

IMAGE SEGMENTATION METHODS IN MAMMOGRAPHY: A LITERATURE REVIEW**BERNARDO TERNUS DE ABREU**

<https://orcid.org/0000-0003-4081-9590> / <http://lattes.cnpq.br/9234153102615950> / bernardoternus@gmail.com
Universidade do Vale do Rio dos Sinos, UNISINOS, São Leopoldo, Rio Grande do Sul, Brasil

JEAN SCHMITH

<https://orcid.org/0000-0001-7642-3235> / <http://lattes.cnpq.br/7662727778218200> / JSCHMITH@unisinis.br
Universidade do Vale do Rio dos Sinos, UNISINOS, São Leopoldo, Rio Grande do Sul, Brasil



Recebido em: 26/02/2026

Aprovado em: 13/10/2026

Publicado em: 26/10/2026

ABSTRACT

This article aims to present a systematic literature review on modern image segmentation methods in mammography. Since the 1980s, mammography has been a standard examination used to identify nodules and signs of breast cancer, being an accessible and low-cost examination. Currently, breast cancer is the most frequent type of cancer in women in Brazil. As a form of early disease control, imaging exams are an important medical ally, and in recent years technology has been used to identify lesions in images. This work was restricted to research articles that use image segmentation methods involving machine learning and deep learning. The selection adopted was of works published in the last five years in the journals: Elsevier, IEEE, Springer and MDPI. Initially, 550 articles were selected, and of these, only 51 passed the funneling stage. The literature review revealed a predominant use of convolutional neural networks (CNNs), mainly U-Net and ResNet, and free databases, especially CBIS-DDSM and DDSM. In recent years, significant results have been observed with the U-Net architecture and with CNN+ Transformers integrative models. The accuracy of the models exceeds 98% in some studies, and recently, the FDA approved the use of deep learning technologies to support medical decision-making.

Keywords: Segmentation, mammograms, review, methods, deep learning.

RESUMO

Este artigo tem como objetivo apresentar uma revisão sistemática de literatura sobre métodos modernos de segmentação de imagens em mamografias. Desde os anos 1980, a mamografia é um padrão de exame utilizado para identificar nódulos e sinais de câncer de mama, sendo um exame acessível e de baixo custo. Atualmente, o câncer de mama é o tipo de câncer mais frequente em mulheres no Brasil. Como forma de controle precoce da doença, os exames por imagem são um aliado médico importante, e nos últimos anos a tecnologia tem sido utilizada para identificar lesões em imagens. Este trabalho se restringiu aos artigos de pesquisa que utilizam métodos de segmentação de imagens que envolveram aprendizado de máquina e aprendizado profundo. O recorte adotado foi de trabalhos publicados nos últimos cinco anos nos periódicos: Elsevier, IEEE, Springer e MDPI. Foram selecionados inicialmente 550 artigos e, destes, somente 51 passaram pela etapa de funil. A revisão de literatura apontou um uso majoritário de redes neurais convolucionais (CNN), principalmente U-Net e ResNet, e de bancos de dados gratuitos, sobretudo de CBIS-DDSM e DDSM. Nos últimos anos, foram observados resultados expressivos com a arquitetura U-Net e com modelos integrativos CNN+ Transformers. A acurácia dos modelos supera 98% em certos trabalhos e,

recentemente, houve a aprovação pelo FDA para emprego de tecnologias de aprendizado profundo no apoio à decisão médica.

Palavras-chave: segmentação, mamografias, revisão, métodos, aprendizado profundo.

1 INTRODUCTION

In recent years, computational architectures have been developed with potential uses in image segmentation of brain tumors, liver segmentation, retinal vessel segmentation, and breast tumor segmentation. Since the U-Net architecture was proposed in 2015, tests have been carried out by different researchers, using different databases, with images of varying sizes. Gradually, adaptations of the initial architecture emerged, such as H-DenseUNet, published in 2017; or UNet++ (2018), which proposed a densely connected U-Net architecture. In the same year, the nn-U-Net architecture (2018) was published, which consisted of an adaptation of U-Net that achieved great results for image segmentation.

The principle of these U-Net architectures was based on an encoder-decoder architecture, featuring an encoder (contracting path), composed of convolutional layers and max-pooling in the first half of the architecture. The contracting path in U-Net is the left-side encoder responsible for capturing context and extracting features from an input image. It consists of repeated blocks, each containing two 3×3 unpadded convolutions, a ReLU activation, and a 2×2 max-pooling operation with a stride of 2. This process halves the image resolution while doubling the feature channels, effectively reducing spatial information to focus on high-level feature extraction.¹

A traditional CNN contains semantic information and less spatial information, and is called the encoder portion. However, spatial information is also important for segmenting semantic information. The specific information of the decoder portion is fed into the U-Net, where semantic information is extracted from the lowest layer of the U-Net network. The decoder portion contains

¹ U-Net proposed integrating the high-level characteristics of the decoder with the low-level characteristics of the encoder. The original idea of the UNet architecture was to add “skip connections” between an encoder network and a decoder network. The network has a symmetrical U-shaped architecture, composed of a subsampling path and an oversampling path. With the introduction of the “skip connection” path, it was possible to recover the spatial information that is lost during the subsampling path due to pooling operations (Baccouche, Garcia-Zapirain, Olea, Elmaghraby, 2021, p. 6).

high-resolution features; these features are extracted from the encoder portion. A ReLU with a data normalization and standardization step is employed in the article by Salama and Aly (2021, p. 4703).

In the second half, the U-Net is composed of a decoder (expansive path), which consists of upsampling layers and convolutional layers that produce segmentations from feature extraction. The expansive path or decoder in a U-Net is the right-side, upward-sampling portion of the network designed to reconstruct spatial resolution from low-resolution features. It uses transposed convolutions to double image size and merges them with high-resolution features from the contracting path via skip connections to enable the image segmentation.

The skip connection combines the characteristics of the preceding and subsequent layers, assisting in the location of information. Tests with the U-Net model were carried out by different researchers, using different databases, with images of varying sizes. Subsequently, adaptations of the U-Net emerged. The U-Net with an attention mechanism obtained better results than the simple U-Net. The U-Net network with ResNet characteristics also has better results in segmenting mammography images than the simple U-Net.

VGG16 was proposed in 2015 and was designed to reduce the number of parameters required for CNN training, aiming to reduce training time. This architecture was tested by some authors in comparison with the U-Net model. VGG16 contributed by demonstrating that deep architectures with simple and homogeneous blocks, that is, stacked 3x3 convolutions, could achieve high performance and easy transfer to other tasks. Its impact was mainly in the standardization and use of pre-trained models.

Another CNN proposed around the same time was ResNet. ResNet was introduced by He, Zhang, Ren, and Sun (2015). In general, using a larger number of layers results in better separability and, consequently, superior classification performance. However, it is observed that accuracy saturates and begins to degrade after a certain number of layers in the network, due to the "gradient vanishing problem". ResNet was proposed to solve this problem and introduced the concept of hop connection or identity mapping. This is a type of replication that allows copying the input to the next layer, providing an alternative path for gradient flow. The input of the first layer is replicated to the output of the second layer. By using residual identity mapping, high-resolution feature maps of the surface layers are added directly to the low-resolution maps of the deeper layers in residual networks

(Sahu, Das, Meher, 2024, p. 4). As more layers are included, training becomes slower and accuracy tends to decrease, making this architecture an alternative model for dealing with this type of situation.

Convolutional neural networks were among the first architectures that attempted to label pixels around objects in segmentation models. Encoder-decoder networks, such as the fully convolutional network (FCN), were developed and used for semantic feature extraction and for mapping target object details. However, with the introduction of skip connections, encoder-decoder architectures were transformed into the U-Net architecture, which became predominant for almost a decade (Baccouche, Garcia-Zapirain, Olea, Elmaghraby, 2021, p. 1).

In this article, we will analyze the distribution of methods used by researchers in articles for the segmentation of mammography images. From the results, it will be evident that methods based on U-Net and ResNet are quite recurrent, with some methods also emerging after them.

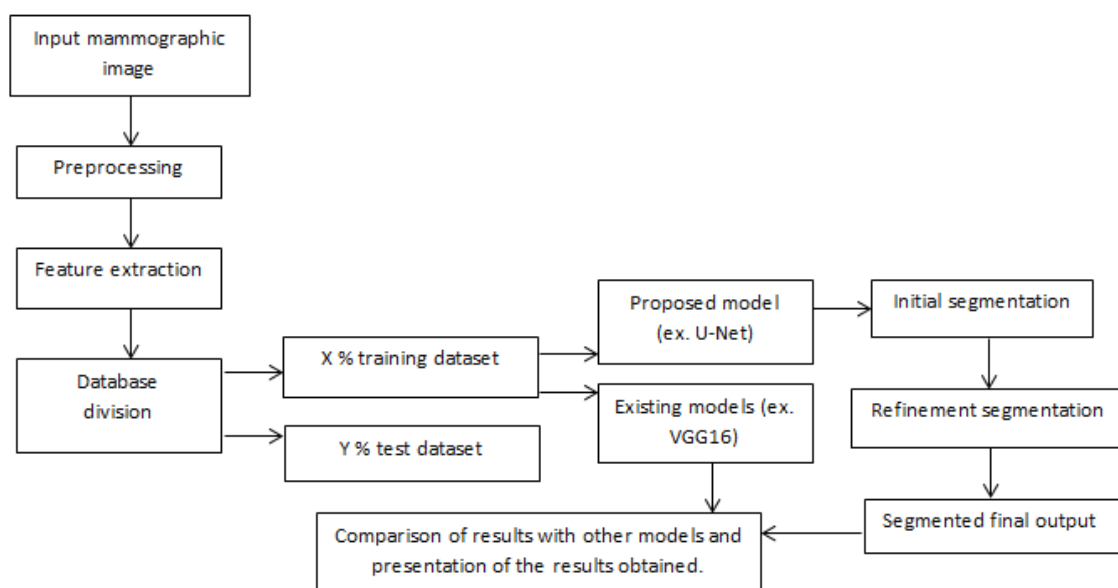
2 THEORETICAL BACKGROUND

Image processing often involves converting one image format to another. According to Nour and Boufama (2025), the steps include: conversion, data anonymization, and image normalization. Image format conversion is sometimes used to simplify processing, improve visualization, reduce image size, or ensure consistency in analysis. "Data anonymization" is used to ensure privacy. In this step, patient identifiers are removed before data processing. "Image normalization" is used to standardize data. Minimum-maximum normalization is applied to standardize pixel intensity values (Nour and Boufama, 2025, p. 9).

The steps involved in creating an image segmentation model can be expressed in a simplified framework. Initially, a set of images is received and pre-processed for better quality in the final segmentation. The database is divided into a training dataset and a test dataset. In the training step, the proposed model is trained, and simultaneously, tests are performed with existing models for the same database. Finally, the output results of the proposed model are compared with existing models

to verify whether the results of the proposed model are satisfactory or not, using the other models as a comparison parameter.²

Figure 1 – Steps in breast cancer image segmentation

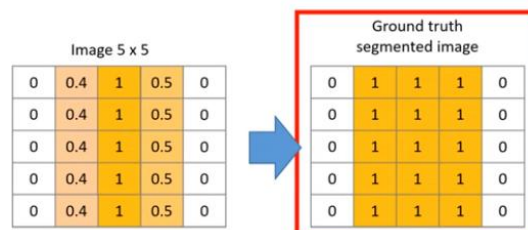


Source: adapted from Nour and Boufama (2025, p. 9)

Some concepts about U-Net networks can be used to understand a segmentation model based on modern methods. The algorithm learns to identify patterns from manual markings on images, called "ground truth". Once the images are marked, we add them to the set of images that the algorithm will receive during execution and let the network predict the mask. To explain how a U-Net network works, we can use an image with 25 pixels and train the network to use the mask with "ground truth".

Figure 2 – Ground truth segmented image

² When the training information is sufficient for the algorithm to perform the detection, we have an algorithm that is in the appropriate operating region, at least in principle, having greater or lesser accuracy, which may or may not be sufficient for the quality requirements. On the other hand, a situation may occur in which the model has good performance applied to the training dataset, but proves ineffective in predicting results when applied to the test set. In this case, we have an overfitting scenario (Melo, 2021, p. 4).



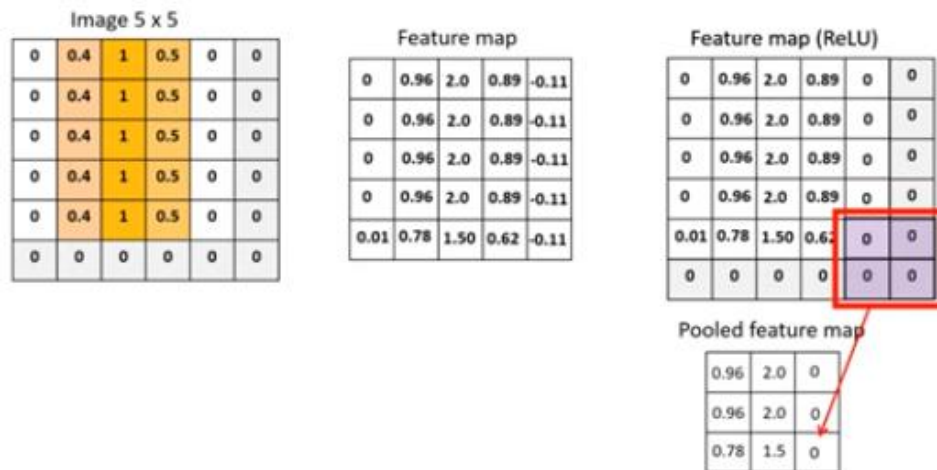
Source: adapted from TileStats Portal (2026)

It can be used a filter known as a "Kernel" filter. This filter has four weights that will be optimized during the training process. The filter is placed on top of the image, where the dot product is performed, meaning that the values are multiplied by the overlapping elements. Then the calculations are performed and the value is plugged into the feature map. In this case, the feature map is expected to be the same size as our image, which can be done using padding. To apply padding, an extra row and column with zeros can be added.

In the next step, it's possible to slide the filter one more step and perform the calculations. And thus, slide the filter until the entire feature map is filled. The architecture depends on the choice of an activation function to be used. In this case, the ReLU function can be chosen, which receives a value and returns the larger value between zero and the value entered in this analysis. If the value -0.11 is entered, the function will return the value zero, because zero is the maximum value between -0.11 and zero. This means that all negative values in the feature map will be set to zero, while positive values will remain unchanged.

Then, pooling is performed using a 2x2 window with a step of 2. At this point, max pooling is used to find the maximum value in that window. The values are inserted into a pooled feature map. Padding can also be performed so that all values in the feature map are used to build the pooled feature map.

Figure 3 – Feature map

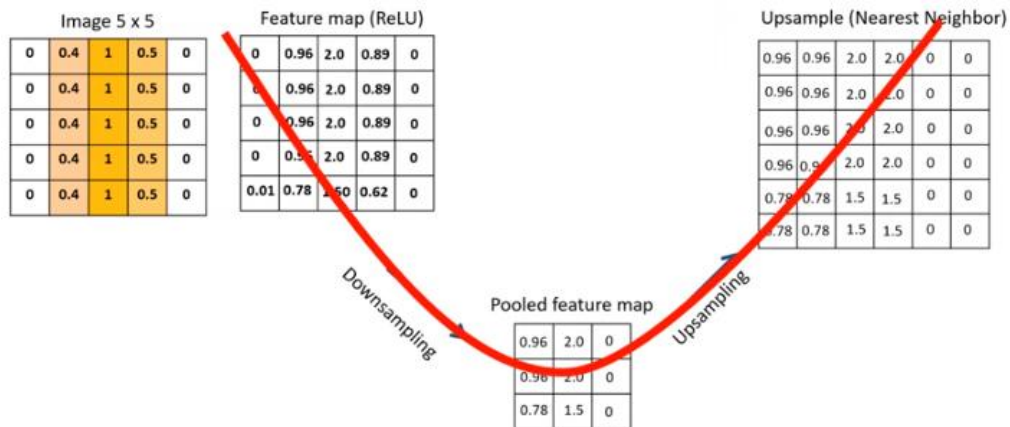


Source: adapted from TileStats Portal (2026)

In the next step, an update sampling is performed using the nearest neighbor method. The first value of the pooled feature map is inserted, and this value is copied to its nearest neighbors. Then, the next value is inserted, and subsequent values are copied to form a complete Upsample (Nearest Neighbor) matrix. Thus, from a "Pooled feature map" matrix, an "Upsample (Nearest Neighbor)" matrix is obtained.

Analyzing the entire process, initially there was a 5 x 5 image. From it, a feature map was created, and then the ReLU activation function was applied. This type of architecture has a subsequent step in which downsampling with max-pooling is performed, resulting in a reduction in the size of the data. Then, upsampling is usually performed, resulting in an increase in the size of the matrix. Observing the architecture design from afar, it looks like a U, which explains its name.

Figure 4 – U-Net architecture



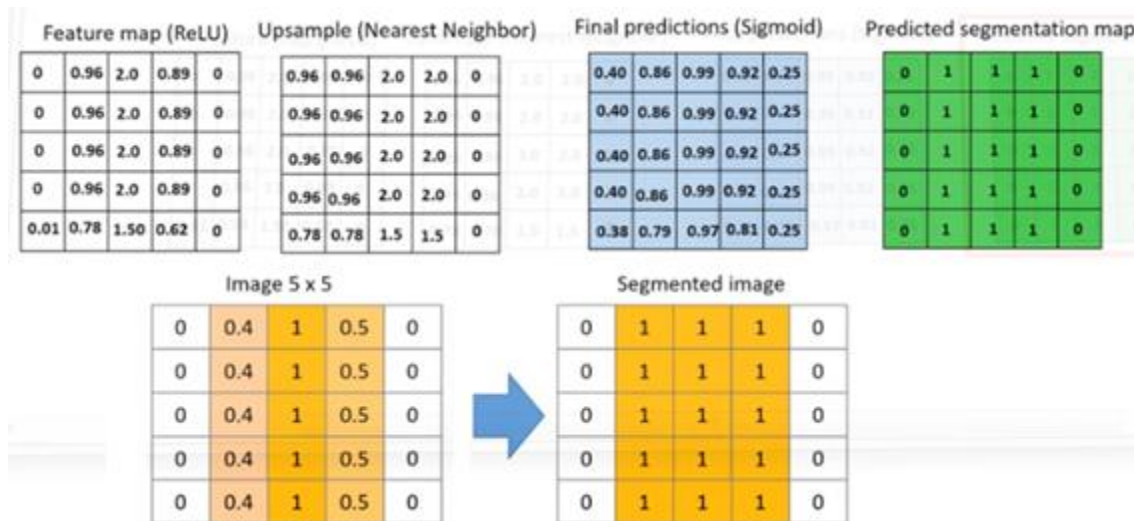
Source: adapted from TileStats Portal (2026)

Another step in U-Nets is performing a skip connection with the upsampled feature map. The upsampled feature map must be clipped so that the two feature maps are the same size. Finally, a final convolution can be calculated using a filter with a height and width equal to one, but a depth equal to two. This occurs due to the filter sliding over the two overlapping feature maps.

After performing some calculations involving the matrix fields, the sum can be inserted into the sigmoid activation function to restrict the output values between zero and one. The value is inserted into the final predicted output matrix, composing a probability map. The filter is slid one step to the right and a calculated value is inserted. These values can be viewed as the probability of the corresponding pixel in the input image belonging to the foreground.

Hypothetically, it is possible to consider the probability of a given pixel belonging to the foreground to be 40%. If a given probability is greater than 0.5, we generally classify the pixel as belonging to the foreground of the image. And if it's less than 0.5, we classify the pixel as belonging to the background. If we apply this rule to these probabilities, we obtain the final predicted segmentation map.

Figure 5 - Each processing step of the matrix



Source: adapted from TileStats Portal (2026)

During U-Net training, the eight weights associated with the filters are optimized. The predicted segmentation map recreates the real image as faithfully as possible. The network attempts to reduce the probability values to zero and 1 by altering the weight values. Previously, the image resolution was increased by copying a value from the grouped feature map and inserting these values into the nearest neighboring elements. A disadvantage of this method is that it does not involve trainable weights. Another existing upsampling method is transposed convolution.

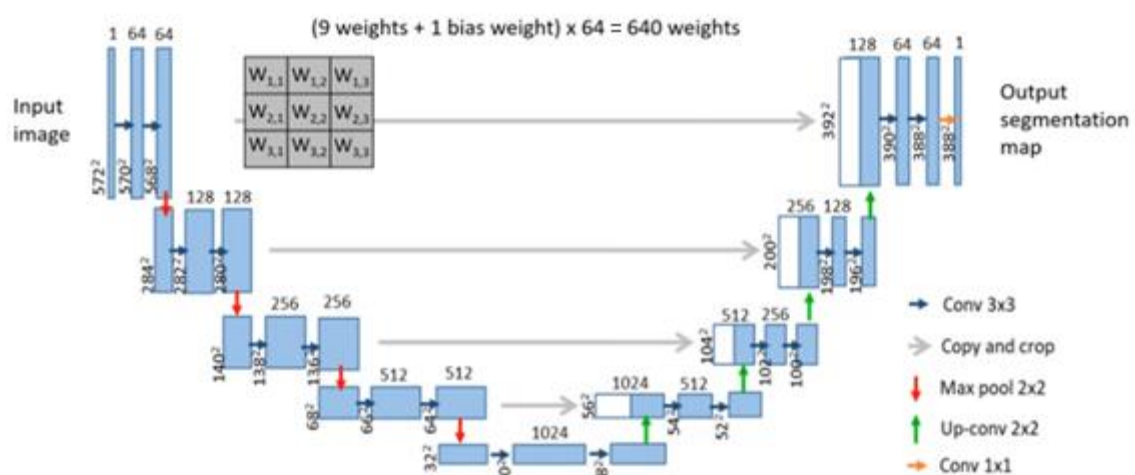
If the transposed convolution method is adopted, other values can be obtained for the grouped feature map. This is possible because the previous weights in the filter will be optimized to other values when additional weights are added to the network. This is the additional filter that can be used to perform upsampling. The filter has weights optimized from the bias weight value.

The matrix values can be optimized, and the bias weight can be added to this product. After filling the upsampled feature map, the following steps are performed: clipping and concatenation. By employing transposed convolution, the network will have five additional weights to adjust during training, compared to the nearest neighbor method.

Regarding the original U-Net model proposed in 2015, some comments can be made (Ronneberger, Fischer and Brox, 2015). The network starts with a grayscale image. The network

begins with only one channel, which consists of a 2D matrix in the case of mammograms. Despite this, U-Nets show good results for color images with three channels. The image is resized to a 2D matrix with 572 rows and 572 columns. The architecture uses 64 different filters with a size of 3 by 3 in the first convolution step. After the first convolution, the network will result in 64 maps with different features. Since each filter has nine weights plus one bias weight, and there are 64 filters of this type, there will be 640 weights to adjust in the first convolution during the training process.

Figure 6 – U-Net architecture stages



Source: adapted from TileStats Portal (2026)

After each convolution, the ReLU activation function is applied. Since no padding was used, the resulting feature maps will have 2 fewer rows and 2 fewer columns than the input matrix. Then, a second convolution of the 64 feature maps can be performed. With a 3D 3x3 filter with a depth of 64, applied to the 64 feature maps. The authors used 64 deep filters of this type, resulting in 36,928 weights to adjust in the second convolution.

Since no padding was used, the new feature maps have 2 fewer rows and 2 fewer columns than the previous ones. Next, Ronneberger, Fischer, and Brox (2015) performed max-pooling using a 2x2 window, with a spacing of two, resulting in 64 feature maps grouped at half the size. In each subsampling step, the system uses twice as many filters, resulting in 128 feature maps at this stage. The smallest feature maps in the bottleneck are 28 by 28 in size. Next, the architecture uses a

transposed convolution with 512 filters for augmentation. Each filter of this type has two rows, two columns, and a depth of 1024. This therefore results in 512 oversampled feature maps.

The augmented feature maps are concatenated with the corresponding feature maps in the resolution reduction phase. Because these feature maps are larger than the others, they are clipped, removing 8 columns and 8 rows. A convolution with 512 “3 by 3” filters, with a depth of 1024, is applied to the concatenated feature maps, followed by the ReLU function. In the upsampling phase, all oversampled feature maps are concatenated with the corresponding feature maps in the downsampling phase, which is called a jump connection.

A skip connection means that the algorithm “skips” some layers, passing feature maps directly from previous layers to subsequent layers. Skip connections help recover spatial information that is lost during the downsampling process due to the clustering process.

In the last step, the model has a 1-by-1 filter with a depth of 64. The architecture has a SoftMax function, which generates a channel for each class, and which is applied when you want to predict more than two classes of results. One channel, or matrix, includes the probability values of a given pixel belonging to the foreground, while the second channel includes probability values of a given pixel belonging to the background. If, instead, a sigmoid activation function is used, only one output channel is obtained, which is sufficient if the required result is binary. The channel will only contain probability values that a given pixel belongs to the foreground.

The architectural image can be divided by a vertical line down the center of the image: the first part of the network, on the left, is called the encoder, which captures important features in the image. The second part, on the right, is called the decoder, and it helps to locate where the elements are situated in the image.

3 MATERIALS AND METHODS

The development of the review research was guided by research questions that structured the article's development. The five research questions are listed below:

Table 1 – Questions proposed in the review article

Number	Research Question
1	What techniques are used for segmenting mammography images?
2	Is it possible to identify methodological preferences within the set of articles that use modern segmentation methods?
3	What databases are used in the literature? What is the quality of the data in these databases? Is the amount of data sufficient considering the wide variation in tumor shapes?
4	Are the results achieved by the published works significant, or is the state of the art still far from being applicable in everyday examinations? Is the research mature enough to approximate real-world applications?
5	What are the gaps in research, lacking publications, in terms of image segmentation research? What are the limitations of the studies in terms of the models used?

Source: the author (2026)

Having presented the questions, the inclusion and exclusion criteria that structured the article selection funnel can be considered. This review focused on recent research articles on mammography image segmentation between 2021 and 2025. The following conditions were primary for establishing selection criteria to conduct a comprehensive analysis on the topic of interest.

Table 2 – Keywords used in the search for articles

Number	Keyword
1	“Breast” AND “Cancer” AND “Mammography”
2	“Breast” AND “Cancer” AND “Segmentation”
3	“Breast” AND “Cancer” AND “Dataset”
4	“Breast” AND “Cancer” AND “Machine” AND “Learning”
5	“Breast” AND “Cancer” AND “Deep” AND “Learning”
6	“Breast” AND “Segmentation” AND “Deep” AND “Learning”

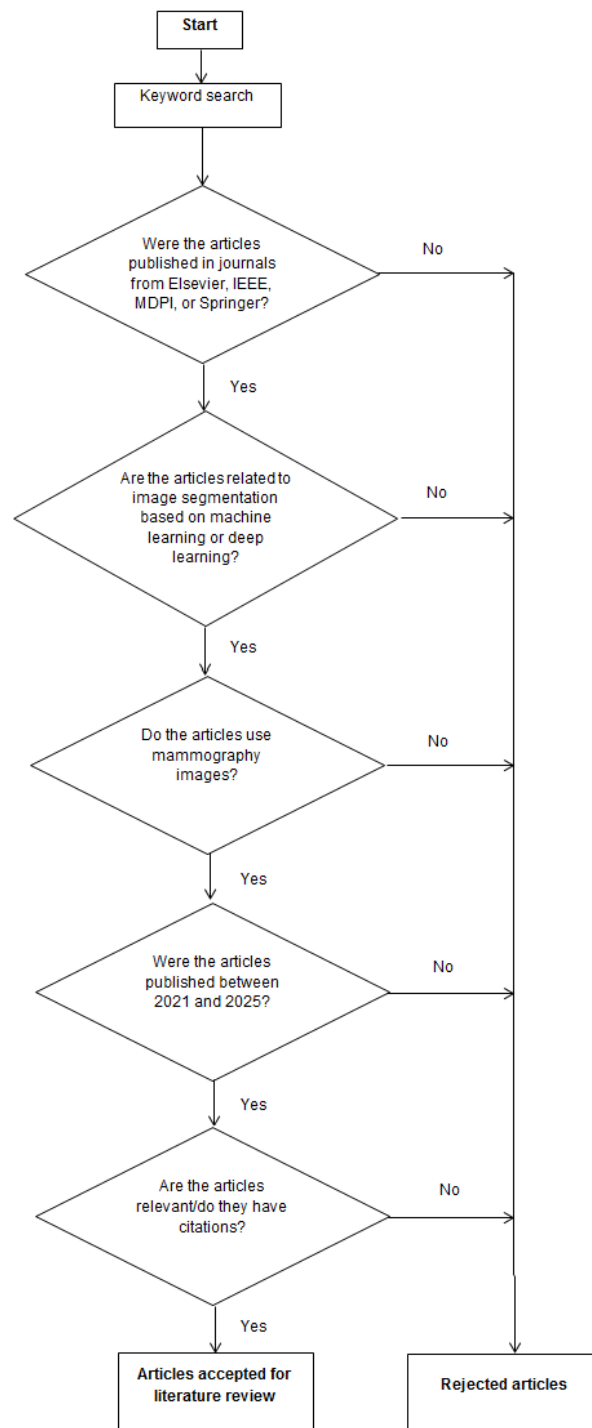
7	"Breast" AND "Segmentation" AND "Machine" AND "Learning"
8	"Mammography" AND "Segmentation" AND "Machine" AND "Learning"
9	"Mammography" AND "Segmentation" AND "Deep" AND "Learning"
10	"Transformer" AND "Segmentation" AND "Mammogram"

Source: the author (2026)

The review conducted was systematic, as it answered specific research questions. A possible improvement would be to use the PRISMA method. PRISMA aims to ensure that the process is described transparently, allowing the reader to understand how the studies were identified, selected, evaluated, and synthesized. For a more thorough review on the topic, the use of this method is recommended.

Some conditions were established for the selection of articles: a) The article must be related to breast cancer; b) Articles related to machine learning and deep learning models must have been published between 2021 and 2025; c) Articles must have been published in the following journals: Elsevier, Springer, IEEE, and MDPI; d) Selected articles must be relevant to the field, which can be indicated by the existence of citations; e) Articles must use mammography images as a database. Articles that use only databases of other types of exams should not be selected. Figure 7 presents a flowchart of the article selection process. Based on the selection criteria, which restricted the scope of articles to four main scientific journals, it was possible to separate a set of articles. Articles with relevance, measured by the number of citations, were prioritized.

Figure 7 – Flowchart of selection criteria



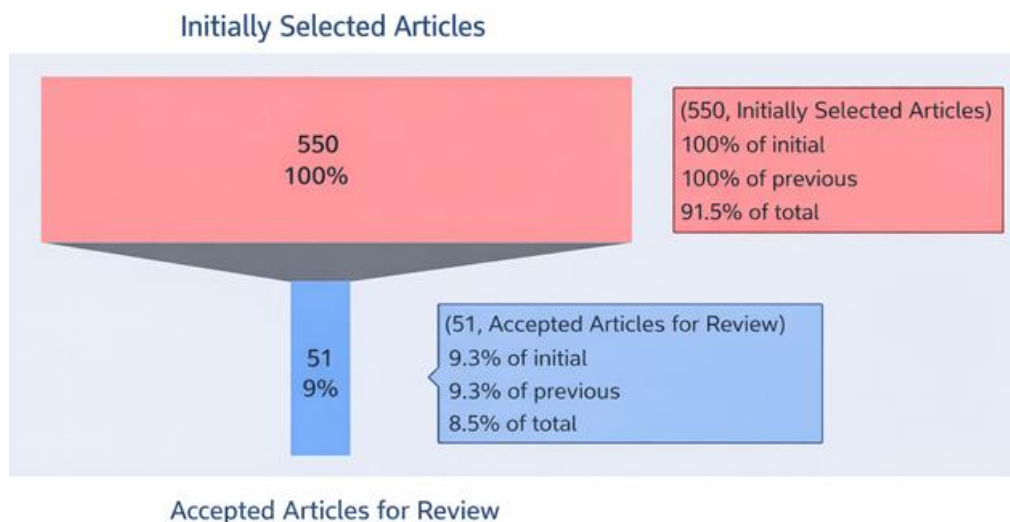
Source: the author (2026)

4 RESULTS AND DISCUSSION

The research articles related to the study were published in the following journals: Elsevier, IEEE, MDPI, and Springer. Articles from other journals were not selected for the literature review. Articles that did not use mammography data were not selected, as well as articles prior to 2021 or articles published after 2021 with a lack of citations. The aim was to restrict the scope to a small set of journals considered reliable, valuing the existence of citations as a criterion, as well as the presentation of metrics and authors' evaluation of their proposed model. Articles containing classic segmentation for mammograms and whose content did not mention machine learning or deep learning were not selected for the literature review.

The table presented in the Appendix section describes the main characteristics of the original articles and the models presented. It is the result of the restrictive funnel presented in the flowchart in Figure 8. Of the 550 files identified in the literature review, only 51 were accepted for analysis.

Figure 8 – Funnel Chart of the Literature Review



Source: the author (2026)

Once the selection and information gathering for the 51 articles has been completed, the research questions can be answered.

Research Question 1: What techniques are used for segmentation of mammography images?

The main methods for segmenting mammograms include several classic and modern approaches. Thresholding converts images into binaries by separating background and object intensities, being simple and fast, but sensitive to noise and variable illumination. Morphological techniques use operations such as erosion and dilation to refine binarized shapes, to remove noise and fill gaps. The zero-crossing method identifies contours where the intensity gradient changes sign, being indicated for smooth edges. The watershed technique models the image as topographic relief and segments by "flooding" to ridge marks, however it can lead to oversegmentation.³

Clustering techniques group pixels by intensity or texture similarity, being effective for homogeneous regions. Classifier-based segmentation methods employ trained models, such as SVM and RF, to decide pixel by pixel or region by region. In turn, the Markov random field models spatial relationships and prioritizes regularity in segmentation, making it suitable for probabilistic models. Active contours adjust deformable curves to fit gradient-guided lesion edges. Level sets make implicit contour adjustments according to minimized energy functions. Texture-based methods use descriptors, such as gray-level concurrency, to discriminate between tumor and normal regions.

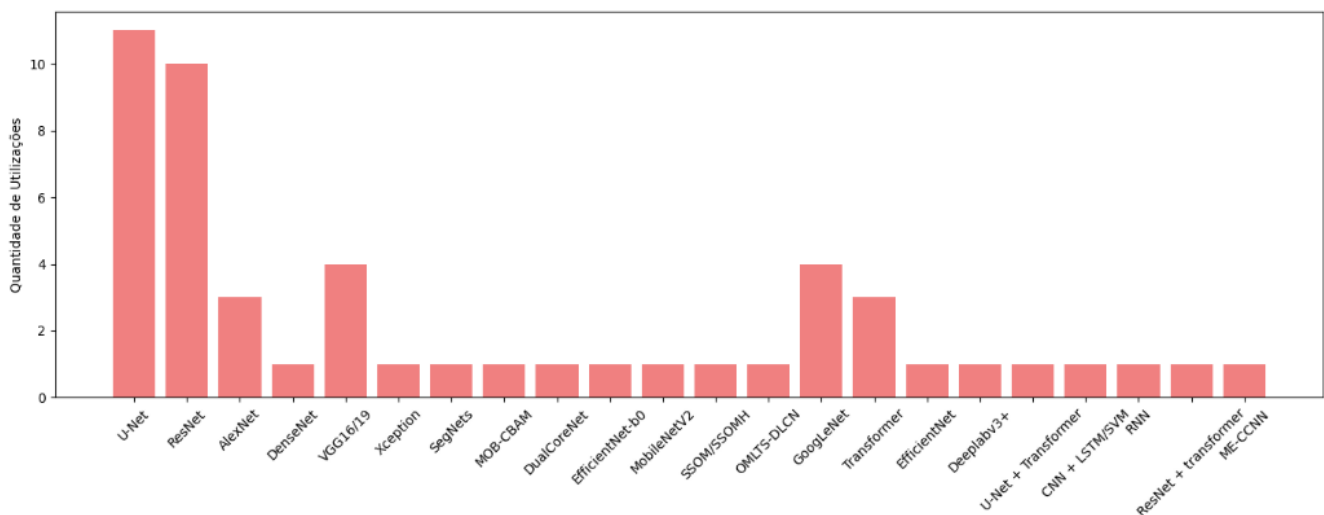
There are also methods based on machine learning and deep learning, such as U-Net networks or CNNs, which learn visual patterns and automatically segment lesions with high precision, especially with large annotated databases. These methods are indicated by the lower need for manual inputs throughout the process, once the model is established.

³ Image segmentation and classification techniques are distinct. Although this work focuses on segmentation techniques, the main classification techniques can be mentioned. Mammographic images are classified using classifiers to identify the classes as: malignant, benign, and normal. For this purpose, classifiers of the following types can be used: Naive Bayes (NB), Decision Tree, Support Vector Machine (SVM), and Random Forest. The Decision Tree (DT) classifier is a supervised machine learning classifier where data is continuously partitioned based on a given parameter. Nodes represent the features extracted for the breast cancer classification problem, and edges represent the test result, interconnecting the next node or leaf. The classification result (benign, malignant, or normal) is present in the node. The Random Forest classifier has a set of decision tree classifiers trained for the testing phase, being more refined than the conventional decision tree. The correct input features need to be obtained to act as nodes. There is a number N of decision tree classifiers, and the features obtained from the input image are sent through each decision tree to obtain the class labels. Finally, a bagging technique is applied to the result obtained from the trees in the previous step. In the case of the Support Vector Machine (SVM) classifier, the method provides accuracy with less computational power. The main goal of SVM is to find the hyperplane in an N-dimensional space to uniquely classify the data points. The value of N represents the number of features, and the hyperplane is a decision boundary used to classify the data points.

Research Question 2: Is it possible to identify methodological preferences in the set of articles that use modern segmentation methods?

Regarding the models or architectures used, it can be noted that the U-Net architecture is predominant in the works, followed by ResNet configurations or those derived from it. The VGG16 and VGG19 architectures are also frequently found, as three distinct articles used this structure. Other configurations, derived from the aforementioned architectures, or new configurations, were found more sporadically in the analyzed articles, such as the transformer model.

Figure 9 – Use of models or architectures

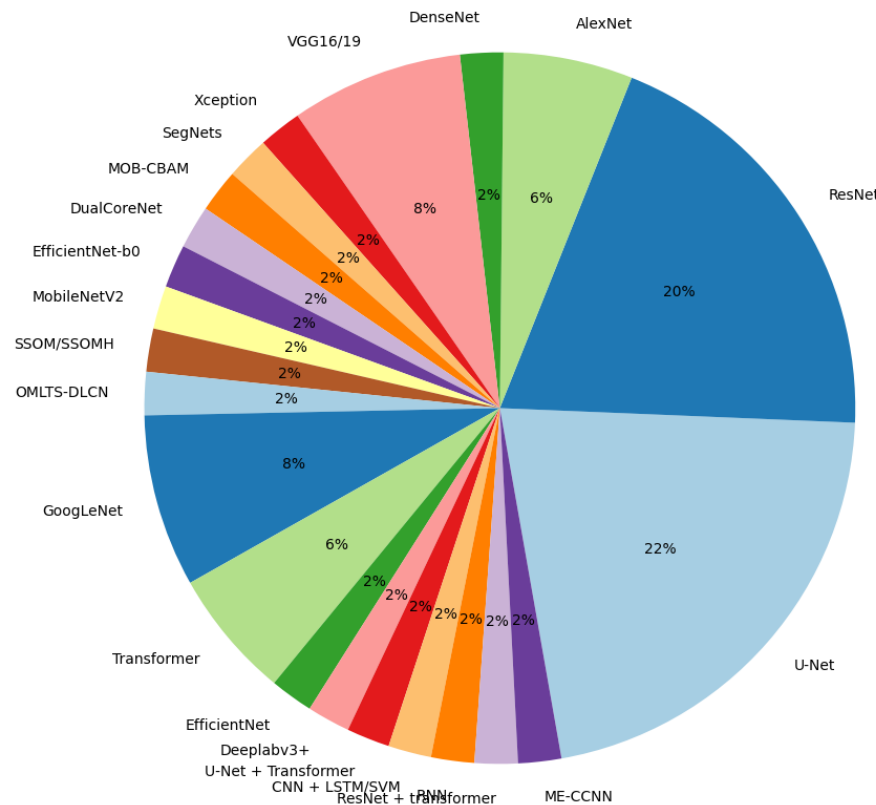


Source: the author (2026)

The analysis showed that U-Net architecture convolutional neural networks were found in at least 22% of the studies, while the ResNet architecture was used in 20% of the articles. Studies using VGG16/19 architecture represented 8% of the articles, while AlexNet was used in 6% of the publications. Studies employing the Transformer model represent 6% of the articles.

Considering that the article by Su et al. (2022) with the transformer model for mammography image segmentation has 125 citations, and theoretical articles on transformers have also been cited by other researchers, the methodology seems to be on an expansion trend for the near future. In this sense, it is possible to project an increase in studies with combined CNN + transformer architectures. The other configurations show lower percentages of use.

Figure 10 – Percentage of use of each model

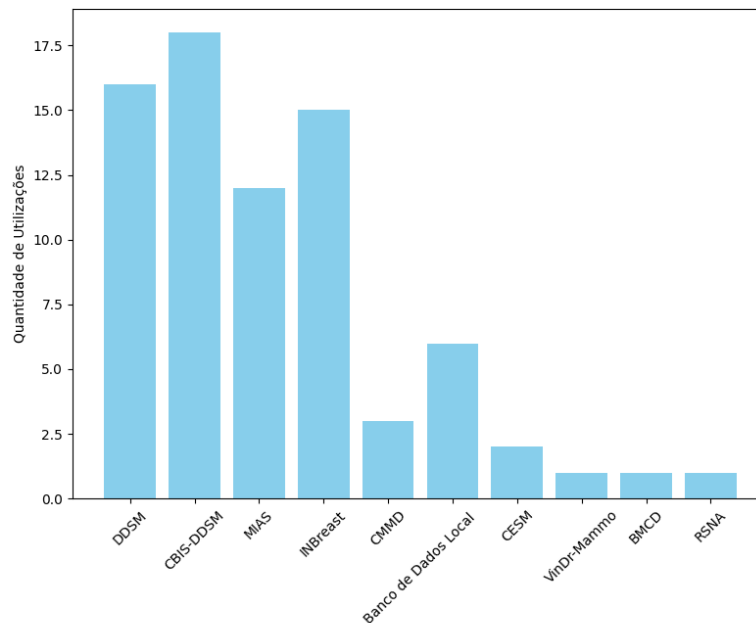


Source: the author (2026)

Research Question 3: What databases are used in the literature? What is the quality of the data in these databases? Is the amount of data sufficient considering the wide variation in tumor shapes?

Based on the data in the table in the Appendix section, several analyses can be performed. The most frequently used databases were CBIS-DDSM and DDSM, followed by INBreast and MIAS. Some studies opted for the use of local databases, indicating the authors' interest in information from specific regions where the studies were conducted.

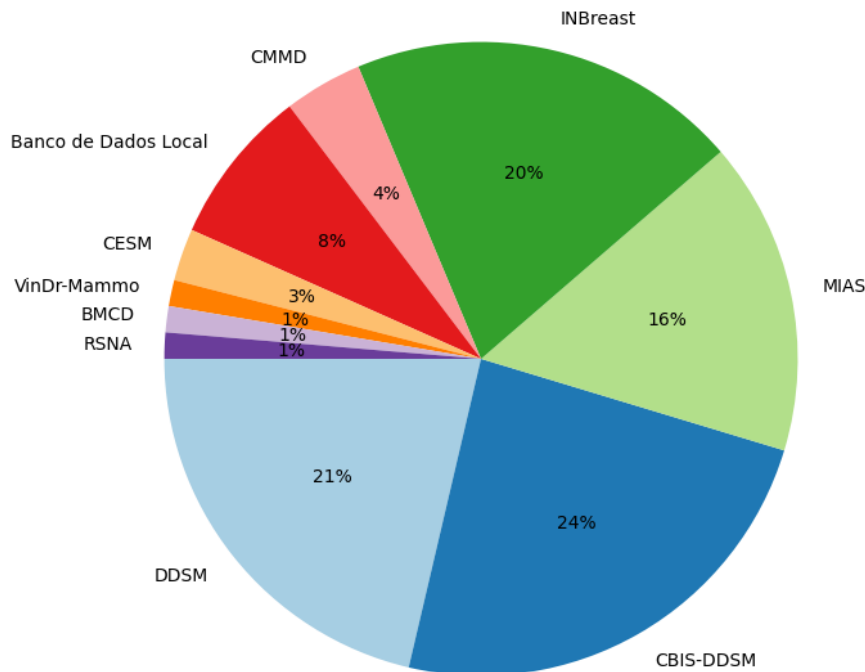
Figure 11 – Use of databases



Source: the author (2026)

Another point that can be made is that practically half of the studies used some database derived from DDSM: either DDSM itself (21%) or CBIS-DDSM (24%). The DDSM is a public dataset provided by the Department of Computer Science and Engineering at the University of South Florida, Sandia National Laboratories, and Massachusetts General Hospital. The CBIS-DDSM is an updated and standardized version of the DDSM. It contains a variety of pathologically verified cases, including malignant, benign, and normal cases. The DDSM is an extremely useful database for the development and testing of computer-aided diagnostic (CAD) systems due to its scale and the basic truth validation it offers. The CBIS-DDSM collection includes a subset of the DDSM data organized by experts.

Figure 12 – Percentage of use of each database



Source: the author (2026)

The article focused on studies that utilize free databases. However, the literature review identified the existence of private databases. Information on private databases is compiled in Table 3.

Table 3 – Private mammography databases

Database	Features					
	Full Name	Year	Region / Locality	Num. Patients	Cost to Use	Reference
CALMA (Mammo Grid)	Computer Assisted Library for Mammography	2001	Italy	3369	For project members only	Marzuli (1999)
LLNL	Lawrence Livermore National Laboratories	2010	EUA	198	Paid (100 dollars)	gdo-biomed.ucllnl.org/pub/mammo-db/
MIDB	Mammographic Image Database	2014	University of California	-	Paid	dblp.org/rec/conf/mammo/Halling-BrownLPWMY14

NMD	National Digital Medical Archive	2005	University of Pennsylvania	100000 0	Paid	nrdr.acr.org/Portal/NMD
OMI-DB	OPTIMAM Mammography Image Database	2008	United Kingdom	200000 0	Paid	medphys.royalsurrey.nhs.uk/omidb/

Source: Adapted from Ansar and Raza (2023, p. 8).

Table 4 lists free databases for analyzing mammography images from randomized patients. These databases are freely accessible.

Table 4 – Free mammography databases

Database	Features						
	Strength	Limitation	Visualization	Image format	Num. Patients	Num. Images	Link
DDSM	Information available on different types of injuries. Solid documentation and widespread use.	The precise location of the injury is not available, and the database has not been updated.	Both MLO and CC	LJPEG	2620	10480	www.eng.usf.edu/cvprg/mammography/database.html
MIAS	Widely used	Small dataset with low resolution.	MLO	Portable Gray Map (PGM)	161	322	www.repository.cam.ac.uk/items/b6a97f0c-3b9b-40ad-8f18-3d121eef1459
Banco-Web	It has a system with selection options for downloading images.	Approximately 25% of the images contain lesions or calcifications. 68% of the data are normal images or images that contain cancer but	Both MLO and CC	ATIFF	320	1400	lapimo.sel.eesc.usp.br/bancoweb/

		have not yet been detected.					
CBIS-DDSM	It provides lesions with different shapes and margins, being a standardized version of the DDSM.	It hasn't been updated since 2017 and doesn't provide user support.	Both MLO and CC	DICOM	6775	10239	www.cancerimagingarchive.net/collection/cb-is-ddsm/
BCDR	It provides a file format with different resolutions indicating the location of the lesion.	Limited size	Both MLO and CC	TIFF	-	1247	bcdr.ceta-ciemat.es/information/about
INBreast	Well-defined outlines are provided in XML format by experts.	The variation in the shape of the lesion mass is limited.	Both MLO and CC	DICOM	115	410	medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database
CMMD	It provides more recent images than some databases.	Absence of pre-existing label masks for the images. Requirement to employ image processing techniques to generate the necessary segmentation masks.	Both MLO and CC	DICOM	1775	5202	www.cancerimagingarchive.net/collection/cmmd/

Source: Adapted from Ansar and Raza (2023, p. 8)

The literature review pointed out that one of the challenges in mammography research is the size of the databases. The lack of databases with larger datasets has considerably affected the testing of existing models. On the other hand, the existence of free databases has contributed to the existence of several studies in a more accessible way for researchers.

Research Question 4: Are the results achieved by the published works significant, or is the state of the art still far from applications in everyday examinations? Is the research mature enough in terms of its approximation to real-world applications?

Recently, the Food and Drug Administration approved the use of deep learning technologies to support medical decision-making, specifically in the automatic segmentation of mammography images. This approval represents an important milestone, as it scientifically validates the use of artificial intelligence (AI)-based systems in image-assisted diagnosis, particularly in digital breast tomosynthesis (DBT), also known as 3D mammography.

Among the approved systems is ProFound AI Version 3.0, from iCAD, authorized in 2021. This model has demonstrated relevant clinical improvements, such as increased specificity and sensitivity, in addition to reducing exam reading time by up to 40%. Another notable system is Saige-Q, developed by DeepHealth, also approved in 2021, which allows for the automatic screening of mammograms with support for both 2D and 3D exams. It was followed by Saige-DX, launched in 2022, focusing on reducing false positives and unnecessary recalls.

Furthermore, MammoScreen, from the French company Therapixel, received FDA authorization for use in breast tomosynthesis, demonstrating clinical benefits such as a higher cancer detection rate and a lower incidence of false alarms. More recently, in 2023, the FDA approved Lunit Insight DBT, further reinforcing the presence of automatic segmentation and classification algorithms in supporting radiological diagnosis.

These technologies stand out for providing precise segmentations of suspicious regions in images, as well as generating heat maps, detection boxes, and risk scores. With this, radiologists can make faster and more informed decisions, increasing the efficiency and accuracy of the diagnosis.

Research Question 5: What are the gaps in studies, lacking publications, in terms of research on image segmentation? What are the limitations of the studies in terms of the models used?

One of the limitations of the articles is that, due to the publication model for scientific articles, which requires a concise presentation of the results obtained and an objective communication of the methodology employed, articles sometimes fail to discuss crucial concepts in the methodology used, such as the constitutive characteristics of U-Nets or structural aspects of the Transformer architecture. These aspects are more commonly found in computer science/technology books and in teaching materials sometimes produced by the same authors.

The gaps in the studies consist of evaluating models with larger datasets, as well as the use of recent methods, especially new architectures published in recent years. Integrative methodologies have been successful, as well as methodologies based on convolutional neural networks.

The final segmentation results depend on factors such as the quality of image preprocessing. Some architectures achieve better segmentation results than others, but defining which is better may not be a trivial task, considering that the result is influenced by preprocessing steps that may have greater or lesser individual interference, such as the choice of which images or data to use to feed the model.

Some descriptive comments on the mentioned articles can be made with a view to presenting general aspects of each model or method. Tiryaki and Kaplanoglu (2022) presented a fully automated cascade model based on deep learning for breast density assessment. In the first stage, segmentation of adipose, fibroglandular, and pectoral muscle tissues from digitized mammograms occurred. The database used was the "Digital Database for Screening Mammography" (DDSM). The method involved several types of U-nets. Features extracted from breast tissue segmentation predictions were used to assess breast density.

Tiryaki and Kaplanoglu (2022) used 66 images for the segmentation systems and 296 images for breast density assessment, from mediolateral oblique mammograms that were selected from the DDSM dataset. The authors used different U-nets with varying numbers of layers and implemented filters, determining the best performance among the options adopted. The performance of the U-nets was investigated using categorical cross-entropy functions, Dice, Tversky, Tversky Focal, and Dice hyperbolic cosine logarithmic loss functions.

It was investigated the response of U-nets with pre-trained weights of the VGG16, VGG19, and ResNet50 networks in the encoding path. The system proposed by Tiryaki and Kaplanoglu (2022) can be used for automatic mammography segmentation into adipose, fibroglandular, and pectoral muscle tissues. The segmentation model allows for the estimation of the fibroglandular-adipose tissue interface, which has recently been found to be an important region for breast cancer investigations.⁴

Mao and colleagues proposed three different CNNs: DenseNet 121, Xception, and ResNet 50. They incorporated backbone architectures and CBAM-type attention blocks. The final models were compared with the diagnostic performance of conventional CNNs, radiomic models, and two radiologists with specialized experience in breast imaging. According to the authors, the best-performing deep learning model was the CBAM-based Xception. The model achieved an area under the ROC curve (AUC) of 0.970, a sensitivity of 0.848, a specificity of 1.000, and a precision of 0.891 on the external test set, superior to that of other CNNs, radiomic models, and radiologists.

Su, Liu, Xi, and Hu proposed a YOLO-LOGO transformer model. The model simultaneously used the YOLO (You Only Look Once) and LOGO (Local-Global) architectures. They adopted YoloV5L6, which is an object detection model, to position and crop the breast mass in high-resolution mammograms. They used YoloV5 to detect the region of interest in the mammogram and cropped the portion directly from the high-resolution image to increase training effectiveness. Subsequently, they used an updated version of the local-global segmentation (LOGO) strategy, which increased the segmentation resolution at the original pixel level (Ayana, Dese, Dereje, Kebede, Choe et al, 2023, p. 4).⁵

⁴ The results were 82.71%, 73.39%, and 95.30% using ResNet50-U-net on the test data. DSC scores were obtained for segmentation of adipose, fibroglandular, and pectoral muscle tissues, respectively. The mammographic segmentation results are 3% to 12% better than the latest generation DSC in the literature, considering all tissues in the foreground simultaneously. A breast density classification accuracy of 76.01% was achieved on a separate mammography dataset, which is comparable to recent studies in the literature.

⁵ To balance training efficiency and segmentation performance, the authors modified the LOGO training strategy to train whole images and cropped images on the global and local transformer branches separately. The two branches were then merged to form the final segmentation decision. The proposed YOLO-LOGO model was tested on two independent mammography datasets (CBIS-DDSM and INBreast). The model showed significantly superior performance compared to previous work. It achieved a true positive rate of 95.7% and an average accuracy of 65.0% for mass detection on the CBIS-DDSM dataset. Their performance for mass segmentation on the CBIS-DDSM dataset was F1-score = 74.5% and IoU = 64.0%. A similar performance trend was also observed on another independent dataset, INBreast. The model proposed by the authors showed greater efficiency and better performance, reduces computational requirements, and improves

Currently, some high-configuration GPUs are designed for convolution computation, which makes CNNs faster. One limitation of CNN models is that they devote a lot of attention to the local square of pixels. In fact, a newly emerged deep learning architecture, called **Vision Transformer** (ViT), has recently been leading its trend of replacing CNNs in computer vision. Compared to traditional CNNs, ViT is powerful at capturing local and global or long-range visual dependencies through its self-attention mechanisms, global attention obtained by course or refined local attention.

In recent years, ViT has been a good option for solving the mammography segmentation problem. The main challenge of the ViT-based medical image segmentation model is that training this model requires significant computational power. Due to GPU limitations and other hardware constraints, current computational resources can only segment low-resolution images.

Chen and colleagues (2021) proposed TransUNet, the first medical image segmentation framework that establishes self-awareness mechanisms from a sequence-by-sequence prediction perspective. To compensate for the feature resolution loss brought about by Transformers, TransUNet employs a hybrid CNN-Transformer architecture to leverage detailed high-resolution spatial information from CNN features and the global context encoded by Transformers. The model was inspired by U-shaped architectural design, and its self-awareness feature is encoded by Transformers and amplified to be combined with different high-resolution CNN features that are ignored from the encoding path. The result is precise localization. The authors assessed that the design allows the framework to preserve the advantages of Transformers while also benefiting medical image segmentation. Compared to other works, TransUNet presents better self-awareness than CNN-based self-awareness methods.

In 2021, Chen and his colleagues presented a Transformer-centric encoder-decoder framework called TransUNet. It was conceived as an alternative architecture to U-Net, given its limitations. The proposed framework tokenizes patches of CNN feature maps, facilitating a richer extraction of global contexts. At the same time, a Transformer decoder was presented, having been designed to refine segmentation regions, leveraging cross-attention mechanisms and U-Net features (Chen et al, 2024, p. 8). Empirical results demonstrate the superior performance of TransUNet in

the versatility and accuracy of computer-aided breast cancer diagnosis. Therefore, it has the potential to allow more assistance to physicians in the early detection and treatment of breast cancer, thus reducing mortality.

tumor segmentation. However, the authors did not study the model with mammography or breast cancer data.

Due to the numerous convolutional tasks across multiple layers of the network, CNNs require high computational capacity when the data volume increases. Furthermore, when analyzing mammograms, CNNs focus on a specific region, which is the region where a tumor is suspected, and the rest of the image is disregarded. This causes CNNs to miss some crucial details that would have been discovered if the entire image were examined at once.

According to Ayana, Dese, Choe et al (2023), the Vision Transformers-based transfer learning approach overcomes the shortcomings of CNN-based transfer learning methods by leveraging the self-awareness approach of Transformers. Because of this, Vision Transformers (ViTs) have come into use as they have overcome the limitations of CNNs for large datasets (Ayana, Dese, Dereje, Choe et al, 2023, p. 3).

Most CNN-based deep learning models use a "patch-based" approach, in which a suspected tumor area on a mammogram is cropped and inserted into the model. This leads to the loss of information from the entire mammogram, resulting in false positives. In addition, the patch-based approach is time-consuming and computationally demanding. The second limitation of the CNN-based approach is that its performance varies depending on the size of the lesions in an image. Thus, the size of the lesion in the region of interest (ROI) affects the performance of CNNs. Thirdly, CNNs require considerable preprocessing to compensate for low image quality.

According to Su, Liu, Xie, and Hu (2022), the TransUNet proposed by Chen and colleagues (Chen, Lu, Yu, Luo, Adeli, Wang, Lu, Yuille, Zhou, 2021) is the first Transformers-based medical image segmentation framework that has outperformed other state-of-the-art segmentation models, being a combination of Transformers and U-Net.

The U-Net consists of a symmetric encoder-decoder network with hop connections to improve detail retention and has become a benchmark in neural networks. However, U-Net architectures generally exhibit low performance, especially for target structures that exhibit large interpatient variation in terms of texture, shape, and size.

To overcome this limitation, existing studies propose establishing CNN feature-based self-attention mechanisms. On the other hand, Transformers are designed for sequence-by-sequence

prediction and may be an alternative architecture as they use dispensable convolution operators and rely exclusively on attention mechanisms.

Recently, there has been work combining the advantages of ViT and CNNs. For example, CvT introduced convolutions into ViT to achieve better performance and efficiency, but it was not designed to solve the segmentation problem. TransUNet is the first medical image segmentation framework based on Transformers that has outperformed other state-of-the-art segmentation models. It is also a combination of CNNs and ViT. Medical Transformer (MedT) proposed a gated axial transformer model framework, composed of convolutional layers and gated axial attention layers. The gated axial transformer blocks are then equipped with a LOGO (Local-Global) training strategy, which can fully utilize medical image data to solve the small sample size problem.

Inspired by the two-stage detection and segmentation workflow of MedT and Jocher et al., Su, Liu, Xi, and Hu proposed a novel breast mass segmentation model. Su, Liu, Xi, and Hu used YoloV5 and its latest model to detect the region of interest of the breast mass in a mammogram, crop it directly from the high-resolution image, and use an updated LOGO training strategy to segment the mass from the cropped images.

Due to reduced visibility, low contrast, low sharpness, and noise, a considerable proportion of abnormalities are misdiagnosed or overlooked. Common preprocessing methods, such as filters, have been suggested to improve image quality by smoothing the image and reducing noise. However, selecting the best method to preprocess mammograms to enhance CNN classification remains a challenge. CNNs perform poorly on unbalanced datasets, that is, those with an inequality between positive and negative classes. Finally, CNNs perform poorly in classifying tumors in mammograms with additional malignancies, as they may ignore data in other image regions besides the primary one.

Almost all CNN-based transfer learning approaches for breast cancer detection in mammograms use a transfer learning approach to compensate for the lack of large datasets and to use an optimized model with prior knowledge of features for new tasks.

The ViT concept is a development of the original text-based transformer concept. With a small adjustment to the code to accommodate the various data modalities, it consists of a transformer applied to the image domain. A ViT employs tokenization and embedding techniques,

and its architecture resembles a transformer. A source image is divided into a collection of image fragments known as visual tokens. The visual tokens are embedded into a collection of fixed-dimensional encoded vectors. The Transformer's encoding network, responsible for processing the text input, receives the position of a fragment in the image along with the encoded vector. The ViT encoder is composed of several blocks, each with three main processing components: the layer norm, the multi-head attention network (MSP), and the multilayer perceptron (MLP).

The model is able to adjust to differences in the training images thanks to the layer norm, which keeps the training process in the desired direction. A network called MSP is responsible for creating attention maps from the visual tokens. These attention maps help the network focus on the most relevant areas of the image. The MLP is a two-layer classification network with a GELU (Gaussian Linear Error Unit) at the end. The last block of the MLP, also known as the MLP head, serves as the Transformer's output. SoftMax can be used in this output to provide image classification labels (Ayana, Dese, Dereje, Kebede, Barki, Amdissa, Husen, Mulugeta, Habtamu, Choe, 2023, p. 4).

The authors tested the state-of-the-art object detection model, YoloV5, to detect breast mass in mammograms and provided the specific clipped local images for further segmentation analysis. According to Su, Liu, Xi, and Hu, the advantage of ViT in considering long-range dependencies was fully utilized in the model. The LOGO framework improved the segmentation resolution at the original pixel level and maintained the positional accuracy of the segmentation result in the original image.

Su, Liu, Xi, and Hu (2022) used two mammography datasets, the DDSM Cured Breast Image Subset (CBIS-DDSM) and the INBreast dataset. These data were used to train and test tumor detection and segmentation models. Each of the two datasets consists of multiple pairs of digitized mammography images on film and their corresponding binary images for breast mass segmentation. CBIS-DDSM has its standardized training-test split.

Su, Liu, Xi, and Hu (2022) used the training portion of the CBIS-DDSM dataset as a training set for both tumor detection and image segmentation. Half of the data was used for testing CBIS-DDSM. To test the model's reliability and versatility, the authors calculated the performance on the INBreast dataset and on the other half of the CBIS-DDSM test data. For each image in both sources, it was possible to find one or more breast masses.

Zhou, Li, and Zhao (2022) present a deep learning-based method for extracting breast regions that combines pre-processing techniques. The techniques include noise suppression using median filters, contrast enhancement using CLAHE, and semantic segmentation using Deep lab v3+ models. The method was trained and evaluated on the mini-MIAS dataset and also evaluated on the INbreasts dataset. The results outperform those of recent studies, highlighting the model's ability to maintain accuracy across different datasets.

Li, Chen, Nailon, Davies, and Laurensen (2022) proposed a novel deep learning framework for mammography image processing. The framework involved mass segmentation and simultaneous prediction of diagnostic outcomes. The authors built a context-quality texture learner, known as a Locality-Preserving Learner, using a convolutional block stack to map regions of interest to class labels at a relatively large scale. Secondly, Li, Chen, Nailon, Davies, and Laurensen (2022) applied a Graph Conditional Learner, which combines graphs and CNNs, to learn correlations in relatively small-scale ROIs and utilize the extracted segmentation features to enhance the final quality classification performance. The "Dual Corenet" framework achieves optimal mammographic segmentation and classification, exhibiting superior segmentation performance at both low and high resolutions.

Li, Zhao, Zhang, Lin, and Wang (2021) proposed a novel network architecture for massive image segmentation in digital mammograms. The architecture combines two modules within the core cGAN framework: a superpixel mean clustering layer and a multiscale input module. These modules provide prior boundary information and scale-invariant features. The model's performance is evaluated on large-scale images from two commonly used datasets, CBIS-DDSM and INBREAST. The model achieves good results in terms of Dice and Jaccard scores, accuracy, specificity, and sensitivity. For the CBIS-DDSM dataset, the Dice and Jaccard scores are 93.37% and 87.57%. For the INBREASTS dataset, the Dice and Jaccard scores are 91.54% and 84.40%, respectively.

Chakravarthy and Rajaguru (2022) presented an enhanced version of the U-Net called SKMAT-U-Net. The model incorporated a selective kernel with an attention mechanism to adaptively adjust the receptive fields of the network. It combined feature maps extracted through extended and standard convolution operations. Chakravarthy and Rajaguru (2022) integrated four attention loss functions based on the traditional cross-entropy loss function to form a U-Net using the Mixed

Attention Loss Function. The proposed model segmented lesions in ultrasound images of mammary glands.

Alkhaleefah, Chang, Tan, Chang, Wang, Chang, and Ma (2022) introduced a deep learning model, called Connected-SegNets, which combines two SegNet architectures with hop connections to improve the segmentation of breast tumors in mammograms. Based on the Connected-UNets concept, the authors proposed a model called Connected-SegNets, and used X-ray images. The authors integrated two SegNet architectures with hop connections between their layers. They replaced the cross-entropy loss function of the original SegNet with the intersection-at-union (IoU) loss function. The authors used two encoder networks and two decoder networks. The first decoder network and the second encoder network are connected with additional hop connections after chaining a second SegNet. This helps to recover the fine-grained features lost in SegNet encoding and apply them to encode high-resolution features by connecting them to previously decoded features.

The upper half of the proposed architecture is similar to SegNet, which uses the first 13 convolutional layers of the VGG16 network as the encoder network. In the decoder network, the last convolutional layer is removed. Each encoder network comprises two convolutional cores, which include 3×3 convolutional layers followed by a rectified linear activation unit (ReLU) and a batch normalization layer (BN). Next, a maximum clustering index operation is applied to the output of each encoder network before passing the information to the next encoder.

Each decoder network consists of a 2×2 transposed convolution unit that is concatenated with the output of the previous encoder, and then the result is fed into two convolution blocks, which consist of 3×3 convolutions followed by a ReLU activation and a BN layer. In addition, a second SegNet is connected to the first SegNet via new hop connections that use information from the first upsampling path. The result of the last decoder block is concatenated with the same result after being fed into a 3×3 convolution layer followed by a ReLU activation and a BN layer. This serves as the input from the first encoder network to the second SegNet.

Jabeen, Khan, Balili, Alhaisoni, Almujaally, Alrashidi, Tariq, and Cha (2023) proposed a deep learning model, EfficientNet-b0, for breast cancer classification from mammography images. The proposed framework comprises image acquisition and classification steps. In the first phase, the authors presented a contrast enhancement technique. The resulting enhanced images were used to

train the deep learning model (EfficientNet-b0). The results indicated that the accuracy of the proposed enhancement technique was better than the analysis without the technique, but the accuracy obtained recently was not achieved. Due to this, the authors proposed a fusion technique.

Ghani, Jumaat, Mahmud, Maasar, Zulkifle, and Jasin (2023) presented two SOM-based machine learning models for mammography image segmentation. The models were named SSOM and SSOMH. They used an idea in the model that consists of spreading the intensity values using a piecewise linear function, which was applied to increase the contrast of the mammography images.

To minimize the energy functions of the SSOM and SSOMH models, Euler-Lagrange equations were established using the calculus of variations. Then, the equations were solved in MATLAB software using the gradient descent algorithm. The efficiency of each model was evaluated in terms of computational time, and the segmentation accuracy was measured by evaluating the JSC, DSC, precision, and error values for each image. After testing with images from the INbreast database, the authors indicated that the SSOMH model was more efficient for segmenting mammography images due to its accuracy compared to the SSOM model.

In the same experiment, SSOMH outperformed the SOMCV, IIS, U-Net, and PD2 models in terms of accuracy. The authors used images from the CBIS-DDSM database. After testing, Ghani, Jumaat, Mahmud, Maasar, Zulkifle, and Jasin (2023) found that the U-Net architecture has a faster test time but a slower training time compared to SSOMH. However, the SSOMH segmentation process was faster than the other models compared (Ghani, Jumaat, Mahmud, Maasar, Zulkifle, Jasin, 2023, p. 17).

Alhussen, Haq, Khan, Mahendran, and Kadry (2025) proposed an innovative architecture that combines accurate detection, ROI segmentation, and interpretable explanation in mammograms. The authors developed a precise and relevant method for breast cancer detection using XAI-based ROI selection from mammography images. The model is called XAI-RACapsNet. It consists of five consecutive processes: (a) data acquisition, (b) two-level preprocessing, (c) ROI segmentation, (d) feature extraction and selection, and (e) classification. Initially, the input image undergoes two-level preprocessing using median filtering and the CLAHE model. The results of CLAHE and median filtering are named ImgCLAHE and ImgMF, respectively. XAI O-Net is then applied to segment the preprocessed image, resulting in ImgROI. After segmentation, features such as Texture (Ftext), Spectral (Fspectral), and Spatial (Fspatial) are extracted from ImgROI using the

Adaptive Feature Module. A Relevance-aware capsule network (RACapsNet) was employed to classify the chosen features and provide the outputs.

Iqbal and Sharif (2023) proposed a hybrid BTS-ST model, composed of a BTS-ST vision transformer-based network for segmentation and classification of breast tumor images. The primary structure of the BTS-ST utilized a dual encoder using U-Net and Swin-Transformer, Spatial Interaction Block (SIB), Feature Compression Block (FCB), and Relationship Aggregation Block (RAB). The BTS-ST model achieved a segmentation accuracy of 0.996, superior to other models such as U-Net (0.984), TransUNet (0.990), and TransFuse (0.992). The hybrid model integrated CNNs and transformers and obtained better results than methods based purely on CNNs, such as U-Net. Currently, the article has 82 citations since 2023, indicating its use by authors in the field. In the model, from input to output, there is a first encoding stage (primary encoder) and a second decoding stage (primary decoder) that follow a U-shaped flow. Because the encoder-decoder structure forms this U, the architecture can be described as a U-Net type. In parallel, it has transformers that complement the model's architecture.

Mohammadi and Livani (2025) proposed a hybrid CNN + transformer model, called Hybrid Transformer U-Net (HTU-net), developed for breast mass segmentation in mammograms. The authors used the databases: CBIS-DDSM and INBreast. The spatial and channel self-attention mechanisms were improved in the model and integrated into the convolution layers in the HTU-Net, creating a hybrid architecture. The architecture combines the strengths of CNNs and ViTs. The authors introduced a multiscale attention mechanism that further enhances the model's ability to merge information from different resolutions, improving the decoder's ability to reconstruct fine details in the segmented output. Mohammadi and Livani (2025) improved the model for capturing local texture-based features and global contextual information, aiming for better segmentation results.

Mahmood, Saba, Rehman, and Alamri (2024) developed an integrated deep learning-based architecture (CNN+LSTM and CNN+SVM) for rapid breast malignancy classification. Lee, Dunnmon, He, Tang, Ré, and Rubin (2021) compared distinct segmentation techniques. The study compared computer-aided diagnostic (CADx) approaches in mammograms, divided into two groups: (1) Segmentation-dependent methods, which require prior segmentation of the mass, followed by feature extraction (shapes, intensities, spiculations) and application of classifiers such as LDA or

small neural networks; (2) Segmentation-free methods, which learn representations directly from the image without prior segmentation, using approaches such as Bag-of-Visual-Words (BoVW) or CNN. They used mammograms from the CBIS-DDSM dataset.

The authors concluded that segmentation-free approaches (CNN) had an AUC of 0.86, outperforming traditional methods that use prior segmentation comparators (AUC between 0.58 and 0.75). Based on these findings, the authors concluded that segmentation-free approaches, especially those based on deep learning, are preferable for diagnostic tasks in mammography, as they offer better accuracy without the need for manual segmentation or intermediate heuristics.

The article by Maqsood, Damasevicius, and Maskeliuna (2022) compared CNN architectures and proposed a new model. They concluded that the GoogLeNet model has the best performance compared to other deep CNN models (88.54% accuracy rate and 11.46% error rate). Next, the best model was ResNet-101 (88.12% accuracy and 11.88% error rate). The next model in the ranking was ResNet-18 (87.68% accuracy rate and 12.32% error rate). The VGG-16 model did not achieve the same results (81.41% accuracy rate and 18.59% error rate).

The article by Maqsood, Damasevicius, and Maskeliuna (2022) has 133 citations, and is therefore used by researchers in the field. Their distinguishing feature appears to have been the performance of a comparative analysis of existing CNN models, presenting the metrics used for evaluation and the informative graphics. In addition, they not only evaluated the different CNNs for the model, but also used them to build a model, called TTCNN (Transferable Texture Convolutional Neural Network), an integrated system for diagnosing breast cancer in mammograms. Based on the set of studies presented, an evolution can be observed from traditional CNN architectures, such as U-Net, ResNet, DenseNet, Xception, and SegNet, to hybrid architectures that combine convolutional networks and Transformers, such as TransUNet, in addition to approaches that incorporate attention mechanisms. CNNs have an advantage in extracting local features, textures, and spatial details, while Transformers, through self-attention mechanisms, can model long-range dependencies and global relationships in the image, an aspect that is particularly relevant in mammograms, where lesions can present great variation in size, shape, and texture.

For a better evaluation of the articles, it would be necessary to reproduce the experiments to compare the metrics with the results presented by the authors, which was not done and would take some time, considering the number of articles reviewed. Without a test

with the algorithms presented by the authors, or without evaluating the quality of the database, it is not possible to make a general comparison of the works, which constitutes a limitation of this review in terms of a comparison of the results between the main articles.

5 FINAL CONSIDERATIONS

In this literature review, 51 original articles published between 2021 and 2025 in the Elsevier, IEEE, MDPI, and Springer journals were selected. Several models have been published in recent years, mainly based on convolutional neural networks. The existence of public databases with a significant amount of data, as well as well-established evaluation metrics, has allowed research groups from different locations to publish works that present comparisons with other models in a clear way.

The choice to focus on the period between 2021 and 2025 led to a selection in which the methodologies found were already based on an architecture of modern methods, such as machine learning, and very rarely on the classic image segmentation methods, which were more commonly used in the early 2000s.

The results indicate that some models achieve accuracy when compared to manual tracings by experts that are greater than 95%. Current challenges consist of the computational costs and processing time of some works, as well as the existence of noise in many images present in the databases, which require refinement. However, the field of image processing for mammography applications has expanded substantially, with publications in high-impact journals from all continents of the planet.

The fact that breast cancer is a disease with visibility and public research interest implies a substantial incidence of original articles that deal with methods to collaborate with the detection of regions of interest by computer. As a consequence, the volume of work on breast cancer is greater than the volume of articles on cancers of other parts of the human body.

In the last decade, with the development of methods based on machine learning and deep learning, modern segmentation techniques have excelled in relation to classical segmentation, which had also been obtaining good results in recent years. In parallel with this, popular interest in the

aforementioned computational methods has also increased, with the dissemination of tools based on generative artificial intelligence, from free websites accessible to the general public.

The development of the field of deep learning applied to medicine is also due to a search for robust methods that perform all image handling steps automatically after an input is provided to the algorithm. This is justified because fewer human interventions throughout the process can mean less variability and greater model confidence, since it has clear inputs and outputs.

At least ten original studies were identified with detailed methodology and concrete results using the U-Net architecture, being the most recurrent type of strategy in the analyzed articles. About eight studies were identified using variations or adaptations of the ResNet architecture, being a widely used option by researchers for the purpose of mammography segmentation. Two studies based on the VGG16 architecture and based on AlexNet were also found.

The most cited article among the selected works was the U-Net model by Salam and Aly, published in 2021 in the Alexandria Engineering Journal, from Elsevier, with 240 citations. Because it was published four years ago, the time since publication is one of the factors for the greater number of citations compared to other more recent works. In recent years, models with transformers have also been successful and cited by other researchers. This is the case of the work by Su, Liu, Xi and Hu (2022), which obtained 124 citations.

In recent years, most deep learning-based image segmentation models use deep convolutional neural networks (CNNs) based on the U-Net architecture to achieve their best performance. Currently, some high-configuration GPUs are designed for convolution computation, which makes CNNs faster and eventually allows CNNs to continuously flourish in the field of computer vision.

However, a major limitation of these CNN models is that they pay too much attention to the local pixel square. In fact, a deep learning architecture called Vision Transformer (ViT) has recently been leading the trend of replacing CNNs in computer vision. Compared to traditional CNNs, ViT is powerful for capturing local and global or long-range visual dependencies through its self-attention mechanisms (global attention obtained by course or refined local attention). These advantages of ViT make it a good option for solving the mammography segmentation problem.

The main challenge of the ViT-based medical image segmentation model is that training this model requires high computational power. Due to GPU limitations and other hardware constraints, current computational resources can only segment low-resolution images. In light of this, CNN strategies are being combined with ViT in some computer vision studies in different areas.

The article “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation”, by Chen and colleagues (2021) presents a model with integrative characteristics, combining U-Net CNN resources with transformers. The publication has 6361 citations and paved the way for other articles that explore transformer architecture.

In the literature review conducted in this article, the DDSM database was the most used, having been employed in at least nineteen works selected for the review. Its large amount of data and free and easy access are factors that contributed to its dissemination among articles on mammographic image processing. Eight articles were found using the INBreast database, indicating its recurrence among studies in the area.

U-Net-based models are widely used for breast lesion segmentation when compared to InceptionV3, DenseNet121, VGG16, ResNet50, and MobileNetV2. The DDSM database provides better results compared to other free databases, such as MIAS and CBIS-DDSM. The data augmentation technique incorporated in the U-Net model achieved 98.87% accuracy according to Salama and Aly (2021).

Classification models are divided into statistical machine learning-based classifier models and deep learning-based classifier models. It is observed that the convolutional neural network (CNN) has been the most widely used model for image segmentation and classification based on deep learning. Recently, the Vision Transformer model has been used as a modern alternative to CNNs. Other classifiers that have been used for longer are: kNN, ANN, random forest, AdaBoost, and kNN.

Until the 2010s, there was a trend towards Support Vector Machine (SVM) methodologies, according to Thakur, Kumar, and Kumar (2023). The main disadvantage of the SVM-type machine learning approach is that features are extracted implicitly. By opting for CNNs, this problem can be avoided, as deep learning models extract and learn high-level features in their own way. In the 2010s

and around 2020, many works based on CNNs and U-Net, the model that led the publications, were published. In the 2020s, alternative works using the transformer methodology are being published.

There are still challenges to overcome in the segmentation of mammography images, and more applications are needed to improve the accuracy of breast cancer detection. One of the challenges is the lack of balance in the available datasets on breast cancer. Medical research can be limited by the lack of information from some informative datasets that are not publicly accessible or annotated. In order to improve accuracy, recent deep learning techniques can be used, such as Transformers, as they have demonstrated improvements in image segmentation. In recent years, some artificial intelligence (AI) tools have been approved by the Food and Drug Administration (FDA) in the United States to assist radiologists in decision-making. The approvals are an indication that the methodologies in the field are mature enough to be applied in real-world contexts.

Science is done collaboratively and its purpose is social well-being. Each article can be useful for other research by someone else. This article did not evaluate each paper very thoroughly, but it may help other future studies.

REFERENCES

- Alhussen, A.; Haq, M.; Khan, A.; Mahendran, R.; Kadry, S. XAI-RACapsNet: Relevance aware capsule network-based breast cancer detection using mammography images via explainability O-net ROI segmentation. *Expert Systems With Applications*, 2025.
- Alkhaleefah, M.; Tan, T.; Chang, C.; Wang, T.; Ma, S.; Chang, L.; Chang, Y. Connected-SegNets: A Deep Learning Model for Breast Tumor Segmentation from X-ray Images. *Cancers*, 2022.
- Ansar, W.; Raza, B. Breast Cancer Segmentation in Mammogram Using Artificial Intelligence and Image Processing: A Systematic Review. *Current Chinese Science*, 2023.
- Ayana, G.; Dese, K.; Dereje, Y.; et al. Vision-Transformer-Based Transfer Learning for Mammogram Classification. *Diagnostics*, 2023.
- Baccouche, A.; Garcia-Zapirain, B.; Castillo Olea, C. et al. Connected-UNets: a deep learning architecture for breast mass segmentation. *npj Breast Cancer*, 2021.
- Chakravarthy, S.; Rajaguru, H. SKMAT-U-Net architecture for breast mass segmentation. *International Journal of Imaging Systems and Technology* (2022).
- Chen, J.; Lu, Y.; Yu, Q.; et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv:2102.04306*, 2021.
- Chen, J.; Mei, J.; Li, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 2024.

- Ghani, N.; Jumaat, A.; Mahmud; et al. Breast Abnormality Boundary Extraction in Mammography Image Using Variational Level Set and Self-OrganizingMap (SOM). Mathematics, 2023.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. arXiv:1512.03385, 2015.
- Iqbal, A.; Sharif, M. BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images. Knowledge-Based Systems, 2023.
- Jabeen, K.; Khan, M.; Balili, J.; et al. BC2NetRF: Breast Cancer Classification from Mammogram Images Using Enhanced Deep Learning Features and Equilibrium-Jaya Controlled Regula Falsi-Based Features Selection. Diagnostics, 2023.
- Lee, R.; Dunnmon, J.; He, A.; et al. Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset. Journal of Biomedical Informatics, 2021.
- Li, H.; Chen, D.; Nailon, W.; et al. Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography. IEEE Transactions Medical Imaging, 2022.
- Li, J. et al. Breast Cancer Segmentation Methods: Current Status and Future Potentials. Biomed Research International, 2021.
- Li, Y.; Zhao, G.; Zhang, Q.; Lin, Y.; Wang, M. SAP-cGAN: Adversarial learning for breast mass segmentation in digital mammogram based on superpixel average pooling. Med Phys, 2021.
- Mahmood, T.; Saba, T.; Rehman, A.; Alamri, F. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. Expert Systems With Applications, 2024.
- Mao, N. et al. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. British Journal of Cancer, 2022.
- Maqsood, S.; Damasevicius, R.; Maskeliunas, R. TTCNN: A Breast Cancer Detection and Classification towards Computer-Aided Diagnosis Using Digital Mammography in Early Stages. Applied Sciences, 2022.
- Marzuli, V. The CALMA Project: Computer Assisted Library for Mammography. Neural Nets WIRN VIETRI-98, 1999.
- Mohammadi, S.; Livani, M. Enhanced breast mass segmentation in mammograms using a hybrid transformer UNet model. Computers in Biology and Medicine, 2025.
- Nour, A.; Boufama, B. Hybrid deep learning and active contour approach for enhanced breast lesion segmentation and classification in mammograms. Intelligence-Based Medicine, 2025.
- Rezaei, Z. A review on image-based approaches for breast cancer detection, segmentation, and classification. Expert Systems with Applications, 2021.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, Springer, 2015.
- Sahu, A.; Das, P.; Meher, S. And eficiente deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. Biomedical Signal Processing and Control, 2024.
- Salama W, Aly, M. Deep learning in mammography images segmentation and classification: automated CNN approach. Alexandria Engineering Journal.2021.

Su, Y.; Liu, Q.; Xi, W.; Hu, P. YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine (Elsevier), 2022.

TileStats Portal. U-Net clearly explained. Available at: <<https://www.tilestats.com/>>. Accessed in: 30. Jan. 2026.

Thakur, N.; Kumar, P.; Kumar, A. A systematic review of machine and deep learning techniques for the identification Multimedia Tools and Applications and classification of breast cancer through medical image modalities. Multimedia Tools and Applications, 2024.

Tiryaki, V.; Kaplanoglu, V. Deep Learning-Based Multi-Label Tissue Segmentation and Density Assessment from Mammograms. IRBM (Elsevier), 2022.

Wang, L. Mammography with deep learning for breast cancer detection. Frontiers in Oncology, 2024.

Zhang, J.; Wu, J.; Zhou, X.; Shi, F.; Shen, D. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. Seminars in Cancer Biology, 2023.

Zhou, K.; Li, W.; Zhao, D. Deep learning-based breast region extraction of mammographic images combining pre-processing methods and semantic segmentation supported by Deeplab v3. Technol Health Care, 2022.