BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION MATHEMATICS AND COMPUTER SCIENCE IN ENGLISH

DIPLOMA THESIS

Binary Mammogram Classification From Images Using Transfer Learning

Supervisor
Prof. Dr. Camelia Chira

Author
Domniț Alexandru Adrian

2025

ABSTRACT

Breast cancer is considered one of the leading causes of death for women worldwide along with being one of the most preventable causes of death for women. Detecting breast cancer as soon as possible in its progression plays a crucial role when talking about improving survival rates and still, mammography remains by far the most effective screening method. Even with this being the case, the manual interpretation of mammograms is a complex task, requiring a lot of time and effort while variability between radiologists and other factors propose challenges in the diagnostic process. In this context, the usage of intelligent systems that are based on deep learning can be of significant help in improving diagnostic accuracy and efficiency.

This thesis takes upon presenting a complete and modular pipeline for the binary classification of mammograms using modern Convolutional Neural Networks (CNNs). The presented approach integrates a preprocessing stage, automatic tensor conversion and transfer learning applied to state-of-the-art architectures. Experiments were conducted on two public and widely available datasets (CBIS-DDSM and Mini-MIAS) that represent both small-scale and large-scale evaluation scenarios. In comparison to current methods, the top-performing models demonstrated competitive performance with F1-scores of 0.77 on Mini-MIAS and 0.74 on CBIS-DDSM. A user-friendly online application was also created to enable users to input mammograms, view preprocessing results, and receive real-time classification predictions. This study establishes the groundwork for future incorporation into clinical decision support systems and demonstrates the usefulness of deep learning in assisting with breast cancer screening.

This study establishes the groundwork for future incorporation into clinical decision support systems and demonstrates the usefulness of deep learning in assisting with breast cancer screening. Such systems may become useful instruments in actual diagnostic procedures with more enhancements and validation on bigger clinical datasets.

This work is the result of my own activity. I have neither given nor received unauthorized assistance on this work.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation and Objectives

Breast cancer remains one of the most prevalent and life-threating cancers among women worldwide. Putting a strong emphasis on the information given by the World Health Organization **(WHO)**, breast cancer is by far the most frequently diagnosed cancer while at the same time, a leading cause of cancer-related deaths worldwide. This ravaging disease affects millions of women yearly while incidence rates increase steadily in both developed and developing countries. Even though advancements in the area have significantly improved the survival rate for many patients, outcomes still depend heavily on how early the disease is discovered. This specific type of cancer is not considered a uniform disease, it varies a lot in terms of its biological behavior, rate of progression and of course, the response to treatment. This factors sum up to the high complexity in treating breast cancer and to the need for effective screening programs. While population-wide efforts such as routine mammography screenings have shown measurable success in the reduction of the mortality rate but disparities in access, diagnostic accuracy and awareness are still considered significant barriers to global health equity[Ora24].

The main objectives of this thesis are:

- To explore different machine learning models and comparing them one to another for identifying and classifying breast cancer utilizing image datasets.

- To develop a deep learning-based approach in order to improve diagnostic performance and reliability.

- To evaluate the proposed approach using standard metrics and publicly available datasets, taking into account reproducibility and scalability.

1.2 Main contributions

This thesis has as the main focus the development of a complete pipeline for the classification of mammograms utilizing deep learning. This approach begins with a image preprocessing stage in order to remove artifacts and enhance contrast followed by converting cleaned images into normalized tensors. A custom CNN is trained afterwards on the processed dataset in order for it to make a distinction between benign and malignant cases.

The main contributions of this thesis are as follows:

- **The development of a custom preprocessing pipeline** tailored to mammograms, including pectoral muscle removal, noise reduction and contrast enhancement using CLAHE (Contrast Limited Adaptive Histogram Equalization). The main advantage of this pipeline consists in the enhancement of tumor visibility for improving downstream analysis.

- **Automated tensor conversion and dataset preparation** as to enable the efficient use of deep learning models by converting grayscale PNG's into normalized tensor formats.

- **Design and implementation of a convolutional neural network (CNN)** for the binary classification (benign/malignant) of mammograms.

- **Integration of preprocessing, transformation, training, and evaluation scripts** that translates into a modular and scalable diagnostic workflow, easily reusable or expandable for any future work.

- **Empirical evaluation** achieved by the usage of a controlled test subset, having the performance measured using classification accuracy across randomly sampled mammograms.

Chapter 2

Scientific and technical context

2.1 The Problem

Breast cancer is a complex and heterogeneous disease, mainly characterized by an uncontrollable cell growth in breast tissue which can eventually spread in the body (metastasis) if not detected and treated early. The most common method to this day for early diagnosis is mammography (a specialized medical imaging technique that consists of using low-dose X-rays in order to detect abnormal growths in breast tissue). While as a practice, mammography has been proven to be effective in reducing mortality rates, there are still significant limitations present. In medical practice, radiologists are required to visually inspect mammograms, which often contain subtle lesions, overlapping dense tissue or even benign abnormalities that could mimic malignancies. Taking out the fact that this process is time-intensive, there also arise the issues of inter-observer variability, cognitive bias and fatigue, factors that may influence the diagnosis and produce missed malignancies or false positives[EJA+09].

Extrapolating the issue at hand, some tumors (particularly in generally younger patients with denser breast tissue) may be obscured or even appear similar to the naked eye to benign conditions, further complicating the detection process. This being taken into consideration, even with experienced radiologists, achieving consistently accurate diagnoses can become a challenge. While the demand for early detection increases, especially into population-wide screening programs, the limitations or even the unavailability of manual analysis in some cases have prompted researchers to dig further into automated and Artificial Intelligence assisted methods for diagnosing breast cancer.

2.1.1 Technical formulation of the problem

This thesis works towards the binary classification of breast tumors utilizing mammographic images. The main objective would be to develop a computational pipeline that can make a distinction between benign and malignant cases based only on image data, by the use of a supervised learning network.

Let:

- $X \subset \mathbb{R}^{1\times H\times W}$ be the set of preprocessed grayscale mammograms (resized to 256×256),
- $Y = \{0, 1\}$ be the binary label space where 0 = benign and 1 = malignant.

The main objective is to learn a function $f_\theta : X \to Y$, parametrized by weights θ, minimizing classification error over a labeled training set. Given that the proposed model outputs unnormalized scores (logits) for two classes, the loss function used would be **categorical cross-entropy loss**:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{z_{i,y_i}}}{\sum_{j=1}^{2} e^{z_{i,j}}}\right)$$

Here:

- $z_{i,j}$ is the logit (unnormalized score) produced by the model for input x_i and class j,
- y_i is the correct label index (0 or 1) for input x_i,
- N is the number of training samples.

This loss encourages the model to assign a higher score to the correct class label. During inference, the predicted class is the one with the maximum logit value:

$$\hat{y}_i = \arg\max_j z_{i,j}$$

2.2 Importance and Challenges

The importance of addressing this problem properly cannot be overstated. Breast cancer affects millions of women globally each year and the likelihood of successfully treating the disease improves drastically if it is detected in its early stages when it is most treatable. Studies show that early detection utilizing imaging and biopsy techniques can reduce the mortality rate of breast cancer by as much as 25-30%[Ora24]. With all that in mind, current diagnostic workflows often suffer from high false positives and false negatives rates. Another problem that can be seen in the presence of a high number of false positives and false negatives is the fact that false positives can lead to unnecessary biopsies and psychological stress while the latter can result in the delay of treatment, reducing survival chance[LKB+17].

Several technical and clinical challenges stay in the way of the development of robust automatic diagnostic systems:

- **Data Imbalance:** Most publicly accessible datasets contain significantly more benign cases than the malignant ones. In a practical scenario, this can lead to the biasing of models into predicting the majority class, reducing the overall sensitivity to cancers.
- **Low Contrast and Image Variability:** Mammographic image quality can be influenced by a variety of factors like imaging equipment, acquisition protocol and patient anatomy, leading to image inconsistency and variability. These variations make models less efficient in generalizing across different datasets.
- **Lack of Annotated Data:** Deep learning models need extensive volumes of high-quality annotated data. In medical imaging, the process of procuring this sort of datasets is expensive, time-consuming and more often than not subject to strict medical privacy regulation[SOPH16].
- **Model Interpretability:** A general challenge for wide adoption would be the fact that clinical adoption requires the AI tools to be able to be transparent and easily explainable, not only accurate. Black-box models that do not offer any insight into their decision-making process are less likely to be trusted and accepted by medical practitioners or regulatory bodies.
- **Boundary Detection:** Delineating tumor edges is a precise endeavour and it is essential for treatment planning, but it remains a significant challenge because of the irregular shapes and subtle cues of some lesions[JA15][IMDA23].

Given all these limitations, any progress towards improving segmentation accuracy, automation or even clinical reliability has the potential of making a meaningful impact in real-world diagnostic environments.

2.3 Applications

The development of automated systems for mammogram classification has direct applications in clinical practice, as well as in large-scale public initiatives. The adoption of artificial intelligence (AI), more precisely deep learning-based models, into diagnostic pipelines can improve accuracy, reduce workload and help with the standardization of breast cancer screening practices across different medical environments.

- **Clinical decision support:** Mainly in radiology settings, automated mammogram classifiers can be used as **decision-support tools** by signaling potentially malignant cases for further review by a certified professional. This comes in handy in order to combat human error or misdiagnosis occurred because of fatigue.
- **Early detection and screening programs:** One of the most promising use case for this sort of approach would be in national screening programs, programs in which AI could process large volumes of mammograms efficiently without the need for human intervention. Of course, this allows for faster triage of cases, ensuring that high-risk patients gain priority in further investigations, especially in regions with limited access to specialized healthcare workers.
- **Telemedicine and remote diagnostics:** In areas with poor access to specialized healthcare or in rural regions, mammogram classification systems could be deployed as part of telemedicine platforms.
- **Education and training:** Deep learning models could also be used in the education process of future radiologists. By putting side to side model outputs with expert annotations, students could better grasp the decision-making process and improve their diagnostic skills thorough guided learning.
- **Research and dataset annotations:** A well-preforming classification system could make a huge difference in semi-automated labeling of large image datasets, speeding up the annotation process. This is incredibly valuable in medical domains because it would translate into lower costs and easier access to the respective datasets.
- **Integration into CAD systems:** Binary classification models such as the one presented in this thesis would be the foundation of Computer-Aided Diagnosis (CAD) systems. This sort of systems often combine multiple tasks such as preprocessing, segmentation, classification and report generation into a unified diagnostic tool that supports clinicians.

Chapter 3

Related work

In recent years, various artificial intelligence driven techniques were developed having the main focus the detection of malignant and benign growths in mammograms. The most common methodology mainly involves leveraging specialized image processing techniques, including but not only machine learning based segmentation and classification, in order to enhance accuracy and detection. These methods offer a wide range of advantages, such as the ability to detect early stage breast cancer, the possibility to reduce the number of false positives and negatives, and the ability to provide a second opinion to radiologists. The following section will describe some of the most effective and commonly used solutions currently available.

3.1 Breast Cancer Images Segmentation using Fuzzy Cellular Automaton

Article [IMDA23] proposed a branching between two approaches for tackling the problem at hand, the first one being a Cellular Automata (a mathematical model where cells update based on local rules) and Fuzzy Logic (a technique that allows flexible classification of pixels rather than binary decisions). The used dataset [ea94] consists of 322 mammograms of which, 188 containing malignant tumors and the algorithm is being tested on 5 sample images and then on the entire dataset. The data was selected in such a manner that each image sits at a dimension of 1024 x 1024 pixels along with annotations about mass locations, abnormality types (benign/-malignant) and the radius of the suspicious regions. Unlike other larger datasets, [ea94] is small enough for allowing faster training, testing and quick experimentation with different segmentation techniques.

The authors of the article started their implementation by preprocessing the input data. The preprocessing stage is achieved by converting the mammograms to grayscale, applying Gausian Blur and Thresholding to remove the noise, detecting and removing the pectoral muscle in order to avoid false positives and extracting the region of interest.

Cellular Automata (CA) is a technique used in image segmentation, particularly for identifying regions of interest (ROI) such as tumors. The algorithm starts from a seed point (ideally the tumor center) and expands outward, grouping pixels based on similarity, intensity and texture. In order to achieve neighborhood definition, the Von Neumann neighborhood is used, where each pixel takes into consideration the four closest neighbors (up, down, left and right). This structure ensures that the expansion of the segmentation process takes place in a controlled manner. Another classification method present in the cited paper would be Fuzzy Logic for improved boundaries where instead of strict binary classification (tumor vs non-tumor), Fuzzy Sets classify pixels with some degree of suspicion. The main advantage that comes from this approach is a greater smoothness of tumor edges, helping with accuracy in medical diagnostics.

The metrics used in evaluating the effectiveness are: intersection over union (IOU), accuracy, precision, recall, Jaccard index and Dice coefficient. As shown in the Tab. 3.1, The Fuzzy Cellular Automaton achieves higher accuracy and better contour detection than standard Cellular Automata helping with the conclusion that the addition of Fuzzy Logic improves tumor boundary detection, making it more reliable in medical use.

Metric	Cellular Automaton	Fuzzy Cellular Automaton
Accuracy	92.46%	98.66%
Precision	52.05%	52.54%
Recall	73.00%	72.87%
Jaccard Index	40.87%	41.18%
Dice Coefficient	53.68%	54.10%

Table 3.1: The results obtained from [IMDA23], using different classification algorithms.

An important observation made by the authors of [IMDA23] is that this type of approaches helps with the preservation of the interpretability of the segmentation process, this step being crucial in real life clinical practice. Additionally, the CA model‘s locality principle makes way for efficient parallel computation, making it scalable for higher-resolution datasets in future applications.

3.2 Mammography Lesion Detection Using an Improved GrowCut Algorithm

Article [MDIA21] presents a novel approach for detecting breast cancer lesions in mammograms using a modified version of the GrowCut algorithm. The dataset used in this study is the Mini-MIAS dataset [ea94]. The authors of the article have implemented a preprocessing stage that includes image enhancement, noise reduction, and pectoral muscle removal. The proposed algorithm is tested on 10 sample images and then on the entire dataset.

The original GrowCut algorithm is a semi-supervised segmentation technique that takes user given seed data for each object in order to be segmented. It iterates over the present pixels in the image until each pixel has a class assigned. Every pixel is being characterized by a label, a ”strength” representing the certainty of its class and a feature vector, based on the image intensity. The original algorithm relies heavily on human input in order to select initial seeds, which can become problematic in mammographic images due to their size and the difficulty in distinguishing tumors from dense tissue.

The authors of [MDIA21] proposed the usage of the Canny edge detection operator in order to reduce noise and improve segmentation accuracy as much as possible. Unlike the original GrowCut algorithm, this article talks about Threshold-based GrowCut (TbGC), a method in which a threshold value is being introduced in order to limit the label changes of pixels based on their ”strength” and limits the algorithm to a fixed number of iterations in order to conserve time.

For the background seed selection, the authors explored three branching variants:

- Generating background seeds inside the breast based on foreground seeds.
- Using the image’s background as initial seeds.
- Not using initial background seeds at all.

The results of the proposed algorithm are evaluated using the Dice coefficient, Jaccard index, and Hausdorff distance. The authors conclude that the TbGC algorithm outperforms the original GrowCut algorithm in terms of segmentation accuracy and computational efficiency. The proposed method achieves a Dice coefficient of 0.87 and a Jaccard index of 0.76, indicating a high level of agreement between the segmented regions and the ground truth annotations.

In the cited article, there is a discussion being made about the potential of integrating Threshold-based GrowCut with deep learning-based classifiers after the segmentation step, in order to refine diagnostic accuracy. This being taken into account, it is safe to assume that for future research, hybrid models could be designed for both segmentation and classification within a unified pipeline.

Furthermore, the use of fixed iterations brings into equation the predictability in computational requirements, which can have a positive impact on real-time diagnostic systems. Taking into account both the algorithmic efficiency and the high segmentation accuracy, there is a deduction to be made that this method would be suitable for low-resource clinical environments.

3.3 An Efficient Method for Automated Breast Mass Segmentation and Classification in Digital Mammograms

Another relevant approach is being presented in [BNF21]. A vast number of traditional CAD systems rely purely on segmentation and hand-crafted features which are not optimal. These methods often use wavelet coefficients and features extracted from wavelet coefficients in order to detect breast abnormalities in mammograms. The main drawback from this sort of approach consists in the fact that these methods have limitations including but not solely high computational cost and the presence of a considerable number of false positives. More recently, deep learning-based approaches, especially convolutional neural networks (CNNs), have been introduced in order to tackle these limitations. However, these methods require large amounts of labeled data for training which, especially in the medical field, can be very expensive and time-consuming to aquire.

This study aims to provide a novel method for automatic segmentation and classification of masses in mammograms in order to assist radiologists in diagnosing breast cancer more accurately. The methodology section of the cited article splits into:

- **Preprocessing:** To improve the visual characteristics of the breast region and raise segmentation accuracy, a variety of image enhancement methods are being utilized, such as median filtering, guided imaging, and contrast-limited adaptive histogram equalization (CLAHE).

- **Suspected Region Localization(SRL):** To identify suspicious mass areas or regions of interest (ROIs), the quincunx lifting scheme (QLS)-based density of discrete wavelet coefficient density (DDWCs) is suggested.

- **Mass Classification:** Mass lesions are divided into four groups (benign, possibly benign, malignant, and certainly malignant) based on morphological form criteria.

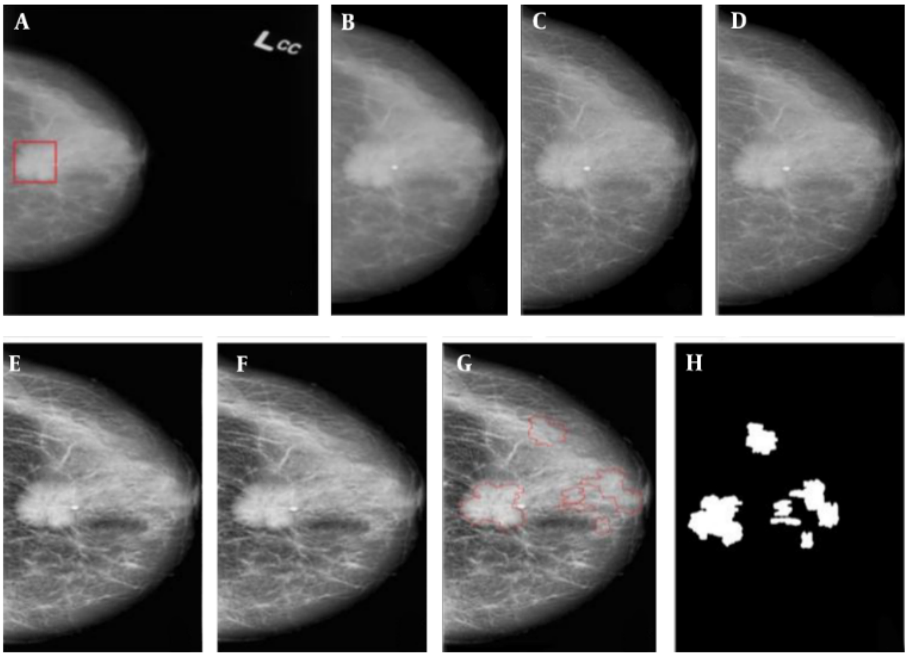


Figure 3.1: An example of applying the suspicious region localization (SLR) method in a mammogram. (Source: [BNF21])

In the experiments section of the article [BNF21], there is a highlight on the robustness of the presented method across variable imaging conditions and subtle variations in the mammogram's quality. The segmentation performance along a classification quality of over 90% for clearly annotated datasets places this method as a competitive option in CAD systems. Another strength of the proposed approach lies in the interpretability of the classification results, which mainly rely on morphological features that are frequently being used in radiological practice. This being the case, the system is more likely to not be classified as only technically effective, but also more likely to gain clinical acceptance.

3.4 An Unsupervised Threshold-based GrowCut Algorithm for Mammography Lesion Detection

The article [MDBAC22] introduces as an approach a fully Unsupervised segmentation technique tailored to the identification of breast cancer lesions using an enhanced GrowCut algorithm. This method builds upon the previously cited work on Threshold-based GrowCut [MDIA21] by removing the need for human-defined seed points, instead introducing automated seed generation through image preprocessing.

Pre-processing takes the role of the foundation stone of this approach, involving artifact removal, contrast enhancement using CLAHE **(Contrast Limited AHE)** and pectoral muscle suppression via Seeded Region Growing **(SRG)**. This ensures improved seed generation and segmentation accuracy.

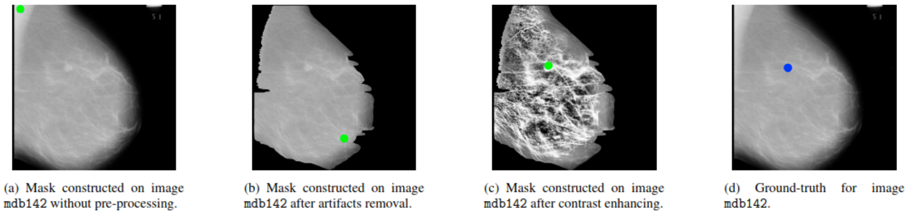

(a) Mask constructed on image mdb142 without pre-processing.
(b) Mask constructed on image mdb142 after artifacts removal.
(c) Mask constructed on image mdb142 after contrast enhancing.
(d) Ground-truth for image mdb142.

Figure 3.2: Example of masks on image mdb142. (Source: [MDBAC22])

This approach of the algorithm automates foreground seed creation by utilizing two sets, from which the most effective one is being selected based on performance. After the rafination process of the seeds is finished, they are integrated into the GrowCut algorithm using a fixed iteration cap and strength-based pixel labeling threshold.

As for the performance metrics, this method has been evaluated on the mini-MIAS dataset [ea94], testing on 50 mammograms. It achieved a foreground seed mask precision of 93.63%, this translating into effectiveness in practical applications. Comparing the presented approach with earlier semi-supervised GrowCut variations, this model shows superior automation while maintaining high accuracy making it suitable for integration into fully automated CAD pipelines.

While this paper displayed strong segmentation results, the authors also emphasized on the importance of tailoring the algorithm such that it would be able to handle the variability of mammographic data, such as broad image resolutions, breast density and acquisition notes. Mainly due to the fact that this approach is unsupervised, it can adapt to a wider range of clinical scenarios without running into the issue of retraining or reconfiguration, which is of upmost importance in real-word screening environments where expert annotations may be limited or inconsistent.

Another key aspect of [MDBAC22] can be found in its detailed evaluation process. The authors get involved with both visual inspection and quantitative analysis, showing not only that the algorithm returns accurate segmentations but also that it preserves lesion boundaries more consistently than previous approaches. This type of boundary refinement is very important in clinical contexts, where precise delineation of tumor edges influences both diagnosing and treatment planning. Extending the previous idea, by incorporating a dual-path seed evaluation mechanism, the proposed method greatly increases it's robustness. Having more flexibility in the seed selection strategy allows it to be able to handle well-defined lesions, as well as low-contrast ones with relatively equal effectiveness, reducing the sensitivity to variation that can occur in contrast and noise levels. Encompassing all this information, this makes the presented approach a promising candidate for future extensions such as but not only integration with post-segmentation classification systems.

3.5 Mammogram segmentation using maximal cell strength updation in cellular automata

The study [JA15] revolves around a completely automatic segmentation method used for identifying suspicious mass regions in mammograms utilizing a Cellular Automata framework augmented with a novel transition rule **(Maximal Cell Strength Updation)**. The main motivation of the presented approach is given by the limitations present in classical **CA** and other segmentation techniques in accurately delineating mass boundaries, more precisely, when dealing with dense tissue and low-contrast abnormalities.

For this article, the segmentation pipeline was split into several key stages. Firstly, a comprehensive preprocessing stage is applied onto raw mammograms, a process which involves removing noise, labels, curvilinear structures and the pectoral muscle (elements known to obscure accurate detection). The authors put a heavy emphasis on the fact that the preprocessing stage is crucial based on the fact that subtle appearances of masses or high intensity of dense tissue can distort or mimic pathological regions. This is achieved through the utilization of a combination between morphological operations and a single-seeded region growing method guided by automated orientation detection of the breast.

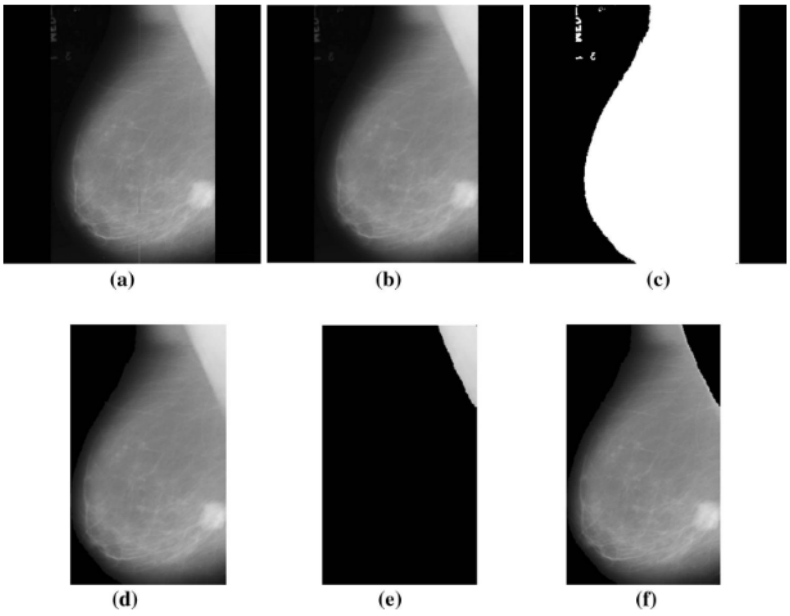


Figure 3.3: Stages of preprocessing applied to mammograms prior to segmentation. (Source: [JA15])

After the preprocessing stage, this method performs coarse segmentation using Histogram Peak Analysis **(HPA)**. This step allows the algorithm to determine a rough region of interest by recognizing dominant intensity peaks from the histogram distribution which are bounded to areas more likely to include abnormalities. Unlike static thresholding approaches, **HPA** adapts dynamically to inconsistencies in image contrast and intensity distribution across different subjects.

The third component introduces automatic seed selection, an important innovation that removes human input. This is constructed upon the sum average **(SA)** feature extracted from the Gray-Level Co-occurrence Matrix **(GLCM)**. The SA implementation highlights with good effectiveness regions with distinct texture patterns characteristic of mass tissue. Afterwards, taking into account the most relevant 32x32 block, the pixel that has the highest intensity is selected as the preferred seed point while surrounding pixels are used in defining the foreground and background.

In [JA15], the authors evaluated the proposed method on 70 mammograms with verified masses from the mini-MIAS dataset ([ea94]), including various mass types (CIRC, SPIC, ARCH, ASYM, MISC). The extracted results demonstrate a sensitivity of 92.25% and an accuracy of 93.48%, metrics critical in coming to the understanding that this method outperforms standard region-growing and active contour methods in accuracy, as well as speed. In addition to this, they provided detailed visual comparisons meant to showcase how the segmented mass closely follows ground truth annotations, even in more complex cases where irregular boundaries or dense surrounding tissue would be an additional challenge.

The highest distinguishing factor that this method brings into the picture would be its computational efficiency. Because of the prioritization of the strongest neighbor during label propagation, the model not only achieves higher convergence but also mentains a high degree of segmentation fidelity. This makes the technique tailored for deployment in low-resource settings or integration into larger computer-aided diagnosis **(CAD)** systems where both speed and accuracy are of upmost importance.

Taking into account all of the facts listed above, [JA15] achieves a strong ballance between precision, automation and efficiency, qualities that would make it especially promising for scaling up breast cancer screening initiatives in clinical practice as well as in population-wide early detection programs.

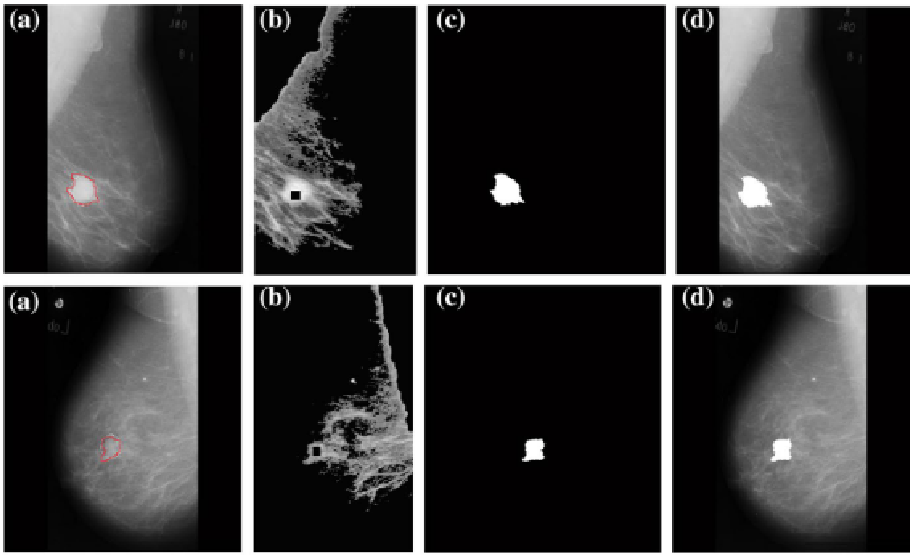


Figure 3.4: Mass segmentation example using MCSU-enhanced CA. (Source: [JA15])

3.6 Conclusions and comparisons between the presented methods

Ref.	Authors	Year	Model	Characteristics	Datasets	Results
[BNF21]	Behrouz Niroomand et al.	2021	DDWC-based	Automated segmentation and classification; uses CLAHE, guided imaging, and median filtering.	CBIS-DDSM	Sensitivity: 100%, FPPI: 6.4±4.5, Accuracy: 85.9%, AUC: 0.901
[JA15]	J. Anitha et al.	2015	MCSU in CA	Automatic seed selection using histogram peak analysis and GLCM-based features.	mini-MIAS	Sensitivity: 92.25%, Accuracy: 93.48%
[IMDA23]	Iulia-Andreea Iona et al.	2023	Fuzzy Cellular Automaton	Combines Cellular Automata with Fuzzy Logic for image segmentation.	mini-MIAS	Accuracy: 98.66%
[MDBAC22]	Moroz-Dubenco et al.	2022	Unsup. TbGC	GrowCut with automatic foreground seed generation and unsupervised segmentation.	mini-MIAS	Precision: 93.63%
[MDIA21]	Moroz-Dubenco et al.	2021	Threshold-based GrowCut	Improved GrowCut with reduced manual intervention and lower computational cost.	mini-MIAS	Results vary with background seed selection.

Table 3.2: Summary of Related Work, putting an emphasis on the main distinctions between the presented approaches.

Chapter 4

Theoretical Background

4.1 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computational models that are heavily inspired by the function and structure observed in biological neural networks found in the brain. Mainly, they consist of layers of interconnected nodes, also called neurons, where each connection has an attachment weight which adjust as the network learns. This networks have as their main purpose to approximate functions that are able to map an input space $X \subset \mathbb{R}^n$ to an output space Y, usually by using a sequence of learnable transformations.

4.1.1 The Structure of an ANN

The composition of an **ANN** is: an input layer, one or more hidden layers and an output layer. Diving deeper, we can see that each layer consists of a multitude of neurons, each neuron being a processing unit that is tasked with performing a weighted sum of its inputs followed by an activation function:

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$a = \phi(z)$$

where:

- x_i are the input features,
- w_i are the corresponding learnable weights,
- b is the bias term,
- ϕ is the activation function (e.g., Sigmoid, Tanh, ReLU),

- a is the neuron's output.

The Rectified Linear Unit (ReLU), defined as:

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

is especially prevalent in deep learning due to its simplicity and ability to mitigate the vanishing gradient problem[LBBH98].

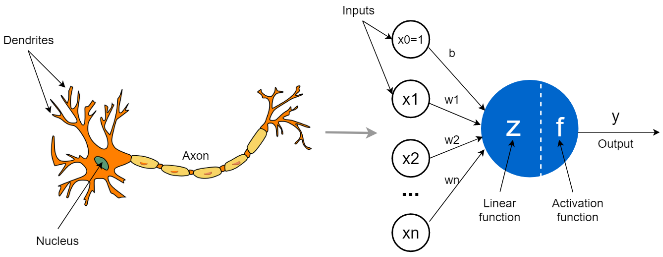


Figure 4.1: Comparison of a biological neuron (left) and an artificial neuron (right). Source: [Pra21]

4.1.2 Forward Propagation and Backpropagation

In the **forward pass**, the inputs are being propagated through the layers of the network in order to produce a prediction while the loss function quantifies the error between the predicted and expected outputs. Learning takes place via **backpropagation**, an algorithm that computes gradients of the loss with respect to each weight using the chain rule of calculus. These gradients are used afterwards in order to update the weights using an optimizer like Stochastic Gradient Descent or Adam, enabling the model to iteratively minimize the loss function and improve performance over time.

The update rule for weight ww using gradient descent is:

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

where η is the learning rate and L is the loss.

4.1.3 Depth and Non-linearity

By continuously stacking multiple hidden layers, ANNs are then capable of learning complex, non-linear relationships in data. This is based on the Universal Approximation Theorem, which states that a feedforward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on a compact input space[HSW89].

With all this in mind, it is important to note that deep networks are not trivial to train. Overfitting can occur in the case where the model is too complex relative to the available data while underfitting arises when the model lacks sufficient capacity or when the optimization is not ideal. There also worth mentioning that techniques such as early stopping, batch normalization and dropout are often used to improve generalization.

4.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized type of deep neural network which are designed for the processing of data with grid-like structure, such as images for example. Introduced by Yann LeCun in the context of handwritten digit recognition [LBBH98], CNNs have become a dominant architecture for visual data mainly due to their efficiency and accuracy in tasks like classification, detection and segmentation.

Different from fully connected network in which every neuron is connected to every input, CNNs are leveraging local spatial correlations by using convolutional layers. Each convolutional layer applies a series of kernels that slide over the input image and in turn, produce feature maps. These filters are meant to capture local features (eg. edges, textures), in earlier layers, while abstract concepts (eg. shapes, objects) in deeper layers[Kar16].

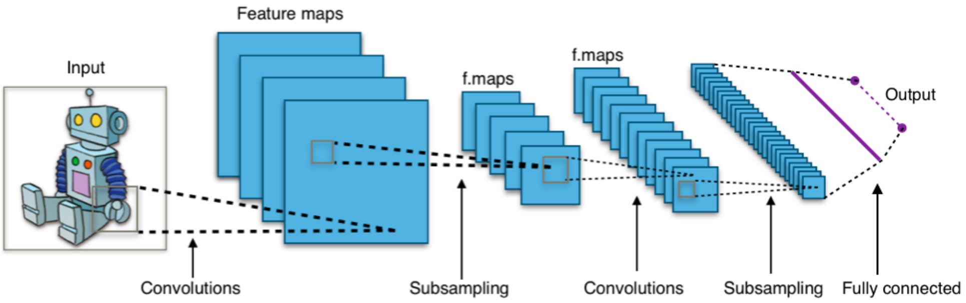


Figure 4.2: Typical CNN architecture. Source: Wikimedia Commons.

After convolution, a non-linear activation function is being applied, usually the Rectified Linear Unit (ReLU), also defined as:

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

ReLU is often preferred over older activation functions like Sigmoid and Tanh due to its simplicity and more importantly, due to its effectiveness in mitigating the vanishing gradient factor[NH10].

In order to reduce the dimensionality of feature maps and increase spatial invariance, CNNs use pooling layers such as max pooling or average pooling. Altogether, these layers summarize small neighborhoods in the feature map in order to reduce overfitting and computation[LBBH98].

Towards the end of the network, the output produced by the final convolutional block is flattened and passed through one or more fully connected layers. These layers are the decision-making gear in the machine, converting the learned spatial features into probabilities. For the purpose of improving generalization, CNNs more often than not integrate regularization techniques such as dropout, which randomly disables a fraction of neurons during training[SHK$^+$14] and batch normalization which stabilizes the learning process for allowing higher learning rates[IS15].

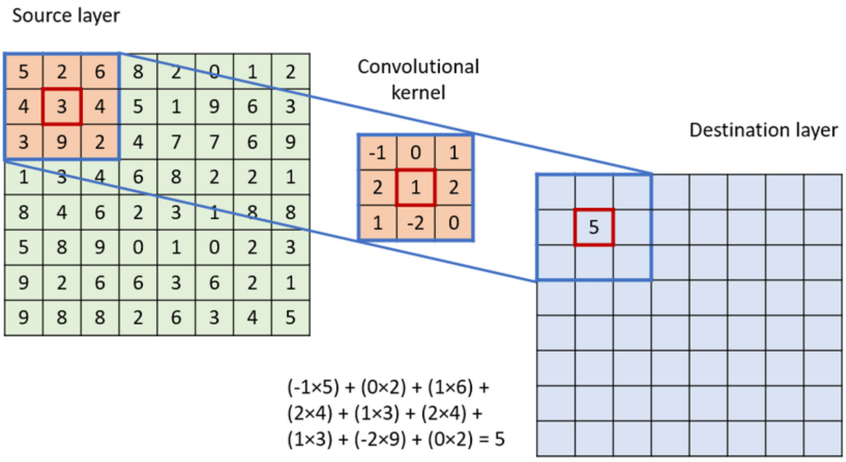


Figure 4.3: Schematic illustration of a convolutional operation. The convolutional kernel shifts over the source layer, computing the dot product at each position to fill the destination layer. Source: [WPCA19].

4.3 Transfer Learning

The notion of transfer learning is associated with a machine learning technique in which a model that has been developed for a certain task is reused as the starting point for another related task. This sort of approach is particularly effective in deep learning, where extensive neural networks are usually trained on large datasets and can be adapted onto tasks where labeled data is scarce.

The motivation for transfer learning originates from the fact many learned features from large-scale datasets such as ImageNet, are generic in early layers of convolutional networks. The main functionality comes from the fact that these layers often capture edges, textures and basic shapes, which reusable across visual domains. By reusing these pretrained weights and by applying fine-tuning on the network for a specific task, the training process becomes even more efficient which in turn means better generalization yield, especially when training data is limited.

The approaches to transfer learning mainly split into two branches:

- **Feature Extraction:** The pretrained model is used for extracting features, where the convolutional base is frozen and only a new classification head is trained on the target dataset.
- **Fine-tuning**: The given pretrained model is then initialized with learned weights but is then, fully or partially retrained on the targeted dataset giving it the ability to adapt to domain-specific patters.

Referencing medical imaging, transfer learning leaves a huge impact due to the fact obtaining annotated datasets is incredibly difficult. Annotating medical imaging often requires the intervention of human expertise, is time-consuming and is subject to privacy constraints. Transfer learning gives researchers and practitioners the ability to use robust pretrained architectures for tasks like organ segmentation, disease classification, anomaly detection and tumor localization[LKB+17][TSG+16].

Various research has repeatedly shown that pretrained convolutional neural networks outperform by a considerable margin randomly initialized models when applied to medical imaging datasets, fact which is intuitive by itself. Studies have shown even in the absence of full fine-tuning, using features from models pretrained on non-medical datasets, like natural images for example, can yield a strong performance in their prediction due to the generality of low-level and mid-level features[SRG+16].

There is the possibility of going even beyond classification where transfer learning has been used for segmentation, detection and even reconstruction tasks in medical AI workflows. Moving forward, the practice is in a continuous evolution, accentuated by the support towards cross-modal transfer (eg. from CT to MRI) and domain adaptation techinques.

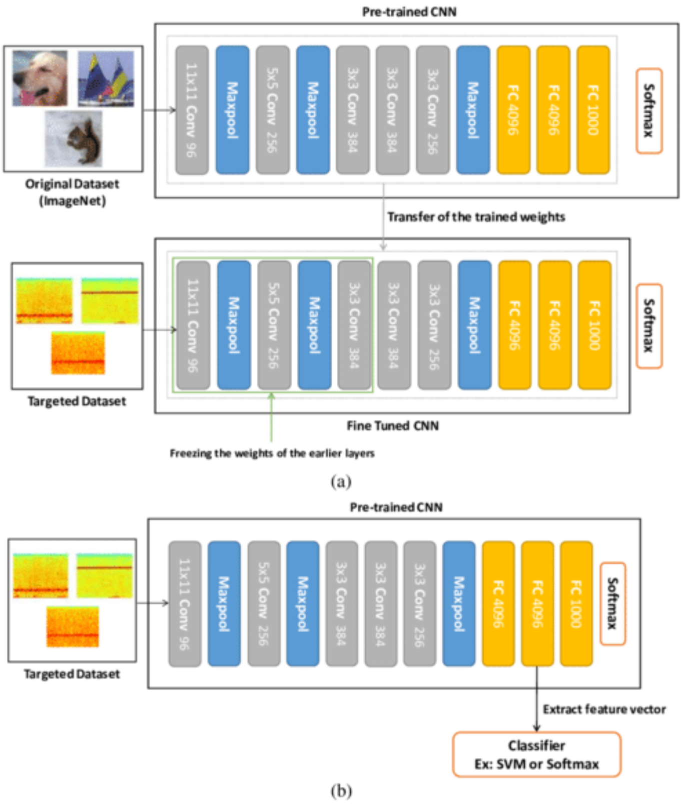


Figure 4.4: Illustration of two common approaches in transfer learning: (a) Fine-tuning, where some layers of the pre-trained model are retrained on the new dataset; (b) Feature extraction, where the pre-trained model is used as a fixed feature extractor. Source: [AAMTW18]

4.4 Loss functions

In a supervized learning, a loss function takes the form of a mathematical expression whose main purpose is to quantify the difference between the target value and the predicted output of the model. It serves as the objective that the learning algorithm looks to minimize in the process of training, guiding further the parameter optimization. The choice in loss function has a direct impact towards how the model interprets its prediction errors and in turn, how it adjusts its weights.

Particularly in classification tasks, the loss function that pops out the most cross-entropy loss (log loss). Cross-entropy loss measures the discrepancy between the predicted probability distribution and the true distribution, usually given as one-hot encoded vector. In the case of binary classification problems, in which the model has as its output a single probability $\hat{y} \in [0, 1]$ and the target label $y \in \{0, 1\}$, the binary cross-entropy loss is defined as:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

where:

- N is the number of training examples,
- y_i is the true label for sample i,
- $\hat{y}_i$ is the predicted probability of the positive class,
- θ represents the model parameters.

This loss function penalizes wrong predictions more harshly when the model has high confidence towards the wrong answer, this encouraging probabilistic outputs to become more accurate. Cross-entropy loss is differentiable and well-tailored for usage with softmax or sigmoid output activations, making it standard in neural networks for classification. However, in multi-class classification, cross-entropy generalizes to compare the predicted and true distributions over multiple classes, often used with the softmax function applied to the output layer.

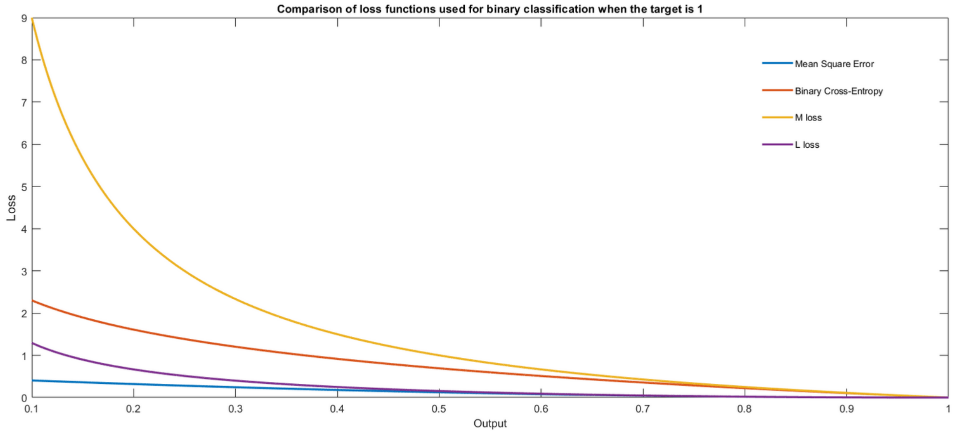


Figure 4.5: Plot shows different loss functions that can be used to train a binary classifier. Source: Wikimedia Commons.

Loss functions are also able to handle class imbalance, a common factor in real-world datasets such as those in medical imaging, in which positive cases may become the subject of underrepresentation. A common sollution to this issue would be weighted cross-entropy, where a higher loss weight is assigned to the minority class:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [w_1 y_i \log \hat{y}_i + w_0(1 - y_i) \log(1 - \hat{y}_i)]$$

w_1 and w_0 are weights that are class-specific and are balancing the contribution that each class has on the overall loss. Other strategies have into their composition focal loss, which weights down the easier examples while focusing training on more strenuous examples, misclassified examples, especially in highly skewed or noisy datasets.

The choice of loss function plays a fundamental role in optimization but also in model fairness and robustness, particularly in more sensitive domains.

4.5 Optimization Algorithms

Algorithms used in the optimization of a given solutions, also called optimizers, are fundamental concepts in the training of neural networks. The goal in implementing an optimization algorithm is to iteratively update the parameters of the model in such a way that minimizes as much as possible the loss function, with the eventual objective of improving the performance of the model on unseen data. Particularizing on the context of deep learning, the process of optimization is performed via variants of **gradient descent**, which primarily utilizes the gradient of the loss function with respect to the model parameters in order to adjust the weights in such a direction that it reduces error.

4.5.1 Gradient Descent

If we want to look for the most basic form of optimization, Stochastic Gradient Descent (SGD) comes to mind. In SDG, model parameters θ are being updated after each mini-batch of training examples, using:

$$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$$

where:

- η is the learning rate,
- $\nabla_\theta L(\theta)$ is the gradient of the loss with respect to the parameters.

Even tho SDG is considered simple and effective, it is also sensitive to the choice of learning rate and can be slow to converge, especially when we are talking about complex, high-dimensional loss surface prevalent in deep networks.

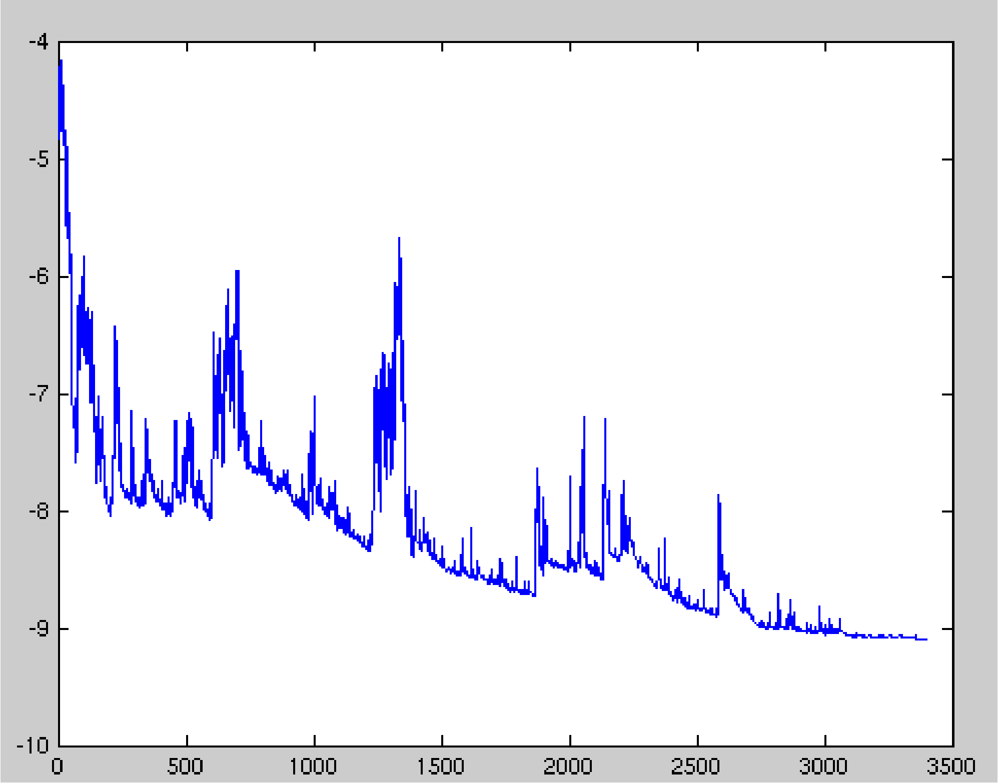


Figure 4.6: SDG fluctuation. Source: Wikimedia Commons.

4.5.2 Adaptive Methods

In order to address the limitations of barebone SDG, a multitude of adaptive learning algorithms have been developed in this direction. Among the most widely used ones, one algorithm that stands out is Adam (Adaptive moment estimation), an algorithm which combines ideas from momentum as well as from RMSProp optimizers. Adam keeps an exponentially decaying average of past gradients (momentum) and squared gradients (adaptive learning rate), making it able to adjust the learning rate for each parameter individually.

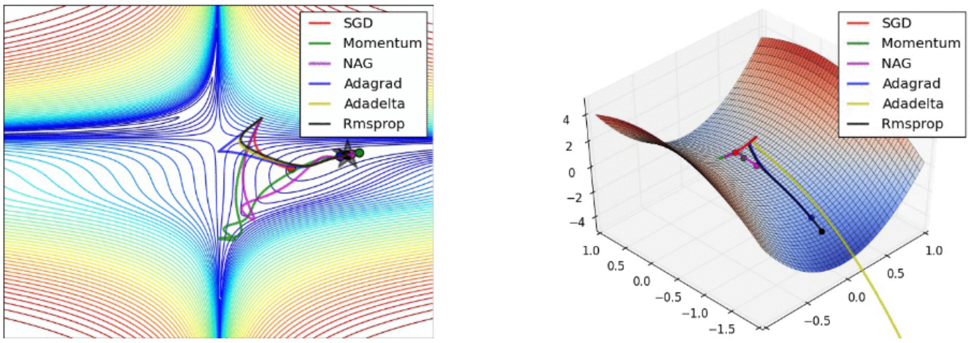
The update rule for Adam is as follows:

$$\theta \leftarrow \theta - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where:

- $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moments of the gradient,
- ϵ is a small constant added for numerical stability.

If we are looking for areas where Adam is especially useful, problems with sparse gradients or noisy data are a prime candidate. This makes it a popular choice in medical image analysis where datasets are limited and training might be unstable.



(a) SGD optimization on loss surface contours

(b) SGD optimization on saddle point

Figure 4.7: Visualization of optimization algorithms. Source: Alec Radford.

4.5.3 Optimizer Selection in Practice

The task of choosing the right optimizer must always be adapted to the specific problem domain, dataset characteristics and of course, model complexity. While SDG with momentum remains a strong baseline foundation, adaptive methods such as Adam are often preferred in order to achieve fast convergence and minimal tuning. Even with all this being said, recent studies have shown that with proper learning rate scheduling and regularization, SDG is able to outperform Adam in specific vision tasks[WRS+17]. In modern practice, most training pipelines rely on Adam for its robustness and simplicity, especially when dealing with research or prototyping scenatios.

4.6 Regularization Techniques

Mainly because of their high capacity and complex architectures, deep neural networks are inclined to overfitting (a phenomenon in which the model preforms very well on training data but fails on unseen, new data). Overfitting is very common in domains with limited training samples (eg. medical imaging). In order to mitigate this recurring issue and improve generalization, some regularization techniques are applied during training.

4.6.1 Dropout

Dropout stand as one of the most widely used regularization techniques in deep learning. Introduced by Srivastava et al. [SHK+14], the mechanism behind it works by randomly "dropping out" a subset of neurons in a layer during each training iteration. This approach makes sure that the model being implemented is not becoming overly reliant on specific neurons, this in turn encouraging the network to learn more robust representations that are also more distributed.

Mathematically, for each unit, dropout applies a binary mask $m_i \sim \text{Bernoulli}(p)$, where p is the probability of retaining a neuron (often $p = 0.5$):

$$\tilde{x}_i = m_i x_i$$

This results into a different network undergoing the training process for each iteration and during inference, all neurons are being used with scaled outputs.

4.6.2 Batch Normalization

Batch normalization is another technique that has significant use in the domain because of its addressing of internal covariate shift, in which the distribution of inputs to each layer undergoes changes during training. Firstly proposed by Ioffe and Szegedy in [IS15], batch normalization normalizes the inputs for each layer such that they will have a mean of 0 and standard deviation of 1 across the entirety of the mini-batch. This type of processing upon the learning process allows for faster convergence, higher learning rates and the minimization of sensitivity to initialization.

In practice, this process also has the effect of regularization by introducing a small amount of noise, which helps a ton in reducing overfitting in the absence of dropout.

4.6.3 Weight Regularization (L2 Regularization)

Another method that is usually found when investigating the domain is **L2 regularization**, also known as weight decay. This approach punishes large weights by incorporating a term to the loss function:

$$L_{\text{reg}}(\theta) = L(\theta) + \lambda \sum_i \theta_i^2$$

where λ takes on the role of a regularization coefficient whose purpose is controlling the strength of the punishment. The integration of this term favours the network to prefer smaller weights, resulting into simpler models with better generalization.

4.6.4 Relevance in Medical Imaging

While not all applications are like this, in medical imaging there is a strong presence of small or imbalanced datasets and because of this, regularization presents itself as a vital role in preventing overfitting. Dropout and batch normalization are very useful in CNN-based architectures, improving test-time performance and training stability. The techniques are usually used alongside data augmentation approaches in order to enhance model robustness in clinical scenarios[LKB$^+$17].

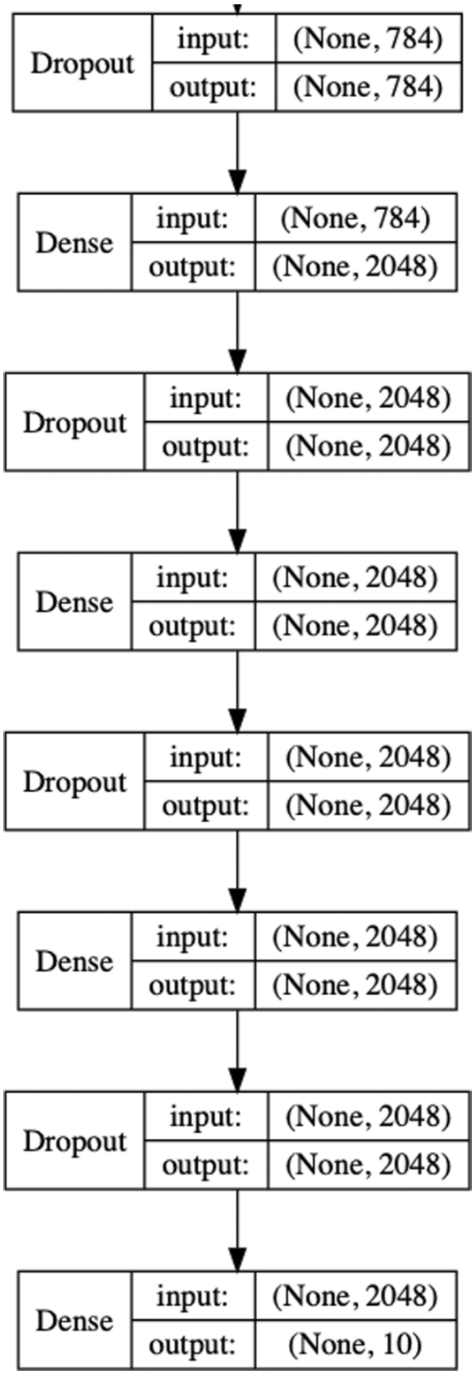


Figure 4.8: Sample dropout MLP network architecture used in the tests (with dropout but without batch normalization). Source: [GZM20]

4.7 Image Preprocessing Techniques

When we are talking about medical image analysis, image preprocessing is the baseline for enhancing data quality and for tailoring the input for deep learning and machine learning models. Medical images, such as mammograms, more often than not contain artifacts, noise and irrelevant anatomical structures that could hinder the proper development of the targeted models. With all of this in mind, a preprocessing pipeline that has also proved itself effective is essential in ensuring that the learning algorithms are focused on the relevant features.s

4.7.1 DICOM to Grayscale Conversion

A lot of the times, medical imaging is being stored in DICOM (Digital Imaging and Communications in Medicine) format, whose role is to preserve extensive metadata alongside high-resolution imaging data. Before this graphical data can be properly used for computational modeling, they are usually converted into grayscale image formats such as PNG or TIFF. This conversion makes sure that the input dimensionality remains simple while at the same time, preserving diagnostically relevant features, as most radiological modalities are inherently grayscale.

4.7.2 Resizing and Input Standardization

The usual image input for neural networks needs to be standardized (consistent dimensions). This in turn necessitates a resizing step, where all images are uniformed in size. For CNNs, common input sized would be 224x224, 512x512 or even higher resolutions, of course depending on model complexity and hardware capabilities. Standardization is not only important for batch processing but also aligns with the architectural requirements of pretrained networks.

4.7.3 Breast Region Segmentation and Artifact Removal

In cases where the input data would consist of mammograms, the presence of irrelevant regions such as air space, background and even the pectoral muscle can be detrimental in model training. Preprocessing might include techniques such as binary thresholding, contour detection and masking approaches in order to isolate as much as possible the breast tissue. Additionally, certain anatomical features such as the pectoral muscle in mediolateral oblique (MLO) views can be manually or algorithm-assisted removed in order to avoid bias in region-specific classifiers.

4.7.4 Contrast Enhancement using CLAHE

Low contrast on medical images is a known and a very common issue, especially in mammograms, where some defects or lesions could blend with dense tissue. Contrast Limited Adaptive Histogram Equalization (CLAHE) is a technique that has been adopted by most researchers in the domain in order to enhance local contrast while at the same time, prevent the amplification of noise. CLAHE works on image tiles that are reduced in size and redistributes intensity values, with the goal of improving the visibility of abnormalities that are subtle and would not be picked up otherwise, without the need to overexpose homogeneous areas. This technique works especially well in highlighting structures that have low intensity like microcalcifications[PZH+98].

4.7.5 Normalization and Tensor Conversion

Before even beginning the training process, most of the time, pixel values are normalized to a consistent range in order to stabilize the learning process and match the expectations of activation functions. If we are talking in the context of deep learning frameworks, images are also being converted into tensor formats, formats which represent multi-dimensional arrays and are optimized in order to get better GPU processing. While not entirely necessary, efficient tensor conversion and storage enables the streamlining of data loading, batch training and even augumentation.

Chapter 5

Proposed Solution

5.1 Overview of The Proposed Approach

The proposed solution of this work mainly consists of a modular and highly scalable diagnostic pipeline for mammographic image binary classification utilizing the power of deep convolutional neural networks (CNNs). The pipeline was designed such that it would be able to work as a fully automated end to end system, beginning with raw medical imaging and eventually leading into a binary diagnostic output (benign vs. malignant). The approach presented in this thesis was constructed with careful attention to practical limitations found in clinical datasets, as well as to architectural constraints of modern deep learning models.

The three main challenges that were present in the development of the proposed solution would be:

- **The constant need for high-quality and standardized input data**, especially taking into consideration the variability in image formats, resolutions or acquisition protocols in public datasets.
- **The fundamental need of preserving diagnostically relevant features** while at the same time, removing artifacts or structures that would hinder the training process.
- **The difficulty of training generalizable models** because of small or heavily imbalanced datasets.

In order to combat the presented difficulties, the pipeline was structured into three distinct stages:

- **Preprocessing and tensor preparation**, the module which is responsible with cleaning and formatting the input data (mammograms) into consistent, high-information outputs.

- **Model-based classification**, made up of two architectures that run parallel one to another: a custom CNN built entirely from scratch and a pretrained EfficientNet-B1 model that has been adapted through the usage of transfer learning.
- **Evaluation and analysis**, in which the model performance is being tested on unseen examples and the prediciton outputs are interpreted with respect to clinical relevance.

This explicit architecture is chosen based on its experimental effectiveness but also, because of its practical modularity (each stage of the pipeline can be modified, evaluated or even replaced without interfering with others). Unlike traditional handcrafted approaches, approaches which heavily depend on feature extraction, this system takes advantage of the hierarchial representation features of CNNs to autonomously extract features that are considered relevant. Furthermore, utilizing transfer learning on a clinically diverse and large dataset (CBIS-DDSM), this specific pipeline is generally better equiped to generalize across image sources.

5.2 Algorithm and Method

The diagnostic pipeline that is being proposed in this thesis is made up of a series of deterministic and learnable stages, each of those stages having as target a specific function in the wider task of binary classification of mammographic images. The algorithm can be broken down into 5 major stages: data cleaning and preparation, preprocessing, tensor conversion, model training and of course, inference. Each one of these stages is implemented in a modular way, allowing future adaptation and extension.

5.2.1 Step-by-Step Description

1. DICOM Conversion and Image Cleaning

Mammograms stored in DICOM format are being converted at first into grayscale image formats, a crucial step in preserving diagnostic fidelity while also ensuring compatibility with image processing libraries. While this conversion takes place, pixel intensity values are being normalized to a range that can be displayed and this is done by using the window center and width parameters embedded into the DICOM header.

2. Preprocessing pipeline

Each item undergoes a deterministic preprocessing pipeline, pipeline that is designed to eliminate irrelevant features and in turn, highlight important structures. The sequence of operations includes:

- Resizing to a standard resolution (eg. 1024x1024 pixels)
- Applying a Gaussian blur in order to suppress noise
- Otsu thresholding and contour detection are being used to isolate the breast region
- Masking of the pectoral muscle
- Cropping to the minimal bounding box enclosing breast tissue
- Applying CLAHE to enhance local contrast

These steps are of utmost importance in removing anatomical bias and for improving the consistency of data passed to the model.

3. Tensor Conversion and Normalization

Preprocessed images are sent afterwards in order to be transformed into tensors, tensors which are suitable for PyTorch-based deep learning. Every grayscale image gets reshaped into $1 \times H \times W$, converted to a floating-point representation and then, normalized to a range of either [0,1] or centered at 0 using a z-score-like scaling. The resulting tensors are serialized as .pt files to reduce loading overhead and improve training throughput.

4. Model Initialization and Training

Two distinct model architectures are employed:

- A **custom CNN**, implemented from scratch with three convolutional layers, ReLU activations, max pooling, and two fully connected layers.
- A **transfer learning model**, based on EfficientNet-B1 pretrained on ImageNet, adapted for single-channel input and fine-tuned on the mammographic dataset.

Each model is trained using cross-entropy loss, the Adam optimizer, and learning rate scheduling. To compensate for dataset imbalance, class weights are applied dynamically during training.

5. Inference and Evaluation

Once trained, models are evaluated on unseen images. During inference, each image follows the same preprocessing and tensorization pipeline and is then passed through the trained network to obtain a softmax probability distribution. The output class is assigned based on the highest score, and performance is assessed via accuracy, precision, recall, and confusion matrix analysis.

5.2.2 Algorithm Representation

In order to formalize the entire pipeline, Algorithm 1 presents the foundational steps required for converting raw mammograms into a binary classification output. The algorithm consists of preprocessing, tensor conversion and model training utilizing a modified EfficientNet-B1 architecture. This structure reflects the actual implementation used in the proposed system.

Algorithm 1 Breast Cancer Classification Pipeline

```
Input: Raw DICOM images
Output: Benign or Malignant label per image
for each DICOM image do
    Convert to PNG (grayscale, normalized)
    Apply preprocessing:
        • Resize to 1024x1024
        • Apply Gaussian blur
        • Otsu thresholding
        • Segment breast region
        • Remove pectoral muscle
        • Apply CLAHE
    Save preprocessed PNG
end for
Convert all preprocessed PNGs to PyTorch Tensors
Initialize EfficientNet-B1 model
Train model:
        • Feature extraction phase — freeze pretrained layers
        • Fine-tuning phase — unfreeze last N layers
        • Use class-balanced cross-entropy loss
Apply early stopping
Evaluate model on test set
```

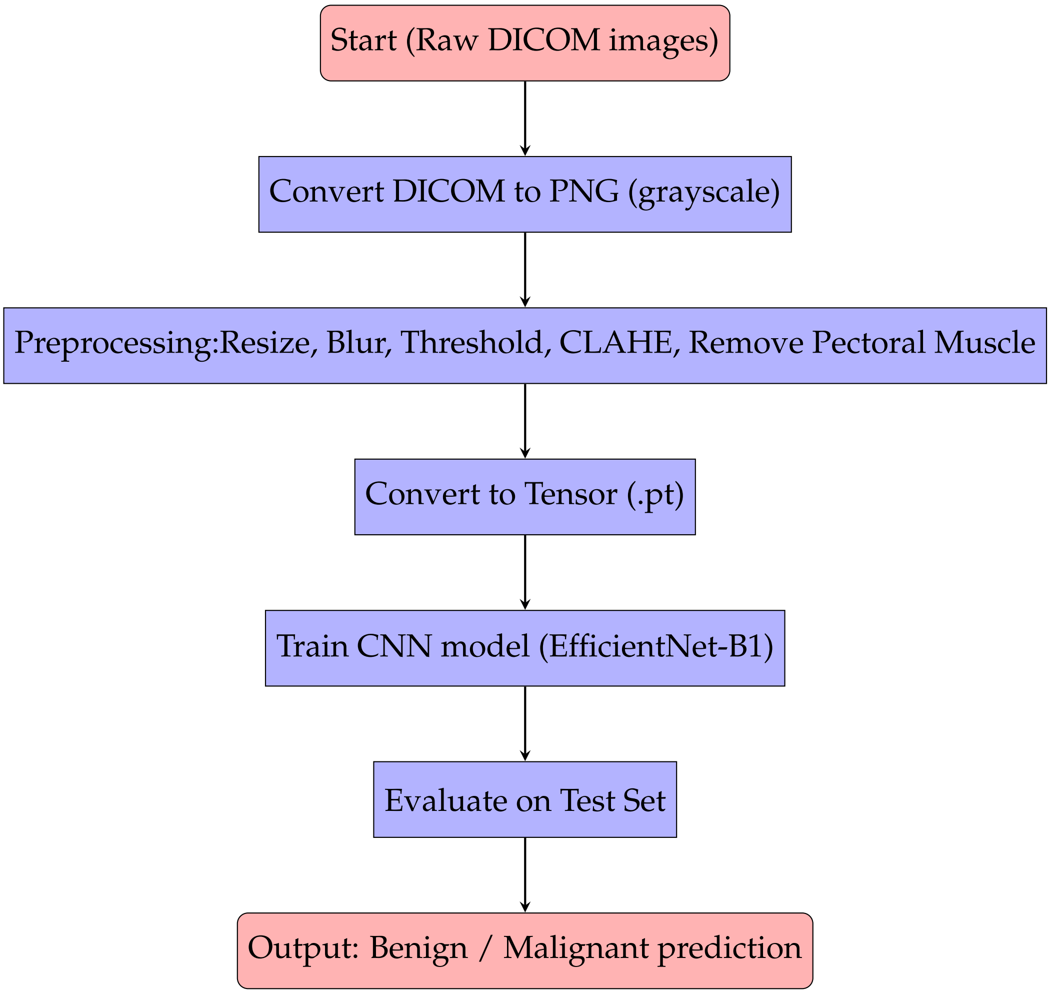

Start (Raw DICOM images)
Convert DICOM to PNG (grayscale)
Preprocessing:Resize, Blur, Threshold, CLAHE, Remove Pectoral Muscle
Convert to Tensor (.pt)
Train CNN model (EfficientNet-B1)
Evaluate on Test Set
Output: Benign / Malignant prediction

Figure 5.1: Flowchart of the Proposed Breast Cancer Classification Pipeline

Chapter 6

Computational experiments and results

6.1 Overview

This chapter serves the role of presenting the experimental evaluation of the proposed approach. Several deep learning models were trained in the making of this thesis and compared on two publicly available datasets: Mini-MIAS ([ea94]) and CBIS-DDSM ([SLGHR16]). The experiments were designed such that they would test both the effectiveness and generalizability of different CNN-based architectures under consistent preprocessing and training protocols.

The goal of this evaluation is split into:

- To assess the impact of image preprocessing and data representation on classification performance
- To compare the accuracy and training behavior of multiple model architectures, including both custom and pre-trained networks

The previously stated experiments follow a standardized setup which involves class-balanced loss functions, data augmentation and of course, early stopping in order to avoid overfitting as much as possible. Evaluation metrics like accuracy, validation loss and even training curves are used to compare models in a reproducible and interpretable manner.

Models were tested across both Mini-MIAS, a smaller and homogeneous dataset, and CBIS-DDSM, a very complex and clinically realistic dataset. The performance differences observed across models and datasets highlight the strengths and limitations of each architecture in practical diagnostic settings.

6.2 Datasets and Setup

For the evaluation of the proposed classification pipeline, two publicly available and frequently used datasets in the field of mammography were chosen: **Mini-MIAS** and **CBIS-DDSM**. The advantage of adopting this sort of approach lies in the fact that these datasets provide complementary characteristics, enabling a comprehensive assessment of model performance and integrability on both small-scale and large-scale data.

6.2.1 Mini-MIAS Dataset

The Mini-MIAS database [ea94] is a widely referenced benchmark in breast cancer image analysis research. It contains 322 digitized film mammograms from the UK National Breast Screening Programme. Each image is stored at a resolution of 1024×1024 pixels and is annotated with metadata indicating the lesion type, position, and severity (benign or malignant).

This dataset is relatively small and homogeneous, making it ideal for preliminary evaluation of model architectures and preprocessing effectiveness. It was used in this thesis to validate that the proposed classification pipeline performs well under controlled conditions.

6.2.2 CBIS-DDSM Dataset

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [SLGHR16] is a large-scale, clinically realistic mammography dataset. It consists of thousands of high-resolution mammographic images derived from the original DDSM dataset, curated to include annotations for breast lesions such as masses and calcifications.

CBIS-DDSM poses a more challenging problem for automatic classification due to the variability in image acquisition protocols, tissue density, and annotation consistency. It also includes a greater number of samples and exhibits class imbalance, which was addressed using class-weighted loss functions and stratified dataset splits.

The CBIS-DDSM dataset was used to train and evaluate all models described in this thesis, including both custom CNNs and transfer learning-based architectures. It enabled a comprehensive analysis of model performance in realistic diagnostic scenarios.

6.2.3 Dataset Splitting and Augmentation

Both datasets were split into into training (80%) and validation (20%) subsets, utilizing a fixed random seed for reproducibility. It is very important to note that absolutely no images from the training set were used in the validation set (the models were evaluated on unseen data).

Data augmentation has been applied in a dynamic manner during training:

- Random horizontal and vertical flips
- Random rotations (up to ±20°)
- Affine transformations
- Contrast jittering

These augmentations were used to improve the robustness of the presented models, robustness which is especially necessary in smaller datasets like Mini-MIAS.

6.2.4 Model Training Environment

All models were trained using `PyTorch` on either GPU (CUDA) or CPU depending on hardware availability. Consistent training settings were used:

- **Batch size**: 8
- **Optimizer**: Adam with learning rate of 0.0005 or 0.001
- **Early stopping**: based on validation loss (patience = 7 epochs)
- **Loss function**: Weighted cross-entropy or Focal Loss for class imbalance (in CBIS-DDSM runs)

The trained models include:

- Custom CNN (3-layer architecture)
- EfficientNet-B1
- EfficientNet-V2
- ResNet18, ResNet34, and ResNet50

All models were trained from scratch or fine-tuned, depending on the experiment.

6.3 Training Protocol and Evaluation Metrics

6.3.1 Training Configuration

To maintain a consistent comparison framework across all architectures, each model was trained following a uniform training strategy. A maximum of 50 epochs was allowed for each run, with an early stopping mechanism activated if no improvement in validation loss was observed for 7 consecutive epochs. This approach ensures that overfitting is mitigated while also optimizing training efficiency.

The optimization algorithm selected for all experiments was `Adam`, chosen for its adaptive learning rate behavior. Two different learning rates were used:

- `0.001` for models trained from scratch, such as the custom CNN.
- `0.0005` for fine-tuning pretrained networks, including EfficientNet and ResNet variants.

To address class imbalance—particularly present in the CBIS-DDSM dataset—custom class weights were computed and incorporated into the loss function. Two types of loss functions were tested:

- **Weighted Cross-Entropy Loss**, based on the label distribution in the training set.
- **Focal Loss**, which dynamically scales the loss contribution of well-classified versus hard-to-classify samples.

Each training session used a batch size of 8, and every model was trained using tensors generated from preprocessed grayscale images, as outlined in Section **??**. Data augmentation techniques such as random rotations and flips were applied to improve model generalization.

6.3.2 Evaluation Strategy

The performance of each model was evaluated using a suite of metrics appropriate for binary classification tasks:

- **Accuracy** – the proportion of correctly predicted samples out of the total.
- **Validation Loss** – the loss value on the validation set, used for model checkpointing and early stopping.
- **Confusion Matrix** – a tabular visualization of prediction distribution across classes, helpful for understanding model bias or class-specific trends.

- **Precision and Recall** – analyzed where relevant to measure false positive and false negative sensitivity, especially important in the medical context.

Throughout the training process, both training and validation curves (for loss and accuracy) were logged to facilitate performance visualization. The best-performing version of each model was selected based on the minimum validation loss achieved during training.

6.3.3 Reproducibility and Fairness

To ensure reproducibility across experiments, the following practices were adopted:

- All dataset splits were generated using a fixed random seed.
- Data preprocessing and transformation steps were identical across all training runs.
- Experiments were executed within modular and version-controlled scripts to allow traceability and parameter tuning.

This consistent training methodology supports a fair comparison between models and ensures that observed differences in performance are due to architecture or training dynamics, not procedural variance.

6.3.4 Results and Discussion

This section presents and discusses the results obtained by the various trained models on both the Mini-MIAS and CBIS-DDSM datasets. The models were evaluated using standard classification metrics: **accuracy**, **precision**, **recall**, and **F1-score**, alongside **confusion matrices**, which offer a more thorough understanding of model behavior.

Mini-MIAS Results

The ResNet34 model, with an accuracy of 83% and an F1-score of 0.77, achieved the best overall performance on the Mini-MIAS dataset. The SimpleCNN and ResNet18 models frequently overpredicted the malignant class and had a tendency to provide high recall at the expense of precision. This suggests that deeper models (ResNet34 and ResNet50) are better equipped to discern subtle variations in mammographic patterns, particularly in small datasets such as Mini-MIAS.

Model	Accuracy	Precision	Recall	F1-score
SimpleCNN	53%	0.49	1.00	0.66
EfficientNet-B1	54%	0.47	0.64	0.54
ResNet18	56%	0.49	0.80	0.61
ResNet34	83%	0.81	0.73	0.77
ResNet50	68%	0.63	0.44	0.52

Table 6.1: Results on Mini-MIAS test set.

CBIS-DDSM Results

Across the CBIS-DDSM dataset, the models based on EfficientNet-B1 and ResNet50 achieved the highest F1-scores (0.74) and strong recall values (0.91–1.00), which is particularly advantageous for medical screening tasks where minimizing false negatives is critical.

The addition of Focal Loss significantly enhanced the performance of EfficientNet-B1 by better managing class imbalance.

Model	Accuracy	Precision	Recall	F1-score
SimpleCNN	51%	0.50	1.00	0.70
EfficientNet-B1	74%	0.61	0.91	0.73
EfficientNet-B1 (Focal Loss)	76%	0.63	0.91	0.74
ResNet18	66%	0.55	0.80	0.65
ResNet34	68%	0.55	0.65	0.59
ResNet50	69%	0.58	1.00	0.74

Table 6.2: Results on CBIS-DDSM test set.

Discussion of Confusion Matrices

The confusion matrices provide further insights into model behavior. On both datasets, the SimpleCNN and ResNet18 models exhibited a clear tendency to overpredict the malignant class, resulting in excellent recall but subpar precision. While high recall is desirable in medical contexts, excessive false positives can lead to unnecessary follow-up procedures, highlighting a trade-off that must be balanced.

Conversely, deeper architectures such as ResNet34, ResNet50, and EfficientNet models produced more balanced predictions. On CBIS-DDSM, the ResNet50 model achieved perfect recall (1.00), but this came at the expense of more false positives, indicating that the model prioritizes sensitivity. EfficientNet-B1 with Focal Loss provided a better balanced trade-off between sensitivity and specificity, achieving both strong recall and precision.

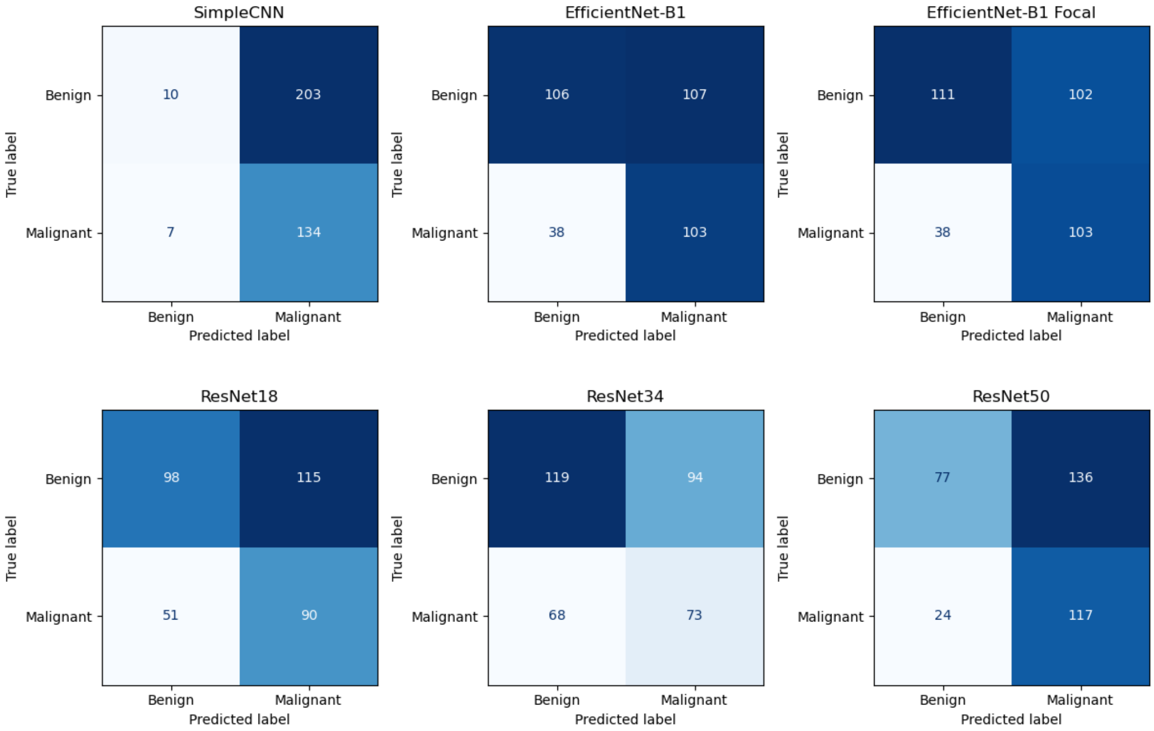


Figure 6.1: Confusion matrices of the evaluated models on the CBIS-DDSM test set.

Cross-Dataset Observations

The relative performance ranking of models remained consistent across both datasets, with ResNet34 and EfficientNet variants outperforming simpler architectures such as SimpleCNN. However, absolute performance was lower on CBIS-DDSM, which is expected due to its greater image variability, higher clinical realism, and pronounced class imbalance.

Importantly, models trained on CBIS-DDSM generalized better overall, while models trained solely on Mini-MIAS were less robust. This underlines the fact that training data that is larger and more diverse significantly contributes to improved model resilience.

Limitations and Observations

Several limitations emerged from the experiments:

- Training efficacy may have been constrained by small batch sizes and limited computational resources.
- No hyperparameter optimization was conducted; further fine-tuning might yield better results.

- The binary classification task remains inherently difficult due to class imbalance and subtle visual differences between classes.
- While helpful for baseline comparison, the SimpleCNN architecture is too simplistic for this domain and shows substantial limitations compared to more modern designs.

Overall, the findings confirm that modern transfer learning architectures, such as ResNet50 and EfficientNet, provide superior performance for mammogram classification. Moreover, careful preprocessing and class balancing are crucial for improving generalization across datasets. Future work should focus on optimizing deeper architectures with refined augmentation strategies and hyperparameter tuning to further enhance clinical relevance.

6.4 Comparison with Related Work

The top-performing models from this thesis obtained competitive accuracy and F1-scores in comparison to current research, as shown in Table 6.3. While segmentation-based techniques such as MCSU and Fuzzy Cellular Automaton report higher accuracy values, they frequently prioritize segmentation quality rather than end-to-end classification performance.

With F1-scores of 0.77 and 0.74, respectively, the proposed ResNet34 and EfficientNet-B1 models demonstrated excellent classification performance on the Mini-MIAS and CBIS-DDSM datasets. These results are particularly encouraging, given that CBIS-DDSM is a more complex and clinically realistic dataset compared to Mini-MIAS.

It is important to note that variations in experimental design, task definitions (segmentation vs classification), and data preparation may limit the ability to make direct comparisons. However, the proposed approach demonstrates that modern deep learning architectures can provide the flexibility of end-to-end learning while matching or surpassing the performance of traditional methods.

Method	**Dataset**	**Accuracy**	**F1-score**
Fuzzy Cellular Automaton [IMDA23]	Mini-MIAS	98.66%	—
MCSU in Cellular Automata [JA15]	Mini-MIAS	93.48%	—
DDWC-based CNN [BNF21]	CBIS-DDSM	85.9%	—
ResNet34 (this work)	Mini-MIAS	83%	0.77
EfficientNet-B1 (Focal, this work)	CBIS-DDSM	76%	0.74

Table 6.3: Comparison of results with selected related work.

Chapter 7

Software Application

7.1 Overview

As the final technical part of this thesis, a web-based software solution was developed in order to demonstrate the usage of trained mammogram classification models in a practical manner. The actual goal of this application is to provide an intuitive interface where users (also taking into account non-technical users) can upload mammographic images and obtain a binary classification result.

The described application integrates a full preprocessing pipeline, model inference using a fine-tuned EfficientNet-B1 neural network and a responsive web interface built upon Flask and Bootstrap. It supports various input formats, these including common image formats like PNG, JPG and JPEG but also, it allows for the input of DICOM (.dcm) medical images. The application gives the user the ability to view the originally uploaded image, the image after it has passed the preprocessing stage and evidently, the predicted classification result, helping to increase transparency and interpretability.

7.2 Architecture

A Flask-based web server, a PyTorch deep learning model, and a basic HTML+Bootstrap user interface are all combined in the application's modular design.

The system's primary parts are:

- **Frontend:** Offers a user interface via which users may examine classification results, initiate preprocessing, and submit mammography images. HTML is used to implement the interface, and the Bootstrap framework is used to style it.
- **Backend:** In addition to handling user requests, a Flask web server prepro-

cesses images, executes model inference, and provides the frontend with the results.

- **Preprocessing Pipeline:** To get pictures ready for classification, the backend has a pretreatment pipeline that uses a number of image processing techniques, including scaling, thresholding, segmentation, and CLAHE enhancement.
- **Deep Learning Model:** The application uses an EfficientNet-B1 model fine-tuned on the CBIS-DDSM dataset. The model is loaded once at application startup and used to perform inference on preprocessed images.

The interaction between the components is illustrated in the following workflow:

1. The user uploads a mammogram image through the web interface.
2. The image is converted to PNG format if needed (DICOM support is provided).
3. The user triggers preprocessing; the image is processed and displayed.
4. The preprocessed image is fed into the EfficientNet-B1 model to obtain a classification result.
5. The classification result is displayed in the interface.

This architecture ensures that the system is lightweight, easy to deploy, and responsive for demonstration purposes.

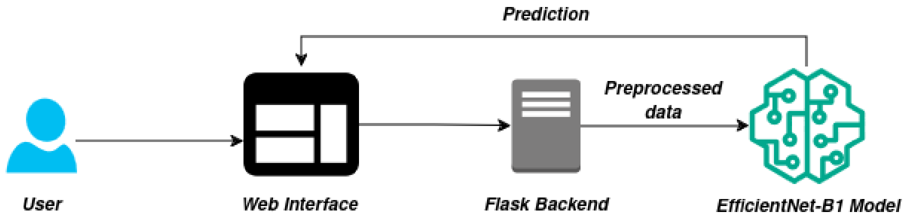


Figure 7.1: Application flow diagram.

7.3 Backend

Using the Flask framework, the application's backend is developed. It controls picture upload and conversion, preprocesses images, handles user interactions, and carries out model inference. The file format is checked by the backend initially when a user uploads an image. For compatibility and presentation, a DICOM (.dcm) image is automatically converted to PNG format if one is supplied. Other widely used image formats, such as PNG, JPG, and JPEG, are directly accepted.

The user can start the preprocessing stage after uploading. Consistency between training and inference is ensured by the backend using the same preprocessing pipeline as utilized during model training. More transparency is made possible by saving and displaying the processed image to the viewer. EfficientNet-B1, a deep learning model optimized on the CBIS-DDSM dataset, is utilized for classification. When the program launches, the model is loaded into memory and is always available for inference. The architecture has been modified to produce class probabilities for benign and malignant cases and to take grayscale mammography pictures.

The preprocessed picture is transformed into a tensor that can be fed into the model and normalized just like the training setup when the user asks a prediction. The final prediction is determined by applying a threshold of 0.75 to the likelihood score that the model generates for the malignant class. The online interface shows the projected class and likelihood. The backend's efficiency, consistency with the training pipeline, and ability to deliver precise predictions in real time are all guaranteed by this design.

7.4 Frontend

The application's frontend offers a straightforward and user-friendly interface for interacting with the mammography categorization system. To guarantee a neat and responsive layout, HTML is used in its implementation, and the Bootstrap framework is used for styling.

The user is shown an interface to submit a mammogram picture when they first use the application. Both DICOM (.dcm) and popular image formats (PNG, JPG, and JPEG) are supported by the interface. An image appears in the interface in its original format once it has been uploaded.

The user is given the ability to launch the preprocessing stage, which applies onto the raw images the same transformation that were described in the model training stage. The processed image that results from this action is also displayed, giving more insight into what ultimately gets passed onto the classification model. Following preprocessing, the application gives you the ability to request a prediction on the selected image and then, after the information gets passed to the model, the result is clearly visible in the interface.

Even non-technical users can engage with the system and comprehend the processing flow and outcomes because to the frontend design's emphasis on transparency and user-friendliness.

7.5 Design

The application's design was informed by the concepts of user accessibility, clarity, and simplicity. The objective was to design an interface that would facilitate simple engagement with the mammography classification process for both technical and non-technical users.

The Bootstrap framework was used to construct the interface layout, guaranteeing a responsive design that works on desktop and mobile devices. The original and preprocessed photos were positioned side by side to promote visual clarity and let people perceive the preprocessing impacts right away.

A clear and straightforward action flow was implemented:

- Upload image
- Trigger preprocessing
- Trigger classification
- View result

A highlighted information box using color cues to enhance reading displays the prediction result. This guarantees that the user may see the crucial output right away.

To prevent visual distraction, the color scheme was kept light and neutral throughout, and the use of icons and space gives the design a contemporary, polished look. Because of its straightforward design, the application may be used as a demonstration tool and as a springboard for the creation of more sophisticated clinical applications.

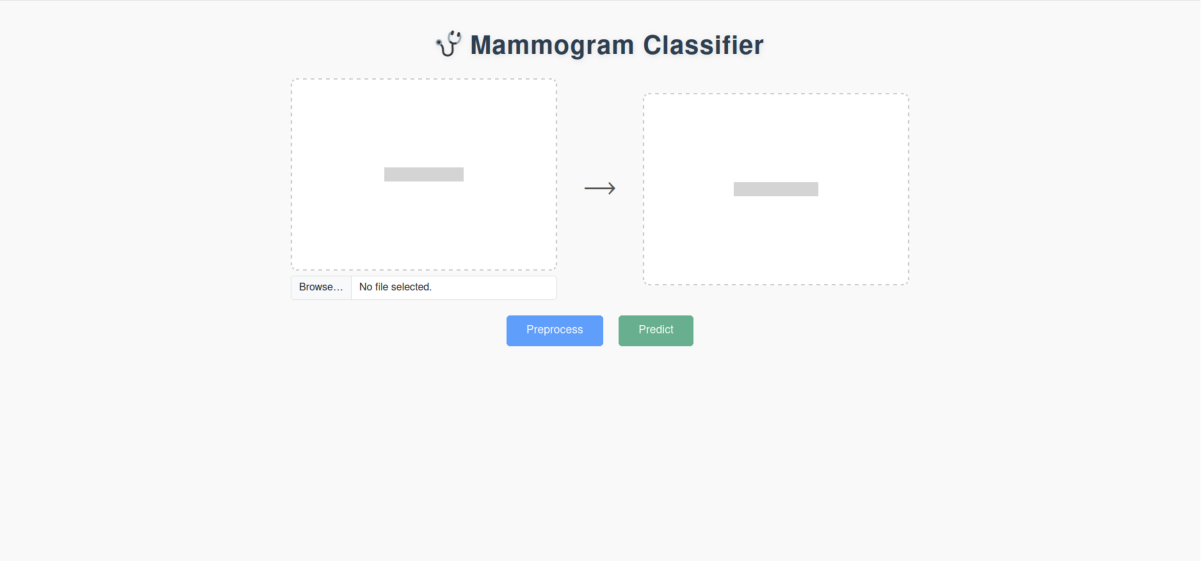
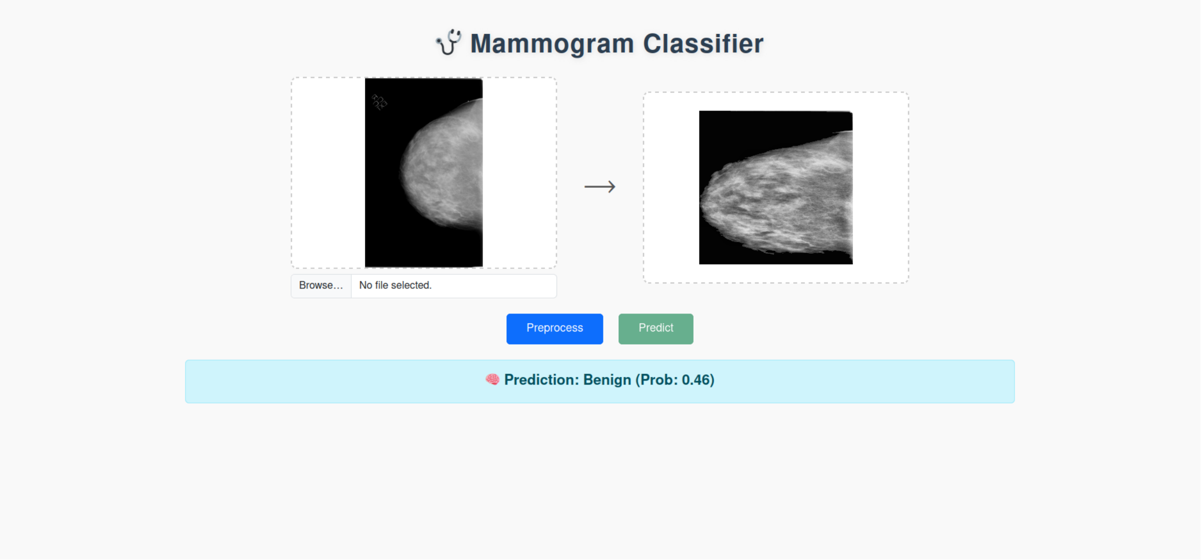


Figure 7.2: Application in the incipient stage.



Prediction: Benign (Prob: 0.46)

Figure 7.3: Application state after the full software flow.

Chapter 8

Conclusions and Future Work

8.1 Contributions and Results

In the light of all that was presented, the main goal of this thesis was to develop a fully automated system for the detection and prevention of breast cancer by utilizing binary classification of mammographic images using deep learning techniques. The project has as its result a complete and modular pipeline.

Several models were trained and evaluated on two widely available and used datasets by strictly following consistent training and evaluation protocols. Overall, the experimental results proved that modern deep transfer learning models such as EfficientNet-B1 and ResNet34 significantly outperformed the baseline CNN on both datasets regardless of changes applied to it. ResNet34 produced the greatest results on Mini-MIAS, demonstrating the efficacy of deeper designs even on tiny and homogeneous datasets with an accuracy of 83% and an F1-score of 0.77. With an accuracy of 76 percent and an F1-score of 0.74 on the more difficult CBIS-DDSM dataset, EfficientNet-B1 trained with Focal Loss produced competitive results when compared to current research.

The preprocessing pipeline that was established, which includes methods like artifact removal and CLAHE enhancement, was crucial for enhancing the quality of the input images and facilitating greater model generalization. This preprocessing step's inclusion increased the classification pipeline's overall robustness.

When we are looking at the proposed approach, one of its main strengths lies in its full automation, giving the user the ability to generate classification results with minimal intervention required. The performance that was obtained across diverse datasets demonstrates even further the adaptability of the solution. The results were competitive in regard to those reported in the related literature, giving further confirmation to the viability of the approach. This project also resulted in lightweight and interpretable web-based application that demonstrates the practical potential of deploying deep learning models in real-world scenarios.

Even with all this being said, there are some limitations to the proposed approach that are worth taking into consideration. Hyperparameter optimization was not used in this work and also, more advanced tuning might produce better outcomes and model performance. The SimpleCNN baseline model gave early signs of clear limitations when put side to side with more advanced transfer learning architecture, putting a strong emphasis on the importance of selecting the right architecture for this respective domain. Additionally, the models showed susceptibility to class imbalance, which Focal Loss helped to mitigate to some extent, while other approaches might be investigated. Lastly, hardware constraints limited the number of training epochs and batch sizes for training and evaluation.

8.2 Future Work

The next logical step when we are talking about future research and development is uncertain since, there are a lot of areas and paths than can be explored in this direction. One important area is Hyperparameter optimization. Model performance could be further improved by carrying out methodical searches to maximize learning rates, batch sizes, and other training parameters. Investigating ensemble approaches, which combine the advantages of several models to increase accuracy and robustness, is another exciting avenue.

The continuous improvement of the interpretability of the system is also an important next step. Adding explainability techniques into the equation such as Grad-CAM visualizations would allow users to better grasp and trust model predictions, a crucial step in medical applications. Another area of interest for future work involves the extension of the system in order to support multi-class classification or even the grading of lesion severity, moving away from the concept of binary classification and more towards full diagnostic capabilities. Additionally, the system's clinical relevance would be enhanced by including it into extensive computer-aided diagnostic (CAD) pipelines that incorporate segmentation, detection, and reporting features.

Nevertheless, collaborating with medical professionals in order to conduct clinical validation of the system onto private datasets and within real diagnostic workflows is a critical step if we even want to think about practical applications and clinical adoption.

To sum up, this thesis establishes a strong basis for a useful and efficient mammography categorization system. Such systems have a lot of promise to help radiologists and enhance breast cancer early detection with more development, optimization, and clinical validation.

Bibliography

[AAMTW18] Esra Alhadhrami, Maha Al Mufti, Bilal Taha, and Naoufel Werghi. Transfer learning with convolutional neural networks for moving target classification with micro-doppler radar spectrograms. pages 148–154, 05 2018.

[BNF21] Madjid KhalilianMadjid Khalilian Behrouz Niroomand Fam, Alireza Nikravanshalmani. An efficient method for automated breast mass segmentation and classification in digital mammograms. *IJ Radiology*, 18(e106717), 2021.

[ea94] J. Suckling et al. The mini-mias database of mammograms, 1994. Accessed: 2025-03-25.

[EJA+09] Joann G. Elmore, Sarah L. Jackson, Lynn Abraham, Diana L. Miglioretti, Patricia A. Carney, Berta M. Geller, Bonnie C. Yankaskas, Karla Kerlikowske, Tracy Onega, Robert D. Rosenberg, Edward A. Sickles, and Diana S. Buist. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, 253(3):641–651, Dec 2009. Epub 2009 Oct 28.

[GZM20] Christian Garbin, Xingquan Zhu, and Oge Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79:1–39, 05 2020.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[IMDA23] Iulia-Andreea Ion, Cristiana Moroz-Dubenco, and Anca Andreica. Breast cancer images segmentation using fuzzy cellular automaton. *Procedia Computer Science*, 225:999–1008, 2023.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In

Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[JA15] Peter J. Dinesh J. Anitha. Mammogram segmentation using maximal cell strength updation in cellular automata. *Medical & Biological Engineering & Computing*, 53:737–749, 2015.

[Kar16] Andrej Karpathy. Cs231n: Convolutional neural networks for visual recognition. `http://cs231n.stanford.edu/`, 2016. Accessed: 2025-04-15.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LKB+17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[MDBAC22] 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.

[MDIA21] Cristiana Moroz-Dubenco, Iulia-Andreea Ion, and Anca Andreica. Mammography lesion detection using an improved growcut algorithm. *Procedia Computer Science*, 192:308–317, 2021.

[NH10] Vinod Nair and Geoffrey Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. volume 27, pages 807–814, 06 2010.

[Ora24] World Health Oranization. Breast cancer. `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`, 2024. Accessed: 15.04.2025.

[Pra21] Rukshan Pramoditha. The concept of artificial neurons (perceptrons) in neural networks. `https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-netw` 2021. Accessed: 15.05.2025.

[PZH^{+}98] Etta D. Pisano, Shuquan Zong, Bradley M. Hemminger, Marla DeLuca, R. Eugene Johnston, Keith Muller, M. Patricia Braeuning, and Stephen M. Pizer. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, 11(4):193, 1998.

[SHK^{+}14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[SLGHR16] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm) (version 1) [data set]. The Cancer Imaging Archive, 2016.

[SOPH16] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[SRG^{+}16] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Ivan Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, May 2016. Epub 2016 Feb 11.

[TSG^{+}16] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, Robert T. Hurst, C. Brandon Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016. Epub 2016 Mar 7.

[WPCA19] Volker Weinberg, Damian Podareanu, Valeriu Bogdan Codreanu, and Sandra Aigner. Best practice guide - deep learning. Technical report, ResearchGate, 2019.

[WRS^{+}17] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in*

Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.