BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Interpretable Automated Breast Cancer Detection and Diagnosis System

– PhD Thesis Summary –

Scientific Coordinator
Prof. Dr. Anca Andreica

Student
Cristiana Moroz-Dubenco

2025

Keywords: Mammogram Analysis, Breast Segmentation, Lesion Classification, Computer-Aided Detection and Diagnosis, Machine Learning, Interpretability

Contents

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 how the decisions are made. 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 five feature extraction techniques, both keypoint-based and textural, and four methods for feature selection. To facilitate interpretation, we employ classical machine learning algorithms for benign/malignant classification and experiment with eight different methods. Our system reaches accuracy scores between 95% and 97% when tested on images from the mini-MIAS, mini-DDSM and RDBMC datasets, while also offering its users the possibility to analyze each of the steps. Moreover, we take the first steps towards a multi-modal system, analyzing the possibility of breast cancer diagnosis from biofluids and mammography reports.

List of Publications

The evaluation standards used are those valid as of October 1st, 2018. The list of conferences, as well as their categorization, is based on the international CORE classification [1]. The journal list is the one used by UEFISCDI for awarding articles published in international scientific journals [2].

1. ***Comparison of Gradient-Based Edge Detectors Applied on Mammograms*, Cristiana Moroz-Dubenco [1]**, published in *Studia Universitatis Babeș-Bolyai Informatica* (2021): pp. 5-18, rank D: 1 point;

2. ***Mammography Lesion Detection Using an Improved GrowCut Algorithm*, Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [2]**, presented at the *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2021), rank B: 4 points;

3. ***SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine?*, Ștefania D. Iancu, Ramona G. Cozan, Andrei Ștefancu, Maria David, Tudor Moișoiu, Cristiana Moroz-Dubenco, Adél Bajcsi, Camelia Chira, Anca Andreica, Loredana F. Leopold, Daniela Eniu, Adelina Staicu, Iulian Goidescu, Carmen Socaciu, Dan T. Eniu, Laura Dioșan, Nicolae Leopold** [3], published in *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* (2022): 120992, rank B: 0.267 points;

4. ***An Unsupervised Threshold-based GrowCut Algorithm for Mammography Lesion Detection*, Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, Camelia Chira [4]**, presented at the *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2022), rank B: 2 points;

5. ***Towards an Unsupervised GrowCut Algorithm for Mammography Segmentation*, Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [5]**, presented at the *International Conference on Computer Vision Systems* (2023) and published in *Lecture Notes in Computer Science* (2023): pp. 102-111, rank C: 2 points;

6. ***Generalizing an Improved GrowCut Algorithm for Mammography Lesion Detection*, Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [6]**, presented at the *International Conference on Hybrid Artificial Intelligence Systems* (2023) and published in *Lecture Notes in Computer Science* (2023): pp. 709-720, rank C: 2 points;

[1] `http://portal.core.edu.au/conf-ranks/`

[2] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole

7. ***Linear Discriminant Analysis Tumour Classification for Unsupervised Segmented Mammographies*, Cristiana Moroz-Dubenco, Anca Andreica [7]**, presented at the *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2023), rank B: 4 points;

8. ***Machine-Learning-Based Approaches for Multi-Level Sentiment Analysis of Romanian Reviews*, Anamaria Briciu, Alina-Delia Călin, Diana-Lucia Miholca, Cristiana Moroz-Dubenco, Vladiela Petrașcu, George Dascălu [8]**, published in *Mathematics* 12, no. 3 (2024): p.456, rank C: 0.5 points;

9. ***Towards an Interpretable Breast Cancer Detection and Diagnosis System*, Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, Camelia Chira [9]**, published in *Computers in Biology and Medicine* 185 (2025): 109520, rank A: 4 points.

Total points from publications: 19.767

Chapter 1

Introduction

1.1 Motivation

According to the World Health Organization, 2.3 million women were diagnosed with breast cancer in 2020, while another 7.8 million women were diagnosed between 2015 and 2020. This disease arises in the glandular tissue of the breast, but may invade surrounding breast tissue and spread to nearby or distant organs. Women who die from breast cancer, die because of the metastasis. Therefore, if the cancer is discovered early and treated adequately, its growth and spreading can be prevented and the life of the patient can be saved.

The most effective way to detect breast cancer in an early stage is through regular screening exams. Mammography is one of the most used screening methods, consisting of an X-ray picture of the breast which is analyzed by doctors, looking for early signs of breast cancer. In Europe, breast cancer screening protocols typically involve double-blind readings. For this reason, more and more radiologists use computer-aided detection (CADe) and diagnosis (CADx) systems as a second opinion, to reduce the workload and to improve the predictive accuracy.

Furthermore, automated systems can assist in analyzing previous mammographic reports to assess changes in a lesion between consecutive mammograms, or can integrate medical imaging with supplementary information, such as biomarkers from biofluids, to enhance diagnosis accuracy.

Artificial intelligence (AI) is widely used in such systems. AI models have the potential to assist radiologists, leading to an increased diagnosis accuracy, while also reducing the workload and the risk of overdiagnosis. However, in critical fields such as healthcare, where lives are at stake, the interpretability of these systems is crucial. For this reason, we focus on interpretable models.

1.2 Objectives

The objectives of this thesis can be summarized as follows:

1. Build an interpretable automated detection and diagnosis system for breast cancer from mammographies

   (a) find a pre-processing technique that removes all unwanted information in a robust manner, while being capable of visual explanations;

   (b) starting from an existing interpretable segmentation technique, remove the need for human intervention, increase its performance and decrease its computational time;

   (c) find the best-suited feature extraction method to work in conjunction with the newly-proposed segmentation method, generating features that are easy to understand;

   (d) choose a feature selection algorithm to reduce the dimensionality in a transparent manner;

   (e) select a classification technique that best balances the trade-off between interpretability and predictive performance.

2. Analyze other types of medical data that could be used together with mammographies to increase the confidence of the diagnosis

   (a) find the classifier that is best-suited to be used for SERS analysis of biofluids;

   (b) choose the representation and classification methods to be employed for extracting a diagnosis from medical reports in two languages (Romanian and English).

Aside from these main objectives, our work intends to achieve the following secondary goals:

- analyze the impact of a dataset's characteristics on the detection and diagnosis processes,
- study the influence of the number of seeds on a region-based segmentation method,
- examine the effect that breast density has over a system's performance,
- create a methodology for evaluating the interpretability of a system,
- create a dataset that contains the textual description of the mammographies, along with the actual images,
- investigate to what extent does language influence the classification results when it comes to medical reports,
- evaluate the relation between a patient's age and their diagnosis.

1.3 Original Contributions

As a means to achieve our objectives, our original contributions are detailed below:

1. *The Threshold-based GrowCut (TbGC) algorithm*, an improved version the existing semi-supervised GrowCut segmentation method. The proposed approach incorporates a constraint on the number of iterations, thereby significantly accelerating the convergence process, making it more suitable for real-time or high-throughput applications. Moreover, by incorporating a thresholding mechanism that refines the labeling process, it improves the segmentation accuracy and maintains a robust lesion delineation in spite of the reduced computational time.

2. *Three methods for automating the generation of initial background seeds*, that can be used in conjunction with TbGC, but also for any seed-based segmentation technique. These methods aim to reduce user dependency by automatically identifying reliable background regions, enhancing segmentation consistency, and improving the accuracy, particularly in cases where manual seed selection is challenging or prone to variability.

3. *Two methods for generating the initial foreground seeds in an unsupervised manner*, which can also be employed for any seed-based technique. With these methods, the need for human intervention can be completely eliminated, and potential lesions can be automatically identified. By ensuring a more consistent and objective initialization, these approaches enhance segmentation accuracy while reducing variability introduced by human intervention.

4. *The integration of TbGC (with automatically generated seeds) in a complete CAD system.* By combining the enhanced segmentation capabilities of TbGC with the proposed automated seed generation methods, the system achieves greater robustness, efficiency, and consistency in lesion detection and delineation, enabling fully automated region-of-interest identification. The incorporation of this segmentation approach in a system for breast cancer detection and diagnosis provides a streamlined pipeline for mammographic analysis, in which each step can be visualized and examined.

5. *The Romanian Dense Breast Mammography Collection*, a novel dataset designed to support research on breast cancer detection in dense breast tissue. The dataset comprises both 2D and 3D mammographic images, accompanied by histopathology reports and radiological assessments, making it a valuable resource for developing and evaluating advanced CAD systems. The inclusion of comprehensive clinical annotations and patient metadata enables in-depth studies on the relationship between imaging characteristics and diagnostic outcomes. To the best of our

knowledge, there is no other publicly available collection focusing exclusively on patients with dense breast tissue, thus RDBMC addresses a critical gap in breast cancer imaging research.

6. *The evaluation of our proposed CAD system on the new RDBMC dataset*, which only contains patients with dense breasts. Since dense breast tissue poses a significant challenge for lesion detection due to its reduced contrast and increased likelihood of false positives, this evaluation provides insights into how well the system generalizes to complex real-world scenarios. The results are compared against radiologist interpretations and existing automated methods, highlighting the system's strengths in handling challenging cases, improving detection rates, and reducing misdiagnoses.

7. *A methodology for evaluating a system's interpretability*, based on a step-by-step rating framework that quantifies how understandable and transparent the system's decisions are to human experts. The methodology evaluates each stage of the CAD pipeline, from pre-processing to final classification, assigning interpretability scores based on expert feedback. This structured approach provides a means to compare different CAD systems in terms of interpretability, facilitating the development of more transparent and trustworthy AI-driven diagnostic tools.

8. *The diagnosis of breast cancer and BI-RADS scoring from textual mammography reports in Romanian*, along with an analysis on the influence of language, age and representation on the classification results. The proposed approach predicts malignancy and detects the BI-RADS score, handling medical terminology and variations in reporting styles. Additionally, the analysis conducted to determine how patient demographics and linguistic nuances affect model accuracy ensures the system's robustness. This contribution addresses the under-representation of non-English medical datasets and highlights challenges in automated diagnosis based on free-text reports.

9. *An analysis on automated breast cancer diagnosis from biofluids.* With this study, we aim to evaluate the performance of serum and urine in breast cancer liquid biopsy using label-free SERS analysis. The proposed methodology could be further integrated in a multi-modal CAD system for breast cancer, increasing diagnosis accuracy.

Chapter 2

Mammography Lesion Detection and Diagnosis

Breast cancer is the most frequently diagnosed cancer in women and the second leading cause of cancer-related deaths, following lung cancer [10]. In most cases, breast cancer can be cured 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. One of the primary methods used for this purpose is mammography, a medical imaging technique that allows for detailed examination of breast tissue.

Mammography interpretation has two main goals: lesion detection and diagnosis. These two distinct, yet interrelated, processes are essential for breast cancer screening and assessment. Lesion detection refers to the initial identification of abnormal regions within a mammographic image, aiming to highlight areas that may require further evaluation. Diagnosis, on the other hand, involves determining the nature of the detected lesion – whether it is benign or malignant. While detection focuses on ensuring that no potentially harmful lesion is overlooked, diagnosis is crucial for minimizing unnecessary biopsies and optimizing patient management.

Mammographies are obtained by using a low-dose X-ray system [11], characterized by a number of special features, which can pose challenges in detecting breast lesions [12]:

- they are gray-scale images;
- they can contain weak boundaries;
- they exhibit Gaussian noise;

- they can contain a different type of background noise: artifacts such as medical labels;
- they can have low quality, low contrast and poor illumination, depending on the machine used;
- dense tissue, like connective or glandular tissue, appears brighter, making it difficult to be differentiated from tumors, which are also made up of dense tissue;
- a mass might be not only a tumor, but also a cyst or a fibroadenoma;
- the contour of a mass, especially a malignant one, is not always well-defined.

Moreover, even if detection is performed successfully, additional challenges might arise when it comes to diagnosis, due to lesions possibly varying significantly in size, shape, density, and texture [13]. Malignant lesions can appear subtle, while certain benign abnormalities mimic cancer [14]. High rates of false positives and false negatives further complicate the process, leading to unnecessary biopsies or delayed treatment [15].

Computer-aided detection and diagnosis systems are meant to assist doctors in the interpretation of medical images, thus applying for mammographies as well, by providing a secondary opinion to their judgment [11]. These systems process the images searching for conspicuous sections and structures, being based on highly complex pattern recognition. Medical images and, in particular, mammograms, are served to the system and analyzed in several steps:

1. *Preprocessing* – enhancing image quality by reducing noise, normalizing contrast, and removing artifacts to improve lesion visibility;
2. *Segmentation* – identifying and isolating regions of interest (ROIs) that may contain abnormalities, such as masses or microcalcifications;
3. *Feature extraction* – analyzing ROIs to extract relevant features, including shape, texture, and density, which help differentiate normal from abnormal tissue;
4. *Feature selection* – selecting the most significant features;
5. *Classification* – assigning detected lesions to a specific category – benign or malignant – based on extracted features and learned patterns.

2.1 Pre-processing

Pre-processing is a critical step in computer-aided detection and diagnosis systems, designed to enhance the quality of mammography images for more accurate and reliable analysis. This stage involves a series of image processing techniques aimed at improving consistency, reducing noise, and enhancing visibility of potential abnormalities.

The process begins with normalization, which compensates for variations in image acquisition parameters, such as exposure levels, contrast differences, and scanner settings. This step is essential to standardize image characteristics, ensuring uniformity across different mammograms and facilitating consistent analysis.

Following normalization, image denoising is applied to remove noise artifacts that could obscure important details or interfere with feature extraction. Various denoising techniques, such as median filtering, Gaussian smoothing, or wavelet transform, may be utilized depending on the type and intensity of noise present in the image.

Once the noise is minimized, image enhancement techniques are employed to improve the clarity and contrast of mammographic images, making subtle abnormalities more distinguishable. Methods such as contrast stretching, histogram equalization, and adaptive histogram equalization are commonly used to amplify important visual details, particularly in dense breast tissue where lesions may be more difficult to detect.

By applying these pre-processing techniques, CAD systems can optimize image quality, ensuring that subsequent steps – such as segmentation, feature extraction, and classification – are performed with higher accuracy and reliability, ultimately improving breast cancer detection and diagnosis. Depending on the characteristics of the images and the desired outcome, the order in which these techniques are applied can be adjusted, or certain steps may be omitted.

2.2 Segmentation

Segmentation is the partitioning of an image into sets of pixels, named areas of interest, according to certain criteria. It is used to recognize, extract or identify objects in images. The goal is to simplify the representation into something easier to process or analyze. In other words, image segmentation is the process of assigning labels to pixels so that all the pixels with a particular label share certain characteristics.

Image segmentation is used in many different areas, such as: machine vision, surgery planning, traffic control systems, face recognition, brake light detection etc.. In this thesis, we are focusing on medical image segmentation and, precisely, on tumor detection in mammographic images. That is, identifying the region of interest in order to further analyze and classify it either as benign or malignant.

Accurate segmentation is an essential step in mammogram interpretation, as the shape of a mass is one of the factors to differentiate between benign and malignant masses [16], while the size of a tumor is an important factor when deciding if surgery can be performed. Yet, due to the nature of breast tissue, mammography segmentation can prove to be a rather difficult task. Tumors can have different shapes and sizes, they can differ in density and localization, the breast can contain other abnormalities, such as cysts, that can be mistaken by tumors. Moreover, the contour of a mass, especially when talking about malignant masses, is not always well-defined. If we take into consideration the possibility of having low contrast images, low image quality, high noise levels or poor illumination [17], we can state that mammography segmentation might prove challenging.

Various techniques of mammography segmentation have been proposed so far, from which we name just a few:

1. *Thresholding* – turns a gray-scale image into a binary one;

2. *Clustering* – divides the pixels into groups such that the pixels in a group are more similar to one other than to the pixels from other groups;

3. *Histogram based* – computes a histogram from all the pixels of an image and uses the peaks and valleys from the histogram to locate the clusters;

4. *Region growing* – compares one pixel with its neighbors and, if a similarity criterion is met, it is set to belong to the same cluster as one or more of its neighbors.

Out of these methods, region-based segmentation is widely regarded as more interpretable than the others, due to its ability to preserve spatial coherence and provide meaningful object representation. Unlike thresholding and histogram-based approaches, which rely solely on pixel intensity values and may struggle with uneven lighting or noise, region-based methods aggregate pixels based on shared characteristics, ensuring that segmented regions correspond to coherent structures within the image [18]. This spatial consistency is particularly valuable in applications like medical imaging, where anatomical structures must be accurately delineated for diagnosis and treatment planning. Additionally, region-growing techniques group pixels in a way that directly aligns with real-world objects, making the segmentation results more intuitive and easier to interpret. Compared to clustering methods, which segment images based on statistical similarity but may disregard spatial relationships, region-based segmentation produces more robust boundaries by considering local pixel interactions [19]. These advantages make region-based segmentation a preferred choice in applications requiring high interpretability and precision, such as biomedical imaging and remote sensing. Therefore, in alignment with our interpretability goal, we focus on region-based segmentation.

The region growing segmentation techniques require the selection of initial seed points – a number of pixels labeled prior to the beginning of the algorithm. Selected based on user criteria, the initial seeds' locations are considered the initial regions. Then, the regions are grown to adjacent points depending on the similarity between a pixel and the pixels from a particular region.

2.3 Feature Extraction

Feature extraction involves identifying and extracting relevant features from the segmented mammography images, which are subsequently used for classification and identification of potential malignancies. Effective feature extraction can improve the accuracy and reliability of breast cancer diagnosis, ultimately leading to better patient outcomes.

In breast cancer CAD systems, feature extraction is typically performed using image analysis techniques that can identify various properties of potential breast cancer lesions. These characteristics can include texture, shape, and intensity measures, among others. The choice of feature extraction technique depends on the specific requirements of the CAD system and the characteristics of the breast cancer lesions being analyzed.

Texture analysis is a common feature extraction technique used in breast cancer CAD systems. Texture features are based on the spatial distribution of intensity values in the mammography images and can be extracted using techniques such as Gray-Level Co-occurrence Matrix (GLCM) analysis and Local Binary Patterns (LBP) analysis.

Keypoint-based feature extraction plays an important role in breast cancer diagnosis from mammograms, enabling the identification of distinctive patterns within the images. This process involves detecting and describing localized features, such as edges, textures, or intensity variations, which are essential for characterizing suspicious regions. Common techniques, including Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Histogram of Oriented Gradients (HOG), enhance the system's ability to differentiate between normal and abnormal tissue structures. By capturing invariant and robust features, key-point extraction helps improve lesion detection, classification, and overall diagnostic accuracy.

Shape analysis is another important feature extraction technique in breast cancer CAD systems. Shape features are based on the geometric characteristics of the potential breast cancer lesions and can be extracted using techniques such as boundary or morphological analysis. Boundary analysis involves measuring the contours of the potential lesions, while morphological analysis involves analyzing the size, shape, and location of the lesions in relation to surrounding tissue.

Intensity measures are also commonly used as features in breast cancer diagnosis. These features are based on the brightness and contrast of the mammography image and can be extracted using techniques such as histogram analysis and wavelet analysis. Histogram analysis involves measuring the distribution of pixel intensity values in the mammography image, while wavelet analysis involves decomposing the image into multiple frequency bands.

Textural and keypoint-based feature extraction methods are preferred over shape- and intensity-based approaches in CAD systems for breast cancer detection and diagnosis, as they capture complex patterns that may not be easily discernible to the human eye. While radiologists can directly assess shape and intensity from mammographic images, texture and keypoints provide additional, quantifiable information that enhances automated analysis and supports more informed decision-making.

2.4 Feature Selection

Feature selection aims to identify the most relevant features from the set of extracted features obtained from the segmented mammography images. The selected features are subsequently used for the classification and identification of potential malignancies, which can aid radiologists in accurately diagnosing breast cancer. The feature selection techniques can be broadly classified into filter, wrapper, and embedded methods.

Filter methods involve evaluating each feature independently and ranking them based on their discriminatory power using statistical measures such as t-tests, ANOVA, or mutual information. The top-ranked features are then selected for use in the subsequent classification stage. Filter methods are computationally efficient but do not take into account the interaction between features.

Wrapper methods evaluate subsets of features by training and testing a classifier using different feature subsets. These methods evaluate the performance of the classifier on each subset of features and select the subset that produces the best classification accuracy. Wrapper methods can better account for the interaction between features but can be computationally expensive.

Embedded methods incorporate feature selection as part of the classifier training process. The most common embedded method is regularization, which involves adding a penalty term to the objective function of the classifier to discourage the use of irrelevant features.

Dimensionality reduction methods, such as Principal Component Analysis, aim to reduce the number of features in a dataset while preserving as much relevant information as possible. While these techniques do not perform feature selection in the traditional sense, they effectively reduce dimensionality by selecting the most informative components, thereby mitigating the curse of dimensionality and improving model generalization. They are usually preferred in CAD systems because they transform the original feature space into a lower-dimensional representation that preserves the most relevant information, whereas classical feature selection methods merely discard less important features, potentially losing valuable diagnostic insights [20].

In addition to the above techniques, feature selection can also be guided by domain knowledge, such as the characteristics of breast cancer lesions and the features that are known to be relevant in previous studies.

2.5 Classification

The classification step is a crucial component of any automated computer-aided diagnosis (CAD) systems. It involves using the selected features obtained from the segmented mammographies to classify anomalies as either benign or malignant.

Various machine learning algorithms, both supervised and unsupervised, can be used for classification in breast cancer CAD systems. Supervised methods require labeled training data to train the classification model, while unsupervised methods attempt to discover patterns in the feature space.

Supervised classification methods include Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Naive Bayes (NB), among others. Unsupervised classification methods include clustering techniques such as K-Means, hierarchical clustering, and Gaussian mixture models. These methods attempt to group similar data points together based on their feature representations. Unsupervised methods can be used to identify subtypes of breast cancer and can aid in the discovery of previously unknown patterns in the data.

The choice of the classification algorithm depends on the specific requirements of the CAD system and the characteristics of the breast cancer lesions being analyzed. Performance evaluation of the classification model is typically done using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

In addition to binary classification (benign vs. malignant), CAD systems can also perform more nuanced classification tasks, such as identifying the specific subtype of breast cancer or predicting the likelihood of malignancy. These tasks require more complex classification models and may require additional features or imaging modalities.

Chapter 3

A Novel Mammography Lesion Detection and Diagnosis System

We propose a complete, fully automated computer-aided detection (CADe) and diagnosis (CADx) system for mammography analysis. The system receives an abnormal mammogram (i.e. a mammogram containing a lesion) as input and outputs a binary value, representing whether the lesion is malignant or benign. As a means to that, the mammography is first pre-processed in order to remove any redundant information. The resulting image is then segmented and features are extracted, from which the most relevant one are selected. Finally, these features serve as input for a classifier, which outputs the diagnosis.

3.1 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.

We use the pre-processing method proposed by Bajcsi et al. [21] and generalized in [22]. This consists of image enhancement, external artifact removal and internal artifact removal. By external artifacts we refer to information outside of the breast and by internal artifacts we refer to information which is included in the mask of the breast. To enhance the image, morphological opening (for noise reduction) and histogram equalization (for contour emphasizing) are employed.

3.2 Segmentation

3.2.1 Threshold-based GrowCut

Our approach, the Threshold-based GrowCut algorithm, is an improvement of the GrowCut algorithm [23], a semi-supervised region-growing segmentation technique. It is meant to perform multi-label segmentation using a Cellular Automaton, with the image being the space of cells and the pixels being the cells. A pixel p from image P is characterized by a triplet consisting of its label l_p, strength θ_p – the certainty that the pixel belongs to the l_p class –, and feature vector $\vec{C}_p$. It starts with a number of user-labeled pixels, which are assigned a strength of 1, and iterates over the space of cells, updating their labels and strengths until no cell is updated during an iteration, which means that the automaton converged to a stable state.

Given that the algorithm stops only when the automaton converges, the computational time can rapidly increase. Therefore, with the TbGC method, we aim to reduce the computational time while persisting a high level of accuracy. As a means to this goal, the following changes are brought to the original method:

- change the cell evolution rule to update a pixel's label only if the new "strength" is higher than a threshold value, chosen experimentally;
- limit the algorithm to a fixed, experimentally chosen number of iterations, thus obtaining a result either when the automaton converges or when the maximum number of iterations is reached – whichever happens first.

With these modifications, the overall time complexity of the TbGC algorithm remains the same as that of the original, as the introduction of an additional condition to the cell evolution rule does not alter the fundamental computational complexity. However, by enforcing a fixed upper limit on the number of iterations, our improved version prevents excessive iterations in cases where changes continue to be detected but diminish in significance, leading to a reduction in execution time without impacting the core complexity of the method.

3.2.2 Automated Seeds Generation

Along with the TbGC algorithm, we propose three methods for automatically generating the initial background seeds, and two methods for the initial foreground seeds. These methods, although evaluated only in conjunction with TbGC, can be adapted and used with any region-growing segmentation technique.

3.2.2.1 Background Seeds Generation

Unprocessed mammographies contain not only the breast, but also a background, thus raising the question: should the initial background seeds be selected inside the breast or outside of it? If the initial background seeds are selected within the breast, they could overlap connective or glandular tissue which, as stated before, have the same properties as tumors. Also, one thing worth mentioning is that masses can be located very close to the edge of the breast, making it more difficult to manually indicate background seeds within the breast and outside the mass. On the other hand, if the initial background seeds are selected outside the breast, for dense breasts, the entire breast might be segmented as region of interest, with the algorithm not being able to differentiate between masses and normal, healthy, dense tissue.

These being said, in order to reduce the need for human intervention and to solve the problem of choosing initial background seeds, we propose three possible solutions:

1. Generate the initial background seeds inside the breast;

2. Use initial background seeds outside the breast;

3. Use only initial foreground seeds.

For the first variant, we use a square enclosing the ground truth, while for the second variant, we employ the darkest pixels from the image as background seeds.

3.2.2.2 Foreground Seeds Generation

Starting from the idea that the center of the abnormality has the highest chance to be correctly labeled by the user, we intend to construct an optimal set of foreground seeds in an automated manner, starting from the center of the tumor. In order for this to happen, we choose the foreground seeds as a circle with the center corresponding to the center of the tumor and the radius chosen experimentally.

After deciding on the best-suited radius value, which will be further referenced as r, we suggest a method for automatically generating the foreground seeds as a circle with a radius of r pixels. Taking into consideration the fact that, in a mammography, the tissue that composes a mass appears brighter than the rest of the breast, we aim to find the brightest circle with a radius of r pixels inside the breast.

In order to achieve this objective, we iterate over the image, considering each pixel at a time as the center of a circle of r pixels radius and compute the brightness of the circle, by summing the intensities of all the pixels that are inside the circle. Subsequently, we compare the obtained value with the ones previously obtained and retain the coordinates of the pixel that represents the center of the circle that yields the highest sum. After the entire image is parsed, we are left with the coordinates of the center of the brightest circle from the image. We use these coordinates to construct two types of circle with a r-pixel radius: (1) a hollow circle and (2) a filled circle, in order to analyze the impact of the number of seed points on the segmentation results and choose the variant that leads to a better segmentation.

3.3 Feature Extraction

Feature extraction plays a key role in automated breast cancer detection systems. We experiment with two types of features: keypoint- and texture-based. For the first category, we employ two methods: Oriented FAST and Rotated BRIEF (ORB) [24] and Scale Invariant Feature Transform (SIFT) [25], while for the second one, we compare three techniques: Local Binary Pattern (LBP) [26], Gray Level Run-Length Matrix (GLRLM) [27], and Gray Level Co-occurence Matrix (GLCM) [28].The methods were chosen according to our interpretability goal: the resulting features can either be visualized or easily computed.

3.4 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 four techniques, namely: Principal Component Analysis (PCA) [29, 30], Kernel Principal Component Analysis (KPCA) [31], Incremental Principal Component Analysis (ICPA) [32], and Singular Value Decomposition (SVD).

3.5 Classification

The aim of this thesis is to construct an interpretable CAD system. Hence, we select the methods for classification according to this goal: we leverage transparent, easy to understand algorithms, which are characterized by their comprehensibility by humans. With this goal, we experiment with the following algorithms: (1) Decision Tree (DT) [33], (2) Random Forest (RF) [34], (3) Gaussian Naive Bayes (GNB) [35], (4) K-Means [36], (5) K-Nearest Neighbors (KNN) Fix [37], (6) Linear Discriminant Analysis (LDA) [38], (7) Quadratic Discriminant Analysis (QDA) [39], and (8) Logistic Regression (LR) [40].

Chapter 4

Evaluation of the Proposed Mammography Lesion Detection and Diagnosis System

4.1 Detection

In order to clearly evaluate the segmentation method described in this thesis, we evaluate the TbGC algorithm with automated seeds on mini-MIAS [41] and mini-DDSM [42]. As a means of properly evaluating our improvements, we compare our approach to the original GrowCut method.

4.1.1 Mini-MIAS

For mini-MIAS, in terms of automated background seeds generation, TbGC yields very similar results for all three alternatives, meanwhile the results produced by the original algorithm are getting worse when using background seeds outside the breast or not using background seeds at all. For the classical GrowCut, the segmented area extends up until the mammogram's background, as we set the background seeds, while in the lack of background seeds, it is extended to the mammogram's background as well. Although once we increase the number of iterations, the segmented area's center, obtained with the original approach, is getting closer to the center of the mass, its area enlarges, thus yielding false positives. For our proposed approach – TbGC –, the area of the segmented area also increases, but the difference is almost unnoticeable. Therefore, although the results obtained with the classical approach are better than our results for well-defined background seeds, inside the breast, there is an incontestable improvement in the results obtained with TbGC for the other two cases.

When generating also the foreground seeds in an automated manner, the best variant towards an unsupervised TbGC is the combination between background seeds outside the breast and foreground seeds generated as a circle with the center corresponding to the center of the mass and a radius equal to 25. With this configuration, our approach attains 98.52% accuracy, 67.76% precision and 57.36% recall.

4.1.2 Mini-DDSM

For mini-DDSM, when only the background seeds are generated automatically, TbGC obtains better results for all three variants, except for the recall value. We emphasize the fact that a higher value for the maximum number of iterations would lead to the entire mass being segmented by the Threshold-based GrowCut algorithm, and, thus, to an increased recall value. We also want to note that the accuracy obtained by TbGC is over 0.9 for all three cases, while for GrowCut, we can see a difference of almost 0.39 between the first and the third cases. Although the values obtained for the other metrics are not satisfactory, they are obviously better for the Threshold-based GrowCut algorithm (except for the recall value, as already highlighted). We can conclude that the existence and the localization (inside or outside the breast) of the background seeds do impact the segmentation results, but, since the differences in the results obtained with TbGC for the three variants are much lower than the ones in the results obtained with the original GrowCut, we can aver that TbGC is a flexible algorithm in comparison to the original GrowCut.

For the unsupervised version of our approach, an accuracy score of 97.12% is obtained, slightly lower than the one obtained on mini-MIAS. From the 66.43% precision and 51.01% recall scores, we can state that the segmented ROI does not coincide with the ground truth, being either smaller (leading to a decreased recall value) or larger (leading to a decreased precision value), yet containing at least half of the pixels of interest. However, we re-iterate the fact that our proposed approach is completely automated, not requiring any human intervention and, thus, given that TbGC is a region-growing technique and the initial seeds are automatically generated, we consider the results to be satisfactory.

4.2 Diagnosis

After obtaining the ROIs from the pre-processed mammographic images, the next step is to extract the features, select the most relevant ones, and perform classification. Because some of the employed methods are parameterized, we need to find the values that lead to the best classification results. In order for this to happen, we perform an exhaustive grid search, analyzing all possible combinations between the parameters values used for feature extraction, feature selection and classification. The total number of different configurations used in our experiments can be computed as follows: 5800 configurations for feature extraction · 280 configurations for feature selection · 267 configurations for classifications = 433608000 parameters combinations that are compared in order to obtain the best possible classification.

4.2.1 Mini-MIAS

From the results obtained on the mini-MIAS dataset, a few observations arise. First of all, all the classifiers employed in our experiments achieve the best performance when applied on LBP features. Similar performance is attained with GLCM features for DT, GNB, KNN and LDA. This leads to the first conclusions:

- textural features are better-suited for our system than the keypoint-based ones,
- LBP feature extraction works best for ROIs extracted using the unsupervised TbGC algorithm.

Following, examining the results from a feature selection perspective, PCA with its variants (KPCA and IPCA) attains a better overall performance than SVD. However, when it comes to choosing one particular variant, the individual scores must be taken into account: the usage of KPCA is the only one that leads to a 95% accuracy score, while with PCA and IPCA, the highest accuracy achieved is of 90%. Therefore, we consider the Kernel Principal Component Analysis to be the most efficient feature selection method employed in our experiments.

In terms of classification, four methods achieve 95% accuracy and F1-score: DT, RF, QDA and LR, thus proving more potential when it comes to being integrated in our proposed system. However, some methods obtain 100% precision and specificity scores (DT with GLCM and KPCA, RF with LBP and KPCA, QDA with LBP and KPCA and LR with LBP and KPCA), while others, perfect recall (DT with LBP and KPCA and QDA with LBP and SVD). Therefore, before deciding on a classifier, another decision needs to be taken: precision or recall? While higher recall 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 recall and precision, and we consider that the end-users should decide which one is more important.

Following, we base our choice on yet another trade-off: diagnosis accuracy versus transparency (achieved either through interpretability or explainability). We consider this to be an important consideration when choosing a model, especially for our goal, which assumes that the system will be used by people with no AI-related technical knowledge. While all the methods employed in our experiments provide some level of interpretability, we decide to use Random Forest as a means to balance the trade-off between transparency and performance. In conclusion, we move forward using LBP for feature extraction, KPCA for feature selection, and RF for classification. For this configuration, we also computed the AUC-ROC score, obtaining a value of 0.93.

4.2.2 Mini-DDSM

In order to prove the robustness of our proposed approach, we apply it to images from the mini-DDSM dataset. In the experiments, we opt for the methods achieving the best results on mini-MIAS. 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.

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, *F1-score* – 0.97, *AUC-ROC score* – 0.98.

These results prove the robustness of our proposed approach. The system trained on mini-DDSM surpasses the system trained on mini-MIAS in terms of accuracy. 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 is wrongly classified as malignant. When applying our system on the mini-MIAS dataset, 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.

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 a high performance.

4.2.3 RDBMC

As a final validation of our proposed approach, we evaluate it on the RDBMC dataset, which contains only dense breast mammographies. According to the American College of Radiology [43], women with dense breast not only have an increased risk of developing breast cancer, but it is also more difficult to be identified, due to lesions mimicking dense tissue on mammographies.

For this reason, a CAD system with satisfactory performance on dense breast tissue is highly desirable. As for mini-DDSM, we use the same methods that resulted in the best performance on mini-MIAS, and fine-tune their parameters. Although RDBMC does not provide a ground truth for segmentation, we consider this step to be essential towards a good classification, and, thus, we still apply it, even if we cannot individually assess its accuracy.

This combination of methods yields 96% accuracy, 100% precision and specificity, 88% recall and 93% F1-score. The perfect precision and specificity, identical to the results obtained on mini-MIAS, show that our proposed approach is a cautious one, leveraging certain decisions over doubts. The decrease in sensitivity, caused by two malignancies being wrongfully labeled, can be attributed to the rather imbalanced data used for both training and testing, leading to a slight bias in favor of benign classifications.

We consider that these results prove that our system is capable of generalization and can be employed regardless of the machine used for the mammogram exams, obtaining satisfactory performance on images with different technical characteristics. Moreover, evaluating the system on a dataset containing only dense breasts and achieving an accuracy score of 96% shows its potential for the usage in a real-life clinical environment.

4.3 Clinical Viewpoint

In order to asses our proposed system's alignment with the end-user, the results obtained for 25 images from the RDBMC dataset, randomly picked from the test data subset, were manually reviewed by an experienced radiologist, both in terms of detection and diagnosis. Out of the 25 images included in this study, an overlap between the automated segmentation result and the actual tumor was identified for 19 images. From the remaining 6 images, the expert could not identify the lesion in 2 of them, thus we cannot say if the system correctly identified the lesion or not. Excluding these 2 images, the percentage of images for which the segmentation result overlaps the lesion is of 82.61%. Out of the 19 images for which the detected ROI intersects the tumor, 2 show a perfect alignment, 3 lesions are over-segmented, and the rest are under-estimated.

Using the same 25 images used for validating ROIs detection, the radiologist provided diagnosis for 22 images, omitting the 2 mammographies where they could not be identify the tumor. An additional mammography was disregarded, for which the tumor was overlapping the pectoral muscle, causing the automated segmentation to wrongfully delineate the ROI. For the remaining cases, there were 16 matches between the radiologist's diagnosis and the outcome of our proposed CAD system. One of these matches was a misdiagnosis. From the images for which the radiologist provided a different diagnosis than the system, 5 contain benign lesions labeled as malignant, and one contains a malignant lesion labeled as benign. All these six images were correctly classified with our proposed approach.

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 radiology expert to evaluate the system from an interpretability point of view as well.

The pre-processing methodology is absolutely clear, as the radiologist does not only understand 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.

Chapter 5

Towards a Multi-Modal Breast Cancer Detection and Diagnosis System

In the process of detecting and diagnosing breast cancer, various information can play an important role. While medical imagining usually provides a clear enough view over a patient's breasts, some abnormalities can still "hide" in the dense tissue, making detection nearly impossible, even for the most experienced professionals. Therefore, especially for patients with dense breasts, combining mammographies with additional information might prove very helpful. For this reason, we analyze two types of data that can be further integrated into a multi-modal breast cancer detection and diagnosis system, along with the proposed system for mammographies.

5.1 Biofluids

There is a great need for new approaches in medical screening or diagnostics that do not require specialist operation and have a high level of accessibility, sensitivity and specificity. The attention is directed to fast, minimally invasive and easy-to-use techniques capable of early identification of diseases, of which label-free surface-enhanced Raman scattering (SERS) won great interest. SERS liquid biopsy – an approach in which machine learning algorithms enable classification of patient and control samples based on the SERS signal of biofluids –, has shown intense advancements towards the translation into the clinical setting over the past decade.

Despite the growing number of research articles in the field of SERS-based liquid biopsy for medical diagnosis, there are still many improvements to be made and questions to answer. Our goal is to assess the performance of serum and urine biofluids in detecting breast cancer.

As such, we compare four classification models – namely, DT, RF, GNB and LDA –, on a private dataset composed of 60 serum and urine samples obtained from the same patients, from which 39 were from patients with breast cancer (confirmed by biopsy) and 21 samples were from control subjects. The information is then divided into two sets, which are further used to train and validate the classifiers.

5.1.1 Experimental Results

LDA is the most common classifier for SERS-based diagnosis and indeed, for serum SERS spectra, LDA yielded the best classification results – 0.83. However, the highest classification accuracy was obtained by DT (0.89), based on the SERS spectra of urine, while LDA yielded an overall accuracy of 0.78, as well as RF and GNB.

To conclude, we tested the relative performance of serum and urine SERS samples for the classification of breast cancer and control samples using four different machine learning algorithms. Similar overall performances in the range of 61–89% indicate both, serum and urine, as candidates for SERS liquid biopsy for breast cancer detection. Slightly higher classification accuracies were achieved using urine samples compared to serum samples, despite the higher variability in the urine SERS spectra, which we link to variations in pH. By identifying the SERS metabolic signatures of serum and urine samples, the misclassified samples were correlated to specific imbalances in the metabolic profile of the respective samples.

We consider that the results of this SERS analysis can be combined with the mammographic exam for a more accurate prediction of breast cancer.

5.2 Medical Reports

Mammography reports, which contain detailed textual descriptions of imaging findings, can be processed using document-level analysis (DLA) techniques to automatically classify them. However, a main disadvantage of DLA refers to the assumption that a document, regardless of its length, belongs to a single class. In our case especially, a mammography report usually describes all the abnormalities present in a breast. This means that, if a breast contains both benign and malignant lesions, they all will be described within the same report. Moreover, if the exam was done for both breasts at once, the report will contains information about both breasts.

We propose two types of classification. The first one refers to the overall assessment of a patient's condition in terms of *healty*, *benign* or *malignant*. The second one is with regard to the gravity of the lesions, according to the BI-RADS classification.

For the experiments, we use both the original report, in Romanian, as well as its translation, as provided with the dataset. Moreover, in order to assess a patient's age connection to their diagnosis, we concatenate the age (also provided with the dataset) at the end of the report.

Following, we perform a pre-processing step, which involves the transformation of the text to lowercase. As for stopwords, experiments are run both with and without removing them, to asses their impact on the model performance. After the data is pre-processed, the text can be embedded. With this aim, we experiment with two methods: *Term Frequency-Inverse Document Frequency (TF-IDF)* and *Latent Semantic Indexing (LSI)*.

To assess the relevance of the TF-IDF and LSI-based embeddings when it comes to the automatic classification of medical reports written in Romanian and translated in English, we train and evaluate four standard machine learning classification models: DT, RF, NB and LR.

5.2.1 Experimental Results

For the first type of classification, the best-suited combination of algorithms (for embedding and classification) for our data is Random Forest classifier and TF-IDF embeddings. For Romanian, the best results are obtained on the reports alone, excluding stopwords: 80% accuracy, precision, recall and F1-score. The same configuration yields the best performance for English also, with an increase of 2% in accuracy and recall (82%), 4% in precision (84%) and 1% in F1-score (81%).

For assessing the BI-RADS score of a patient, the best performance metrics for Romanian are achieved with the same combination as for the three-class classification: RF applied on TF-IDF representations of reports (without the age), excluding stopwords: 73% accuracy, 77% precision, 73% recall and 67% F1-score. Due to the fact that the class corresponding to BI-RADS score 6 is under-represented, there is a decrease in performance compared with the 3-class classification. For English, however, the best combination seems to be between NB and LSI embeddings of reports and ages, without excluding the stopwords (75% accuracy and recall, 74% precision, 73% F1-score).

Overall, we analyze and compare four classifiers and two embedding methods on 24 types of data – reports and reports concatenated with age, in Romanian and English, excluding and preserving the stopwords – using the diagnosis or the BI-RADS score as the class, obtaining a 3-class and a 5-class classification. The experimental results have shown better performance for the 3-class classification, as well as the robustness of RF algorithm, obtaining the highest accuracy on both types of classification for Romanian.

We consider that the selected combinations can be used for classifying mammography reports and further integrated into a multi-modal system that accounts for current and previous imagining data.

Chapter 6

Conclusions and Future Work

In this thesis, 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 Kernel Principal Component Analysis. These features serve as input to a Random Forest classifier, which yields the final result: benign or malignant lesion.

Our approach 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, by using the Threshold-based GrowCut algorithm, an iterative segmentation method, the user can verify every step of the detection.

To obtain the presented CAD, we developed an unsupervised segmentation method, analyzed five feature extraction methods, four feature selection methods and eight classification methods, tested all the variants on 57 images from the mini-MIAS dataset [41] and chose the one that obtained the best results – 95% accuracy, 100% precision and specificity and 90% sensitivity. Moreover, we validated the proposed approach on a different dataset, namely mini-DDSM [44], obtaining 97% accuracy and 100% sensitivity. We also proposed a novel dataset – Romanian Dense Breast Mammography Collection – containing only mammographies of breasts with dense tissue, and tested our system on this dataset as well, obtaining 96% accuracy and 100% precision and specificity.

Additionally, we have already taken the first steps towards a multi-modal system for breast cancer diagnosis, by performing a preliminary analysis on the potential use of biofluids and medical reports in conjunction with mammographic images for an increased level of certainty.

While the system proposed in this thesis demonstrates strong performance and interpretability, there are several directions for future research that could enhance its accuracy, applicability, and integration into clinical workflows. To begin with, we intend to continue our efforts of integrating multiple types of data (textual and numerical) into a multi-modal system for breast cancer detection and diagnosis. This integration could enhance diagnostic accuracy by leveraging complementary information from mammography images, radiology reports, and biofluid-based SERS analysis. Additionally, we aim to explore the inclusion of genomic and patient history data, enabling a more comprehensive risk assessment framework.

Also as a means to improve diagnostic reliability, we intend to develop methods that aggregate results obtained from multiple mammographic images of the same patient. This includes combining findings from craniocaudal (CC) and mediolateral oblique (MLO) views, as well as from different imaging modalities such as tomosynthesis. This would allow the CAD system to reduce false positives and negatives by identifying consistent patterns across multiple images.

Since this thesis introduced a novel dataset focused exclusively on dense breasts, addressing a critical gap in breast cancer imaging research, as a next step, we also plan to introduce another dataset comprising only fatty breast mammograms. This will enable comparative analyses to assess how breast density impacts CAD system performance, lesion detectability, and interpretability. By performing cross-dataset evaluations, we aim to refine our model to ensure generalizability across different breast tissue types.

To further validate the system, we plan to conduct user studies to evaluate its interpretability at a larger scale. While initial evaluations have demonstrated the system's transparency, a broader study involving multiple radiologists and clinicians would provide more robust insights. With this aim, we plan to design a controlled study where experts interact with the system and provide feedback on how well each step aligns with their clinical reasoning.

Finally, in order to facilitate clinical adoption, we plan to integrate the CAD system into existing PACS (Picture Archiving and Communication System) infrastructures. This integration would enable radiologists to seamlessly access CAD-generated insights within their workflow. Furthermore, we intend to implement automated report generation to summarize key findings, highlight suspicious regions, and suggest follow-up recommendations.

Aside from possible improvements to the system proposed in this thesis, we also intend to develop predictive models that assess a patient's risk of developing breast cancer based on historical imaging data. By analyzing longitudinal mammographic datasets, such models could identify subtle changes over time that may indicate an increased risk of malignancy. Leveraging temporal patterns in imaging data could enhance early detection efforts, providing clinicians with valuable insights into disease progression and supporting more informed decision-making regarding screening frequency and preventive measures.

Another promising direction for future work is to extend the proposed methodology beyond breast cancer by adapting it for the detection and diagnosis of other cancer types, such as lung, prostate, or brain cancer. This would involve retraining the system on CT, MRI, or PET scans, adjusting the methods to accommodate the differences in imaging characteristics between mammography and other modalities.

Moreover, by adapting our methodology to different imaging modalities, we can explore the possibility of extending the proposed CAD system to a multi-organ analysis framework, enabling the simultaneous assessment of the primary organ and the organs where cancer is most likely to metastasize. This approach would enhance early detection of secondary tumors, improve treatment planning, and contribute to a more comprehensive understanding of cancer progression across multiple anatomical sites.

Bibliography

[1] C. Moroz-Dubenco. Comparison of gradient-based edge detectors applied on mammograms. *Studia Universitatis Babes-Bolyai Informatica*, 66(2):5–18, 2021.

[2] Cristiana Moroz-Dubenco, Laura Dioşan, and Anca Andreica. Mammography lesion detection using an improved growcut algorithm. *Procedia Computer Science*, 192:308–317, 2021.

[3] Stefania D Iancu, Ramona G Cozan, Andrei Stefancu, Maria David, Tudor Moisoiu, Cristiana Moroz-Dubenco, Adel Bajcsi, Camelia Chira, Anca Andreica, Loredana F Leopold, et al. Sers liquid biopsy in breast cancer. what can we learn from sers on serum and urine? *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 273:120992, 2022.

[4] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, and Camelia Chira. An unsupervised threshold-based growcut algorithm for mammography lesion detection. *Procedia Computer Science*, 207:2096–2105, 2022.

[5] Cristiana Moroz-Dubenco, Laura Dioșan, and Anca Andreica. Towards an unsupervised growcut algorithm for mammography segmentation. In *International Conference on Computer Vision Systems*, pages 102–111. Springer, 2023.

[6] Cristiana Moroz-Dubenco, Laura Diosan, and Anca Andreica. Generalizing an improved growcut algorithm for mammography lesion detection. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 709–720. Springer, 2023.

[7] Cristiana Moroz-Dubenco and Anca Andreica. Linear discriminant analysis tumour classification for unsupervised segmented mammographies. *Procedia Computer Science*, 225:2951–2960, 2023.

[8] Anamaria Briciu, Alina-Delia Călin, Diana-Lucia Miholca, Cristiana Moroz-Dubenco, Vladiela Petrașcu, and George Dascălu. Machine-learning-based approaches for multi-level sentiment analysis of romanian reviews. *Mathematics*, 12(3):456, 2024.

[9] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, and Camelia Chira. Towards an interpretable breast cancer detection and diagnosis system. *Computers in Biology and Medicine*, 185: 109520, 2025.

[10] ZiQi Tao, Aimin Shi, Cuntao Lu, Tao Song, Zhengguo Zhang, and Jing Zhao. Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, 72:333–338, 2015.

[11] Ghada M El-Banby, Nourhan S Salem, Eman A Tafweek, and Essam N Abd El-Azziz. Automated abnormalities detection in mammography using deep learning. *Complex & Intelligent Systems*, 10 (5):7279–7295, 2024.

[12] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic. A survey of image processing algorithms in digital mammography. *Recent advances in multimedia signal processing and communications*, pages 631–657, 2009.

[13] Heang-Ping Chan, Ravi K Samala, and Lubomir M Hadjiiski. Cad and ai for breast cancer—recent development and challenges. *The British journal of radiology*, 93(1108):20190580, 2019.

[14] Nicholas Konz, Mateusz Buda, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłędowski, Jungkyu Park, Jan Witowski, Krzysztof J Geras, Yoel Shoshan, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA network open*, 6(2):e230524–e230524, 2023.

[15] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

[16] J Anitha and J Dinesh Peter. Mammogram segmentation using maximal cell strength updation in cellular automata. *Medical & biological engineering & computing*, 53(8):737–749, 2015.

[17] Valliappan Raman, Putra Sumari, HH Then, and Saleh Ali K Al-Omari. Review on mammogram mass detection by machinelearning techniques. *International Journal of Computer and Electrical Engineering*, 3(6):873, 2011.

[18] Asad Mohammed Khan and S Ravi. Image segmentation methods: A comparative study. *International Journal of Soft Computing and Engineering*, 3(4):84–92, 2013.

[19] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10):1034, 2024.

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

[21] Adél Bajcsi, Anca Andreica, and Camelia Chira. Towards feature selection for digital mammogram classification. *Procedia Computer Science*, 192:632–641, 2021.

[22] Adél Bajcsi, Camelia Chira, and Anca Andreica. Extended mammogram classification from textural features. *Studia Universitatis Babeș-Bolyai Informatica*, 67(2):5–20, 2023.

[23] Vladimir Vezhnevets and Vadim Konouchine. Growcut: Interactive multi-label nd image segmentation by cellular automata. In *proc. of Graphicon*, volume 1, pages 150–156. Citeseer, 2005.

[24] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE, 2011.

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

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

[27] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.

[28] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, SMC-3(6):610–621, 1973.

[29] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[30] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[31] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[32] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International journal of computer vision*, 77:125–141, 2008.

[33] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[34] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[35] Hajer Kamel, Dhahir Abdulah, and Jamal M Al-Tuwaijari. Cancer classification using gaussian naive bayes algorithm. In *2019 International Engineering Conference (IEC)*, pages 165–170. IEEE, 2019.

[36] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28 (2):129–137, 1982.

[37] Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.

[38] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[39] Alaa Tharwat. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2):145–180, 2016.

[40] Jan Salomon Cramer. The origins of logistic regression. Technical report, Tinbergen Institute, 2002.

[41] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, volume 1069, pages 375–378, 1994.

[42] M Heath, K Bowyer, D Kopans, R Moore, and P Kegelmeyer. The digital database for screening mammography, iwdm-2000. In *Fifth International Workshop on Digital Mammography., Medical Physics Publishing*, pages 212–218. Medical Physics Publishing, 2001.

[43] American College of Radiology. Acr statement on reporting breast density in mammography reports and patient summaries, 2017. URL `https://www.acr.org/Advocacy/Position-Statements/Reporting-Breast-Density`. Accessed 12.11.2024.

[44] Charitha Dissanayake Lekamlage, Fabia Afzal, Erik Westerberg, and Abbas Cheddad. Mini-ddsm: Mammography-based automatic age estimation. In *2020 3rd International Conference on Digital Medicine and Image Processing*, pages 1–6. Association for Computing Machinery, 2020.