Computers in Biology and Medicine 185 (2025) 109520

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Towards an interpretable breast cancer detection and diagnosis system

Cristiana Moroz-Dubenco *, Adél Bajcsi, Anca Andreica, Camelia Chira

Babeş–Bolyai University, Mihail Kogălniceanu 1, Cluj-Napoca, 400084, Cluj, Romania

ARTICLE INFO

Keywords:
Digital mammogram processing
Breast tissue segmentation
Lesion classification
Interpretability

ABSTRACT

According to the World Health Organization, breast cancer becomes fatal only if it spreads throughout the body. Therefore, regular screening is essential. Whilst mammography is the most frequently used technique, its interpretation can be challenging and time-consuming. For this reason, computer-aided detection and diagnosis systems are increasingly being used for second opinion. However, in order for doctors to trust such systems, they need to understand their decisions. We propose an automated and interpretable system for the detection and diagnosis of breast cancer, encompassing five steps. After a robust pre-processing and an unsupervised segmentation, we analyze four feature extraction techniques, both textural and shape-based, and three methods for feature selection. To facilitate interpretation, we employ the Decision Tree algorithm for benign/malignant classification and experiment with different methods to avoid overfitting: pre-pruning, post-pruning, and ensemble-based (Random Forest classifier). Our system reaches a maximum accuracy of 95% and 100% precision and specificity when tested on images from the mini-MIAS dataset, while also offering its users the possibility to analyze each of the steps.

1. Introduction

Breast cancer is a major concern around the world, being the most common type of cancer and the leading cause of cancer deaths among women worldwide. A statistic from the European Cancer Information System reports that the number of new cases increases over the years, reaching 2.74 million in 2022 (13.8% of the total cancer cases) [1].

In most cases, breast cancer can be curable by surgery, radiation therapy, and chemotherapy, if detected in its early stages. When detected later, the probability of metastasis increases significantly, which can lead to a lethal outcome. Therefore, regular screening is crucial for early detection and a better chance of successful treatment.

Mammography is the most frequently used technique for detecting early signs of breast cancer. However, its interpretation can be challenging, even for experienced specialists. Computer-aided detection and diagnosis (CAD) systems have been developed to assist radiologists in detecting breast cancer based on mammograms. These systems serve as a second opinion to enhance human expertise (rather than to replace it).

Artificial intelligence (AI) is widely used in CAD systems. However, in critical fields such as healthcare, where lives are at stake, the interpretability of these systems is crucial [2]. This need has drawn increasing attention to interpretable AI in recent years. Interpretability refers to a system's ability to present the internal processes in a manner that is easily understandable to the end user. The field of interpretable AI is relatively new and rapidly emerging, with only a limited number of studies addressing the need for transparency and understanding in AI-driven decision-making systems.

In general, CAD systems utilize image processing and AI techniques to automatically detect and diagnose lesions that indicate breast cancer. These systems usually consist of five steps:

1. *pre-processing*: enhancing the quality of the mammogram,
2. *segmentation*: defining the possible location of the lesion (region of interest – ROI),
3. *feature extraction*: extracting characteristics from the ROI,
4. *feature selection*: selecting the most descriptive features, and
5. *classification*: deciding on the type of the lesion.

The purpose of CAD systems is to help the radiologists' work. Therefore, an important factor in these systems is their interpretability [2]. The transparency of the system ensures that radiologists can understand and interpret the steps that the CAD system undertakes to reach its diagnosis. The purpose of the current study is to build a fully automated CAD system that detects possible lesions and classifies them into benign and malignant classes with the ability to visually explain the result of every step. To achieve our objective, we propose the use of an unsupervised segmentation method to detect the ROI, the computation and selection of easily explainable features, and the Decision Tree (DT) method with its variants for classification.

* Corresponding author.
E-mail address: cristiana.moroz@ubbcluj.ro (C. Moroz-Dubenco).

https://doi.org/10.1016/j.compbiomed.2024.109520
Received 22 June 2024; Received in revised form 9 November 2024; Accepted 2 December 2024
Available online 16 December 2024

Decision Tree models are preferred over artificial neural networks (ANNs) in applications requiring interpretability due to their inherently transparent structure. Unlike ANNs, which operate as "black-box" models with complex layers of interconnected neurons, DTs offer clear rule-based paths that can be easily visualized and understood by humans. Although there are initiatives to improve the explainability of ANNs [3,4], they remain challenging to validate, as the internal representations and decision-making processes of ANNs are often too complex to be fully understood or trusted without extensive analysis.

Explainability and interpretability, while related, refer to distinct aspects of understanding machine learning models. Interpretability is the degree to which a human can comprehend the cause–effect relationships in a model; it focuses on transparency and ease of understanding, especially in simpler models like decision trees. Explainability, on the other hand, refers to the ability to describe how a more complex, often black-box model (such as deep learning) arrives at a specific decision or prediction, typically through post-hoc methods like feature importance or visual heatmaps. Simply put, interpretability is centered on the inner workings of a model, while explainability, on the decisions made. In the current work, we are focusing on interpretability, which provides a greater level of detail.

For pre-processing, we employ the technique proposed in [5,6], while segmentation is done using the Threshold-based GrowCut (TbGC) algorithm [7,8]. For feature extraction, we analyze four methods: Local Binary Pattern (LBP), Gray-Level Run-Length Matrix (GLRLM), Geometrical, and Contour-based. For the selection step, we experiment with Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Linear Discriminant Analysis (LDA). Finally, we explore three methods to avoid the overfitting of the Decision Tree method, namely, pre-pruning, post-pruning, and ensemble-based classification (also known as Random Forest - RF). We evaluate all the possible parameter combinations on images from the mini-MIAS (Mammographic Image Analysis Society) dataset [9], and choose the one that produces the best results. Additionally, the resulting system is tested on the mini-DDSM dataset [10] in order to prove its ability of generalization and robustness.

The novelty of the proposed system lies in its interpretability, which is achieved by visually explaining the outcome of each step in the process. In healthcare, datasets are often limited in size, but our system can be easily and efficiently fine-tuned to accommodate various datasets. Furthermore, we employ a novel segmentation method [7,8], used for the first time in combination with the aforementioned methods for feature extraction, feature selection, and classification, to detect the ROI in mammograms. By using TbGC, our system becomes fully automated. The proposed system classifies lesions into benign and malignant categories with an accuracy of 95.0% and a precision of 100% on the mini-MIAS dataset [9].

1.1. Background

Various systems for the detection and diagnosis of breast cancer have been proposed so far. In the following, we highlight key contributions related to breast cancer classification presented in the literature.

The authors of [11] introduce a CAD system for the classification of benign/malignant lesions in 2018. Pre-processing is done through median filtering, and the suspicious region is cropped and segmented using Otsu's thresholding method. Textural features are then extracted from the Gray Level Co-Occurrence Matrix (GLCM), and the most relevant ones are selected with the K-Means clustering technique. Using a Decision Tree algorithm for classification, the system achieves a 95% accuracy score when tested on images from the MIAS dataset.

In [12] (2018), a five-step system is proposed to classify mammograms into benign and malignant classes. Starting by employing adaptive median filtering to remove the noise, the image is further segmented with the Fuzzy C-Means clustering method and post-processed using morphological techniques in order to determine the boundary of the lesion. Statistical and textural features are extracted from the ROIs and classified with the Random Forest algorithm. The proposed approach was tested on 100 images from MIAS and reached an accuracy of 90.47%.

Boudraa et al. (2020) [13] proposed a CAD system for breast cancer classification based on a two-step approach: feature extraction and classification. The authors used the ground truth provided in the dataset to define the ROI, followed by super-resolution in order to enhance image quality and obtain more visual details. The feature extraction combined statistical and texture features (GLRLM). Nine classification algorithms, including Random Forest, were included in the experiments. The best results were obtained using logistic regression-based classifier, with an accuracy of 96.7% and a sensitivity of 94.7%, based on 93 images from the MIAS dataset, while the Random Forest-based approach achieved an accuracy of 94.6% and a 94.7% sensitivity score.

[14] (from 2021) presents a breast cancer diagnosis system which relays on a nature-inspired algorithm for feature extraction and employs both DT and Gradient Boosting algorithms for classification. The features are extracted through the Particle Swarm Optimization technique, using a validation set's accuracy as fitness function. Afterward, the features are fed to the classifiers, obtaining 85% accuracy score for DT and 92% for Gradient Boosting.

Paper [15] proposed a CAD system for the segmentation and classification of benign and malignant lesions from mammograms in 2022. The pipeline starts with manually cropping the images to obtain approximate ROIs of 700 × 700 pixels, which are subjected to contrast adjustment and noise reduction. From these ROIs, exact breast masses are segmented using a Region Growing technique, which uses the pixel with the highest intensity as the initial seed. Subsequently, multiple types of feature are extracted from the segmented region, both geometrical and textural, from which the top twenty most exclusionary (based on their discriminating power) are selected using the Relief-F algorithm. These features are then classified with nine algorithms: K-Nearest Neighbor, Support Vector Machine, Gaussian Mixture Model, Multi-class Support Vector Machine, Decision Tree, Discriminate Analysis, Naive Bayes, Random Forest and Ensemble Tree. The proposed system is tested on 317 mammograms from the IRMA version of the DDSM LJPEG dataset and the best results are obtained with the Support Vector Machine classifier - 75.8% accuracy, 72% sensitivity and 79% specificity -, while the Decision Tree algorithm reaches a maximum accuracy of 67.3% and the Random Forest algorithm, an accuracy score of 70.7%.

Bukhori et al. [16] employed a computer vision method to classify breast cancer lesions from mammography and published their results in 2023. The proposed methodology is tested on images from the MIAS dataset. After rotating and cropping the images to a ratio of 7:5, contrast stretching and Gaussian filtering are applied. Textural features are extracted from the GLCM. The dataset is then split into training and testing sets using different ratios, and classification is performed using the RF method with various numbers of trees. The best results were obtained for the 80:20 training-testing data ratio and 600 trees: 70.8% accuracy, 85.7% precision, and 25% recall.

Studies from recent years proposed the use of explainable AI. Balve and Hendrix [17] published an article in 2024 comparing the heathmaps of different post-hoc methods for prediction explanation. After building a custom CNN model and training it on MIAS [9] dataset (achieving 77% accuracy), they used LIME [18], Kernel SHAP [19], and Grad-CAM [3] to generate heatmaps for the predictions. The authors concluded that Grad-CAM is the most time efficient and also generates the most accurate heatmaps. Lampour et al. [20] trained five frequently used CNNs (VGG16, VGG19, ResNet50, Xception, and MobileNet) on Chinese Mammogram Dataset [21] and reported 80.02%, 84.79%, 78.15%, 79.09% and 74.95% accuracy, respectively. Additionally, the authors applied Grad-CAM [3] to generate heatmaps for the predictions and concluded that the attention-maps generated for Xception

are the most accurate (achieving 96.68% intersection over union on the InBreast dataset [22]). Lou et al. [23] proposed an improved ReNet50 architecture by extending the original model with an attention module (ECA). The model achieved 96.9% accuracy on the InBreast dataset [22]. The authors also used Grad-CAM [3] to generate heatmaps for the predictions to prove the efficiency of the proposed model.

Another approach to explainable AI is the development of self-explainable models. Chen et al. [4] proposed a patch-based self-explainable model named ProtoPNet. In [24], the authors trained ProtoPNet on the CBIS-DDSM dataset [25] and reported a test accuracy of 68.5%.

In this study, we propose a fully automated and interpretable CAD system for the detection of breast cancer. The methods used are selected such that they can be visualized by the user for validation. The mammograms from the mini-MIAS dataset are used and a comprehensive analysis is presented.

The proposed system is compared to existing CAD systems, leading to the conclusion that our approach can surpass other techniques in terms of performance, whilst being interpretable and completely independent of human intervention.

The rest of the paper is structured as follows. Section 2 describes the current approach. Section 3 presents the results of the conducted experiments, and Section 4 holds the comparison to existing approaches, followed by the conclusions and future work in Section 5.

2. Materials and methods

In the current paper, our aim is to create a CAD system to help doctors in the diagnosis of breast cancer in an early stage. As mentioned in Section 1, support systems generally consist of five steps. Fig. 1 details the structure of the proposed system. . In the following subsections, the methods used are detailed for every step.

2.1. Pre-processing

Mammograms are X-ray images of breast tissue. These images usually have low quality, and their intensity can vary due to the machinery utilized. Therefore, the objective of pre-processing is to enhance the mammograms and prepare for segmentation. In the following paragraphs, we detail the steps of the pre-processing, which are illustrated in Fig. 1a.

In the proposed system, pre-processing starts with the removal of noise and artifacts, followed by enhancing the contrast of the underlying tissue structures as presented in [8]. First, labels are discarded using thresholding. After the preliminary experiments presented in [26], we decided to set the threshold to 50 for each image. The result of the thresholding is presented in Fig. 1aI. The largest component of the binary image is used as the breast mask, as shown in Fig. 1aII. In the next step, the pectoral muscle is removed. In [26], various methods were tested, and the Seeded Region Growing (SRG) algorithm produced the best results. Given its strong performance and ease of understanding, we chose to incorporate SRG into our system. The result of this step holds the breast tissue on the mammogram as presented in Fig. 1aIII. To further enhance the images, morphological operations such as dilation and erosion are applied to remove noise, particularly microcalcifications. After noise removal, contrast-limited adaptive histogram equalization (CLAHE) is applied to increase the contrast of the tissue (Fig. 1aIV).

We highlight the fact that, for better understanding, the pre-processing technique can be displayed step by step, as shown in Fig. 1a.

2.2. Segmentation

In order to identify the region of interest from the pre-processed mammograms, we use the Threshold-based GrowCut (TbGC) segmentation method proposed in our previous work [7,8], which is an improved and unsupervised version of the GrowCut algorithm [27]. In the following paragraphs, we detail the steps of the segmentation, which are illustrated in Fig. 1b.

Being provided with initial foreground and background seeds, characterized by their feature vectors, labels, and strengths (i.e. the confidence that their label is correct), the GrowCut algorithm iterates over the image until all the pixels are labeled. TbGC improves the original algorithm in four aspects:

1. Background seeds selection — the darkest pixels from the image are selected as background seeds, hence discarding the need for manual seed selection.
2. Foreground seeds selection — because abnormal tissue appears brighter on mammograms, the 25-pixel radius (as experimentally chosen in [28]) circle with the highest sum of pixels' intensities is identified, and all the pixels within this circle are used as foreground seeds, completely removing the need for human intervention.
3. Computational time — the algorithm uses Cellular Automation and, thus, yields a result when the automation converges. Taking into consideration the fact that medical images are typically larger than regular images, in order to reduce the computational time, the number of iterations of the GrowCut algorithm is limited to a fixed number (θ_i). For mini-MIAS, 5 iterations are enough to obtain satisfactory results [7]. This limitation not only drastically increase the computational efficiency, by iterating over the image 5 times instead of approximately 100 times, which is how much the automaton needs to converge [29], but it also increases the accuracy, since a complete run leads to more false positives, as shown in [7,29].
4. Performance — to maintain a high level of accuracy despite all the changes made to the original GrowCut algorithm, the cell evolution rule is modified such that a pixel's label is updated only if its new strength exceeds a certain threshold (θ_t). From the experiments presented in [7], it can be concluded that this value should be equal to 0.5.

As detailed in [7,8], TbGC converts the semi-supervised GrowCut (requiring human expert intervention for initial seed selection) into completely unsupervised, by generating the seeds in an automated manner. Furthermore, TbGC also brings an increase in performance, due to the change in the cell evolution rule, obtaining an accuracy score 22% higher than the one obtained with the original GrowCut.

Towards our goal of building an interpretable system, we emphasize the fact that the rule for background seeds selection – using the darkest pixels from the image – can be easily identified using a pre-processed mammogram, where all the pixels aside from the breast are turned into black. The foreground seeds selection, on the other hand, can be seen as an iterative process, where the image is parsed pixel-by-pixel, and each pixel is considered the center of a circle with a radius of 25 pixels. The brightness of this circle is computed and compared to the previous maximum, and, if it is higher, it becomes the foreground seeds circle. Fig. 2 presents the process of choosing the foreground seeds step-by-step, with the foreground seeds circles illustrated in blue. Fig. 2(a) depicts the starting point of the algorithm and Fig. 2(h) depicts its result, while Figs. 2(b)–2(g) depict some of the intermediate steps.

Moreover, since the segmentation algorithm is based on a Region Growing technique and limited to a small number of iterations, the result of each iteration can be displayed so that the users can follow and understand the process. Fig. 3 holds the results yielded by TbGC at each iteration. Although the segmentation is performed on the pre-processed images, we chose to output its results on the original images

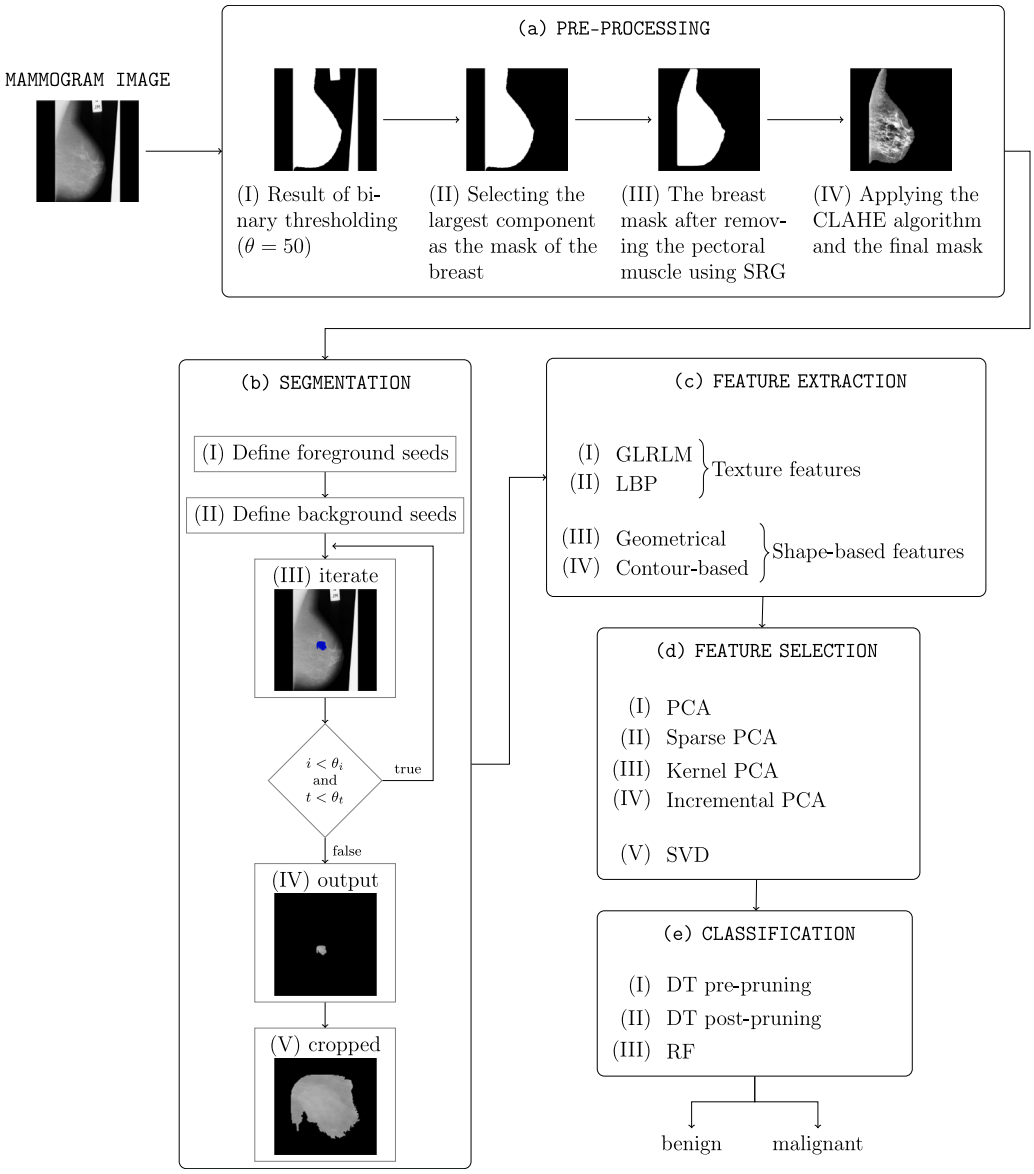


Fig. 1. The structure of the proposed system. The input is a mammogram, which is pre-processed, segmented, and features are extracted. The features are then selected, and the classification is performed. The output is the classification result.

to facilitate interpretation. If required, all of the changes in the pixels' labels could be displayed, thus presenting the entire, complete process of region growing. Fig. 3(f) holds the result of the segmentation step, that is, the region of interest.

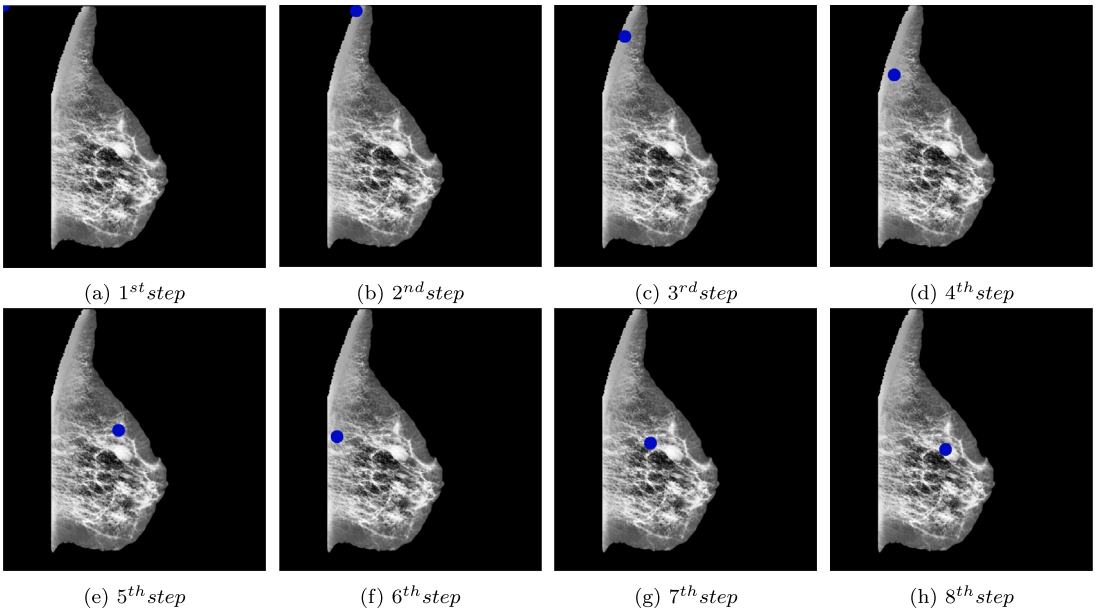
As previously explained, TbGC's execution can be visualized step by step, hence making it a suitable choice for our interpretable CAD system. Moreover, on the mini-MIAS dataset, it yielded an accuracy of 98.52% and a precision score of almost 70% [8], which is comparable to other unsupervised, yet more difficult to understand, segmentation techniques. Therefore, our choice of algorithm was motivated by the robustness, time efficiency, and interpretability of the Threshold-based GrowCut segmentation method.

2.3. Feature extraction

Feature extraction plays a key role in an automated breast cancer detection system. There are two main types of feature extraction methods: texture- (Local Binary Pattern and Gray-Level Run-Length Matrix) and shape-based (geometrical and contour-based). In the current approach, the performance of different features is compared. The methods were chosen according to our interpretability goal: the resulting features can either be visualized or easily computed. This characteristic can prove very useful to the end-user, not only for better understanding the feature extraction process, but also for validating each step separately, such that, if a decision of the system is not in alignment with the medical expert's opinion, they can clearly tell which step of the process caused the difference and decide which outcome (theirs or the system's) is more likely to be correct. In the following paragraphs, the extracted features (also listed in Fig. 1c) are detailed.

2.3.1. Local Binary Pattern

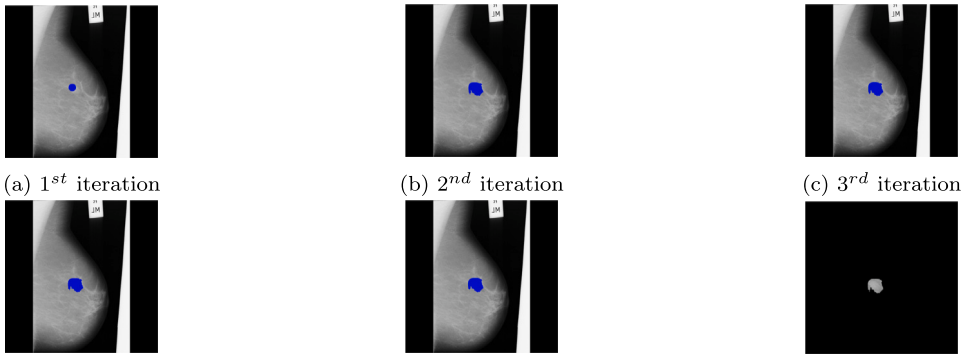
Local Binary Pattern (LBP) [30] compares the intensity of each pixel with its neighboring pixels by thresholding the difference between the intensity of the center pixel and the intensities of its neighbors and considering the results as binary numbers, thus obtaining a binary pattern that describes the local structure of the image. LBP features are extracted from an image by dividing it into small regions and computing the distribution of binary patterns for each region. These patterns are then concatenated, forming a feature vector that describes the texture of the image.



(a) 1^{st} $step$ (b) 2^{nd} $step$ (c) 3^{rd} $step$ (d) 4^{th} $step$

(e) 5^{th} $step$ (f) 6^{th} $step$ (g) 7^{th} $step$ (h) 8^{th} $step$

Fig. 2. Step-by-step selection of foreground seeds.



(a) 1^{st} iteration (b) 2^{nd} iteration (c) 3^{rd} iteration

(d) 4^{th} iteration (e) 5^{th} iteration (f) *Result*

Fig. 3. Step-by-step segmentation.

Table 1
Variants of LBP.

Name	Grayscale invariant	Rotation invariant
default [30]	✓	
ror [31]	✓	✓
uniform [32]	✓	✓
nri_uniform [33]	✓	
var [34]		✓

Table 2
LBP parameters for classification methods.

Pruning	Method	Radius	Multiplier
Pre	default	9	2
Post	default	9	2
Ensemble	default	7	16

Various extensions to the original LBP method have been introduced over the years, introducing rotation invariance, uniform patterns, and neighborhoods of different sizes [31–34]. In our experiments, we evaluate five variants of the LBP technique listed in Table 1.

Other important parameters of the LBP method are the neighborhood radius and the number of neighbors considered for every pixel. In order to find the variant that best fits our needs, we experiment with radii in the set of $[3, 5, 7, 9]$, and set the number of pixels as a multiple of the radius value, employing values in the range of $2^1 - 2^4$ as multipliers. As mentioned in Section 1 and detailed in Section 2.5, we employ the Decision Tree algorithm with three types of pruning for classification, and therefore we need to find the best combination of parameters for each of these variants. As a means of this, we analyze all possible variants, totaling 5 (methods) · 4 (radii) · 5 (multipliers) = 120 different combinations. The variants that produce the best results are shown in Table 2.

As it can be noticed from the first two rows of Table 2, when performing classification with either pre- or post-pruning, the parameters configuration that leads to the highest performance does not change. This means that the same LBP extracted features can be used regardless of the classifier's parametrization. On the other hand, when using ensemble-based pruning — meaning that the output of multiple DTs is combined towards the final classification - a slightly lower radius is employed, while the multiplier is considerably higher, thus leading to more neighbors being considered for each pixel. As will be further

Table 3
Best PCA parameters for classification methods.

Extraction	Pruning	Selection	No. of components	Alpha	Kernel	Gamma	Degree
LBP	Pre	Sparse PCA	12	0.005	–	–	–
	Post	Sparse PCA	10	0.005	–	–	–
	Ensemble	Kernel PCA	6	–	polynomial	0.5	3
GLRLM	Pre	Sparse PCA	6	0.05	–	–	–
	Post	PCA	4	–	–	–	–
	Ensemble	Sparse PCA	8	0.00001	–	–	–
Geomet-rical	Pre	Kernel PCA	2	–	polynomial	$\frac{1}{no.of\ features}$	3
	Post	Incremental PCA	2	–	–	–	–
	Ensemble	Incremental PCA	2	–	–	–	–
Contour-based	Pre	Incremental PCA	2	–	–	–	–
	Post	Incremental PCA	2	–	–	–	–
	Ensemble	Sparse PCA	4	–	–	–	–

Fig. 4. LBP features extracted.

explained in detail in Section 3, the increased number of neighbors proves to be beneficial to the overall performance of the system.

The features extracted with LBP can be displayed to be analyzed by the user, as shown in Fig. 4 for the cropped ROI (presented in Fig. 3(f)), using the parameters from the last row of Table 2, in alignment with our efforts towards an interpretable system.

2.3.2. Gray-Level Run-Length Matrix

Gray Level Run-Length Matrix (GLRLM) [35] represents the spatial distribution of gray level runs, defined as the number of consecutive pixels with the same gray level in an image. The matrix is constructed by specifying the direction of the runs to be considered. For our experiments, we use four angles, corresponding to the horizontal, vertical, first and second diagonal, respectively: 0°, 45°, 90°, and 135°. The values of the features are computed separately for each angle and then the mean is calculated. As we apply the feature extraction only to the region of interest resulting from the segmentation process, which is the brightest area of the breast, we presume that normalizing the pixels to a smaller interval can benefit the classification output. We experiment with seven different levels, ranging from 2^1 to 2^8 (the maximum gray level for 8-bit deep images) for all variants of the classification algorithm. The accuracy of the classification increases until a 2^3 level is used for normalization, then it starts decreasing. Thus, the best results are obtained for a pixel normalization level of 2^3, which will be used in subsequent experiments.

With this configuration, we extract the following textural measures [35]:

1. *Short Run Emphasis (SRE)* $= \frac{1}{N_{runs}} \sum_{j=1}^{N_r} \frac{p_r(j)}{j^2}$,
2. *Long Run Emphasis (LRE)* $= \frac{1}{N_{runs}} \sum_{j=1}^{N_r} p_r(j) j^2$,
3. *Grey Level Non-Uniformity (GLN)* $= \frac{1}{N_{runs}} \sum_{j=1}^{N_g} [p_g(j)]^2$,
4. *Run Length Non-Uniformity (RLN)* $= \frac{1}{N_{runs}} \sum_{j=1}^{N_g} [p_r(j)]^2$ and
5. *Run Percentage (RP)* $= \frac{N_{runs}}{N_{pixels}}$.

Although GLRLM features cannot be analyzed in a visual manner, the matrix can be inspected, and given the formula for each of the features, the process of feature extraction can be easily understood.

2.3.3. Geometrical

The shape features are extracted from the mask of the lesion, independently of the intensity of the pixels. First, geometrical features are calculated: (1) area, (2) perimeter and (3) compactness. These features provide information about the shape of the lesion and can be used as a criterion to differentiate between benign and malignant lesions. The features mentioned above can be easily computed and understood by human experts, making the system interpretable and transparent.

2.3.4. Contour-based

To obtain information on the regularity/irregularity of tumors, features of their contour are extracted. Li et al. [36] proposed the calculation of shape characteristics by fitting an ellipse to the tumor as shown in Fig. 5(b). After the ellipse is defined, the difference between the border of the lesion and the ellipse is calculated (Δd, presented in Fig. 5(a)). Generally, benign lesions have more regular borders, resulting in lower differences, while malignant lesions have more irregular borders, resulting in higher differences. The final features are extracted as follows:

1. *Root Mean Square Roughness* $= \sqrt{\langle \Delta d^2 \rangle - \langle \Delta d \rangle^2}$
2. *Root Mean Square Slope* $= \sqrt{\langle tilt(\Delta d)^2 \rangle}$
3. *Circularity* $= \frac{\langle \Delta d \rangle}{\sigma \Delta d}$,

where $\langle\ \rangle$ denotes the mean of the values. The local tilt is calculated by $tilt(i) = \frac{\Delta d_{i+3} - 9\Delta d_{i+2} + 45\Delta d_{i+1} - 45\Delta d_{i-1} + 9\Delta d_{i-2} - \Delta d_{i-3}}{\Delta d_i}$.

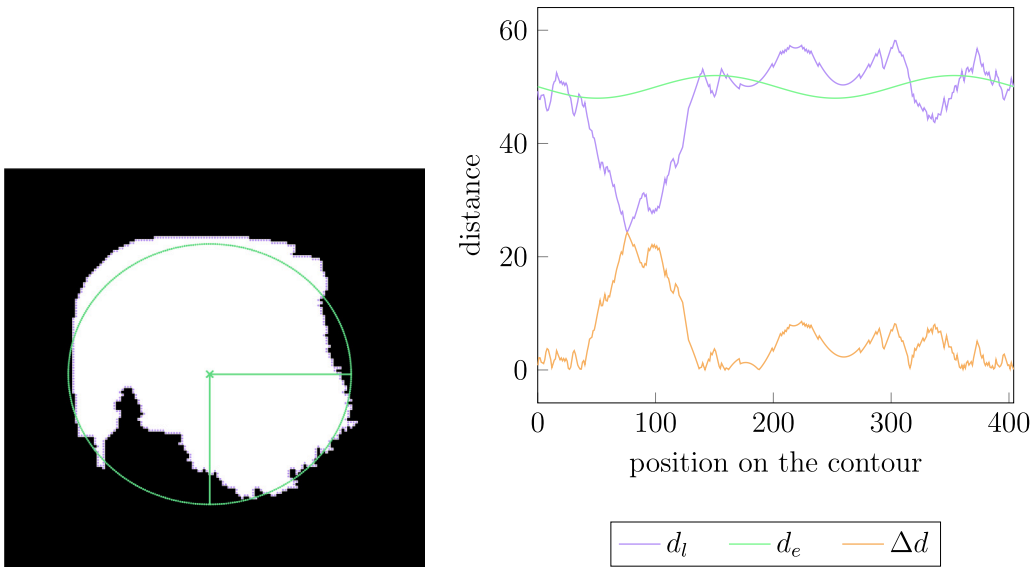
In some cases, investigating a segment of the boundary is enough for a correct decision. Therefore, the previously mentioned features are also computed after splitting the differences into several segments. Based on preliminary results presented in [37], the number of segments is set to 10.

The computed features can be visualized as shown in Fig. 5. To facilitate analysis of the characteristics, the segments can also be shown.

2.4. Feature selection

Feature selection aims to identify the most relevant features from the set of extracted features obtained from the segmented mammographic images. The selected features are subsequently used for the classification and identification of potential malignancies, which can aid radiologists in accurately diagnosing breast cancer. We analyze and experiment with three techniques (Principal Component Analysis — PCA, Singular Value Decomposition — SVD, and Linear Discriminant Analysis — LDA), which will be further detailed (as listed in Fig. 1d).

According to the survey presented in [38], PCA was the most frequently used method to reduce the dimensionality of the feature space extracted from mammographic images between 2018 and 2021, closely followed by LDA and Genetic Algorithm (GA). For this reason, we also employ PCA and LDA in our experiments. However, when it comes to GA, the experiments presented in [5] clearly show that PCA



(a) Fitted ellipse (b) Distances

Fig. 5. Computing featured from the contour of the lesion.

outperforms GA when applied on the mini-MIAS dataset, and therefore we decided not to analyze it further. PCA reduces dimensionality by centering the data and applying singular value decomposition (SVD) to it. To assess the importance of data centering, we also conduct experiments using SVD as feature selection method.

2.4.1. Principal Component Analysis

Principal Component Analysis (PCA) [39,40] is a statistical method that reduces the dimensionality of a dataset while retaining its patterns. It works by projecting the input data into a new coordinate system, whereas much of the variation in the data as possible can be described with fewer dimensions (i.e. features). In order to define the principal components, the eigendecomposition is performed on the covariance matrix, and only the features with the highest eigenvalues are kept. In this manner, PCA increases the interpretability of the data, while preserving only the essential information.

In our experiments, we also employ three variants of PCA: *Kernel PCA*, which utilizes kernels to achieve non-linear dimensionality reduction, *Sparse PCA*, which finds a group of sparse constituents that can efficiently reconstruct the data, and *Incremental PCA*, which performs linear dimensionality reduction using Singular Value Decomposition of the data.

The most important parameter of PCA is the number of components that will be kept. In order to find the best value for the subsequent classification, we experiment with values in the range of $[1, \min(no.\ of\ samples,\ no.\ of\ features - 1)]$ for all variants. For the Sparse PCA variant, we also test different values, ranging from $5 \cdot 10^{-4}$ to 1, for the sparsity control parameter; while for the KernelPCA we experiment with different types of kernels, namely *linear*, *polynomial*, *sigmoid* and *cosine*, different values in the range of $5 \cdot 10^{-4} - \frac{1}{no.\ of\ features}$ for the *polynomial* and *sigmoid* kernels coefficient (denoted as gamma) and with degrees in the interval of $[1 - 5]$ for the *polynomial* kernel. Table 3 presents the combinations for which the highest metrics values were achieved.

For a better understanding of the feature selection process, the ratio of variance $\frac{eigenvalue}{totaleigenvalues}$ can be examined. Fig. 6 holds the explained variance for the training dataset employed in our experiments. The higher the explained variance, the more information is retained in the reduced dataset; however, retaining more features can increase the risk of overfitting. The resulting features are a linear combination of the original features, allowing the system to demonstrate the importance of each feature in the classification process.

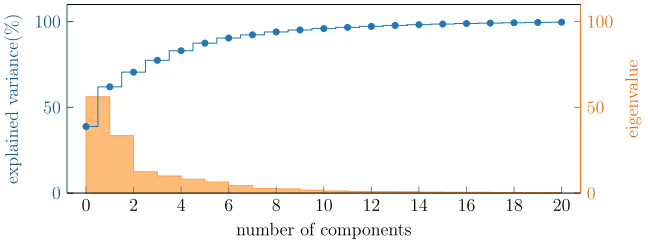


Fig. 6. Explained variance and eigenvalues for training dataset (contour-based features).

2.4.2. Singular Value Decomposition

Singular Value Decomposition (SVD) is a factorization of a matrix that comes from the field of linear algebra. Specifically, a matrix of order $m \times n$ is represented as the product of three matrices: $A_{[m \times n]} = U_{[m \times r]} \Sigma_{[r \times r]} (V_{[n \times r]})^T$, where r is the rank.

In our experiments, we employ a variant of the SVD, namely *Truncated SVD*, which finds a reduced rank approximation by setting all the singular values to 0, except from the k largest ones. In order to find the best suited k value for our needs, we experiment with values in range of $[1, no.of\ features]$. Our findings are presented in Table 4.

2.4.3. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [41] is a generalization of Fisher's linear discriminant [42] which works by finding a linear combination that separates the data into classes by means of statistical measures. It maximizes the ratio of the separation between the classes to the compactness of each class by computing the eigenvectors of the matrix formed by the matrices of between-class variance and data covariance. The highest eigenvalues correspond to a projection of the data onto a lower-dimensional space, which maximizes the separation between the classes. Since the maximum number of components for LDA is equal to $min(no.\ of\ classes - 1, no.\ of\ features)$ and, for our goal, we only have two classes, benign and malignant, the data will be reduced to 1 dimension.

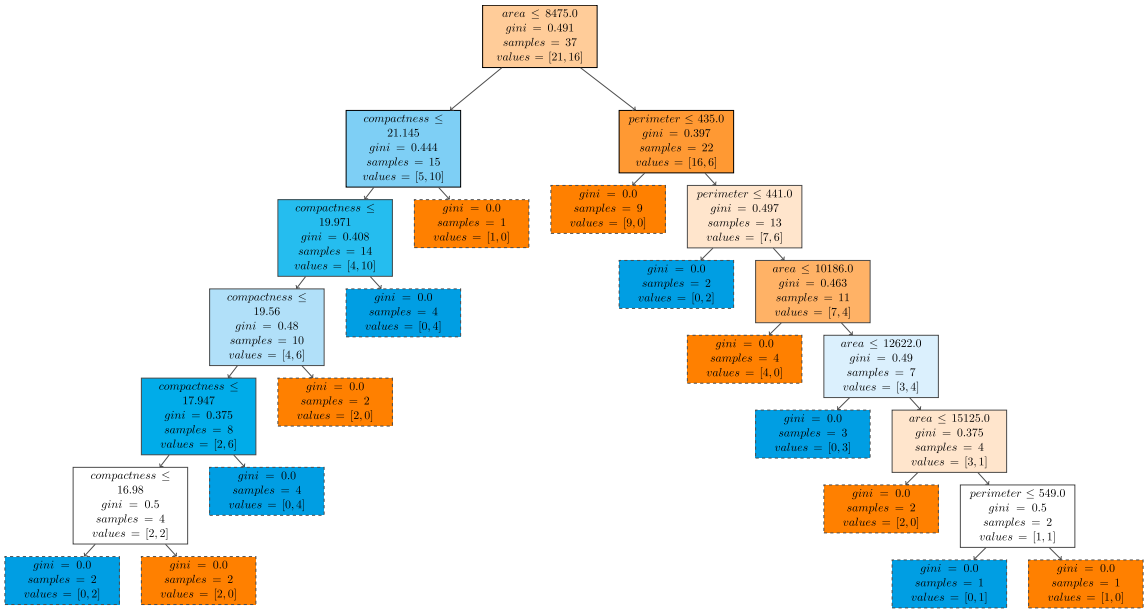


Fig. 7. Constructed DT with pre-pruning using geometrical features.

Table 4
Best Truncated SVD parameters for classification methods.

Extraction	Pruning	No. of components
LBP	Pre	10
	Post	8
	Ensemble	10
GLRLM	Pre	3
	Post	1
	Ensemble	4
Geomet-rical	Pre	2
	Post	2
	Ensemble	2
Contour-based	Pre	2
	Post	2
	Ensemble	15

Table 5
Best DT hyper-parameters.

Features	Depth	Samples-split	Samples-leaf
LBP	5	2	1
GLRLM	5	2	2
Geometrical	6	2	3
Contour-based	2	2	1

2.5. Classification

In the current paper, we propose an explainable CAD system. Hence, Decision Tree-based classifiers (as listed in Fig. 1e) are used due to their comprehensibility by humans. DTs are rule-based models that provide intuitive decision-making criteria and can be easily interpreted (as shown in Fig. 7). In each node, a feature is selected and a criteria is established to split the dataset. The disadvantage of using DTs is their tendency to overfit the data, which can lead to poor generalization. To overcome this issue we propose three methods: (1) pre-pruning, (2) post-pruning and (3) ensemble (RF).

Once the DT is built, it can be displayed such that the user can analyze and understand the entire classification process. Fig. 7 holds an example of a pre-pruned DT built on geometrical features, and such visualizations can be presented regardless of the features used or the tree's configuration.

In Fig. 7, the colors of the nodes represent the majority (orange - malignant or blue - benign) and purity. The more vivid the color, the greater the representation of the associated class. In each node, the first row presents the selected feature (in case of geometrical features: area, perimeter, and compactness) and the split value, the second row indicates the Gini impurity of the node, while the last two rows mark the total number of samples in the node and their distribution.

2.5.1. Pre-pruning

Pre-pruning is the process of stopping the splitting of a node before it reaches a certain depth. This prevents the model from becoming too complex and overfitting the data. This technique implies tuning the hyper-parameters of the model before its training, such that the model would stop early. In order to get a robust model, the maximum depth of the tree, the minimum number of samples required to split an internal node (samples-split) and the minimum number of samples required to be at a leaf node (samples-leaf), should be tuned.

As a means to finding the best set of hyper-parameters, we experiment with values in the range of $[1-9]$ for the maximum depth, also considering the variant of not limiting the depth (denoted as *None*), and with values in the range of $[1-5]$ for samples-split and samples-leaf. The results of the experiments are presented in Table 5.

2.5.2. Post-pruning

Post-pruning, also called backward pruning, involves training the DT to full depth and then removing branches that do not improve the accuracy of predictions on new data. We employ the Minimal Cost-Complexity Pruning algorithm [43], which is based on the cost-complexity measure: $R_\alpha(T) = R(T) + \alpha|\tilde{T}|$, where T is the tree, α is the complexity parameter, $|\tilde{T}|$ is the number of terminal nodes in T and $R(T)$ is the misclassification rate. Therefore, the algorithm is parametrized by α, and the values used in our experiments are presented in Table 6.

2.5.3. Ensemble

Ensemble-based methods involve combining multiple models. Random Forests (RFs) are ensembles of decision trees that combine the output of multiple trees to improve classification performance and minimize overfitting. To make the final decision, the class with the highest number of votes is selected.

Table 6
Best DT cost-complexity parameter.

Features	α
LBP	0.0405
GLRLM	0.0506
Geometrical	0.0
Contour-based	0.0405

Table 7
Structural details of the mini-MIAS dataset.

format	PGM
orientation	MLO
dimensions	1024 × 1024 pixels
resolution per pixel	200 microns
bits per pixel	8

The Random Forest algorithm depends on the number of trees in the ensemble. We experiment with values in the set of $[10, 100, 1000, 10000]$, concluding that 100 estimators are enough to obtain satisfactory results, regardless of the features used.

It is important to mention that RFs, although comprising multiple DTs, do not lose the interpretability characteristic — not only can all the trees be displayed similar to the one in Fig. 7, but they can also be grouped according to the predicted class, such that, once the diagnosis is set based on the majority votes, a single tree from this majority can be displayed and examined by the user.

3. Experimental results

The goal of the current paper is the development of a fully automated CAD system to help radiologists in the interpretation of mammograms. The transparency of CAD systems is important to gain the confidence of experts. Hence, the methods presented in Section 2 are selected by their ability to provide interpretable results. In the following paragraphs, we detail the parameters used, the evaluation metrics, and the results obtained.

3.1. Set up

In the previous section, we proposed four feature extraction methods (LBP, GLRLM, Geometrical and Contour-based), three feature selection methods (PCA, SVD and LDA) and three pruning techniques (pre-pruning, post-pruning and ensemble-based) to avoid overfitting. Experiments were carried out for every combination of these methods. Considering only the methods, their combinations result in 36 experiments. When parameters are included, the number of combinations reaches into millions. Therefore, some parameters were selected based on preliminary experiments, as presented in Section 2, while others were optimized through grid search. The best performing parameters are selected and presented in the following analysis.

3.2. Data

To train the models, the well-known public dataset of Mammographic Image Analysis Society (mini-MIAS) [9,44] is used. The dataset consists of 322 mammograms (left and right breast of 161 patients) distributed as 207 normal, 63 benign and 51 malignant. The structural details for the images in the dataset are provided in Table 7.

In the experiments, we will use images that contain a single lesion. As a result, 57 images are used, 31 with malignant lesion, and 26 with benign lesions. To properly evaluate and compare the variants, the dataset is divided according to a 70%–30% train-test ratio and only the test results are taken into account; that is, the results obtained for new data. Once the best combination is chosen, the same test results are compared with results from the literature.

3.3. Metrics

The performance of the proposed approach is measured with four metrics:

1. *accuracy* - percentage of lesions correctly labeled, presented in Eq. (1),
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$
2. *precision* - percentage of malignant lesions among lesions labeled as malignant, presented in Eq. (2),
$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$
3. *sensitivity* - percentage of malignant lesions correctly labeled, presented in Eq. (3) and
$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$
4. *specificity* - percentage of benign lesions correctly labeled, presented in Eq. (4)
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

where TP is the number of malignant lesions classified correctly, TN is the number of benign lesions classified correctly, FP is the number of misclassified benign lesions as malignant, and FN is the number of misclassified malignant lesions as benign.

3.4. Results

Table 8 details the results achieved with the different feature extraction and feature selection combinations. In the presented approach, two types of features are extracted: texture (LBP and GLRLM) and shape (geometrical and contour-based).

The results show that LBP outperforms GLRLM features both in accuracy and precision, for almost all of the cases. While GLRLM obtains better results when combined with LDA feature selection and ensemble-based pruning, LBP yields higher accuracy, precision and specificity for all other combinations. As for the sensitivity score, GLRLM reaches higher performance when combined with PCA and DT with pre-pruning and RF (ensemble-based pruning), as well as SVD and pre-pruning. However, for these cases, the specificity score achieved with GLRLM is much lower than the one obtained with LBP features, with a difference of at least 30%, while the gain in sensitivity is of 10%. This suggests that using GLRLM features lead to a more imbalanced sensitivity-specificity trade-off. A GLRLM-based system will correctly determine more malignant tumors, but will generate more false positives, which, in the real world, might lead to patients being unnecessarily treated. On the other hand, LBP generates a more balanced trade-off, thus leading to an overall increased performance.

From the presented shape features, contour-based features achieved better performance compared to geometrical features. Based on the metrics presented in Table 8, the best results are achieved by LBP, outperforming the contour-based features. This can be explained by the fact that the segmented ROI is not completely identical with the actual lesion and, thus, the contours may slightly differ.

To reduce the size of the DTs it is essential to decrease the number of input features. Therefore, feature selection is applied to keep the most significant characteristics extracted from the ROI. From the findings outlined in Table 8 arises that PCA outperforms both SVD and LDA.

In our experiments, different pruning methods are used to overcome overfitting of the built models, as mentioned in Section 2.5. Based on the conducted experiments, we can conclude that, in general, using an ensemble for pruning Decision Trees performed the best, achieving the highest accuracy and specificity.

The best performance is achieved using LBP texture features, PCA feature selection, and DTs with ensemble pruning. By employing these methods, the system reached 95% accuracy and 100% precision and specificity. Fig. 8 presents the confusion matrix for this configuration.

Table 8
Evaluation metrics for different combinations.

Extraction	Selection	Pruning	Accuracy	Precision	Sensitivity	Specificity
		Pre	0.85	0.89	0.80	0.90
	PCA	Post	0.85	0.82	0.90	0.80
		Ensemble	0.95	1.00	0.90	1.00
		Pre	0.80	0.88	0.70	0.90
LBP	SVD	Post	0.80	1.00	0.60	1.00
		Ensemble	0.80	0.88	0.70	0.90
		Pre	0.75	1.00	0.50	1.00
	LDA	Post	0.70	0.83	0.50	0.90
		Ensemble	0.45	0.44	0.40	0.50
		Pre	0.75	0.69	0.90	0.60
	PCA	Post	0.70	0.67	0.80	0.60
		Ensemble	0.80	0.70	1.00	0.60
		Pre	0.65	0.62	0.80	0.50
GLRM	SVD	Post	0.50	0.00	0.00	1.00
		Ensemble	0.60	0.62	0.50	0.70
		Pre	0.65	0.80	0.40	0.90
	LDA	Post	0.70	0.70	0.70	0.70
		Ensemble	0.65	0.67	0.60	0.70
		Pre	0.55	0.55	0.55	0.60
	PCA	Post	0.55	0.55	0.55	0.54
		Ensemble	0.65	0.66	0.65	0.50
		Pre	0.50	0.50	0.50	0.40
Geometrical	SVD	Post	0.50	0.50	0.50	0.49
		Ensemble	0.55	0.55	0.55	0.50
		Pre	0.55	0.56	0.55	0.80
	LDA	Post	0.50	0.50	0.50	0.50
		Ensemble	0.50	0.50	0.50	0.50
		Pre	0.75	0.75	0.75	0.70
	PCA	Post	0.75	0.75	0.75	0.70
		Ensemble	0.85	0.88	0.85	1.0
		Pre	0.75	0.75	0.75	0.80
Contour-based	SVD	Post	0.80	0.80	0.80	0.80
		Ensemble	0.85	0.85	0.85	0.90
		Pre	0.80	0.85	0.80	0.60
	LDA	Post	0.80	0.85	0.80	0.60
		Ensemble	0.90	0.91	0.90	0.80

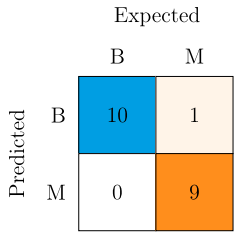


Fig. 8. Confusion matrix for the best-performing configuration on the test images. B stands for benign and M stands for malignant.

4. Discussion

The paper introduces an interpretable and fully automated CAD system for breast cancer detection and diagnosis. Previously, we detailed the system's components and the results obtained with the mini-MIAS dataset. In this section, we discuss the limitations of our system, compare it with related work, and present the results of ablation studies.

4.1. Comparison with related work

In the following paragraphs, we compare the proposed system with existing similar systems. As emphasized throughout the paper, the primary goal of this study is to develop an interpretable CAD system. As detailed in Sections 2 and 3, the proposed system is capable of providing a visual explanation for every step of the system while achieving an accuracy of 95% (with only one misclassification due to the limited number of inputs). There is a trade-off between a system's performance and interpretability. While artificial neural networks typically offer superior performance, their decision-making process is not transparent, which is why they are often referred to as "black-box" models. Our comparison focuses on interpretable systems and evaluates their performance.

Table 9 presents the results obtained by the systems detailed in Section 1.1 when employing Decision Tree for classification. One can easily notice that our approach is the only one to obtain maximum precision and specificity scores. Moreover, it also achieves the highest accuracy, along with the system proposed in [11]. However, if we look at the sensitivity measure, we can conclude that our system can still be improved. It is important to mention that, given the data employed in our experiments, the decrease in sensitivity is due to only one malignant lesion being incorrectly labeled.

From Table 9, we can also note that the usage of multiple trees (i.e. ensemble-based pruning) does not guarantee an increased performance, as the accuracy obtained in [11] using a single DT, equal to the one achieved with our proposed system, is higher than the values obtained with all other RF-based approaches. In terms of sensitivity, [11] also reports the highest value, yet the specificity is lower than both ours and the one obtained with the system proposed in [13], which might come as proof that pruning techniques lead to a more cautious classification.

While higher sensitivity means that no malignancy is overseen, a higher precision means that patients could start treatment directly, without further investigation. Unless a perfect system is created, there is always going to be a trade-off between sensitivity and precision, and we consider that the end-users should decide which one is more important.

When comparing our method to neural networks, the comparison is not straightforward due to the different data used in the training

Table 9
Comparison of the results obtained with different CAD systems.

	Dataset	Pruning	Accuracy	Precision	Sensitivity	Specificity
Proposed	mini-MIAS	Ensemble	0.95	1.00	0.90	1.00
[11]	mini-MIAS	None	0.95	0.94	0.97	0.89
[13]	MIAS	Ensemble	0.94	–	0.94	0.94
[12]	mini-MIAS	Ensemble	0.90	–	–	–
[14]	mini-MIAS	None	0.85	0.88	0.80	–
[16]	mini-MIAS	Ensemble	0.78	0.86	0.25	–
[15]	DDSM	Ensemble	0.71	–	0.61	0.78

Table 10
Results obtained removing each intermediate step of the proposed system at a time.

Removed step	Accuracy	Precision	Sensitivity	Specificity
None	0.95	1.00	0.90	1.00
Feature selection	0.90	1.00	0.80	1.00
Feature extraction	0.40	0.42	0.50	0.30
Segmentation	0.65	0.71	0.50	0.80
Pre-processing	0.65	0.67	0.60	0.70

process. On the other hand, our approach is fully interpretable, while the mentioned neural networks are explainable. There is a fine line between interpretability and explainability as detailed in Section 1. We have chosen to prioritize interpretability, as we believe that it is more important in the medical field. The end-users should be able to understand the whole process of the decisions made by the system.

The trade-off between prediction accuracy and transparency (achieved either through interpretability or explainability) is an important consideration when choosing a model. While neural networks provide very high accuracy, they have a low level of transparency. Decision trees are on the other end, with high interpretability and lower accuracy. However, building an ensemble of decision trees, resulting in what is knows as Random Forest classifier, drastically increases the prediction accuracy, while maintaining the transparency property of the individual decision trees. According to Dam et al. [45], ensemble methods provide the highest prediction accuracy after deep learning methods, and decision trees, the highest explainability. Thus, we consider our choice of using an ensemble of decision trees to be a good approach to balancing this trade-off, by increasing the model's prediction accuracy without affecting its interpretability.

4.2. Ablation studies

As previously described in detail in Section 2, our proposed system consists of five steps: 1. pre-processing, 2. segmentation, 3. feature extraction, 4. feature selection, and 5. classification. Although the last step is the one that returns the output – whether the mammography contains a benign or a malignant mass –, all the intermediate steps are needed for a performant and accurate classification. To prove the importance of each step, we perform an ablation study, keeping the configuration that yielded the best results, as reported in Section 3.4, and removing each transitional step at a time.

Table 10 presents the results of the ablation study. The first row contains the results obtained with the proposed system, using the following configuration, as explained in Section 2:

1. pre-processing: removal of noise and artifacts, pectoral muscle removal, and contrast-limited adaptive histogram equalization;
2. segmentation: Threshold-based GrowCut with automated background and foreground seeds selection;
3. feature extraction: Local Binary Pattern using the 'default' method, a radius equal to 7 and a multiplier equal to 16;
4. feature selection: PCA using a 3rd grade polynomial kernel, with 6 components and a gamma value of 0.5
5. classification: DT with ensemble pruning (i.e. Random Forest) with 100 estimators.

Since classification is the step that returns the final output of the system, it cannot be removed. Thus, rows 2–5 of Table 10 hold the results obtained when removing all the other steps, as follows: feature selection, feature extraction, segmentation, and pre-processing, respectively.

For the first case, when feature selection is no longer applied, the precision and specificity scores are not affected, while sensitivity decreases by 10%, leading to a 5% decrease in accuracy. This proves that selecting the most relevant features leads to more malignant tumors being correctly identified.

When feature extraction is removed from the system, meaning that features are selected directly from the segmented image, the results are strikingly lower. With an accuracy score 55% lower than the one obtained with the entire proposed system, it is clear that this step is the most important, having the highest influence on the performance of the system.

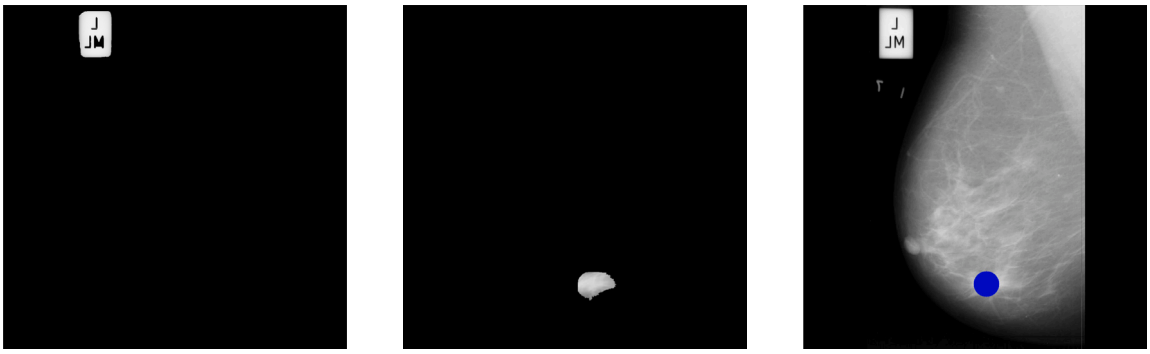
Moving forward, when skipping the segmentation step and, therefore, extracting features from the entire (pre-processed) mammogram, all metrics have significantly lower values not only compared to those obtained with the entire system, but also when the feature selection step is removed (first and second rows of Table 10, respectively). Hence, we can state that, although not as much as feature selection, segmentation plays a very important role in the overall system.

Finally, if we remove the pre-processing step, it means that segmentation is applied on the original mammogram. Comparing the results obtained in this manner to those obtained when segmentation is not applied at all, we can see that the same accuracy score is achieved, while there is an increase in sensitivity and a decrease in precision and specificity. The identical accuracy suggests that performing segmentation without prior processing of the mammogram is not of much help — the segmentation method does not perform well on unprocessed mammographies. As a proof of this statement, Fig. 9 presents a comparison between the segmentation obtained without pre-processing (Fig. 9(a)) and with pre-processing (Fig. 9(b)), with the ground truth depicted in blue in Fig. 9(c). It is easy to notice, from this figure alone, the importance of the pre-processing: without this step, a label from the mammogram is segmented as the ROI, while the result of the segmentation after pre-processing is very close to the ground truth, considering that the ground truth provided with the dataset is an approximation.

Taking into consideration the results presented in Table 10, we can conclude that all the steps of the proposed system are essential towards a good performance, having supportive contributions to the overall model.

4.3. Generalization

In order to prove the robustness of our proposed approach, we applied it to images from the mini-DDSM dataset [10]. This dataset is an improved version of the original DDSM dataset [46], providing easier access, while keeping the original file names and binary masks for the abnormalities. It consists of 1952 cases split into three categories: *Normal*, *Benign* and *Cancer*. Each case encompasses craniocaudal (CC) and mediolateral oblique (MLO) mammographies for a patient's left and right breasts and, if any abnormality is present, its binary mask. Fig. 10

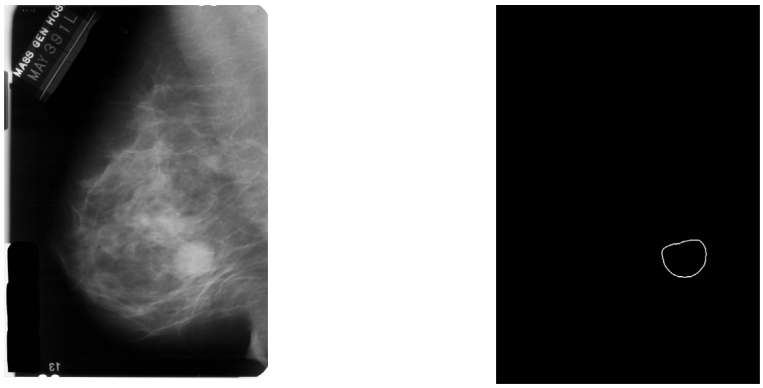


(a) Result of segmentation applied on the original mammogram.

(b) Result of segmentation applied on the pre-processed mammogram.

(c) Ground truth.

Fig. 9. Results of segmentation applied on the original and pre-processed mammogram.



(a) Original mammogram.

(b) Binary mask.

Fig. 10. Mammogram image with its corresponding binary mask.

depicts such a pair, where Fig. 10(a) shows the MLO mammography of a left breast containing a benign abnormality, and Fig. 10(b), its binary mask.

For our experiments, we used 144 MLO images from the *Benign* and *Cancer* categories, which contain a single lesion, to coincide with the criteria used to select the images from the mini-MIAS dataset [9], as explained in Section 3.2.

In the following experiments, we opt for the methods achieving the best results on mini-MIAS: 1. noise, artifacts and pectoral muscle removal and CLAHE for pre-processing; 2. TbGC with automated seeds selection for segmentation; 3. LBP for feature extraction; 4. PCA for feature selection; and 5. RF (DT with ensemble-based pruning) for classification -, However, the image acquisition process differs across datasets, due to the type of machine used. In order to find the best-suited parameter values for every method, another grid search is executed. Therefore, the methods employed for each step remain the same and only the parameters' values are changed. Table 11 presents all of our system's parameters, with their respective values for the application on mini-MIAS and mini-DDSM datasets.

The first step of the proposed approach – that is, pre-processing –, remains the same for both datasets. For the segmentation part, only the number of iterations changes, as the TbGC algorithm requires 20 more iterations ($\theta_i = 25$) for the mini-DDSM images in order to achieve satisfactory results, as shown in [29]. For feature extraction, two out of the three LBP's parameters are modified, using *nri_uniform* variant instead of the *default* one, and without multiplication (the multiplier is equal to 1). For the next step, the same variant of PCA is employed - Kernel PCA -, using also a 3_{rd} degree polynomial kernel; however, for mini-DDSM, a lower value is required for gamma ($\gamma = 0.005$). Moreover, the most relevant information is held in 4 components instead of 6, as for mini-MIAS. Finally, for the classification step, ensemble-based pruning is employed for both datasets, but, for the mammograms for mini-DDSM, 10 trees are enough for a good enough classification.

Since the same methods are employed for both datasets, naturally, the output of each step can be displayed regardless of the dataset, thus maintaining the interpretability property of our system. Fig. 11 presents the results of the intermediate steps, with Fig. 11(a) holding the result of the pre-processing step, Fig. 11(b) holding the result of segmentation, and Fig. 11(c) holding the extracted features.

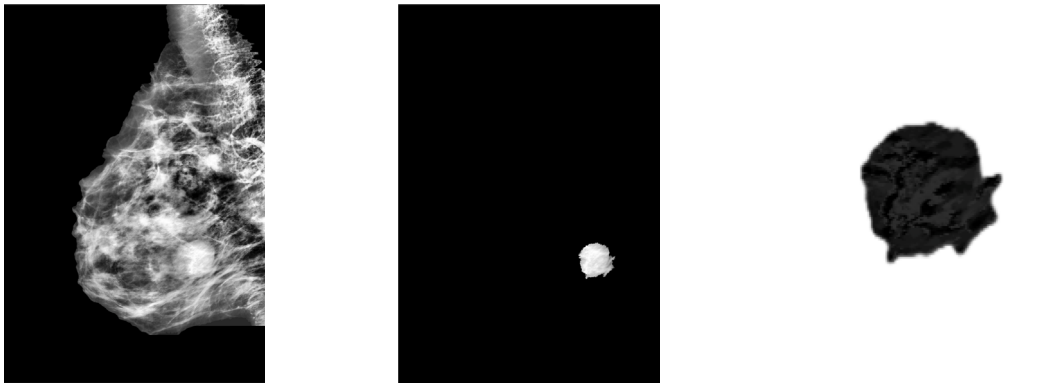
Using the parameters detailed in Table 11, the test results obtained on the images from mini-DDSM, after a 70%–30% train-test split, are as follows:

- *Accuracy*: 0.97,
- *Precision*: 0.95,
- *Sensitivity*: 1.00,
- *Specificity*: 0.95.

These results prove the robustness of our proposed approach. The system trained on mini-DDSM not only surpasses the system trained on mini-MIAS in terms of accuracy, but also exceeds all the other CAD systems presented in Section 1.1. The perfect sensitivity score indicates that all malignant lesions are correctly labeled, while the 95% precision and specificity scores show that only one benign abnormality was wrongly classified as malignant. When applying our system on the mini-MIAS dataset, as discussed in Section 3.4, a malignant abnormality was classified as benign — thus, the other way around. However, as previously highlighted, the trade-off between sensitivity and precision will always be present, and the decision on which of these metrics weighs heavier should belong to the end-users.

Table 11
Parameters of the proposed system with their values for mini-MIAS and mini-DDSM.

Step	Method	Parameter	Mini-MIAS	Mini-DDSM
Pre-processing	Thresholding	Threshold	50	50
Segmentation	TbGC	Foreground seeds circle radius	25	25
		Number of iterations (θ_i)	5	25
		Threshold (θ_t)	0.5	0.5
Feature extraction	LBP	Variant	default	nri_uniform
		Radius	7	7
		Multiplier	16	1
Feature selection	PCA	Variant	Kernel PCA	Kernel PCA
		Kernel	polynomial	polynomial
		No. of components	6	4
		Gamma	0.5	0.005
		Degree	3	3
Classification	DT	Pruning	ensemble	ensemble
		No. of estimators	100	10



(a) Pre-processing result. (b) Segmentation result. (c) LBP extracted features.

Fig. 11. Output of each step for a mammogram image.

We want to emphasize the fact that, although we experiment with a dataset almost triple in size than the one on which our approach was originally validated, it still misclassifies only one image. Therefore, we can conclude that the proposed system can be easily adapted for different datasets by changing only the values of some parameters, while maintaining high performance.

4.4. Clinical viewpoint

While designing the proposed system for breast cancer detection and diagnosis from mammographies, we have collaborated with radiologists from the "Prof. Dr. Ion Chiricuță" Oncology Institute from Cluj-Napoca, Romania, understanding their needs and expectations from such a system. First of all, because the hospital accommodates a very large number of patients every day, a system that can provide a second opinion (instead of another medical expert) would speed up the process of analyzing the mammographies. Such systems are already employed for other types of medical images (CTs, MRIs) and have proven helpful and, thus, a similar approach for mammogram analysis would be useful.

The "perfect" system should have two components (as most CAD systems do): detection and diagnosis. For the first part, it should be able to examine the images without any input from the user, yielding a segmented region of interest in near real-time (NRT), such that the radiologists can inspect the already annotated mammograms. This way, they would just need to validate the output and manually segment the images only if the system missed a lesion or wrongfully detected the ROI.

The second component, which would tell whether a lesion is benign or malignant, would be helpful when deciding if a biopsy is necessary or not. Naturally, medical experts will not rely on the system on such a level that surgery and/or treatment would be followed without a histopathology exam to confirm the disease; however, lesions can have various shapes, dimensions, contours, etc., and, although usually the malignant ones have similar characteristics, they can resemble benign abnormalities enough to deceive even experts. This is where the CAD system comes in: being able to examine a lesion at pixel-level and to discover various patterns and connections that might not be visible to the naked eye.

Although any system with the previously described capabilities would be helpful, due to the different machines being used for taking the mammograms, the need for a generalizable system becomes imminent. Moreover, the computers used for mammography analysis also differ from one hospital to another, raising another need: low resource consumption. And finally, perhaps the most important characteristic of such a system: the human experts need to understand how it works, in order for them to trust its decisions.

Our proposed approach was designed and built to accommodate these requirements. It provides an unsupervised segmentation, which not only does not require user input, but it can also be easily adapted to work on different datasets, as proven in Section 4.3. Due to the limitation of the employed TbGC algorithm to only 5 iterations, the region of interest is detected in NRT, while the fact that the method does not require any prior training reduces the amount of resources needed.

The extraction (and selection) of textural features comes to meet the need for an in-depth analysis of a lesion's structure. In this manner, characteristics that may be otherwise missed by the unaided eye are used for diagnosis.

When it comes to the interpretability of such systems, measuring it is a challenging task due to the lack of standard metrics. [47] proposed

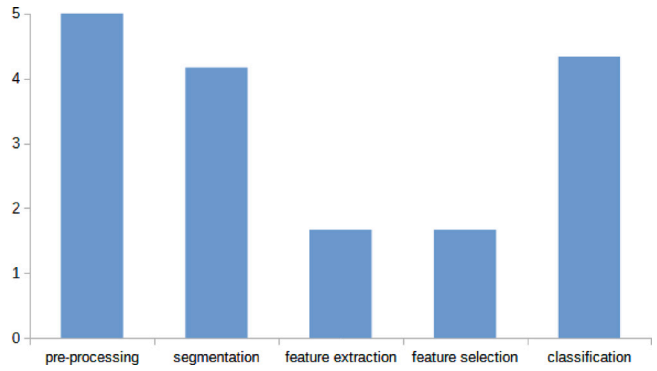


Fig. 12. Interpretability evaluation.

a set of metrics for a synthetic dataset. Unfortunately, it is nearly impossible to adapt real-life data to include the required information. To the best of our knowledge, no mammogram dataset with detailed labeling exists. Yet, as a means to ensure that the radiologists understand how our proposed approach works and, also, that they would feel confident to use it, we asked the coordinator of the Radiology and medical imaging laboratory from the "Prof. Dr. Ion Chiricuţă" Oncology Institute to evaluate the system. As a means to this, we asked them to rate each step of our approach (pre-processing, segmentation, feature extraction, feature selection and classification) according to the following criteria: understanding of the inputs and outputs, understanding of the inner steps (for pre-processing, as depicted in Section 2.1 and segmentation, as shown in Figs. 2 and 3), understanding of the expected output, the ability to identify if any mistake was made, along with the step where it was made and the ability to manually correct a mistake (if applicable). For each question, they were asked to rate their level of understanding on a 1 to 5 scale, with 5 meaning complete understanding. The results for each step are presented in Fig. 12.

As it can be easily noticed, the pre-processing methodology is absolutely clear, as the radiologist not only understands the inner workings but is also confident to manually modify the output if needed. For the segmentation step, some difficulties appear when it comes to the actual region growing process as to how exactly a pixel's label is modified based on its neighbors. For the following two steps, which refer to feature extraction and selection, some understanding problems appear also when it comes to the expected result. However, the classification is better understood, although the expert does not fully comprehend the (numerical) input of this step.

While our proposed approach would certainly benefit from a clearer visual interpretation of the feature extraction and feature selection processes, we find this evaluation encouraging towards an interpretable CAD system.

5. Conclusion and future work

In this paper, we advanced an automated interpretable breast cancer detection and diagnosis system, with the means of serving as a second opinion to doctors who analyze and interpret mammographic images. The system can be divided into five easy-to-understand, yet robust steps, which can be displayed in such a manner that the users can comprehend its decisions without needing a technical background.

The proposed system begins with the removal of redundant information from the raw mammogram and the improvement of its quality. The suspicious masses are then segmented with an unsupervised algorithm, the Threshold-based GrowCut algorithm, thus obtaining the regions of interest. From the respective ROIs, Local Binary Pattern textural features are extracted, and the most relevant ones are selected with the aid of Principal Component Analysis. These features serve as input to a Random Forest classifier, which yields the final result: benign or malignant lesion.

The proposed system is distinguished by its interpretability, achieved through the visual explanation of each processing step. Compared to Artificial Neural Networks, it is easier and quicker to train, which is crucial in healthcare, where datasets are often limited in size. Additionally, the Threshold-based GrowCut algorithm is integrated into the system. By using an iterative segmentation method, the user can verify every step of the segmentation.

To obtain the presented CAD, we analyzed four feature extraction methods, three feature selection methods and three techniques to avoid overfitting, tested all the variants on 57 images from the mini-MIAS dataset [9] and chose the one that obtained the best results — 95% accuracy, 100% precision and specificity and 90% sensitivity. Moreover, we validate the proposed approach on a different dataset, namely mini-DDSM [10], obtaining 97% accuracy and 100% sensitivity, and have a radiologist evaluate it from an interpretability point of view, obtaining a satisfactory outcome. We also compared our proposed system with six other systems from the literature that use Decision Tree-based algorithms for classification, obtaining comparable or even better results.

For future work, we aim to further validate our system using different datasets. Additionally, we plan to extend its capabilities to digital tomosynthesis and to integrate the analysis of numerical tests, such as blood tests, to improve the accuracy of the system. To further validate the system, we plan to conduct user studies to evaluate the system's interpretability.

CRediT authorship contribution statement

Cristiana Moroz-Dubenco: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Adél Bajcsi:** Writing – review & editing, Writing – original draft, Software, Formal analysis, Conceptualization. **Anca Andreica:** Writing – review & editing, Validation, Supervision, Project administration. **Camelia Chira:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

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

Acknowledgments

This study was supported by the Ministry of Research, Innovation and Digitization, as Intermediary Body for the Operational Programme Competitiveness 2014–2020 project code SMIS 2014+ 127725, contract no. 352/390028/23.09.2021, acronym project INSPIRE; and the Ministry of European Investment and Projects (MIPE) as Managing Authority for the Smart Growth, Digitalization and Financial Instruments Programme 2021–2027 and the Ministry of Research, Innovation and Digitalization (MCID) as Intermediary Research Body, project code SMIS 2021+ 324771 contract MIPE no. G-2024-71962/23.10.2024 and contract MCID no.390005/23.10.2024, project acronym INSPIRE-II.

References

[1] EU Science Hub, Joint Research Centre, Cancer cases and deaths on the rise in the EU, 2023, https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/cancer-cases-and-deaths-rise-eu-2023-10-02_en, Published on 02/10/2023. Accessed on 16/05/2024.

[2] Alfredo Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (24) (2020) 18069–18083, http://dx.doi.org/10.1007/s00521-019-04051-w.

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2019) 336–359, http://dx.doi.org/10.1007/s11263-019-01228-7.

[4] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, Cynthia Rudin, This looks like that: Deep learning for interpretable image recognition, 2018.

[5] Adél Bajcsi, Anca Andreica, Camelia Chira, Towards feature selection for digital mammogram classification, Procedia Comput. Sci. 192 (2021) 632–641, http://dx.doi.org/10.1016/j.procs.2021.08.065, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

[6] Adél Bajcsi, Camelia Chira, Anca Andreica, Extended mammogram classification from textural features, Stud. Univ. Babeş-Bolyai Inform. 67 (2) (2023) 5–20, http://dx.doi.org/10.24193/subbi.2022.2.01.

[7] Cristiana Moroz-Dubenco, Laura Dioşan, Anca Andreica, Mammography lesion detection using an improved GrowCut algorithm, Procedia Comput. Sci. 192 (2021) 308–317, http://dx.doi.org/10.1016/j.procs.2021.08.032, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

[8] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, Camelia Chira, An unsupervised threshold-based GrowCut algorithm for mammography lesion detection, Procedia Comput. Sci. 207 (2022) 2096–2105, http://dx.doi.org/10.1016/j.procs.2022.09.269, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022.

[9] John Suckling, The mammographic images analysis society digital mammogram database, in: Exerpta Medica. International Congress Series, 1994, Vol. 1069, 1994, pp. 375–378.

[10] Charitha Dissanayake Lekamlage, Fabia Afzal, Erik Westerberg, Abbas Cheddad, Mini-DDSM: Mammography-based automatic age estimation, in: Proceedings of the 2020 3rd International Conference on Digital Medicine and Image Processing, 2020, pp. 1–6.

[11] J. Kamalakannan, M. Rajasekhara Babu, Classification of breast abnormality using decision tree based on GLCM features in mammograms, Int. J. Comput. Aided Eng. Technol. 10 (5) (2018) 504, http://dx.doi.org/10.1504/ijcaet.2018.094328.

[12] Aleena Johny, Jincy J. Fernandez, Breast cancer detection in mammogram using fuzzy C-means and random forest classifier, Int. J. Sci. Res. Sci. Eng. Technol. 4 (8) (2018) 312–321.

[13] Sawsen Boudraa, Ahlem Melouah, Hayet Farida Merouani, Improving mass discrimination in mammogram-CAD system using texture information and super-resolution reconstruction, Evol. Syst. 11 (4) (2020) 697–706, http://dx.doi.org/10.1007/s12530-019-09322-4.

[14] Sakshi Painuly, Machine learning-based automated diagnosis of breast cancer from mammography images, Math. Stat. Eng. Appl. 70 (2) (2021) 1811–1821, http://dx.doi.org/10.17762/msea.v70i2.2474.

[15] Harmandeep Singh, Vipul Sharma, Damanpreet Singh, Machine learning based computer aided diagnosis system for classification of breast masses in mammograms, J. Phys. Conf. Ser. 2267 (1) (2022) 012141, http://dx.doi.org/10.1088/1742-6596/2267/1/012141.

[16] Saiful Bukhori, Suci Dwi Maysaroha, ADI Januar, Classification of breast cancer tumors using a random forest on mammogram images, Appl. Med. Inform. 45 (1) (2023).

[17] Ann-Kristin Balve, Peter Hendrix, Interpretable breast cancer classification using CNNs on mammographic images, 2024, http://dx.doi.org/10.48550/ARXIV.2408.13154.

[18] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[19] Scott M. Lundberg, Su-In Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777, http://dx.doi.org/10.48550/arXiv.1705.07874.

[20] Charalampos Lamprou, Kyriaki Katsikari, Noora Rahmani, Leontios J. Hadjileontiadis, Mohamed Seghier, Aamna Alshehhi, StethoNet: Robust breast cancer mammography classification framework, IEEE Access 12 (2024) 144890–144904, http://dx.doi.org/10.1109/ACCESS.2024.3473010.

[21] Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, Jinghua Wang, The Chinese Mammography Database (CMMD): An Online Mammography Database with Biopsy Confirmed Types for Machine Diagnosis of Breast, The Cancer Imaging Archive, 2021, http://dx.doi.org/10.7937/TCIA.EQDE-4B16, URL https://www.cancerimagingarchive.net/collection/cmmd/.

[22] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, Jaime S. Cardoso, INbreast: toward a full-field digital mammographic database, Academic Radiol. 19 (2) (2012) 236–248, http://dx.doi.org/10.1016/j.acra.2011.09.014.

[23] Qiong Lou, Yingying Li, Yaguan Qian, Fang Lu, Jinlian Ma, Mammogram classification based on a novel convolutional neural network with efficient channel attention, Comput. Biol. Med. 150 (2022) 106082, http://dx.doi.org/10.1016/j.compbiomed.2022.106082.

[24] Gianluca Carloni, Andrea Berti, Chiara Iacconi, Maria Antonietta Pascali, Sara Colantonio, On the applicability of prototypical part learning in medical images: Breast masses classification using ProtoPNet, in: Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges, Springer Nature Switzerland, Cham, 2023, pp. 539–557, http://dx.doi.org/10.1007/978-3-031-37660-3_38.

[25] Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, Daniel Rubin, Curated breast imaging subset of digital database for screening mammography (CBIS-ddsm)[skup podataka], 2016, The Cancer Imaging Archive.

[26] Adél Bajcsi, Towards a support system for digital mammogram classification, Stud. Univ. Babeş-Bolyai Inform. 66 (2021) 19, http://dx.doi.org/10.24193/subbi.2021.2.02.

[27] Vladimir Vezhnevets, Vadim Konouchine, GrowCut: Interactive multi-label N-D image segmentation by cellular automata, in: Proc. of Graphicon, vol. 1, Citeseer, 2005, pp. 150–156.

[28] Cristiana Moroz-Dubenco, Laura Dioşan, Anca Andreica, Towards an unsupervised GrowCut algorithm for mammography segmentation, in: Computer Vision Systems, Springer Nature Switzerland, 2023, pp. 102—-111, http://dx.doi.org/10.1007/978-3-031-44137-0_9.

[29] Cristiana Moroz-Dubenco, Laura Dioşan, Anca Andreica, Generalizing an improved GrowCut algorithm for mammography lesion detection, in: Hybrid Artificial Intelligent Systems, Springer Nature Switzerland, 2023, pp. 709–720, http://dx.doi.org/10.1007/978-3-031-40725-3_60.

[30] Timo Ojala, Matti Pietikäinen, David Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1) (1996) 51–59.

[31] Timo Ojala, Matti Pietikäinen, Unsupervised texture segmentation using feature distributions, Pattern Recognit. 32 (3) (1999) 477–486.

[32] Timo Ojala, Matti Pietikainen, Topi Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[33] Timo Ahonen, Abdenour Hadid, Matti Pietikäinen, Face recognition with local binary patterns, in: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8, Springer, 2004, pp. 469–481.

[34] Timo Ahonen, Abdenour Hadid, Matti Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.

[35] Mary M. Galloway, Texture analysis using gray level run lengths, Comput. Graph. Image Process. 4 (2) (1975) 172–179.

[36] Haixia Li, Xianjing Meng, Tingwen Wang, Yuchun Tang, Yilong Yin, Breast masses in mammography classification with local contour features, BioMed. Eng. OnLine 16 (1) (2017) http://dx.doi.org/10.1186/s12938-017-0332-0.

[37] Adél Bajcsi, Anca Andreica, Camelia Chira, Significance of training images and feature extraction in lesion classification, in: Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART, SciTePress, INSTICC, 2024, pp. 117–124, http://dx.doi.org/10.5220/0012308900003636.

[38] Dilovan Asaad Zebari, Dheyaa Ahmed Ibrahim, Diyar Qader Zeebaree, Habibollah Haron, Merdin Shamal Salih, Robertas Damaševičius, Mazin Abed Mohammed, Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images, Appl. Artif. Intell. 35 (15) (2021) 2157–2203.

[39] Karl Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572, http://dx.doi.org/10.1080/14786440109462720.

[40] Harold Hotelling, Analysis of a complex of statistical variables into principal components., J. Educ. Psychol. 24 (6) (1933) 417.

[41] Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, Elsevier, 2013.

[42] Ronald A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (2) (1936) 179–188.

[43] Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone, Cart, in: Classification and Regression Trees, Wadsworth and Brooks/Cole Monterey, CA, USA, 1984.

[44] University of Cambridge, Mammographic image analysis society (MIAS) database v1.21, 2015, https://www.repository.cam.ac.uk/items/b6a97f0c-3b9b-40ad-8f18-3d121eef1459. (Accessed on 16 May 2024).

[45] Hoa Khanh Dam, Truyen Tran, Aditya Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, 2018, pp. 53–56.

[46] Michael Heath, Kevin Bowyer, Daniel Kopans, P Kegelmeyer Jr., Richard Moore, Kyong Chang, S Munishkumaran, Current status of the digital database for screening mammography, in: Digital Mammography: Nijmegen, 1998, Springer, 1998, pp. 457–460.

[47] Robin Hesse, Simone Schaub-Meyer, Stefan Roth, FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods, in: ICCV, 2023, pp. 3981–3991, http://dx.doi.org/10.48550/arXiv.2308.06248.