cancer in ultrasound images with semantic and quantitative explanations. IEEE Access 11:79480–79494. https://doi.org/10.1109/ACCESS.2023.3298569
2. Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "An Empirical Evaluation of AI Deep Explainable Tools," In: 2020 IEEE Globecom Workshops GC Wkshps, Taipei, Taiwan, 2020, pp. 1-6, https://doi.org/10.1109/GCWkshps50303.2020.9367541
3. M. A. Anupama, V. Sowmya, and K. P. Soman, "Breast Cancer Classification using Capsule Network with Preprocessed Histology Images," In: 2019 international conference on communication and signal processing (ICCSP), Chennai, India, 2019, pp. 0143–0147, https://doi.org/10.1109/ICCSP.2019.8698043.
4. Peng C, Zheng Y, Huang D-S (2020) Capsule network based modeling of multi-omics data for discovery of breast cancer-related genes. IEEE/ACM Trans Comput Biol Bioinform 17(5):1605–1612. https://doi.org/10.1109/TCBB.2019.2909905
5. R. K and M. S. K, "Breast Cancer Prediction by Leveraging Machine Learning and Deep learning Techniques with Different Imaging Modalities," In: 2022 IEEE 7th international conference for convergence in technology (I2CT), Mumbai, India, 2022, pp. 1-6, https://doi.org/10.1109/I2CT54291.2022.9824749
6. Peta J, Koppu S (2023) Enhancing breast cancer classification in histopathological images through federated learning framework. IEEE Access 11:61866–61880. https://doi.org/10.1109/ACCESS.2023.3283930
7. S. Bose, A. Garg, and S. P. Singh, "Transfer Learning for Classification of Histopathology Images of Invasive Ductal Carcinoma in Breast," In: 2022 3rd international conference on electronics and sustainable communication systems (ICESC), Coimbatore, India, 2022, pp. 1039–1044, https://doi.org/10.1109/ICESC54411.2022.9885314.
8. M. D. Richa, S. A. Ahmed, D. P. Dogra, and P. K. Dan, "Patch Level Segmentation and Visualization of Capsule Network Inference for Breast Metastases Detection," In: 2022 IEEE international conference on signal processing and communications (SPCOM), Bangalore, India, 2022, pp. 1–5, https://doi.org/10.1109/SPCOM55316.2022.9840781.
9. A. K. Titoriya and M. P. Singh, "Few-Shot Learning on Histopathology Image Classification," In: 2022 international conference on computational science and computational intelligence (CSCI), Las Vegas, NV, USA, 2022, pp. 251–256, https://doi.org/10.1109/CSCI58124.2022.00048.
10. D. Chen, H. Zhao, J. He, Q. Pan, and W. Zhao, "An Causal XAI Diagnostic Model for Breast Cancer Based on Mammography Reports," In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), Houston, TX, USA, 2021, pp. 3341–3349, https://doi.org/10.1109/BIBM52615.2021.9669648.
11. M. El-Nakeeb, M. Ali, K. AbdelHadi, S. H. Ahmed Tealab, M. I. Eltohamy, and L. Abdel-Hamid, "Computer-Aided Breast Cancer Diagnosis Using Deep Learning: Malignancy Detection and HER2 Scoring," In: 2023 international mobile, intelligent, and ubiquitous computing conference (MIUCC), Cairo, Egypt, 2023, pp. 1-6, https://doi.org/10.1109/MIUCC58832.2023.10278384
12. Maouche I, Terrissa LS, Benmohammed K, Zerhouni N (2023) An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data. IEEE Trans Biomed Eng 70(12):3321–3329. https://doi.org/10.1109/TBME.2023.3282840
13. E. Mylona et al., "Explainable machine learning analysis of longitudinal mental health trajectories after breast cancer diagnosis," In: 2022 IEEE-EMBS international conference on biomedical and health informatics (BHI), Ioannina, Greece, 2022, pp. 1–4, https://doi.org/10.1109/BHI56158.2022.9926952.
14. T. Brito-Sarracino, M. Rocha dos Santos, E. Freire Antunes, I. Batista de Andrade Santos, J. Coelho Kasmanas, and A. C. Ponce de Leon Ferreira de Carvalho, "Explainable Machine Learning for Breast Cancer Diagnosis," In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Salvador, Brazil, 2019, pp. 681–686, https://doi.org/10.1109/BRACIS.2019.00124.
15. M. Mitu, S. M. M. Hasan, A. H. Efat, M. F. Taraque, N. Jannat, and M. Oishe, "An Explainable Machine Learning Framework

for Multiple Medical Datasets Classification," In: 2023 international conference on next-generation computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 2023, pp. 1–6, https://doi.org/10.1109/NCIM59001.2023.10212821.
16. S. Gengtian, B. Bing, and Z. Guoyou, "EfficientNet-Based Deep Learning Approach for Breast Cancer Detection With Mammography Images," In: 2023 8th international conference on computer and communication systems (ICCCS), Guangzhou, China, 2023, pp. 972–977, https://doi.org/10.1109/ICCCS57501.2023.10151156.
17. S. S. Hossain et al., "Robust AI-enabled Simulation of Treatment Paths with Markov Decision Process for Breast Cancer Patients," In: 2023 IEEE conference on artificial intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 105–108, https://doi.org/10.1109/CAI54212.2023.00053.
18. N. S, S. J, V. M, S. S, and S. G, "SVM-ANN Optimized Algorithm for the classification of breast cancer data as benign and malignant," 2022 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2022, pp. 1-7, https://doi.org/10.1109/STCR55312.2022.10009301
19. M. Ahirwar and A. Agrawal, "Performance Analysis of Deep Learning Models over BreakHis Dataset using Up-Sampling and Down-Sampling Techniques for Classification of Breast Cancer," In: 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2023, pp. 594–599, https://doi.org/10.1109/ICSCC59169.2023.10334935.
20. S. Kabiraj et al., "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1–4, https://doi.org/10.1109/ICCCNT49239.2020.9225451.
21. E. A. Krupinski, "Collaborating across telemedicine specialties for improved cancer care," 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, USA, 2014, pp. 421–422, https://doi.org/10.1109/CTS.2014.6867598.
22. S. Kayikci and T. Khoshgoftaar, "A Stack Based Multimodal Machine Learning Model for Breast Cancer Diagnosis," In: 2022 international congress on human-computer interaction, optimization and robotic applications (HORA), Ankara, Turkey, 2022, pp. 1-5, https://doi.org/10.1109/HORA55278.2022.9800004
23. S. H. P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala and O. Senausera, "LISA : Enhance the explainability of medical images unifying current XAI techniques," In: 2022 IEEE 7th international conference for convergence in technology (I2CT), Mumbai, India, 2022, pp. 1-9, https://doi.org/10.1109/I2CT54291.2022.9824840
24. Hamilton RI, Papadopoulos PN (2024) Using SHAP values and machine learning to understand trends in the transient stability limit. IEEE Trans Power Syst 39(1):1384–1397. https://doi.org/10.1109/TPWRS.2023.3248941
25. F. Xu et al., (2020) "Breast Anatomy Enriched Tumor Saliency Estimation," In: 2020 25th international conference on pattern recognition (ICPR), Milan, Italy, 2021, pp. 2904–2911, https://doi.org/10.1109/ICPR48806.2021.9412593.
26. K. R, S. B and K. V, (2023) "Integrating Explainable AI with Infrared Imaging and Deep Learning for Breast Cancer Detection," In: 2023 OITS international conference on information technology (OCIT), Raipur, India, 2023, pp. 82–87, https://doi.org/10.1109/OCIT59427.2023.10431160.
27. M. S. Ahmed, K. N. Iqbal and M. G. R. Alam, "Interpretable Lung Cancer Detection using Explainable AI Methods," In: 2023 international conference for advancement in technology (ICONAT), Goa, India, 2023, pp. 1-6, https://doi.org/10.1109/ICONAT57137.2023.10080480
28. P. N. Sholapur and I. M, "Explainable AI and Deep Learning techniques for Colon Cancer Detection," In: 2022 4th international conference on advances in computing, communication control and networking (ICAC3N), Greater Noida, India, 2022, pp. 1096–1105, https://doi.org/10.1109/ICAC3N56670.2022.10074383.
29. R. R. Kontham, A. K. Kondoju, M. M. Fouda and Z. M. Fadlullah, "An End-To-End Explainable AI System for Analyzing Breast Cancer Prediction Models," In: 2022 IEEE international conference on internet of things and intelligence systems (IoTaIS), BALI, Indonesia, 2022, pp. 402–407, https://doi.org/10.1109/IoTaIS56727.2022.9975896.
30. P. Shaw, S. Sankaranarayanan and P. Lorenz, (2022) "Early esophageal malignancy detection using deep transfer learning and explainable AI," In: 2022 6th international conference on communication and information systems (ICCIS), Chongqing, China, 2022, pp. 129–135, https://doi.org/10.1109/ICCIS56375.2022.9998162.