Detection of Suspicious Microcalcification Clusters in Mammography: A Patch Classification Approach

Adam Mračko, Aneta Gábrišová, Darina Narova, Ivan Cimrák
University of Žilina
Žilina, Slovakia
Adam.Mracko,Aneta.Gabrisova,Darina.Narova,Ivan.Cimrak@fri.uniza.sk

Lucia Vanovčanová
Oncological Institute of Saint Elizabeth
Bratislava, Slovakia
lucia.vanovcanova@ousa.sk

***Abstract*—Background/Motivation: Breast cancer is the most common type of cancer among women. Integrating machine learning models with mammography holds the potential to improve breast cancer screening and diagnostics by making them more accurate, efficient, and accessible, for example by assessing whether a whole mammogram contains suspicious clusters of microcalcifications. Methods: This study addresses the classification of mammogram patches into two categories: those with suspicious clusters of microcalcifications and those without. Subsequent processing of whole mammograms by sliding window will provide the assessment of the whole mammogram. Using high-resolution patches (674 x 674 pixels) was crucial to preserve the detail necessary for detecting microcalcifications. Dataset: Data were sourced from the CBIS-DDSM and OMI-DB databases, each presenting unique challenges in pre-processing. The study highlighted the importance of manual evaluation to ensure the accuracy of patches, particularly when generating patches without suspicious clusters. Proposed Model: For model training, the ResNet101 convolutional architecture was employed, leveraging transfer learning with pre-trained weights on ImageNet to achieve faster convergence and better performance. Various hyperparameters, including learning rate and weight decay, were optimized. Results: The best model achieved a validation accuracy of 98.2% (F1 score - 0.955, MCC - 0.944, specificity - 99.1%, and sensitivity - 94.6%). Conclusion: The achieved performance and model interpretation demonstrate a strong capability in identifying important radiological features and handling visually challenging microcalcifications. The model has been made publicly available. Despite some incorrect predictions, the model reliably located clusters, suggesting practical utility in clinical settings where high-resolution imaging is essential.**

I. Introduction

Breast cancer has an incidence rate of 46.8 per 100,000 people (age-standardized) [1]. It's the most common type of cancer. Early diagnosis can lead to almost complete recovery, with a 99% 5-year relative survival rate if detected at a local stage [2]. Because of this, many countries have implemented mammography screening programs. Mammography is an effective method for detecting cancer before any symptoms appear. Common abnormalities detectable by mammography include masses, calcifications, architectural distortions, and asymmetries. The presence of these abnormalities may indicate cancer.

This work focuses on the microcalcification findings. The other type of calcifications is macrocalcifications, which are larger and typically benign compared to microcalcifications. Isolated microcalcifications are very common benign abnormalities in the breast. Suspicion of malignancy increases when microcalcifications form a cluster. Classifying these clusters is challenging due to the high variability in their shape, density, size, number, and distribution. Correct classification can lead to the detection of ductal carcinoma in situ (DCIS), a pre-invasive type of cancer that can progress into a more dangerous invasive type if left untreated. Clusters of calcifications are not palpable during a physical breast examination, making regular screening important. DCIS accounts for about 20% to 30% of all breast cancer cases [3], with mammography diagnosing about 80% to 90% of these cases [4].

A screening mammography must be evaluated independently by two radiologists [5]. If both classify the finding as suspicious, a biopsy is recommended. Only a biopsy can provide the most relevant information about the lesion's dignity. Only 15% to 45% of biopsies confirm malignancy [6]. These specifics, along with the volume of examinations involved in screening, highlight the difficulties associated with accurate classification.

Introducing artificial intelligence (AI) models into the examination process could shorten the evaluation time and improve radiologists' accuracy. Currently, convolutional neural networks (CNNs) are one of the best choices for tasks that involve image data. They are suitable for classification, detection, and segmentation tasks. This work focuses on the binary classification of mammogram patches, determining whether a patch contains a suspicious cluster of microcalcifications. Considering that the first step of any examination is the localization of suspicious abnormalities, emphasis will be placed on interpreting incorrect predictions with the Grad-CAM method during model validation [7].

The patch-based approach is designed to function as a medical decision support system, helping the radiologist who has already located a suspicious abnormality. The model provides additional evidence specifically when the physician is experiencing difficulty in binary classification—determining if the localized finding is a suspicious cluster requiring further investigation (such as a biopsy) or if it is a benign formation. By focusing on this specific localized area, the model directly supports the workflow where the radiologist needs a second opinion on a specific finding.

Dense fibroglandular tissue, poor image quality, and overlapping structures can make identifying clusters difficult. Both

microcalcifications and fibroglandular tissue appear white on mammograms.

Normally, fibroglandular tissue undergoes a fatty transformation with age. When it persists, it can obscure other abnormalities, particularly masses, as well as microcalcifications. Therefore these patients are regularly referred for ultrasound examination. However, ultrasound is not sensitive enough for the detection and analysis of microcalcifications. Fibroglandular tissue consists of the following:

- Fibrous tissue: Provides structural support to the breast, giving it shape and firmness. It is made up of connective tissue.
- Glandular tissue: Includes the lobules and ducts involved in milk production and transport. The lobules are the milk-producing glands, and the ducts are the channels that carry milk to the nipple.

A. Related studies

The study [8] focused on solving two tasks using convolutional neural networks (CNNs). The first task was similar to the focus of this study: detecting microcalcifications in mammogram patches. The second task aimed at classifying patches with microcalcifications as either benign or malignant. The authors compared metrics across three different architectures: AlexNet, ResNet18, and ResNet34. They used their own dataset for training and testing the models, consisting of 1986 mammograms from 1000 unique patients. All images were collected from a single institute and annotated by three expert radiologists. Key differences from our study include:

- They used patches of 112 x 112 pixels, whereas our study uses patches of 674 x 674 pixels.
- Several 112 x 112 pixel patches were created from a single cluster (our patches aim to cover the entire cluster or multiple clusters).
- Patches without microcalcifications were taken from mammograms with microcalcification annotations to avoid overlap (we used mammograms from patients without any history of microcalcifications, including patches with other types of abnormalities).

On their test set, they achieved the best accuracy with the AlexNet architecture, reaching 95% accuracy, 98% sensitivity, and 89% specificity.

The second study [9] focused on classifying patches with microcalcifications by developing a custom convolutional architecture. They used the INbreast [10] database, which individually labels calcifications rather than marking a single suspicious cluster. This results in the database having many annotations that either cover too large an area with several clusters or just isolated calcifications. Therefore, manual evaluation of patches by a radiologist was necessary for the created dataset. In this study, smaller patches of 144 x 144 pixels were used. Their custom architecture, with 8,301 parameters, achieved an accuracy of 99.3%. For comparison, the authors also trained a MobileNetV2 architecture with 67,797,505 parameters, which achieved a slightly higher accuracy of 99.8%, but at the cost of significantly more parameters. The authors noted that without augmentation, they obtained 1576 patches with findings. Given that the database only offers 308 mammograms with microcalcifications, it implies that multiple patches were created from each mammogram.

The last study [11] on microcalcification detection used the Categorized Digital Database for Low-Energy and Subtracted Contrast-Enhanced Spectral Mammography (CDD-CESM) [12]. This database includes mammograms from a new imaging modality aimed at improving diagnostic accuracy over standard digital mammography. However, the authors chose to use 212 standard digital mammograms from the database. The study used patches of 224 x 224 pixels and focused on both isolated calcifications and suspicious clusters. All created patches were visually evaluated by four radiologists and categorized as either containing calcifications or not. The study tested three architectures: ResNet18, ResNet50, and ResNet101. All architectures achieved very similar, comparable results, with ResNet50 achieving the highest overall accuracy of 96.4%.

The innovation and contribution of this work stem primarily from our advanced data handling methodology and the scale of input resolution, specifically designed to overcome limitations observed in existing microcalcification detection studies. We present a significant advancement over previous works by leveraging high-resolution image patches and combining diverse clinical data.

The first significant difference in our study is the use of large patches (674 x 674 pixels) that cover entire clusters or multiple clusters. This high resolution is a key methodological contribution, as our previous research confirmed that reducing the resolution significantly decreased the classification accuracy, a critical factor when dealing with tiny microcalcifications.

The second difference is the significantly larger number of mammograms from which the patches were created. Our study worked with thousands of unique mammograms from different patients. By combining CBIS-DDSM (SFM) and OMI-DB (FFDM), we ensured that the model was trained on a wider variety of realistic clinical cases, enhancing reliability compared to models trained on single, smaller databases such as INBreast or CDD-CESM.

A third key contribution is the meticulous and transparent data curation required, especially for generating the negative class. This rigorous process involved manual visual evaluation and the deliberate removal of approximately 550 ambiguous patches, which is often overlooked but crucial for preventing ”trash-in, trash-out” issues.

B. Overview

In Section II we describe two databases of mammography images with insightful details and also we provide the preprocessing steps such as mask adjustments, image inverting, and patch filtering, that were performed in order to prepare the datasets for the training. Section III presents the results of numerous experiments. The best AI model is made publicly available in Supplementary section. For interpretation, the

Grad-CAM method is used to identify which areas in the image were important for the selected class (suspicious cluster or no suspicious cluster). We provide explanations for incorrect predictions, which were consulted with radiologists. Finally, in Section V, we conclude with several remarks.

II. Materials and methods

A. Mammography Data

For training and validation purposes, the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [13] and the Optimam Database (OMI-DB) were used [14].

CBIS-DDSM is a newer, standardized version of the Digital Database for Screening Mammography (DDSM). It is freely accessible to anyone and can be downloaded using specialized software called NBIA Data Retriever. The database provides mammograms in DICOM format, a standard for transmitting medical images. The images were originally created using screen-film mammography (SFM), a now older technology where the breast image is captured on a special physical film. Later they were digitized to create a database. Information on individual findings is available in CSV (comma-separated value) files, and cases include histopathological results. The position and size of each finding are determined using binary masks, which are also stored in DICOM format. These masks are the same size as the corresponding mammogram, with the finding area marked in white and the background (other) area marked in black. The advantage of these masks is the precise delineation of the finding's boundaries, which could allow the use of convolutional networks for segmentation/detection tasks. The database contains two types of findings: masses and calcifications, including both microcalcifications and macrocalcifications. For suspicious microcalcifications, the masks do not mark individual microcalcifications but the boundaries of the cluster. An advantage is the official distribution of data to the training and test set.

Unlike CBIS-DDSM, OMI-DB is not fully open-access. Access is granted to groups with relevant experience affiliated with commercial, non-commercial, or educational organizations. Until 2020, new exams were added annually from several cancer centers across the UK. There are no updates on the number of new exams post-2020. Access is granted to a subset of the exams based on an agreement with the provider. A major advantage of OMI-DB is that the images are created using modern Full-Field Digital Mammography (FFDM), which is the current standard for screening. FFDM produces direct digital images in DICOM format, generally capturing details better than SFM. However, OMI-DB is more complex, and findings are not simply described in CSV files like CBIS-DDSM. Instead, the official Python library "omidb" [15] must be used for data processing. It includes all common types of abnormalities, with masses and suspicious calcifications being the most frequent. The size and position of findings are defined using bounding boxes (two coordinates – bottom left corner and top right corner). A bounding box can contain a combination of multiple types of findings (e.g., calcifications + mass + architectural distortion). The database also includes mammograms of healthy patients with no suspicious abnormalities requiring biopsy. Another unique feature is the inclusion of images from previous exams of the patient before any suspicious abnormality was detected.

Combining these two databases, which use different technologies (FFDM and SFM), proved effective in our previous research [16] focused on the binary classification of patches with microcalcifications into benign or malignant classes. The combination significantly improved classification accuracy and model interpretability. The research showed that a model trained on a single database could not effectively transfer its knowledge to the other, highlighting the benefit of a combined dataset. This improvement is also due to the larger number of training patches and better class balance. CBIS-DDSM provides more benign microcalcifications, while OMI-DB offers more malignant ones. However, a potential drawback of CBIS-DDSM is the number of findings with microcalcifications that form small groups rather than clusters. Generally, a small group of calcifications is not considered suspicious. Specialists debate how many calcifications are needed to be classified as a cluster, usually at least five close together. In comparison, OMI-DB contains more benign clusters that are challenging for radiologists to classify correctly as they closely resemble malignant clusters.

B. Data Pre-processing and Dataset Creation

The goal of this study is to accurately classify image patches into one of two categories: patches with suspicious clusters of microcalcifications and patches without suspicious clusters (see Figure 1). All patches were 674 x 674 pixels in size. High resolution was necessary because microcalcifications are very small, and reducing the resolution could result in the loss of important details. Previous studies [17] have also discussed the importance of not reducing mammogram resolution. Our previous research [18] confirmed that reducing patches to a standard resolution of 224 x 224 pixels decreased classification accuracy. We also observed that patches from FFDM images performed better after resizing compared to those from SFM images. Before creating all patches, the mammograms were normalized to values between 0 and 1.

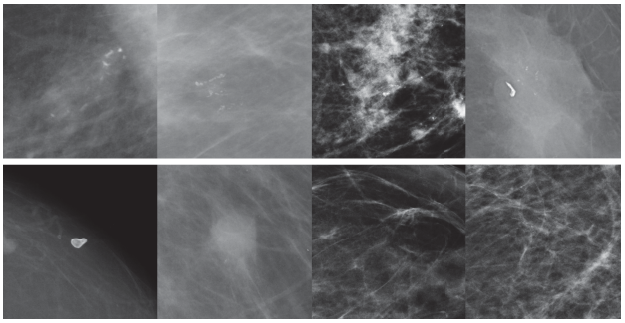

Fig. 1. The upper row displays patches containing suspicious clusters of microcalcifications. In contrast, the lower row presents patches without microcalcification clusters, possibly featuring other types of findings.

Patches with Suspicious Clusters

Creating patches with suspicious clusters of microcalcifications was relatively straightforward. Clusters larger than 674 x 674 pixels were excluded. For smaller clusters, the surrounding area of the mammogram was included, centering the cluster in the patch. If this wasn't possible (e.g., clusters near the edges of the mammogram), the patch was shifted toward the center of the mammogram.

For CBIS-DDSM, around 30 additional adjustments were made. Some masks had different resolutions compared to the original mammogram and were re-scaled. When multiple findings were close together, masks were unified. Some masks were slightly adjusted to cover the finding accurately. If a mask couldn't be linked to a visible finding, it was excluded. Findings with small groups of microcalcifications were retained despite their lesser interest compared to clusters.

In OMI-DB, some inverted mammograms were found (normally, the background is black, but these had a white background). This inversion was likely done by radiologists to better classify abnormalities. The problem was fixed by re-inverting the images. A few images had lower quality with less sharpness and unexpected gray backgrounds. These findings were also retained. Given the database's complexity, suitable findings were filtered using the following criteria:

- Bounding boxes only with suspicious calcifications (no other marked abnormalities).
- Findings clearly linked to histopathological results.
- Histopathology results had to be either malignant or benign.
- Bounding boxes had to have valid coordinates and non-empty content.

Patches without Suspicious Clusters

Creating patches without suspicious clusters was more challenging and brought several non-intuitive obstacles. No public database directly offers such data. The main idea was to include as much variety as possible (other abnormalities + healthy tissue) that the model might encounter on a full mammogram.

From CBIS-DDSM and OMI-DB, patches with masses were manually reviewed, as many contained suspicious microcalcifications. Especially malignant masses typically include other features like microcalcifications, and such cases were excluded. Despite OMI-DB's ability to indicate combinations of findings in one bounding box, many cases did not specify them.

Additional patches came from macrocalcifications in CBIS-DDSM. Macrocalcifications rarely form clusters and are easy for doctors to classify as typically benign. The largest category of added patches was from healthy tissue in OMI-DB. These were from patients without any histopathological records or bounding boxes (no abnormalities noted). One patch was generated from a random location on each mammogram, with at least 70% of the patch overlapping the breast. A total of 9,699 patches were manually reviewed, and approximately 550 patches with groups or clusters of microcalcifications were removed. Due to the frequent presence of unannotated groups/clusters of microcalcifications, the study focused on classifying patches rather than detecting them on full mammograms. Object detection training and validation would be complicated due to insufficient annotations.

Moreover, mammograms would need to be resized to a smaller resolution due to GPU memory constraints. Patches with individually scattered calcifications were retained (Figure 2), as were patches with vascular calcifications (Figure 3), which, while similar to malignant calcifications, are easy for radiologists to diagnose due to their obvious placement along vessels.

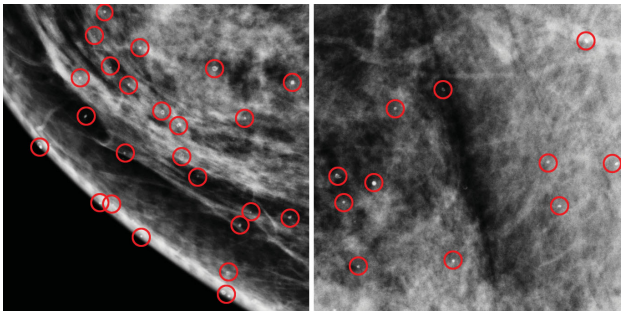

Fig. 2. Examples of diffuse calcifications, individual calcifications are marked with a red circle.

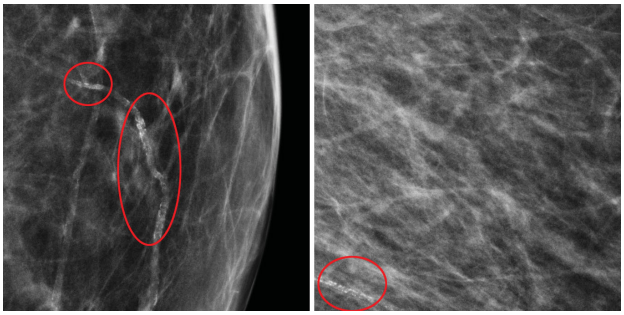

Fig. 3. Examples of vascular calcifications marked with a red circle.

Summary of Created Dataset

For CBIS-DDSM patches, the official distribution into training/validation sets was used. OMI-DB does not provide such a distribution, so the data was split approximately 80:20 into training and validation sets. Care was taken to ensure data independence between sets, which means that patches from a specific patient are never in both sets. All training set patches, except healthy tissue patches, were augmented with three rotations of 90 degrees (90, 180, and 270 degrees). The total number of patches (with augmentation) is provided in Table II-B.

TABLE I. NUMBER OF PATCHES FOR EACH CLASS INCLUDING 90 DEGREE ROTATIONS

Clusters	Train	Validation	Total
Suspicious Clusters	10752	654	11406
Without Clusters	19190	2596	21786
Total	29942	3250	33192

III. EXPERIMENTAL STUDY AND RESULTS

All models were trained using the PyTorch framework [19], running on Ubuntu with an Nvidia GeForce RTX 4080 16GB GPU. The Grad-CAM library [20] was used for model interpretation.

The chosen convolutional architecture was ResNet101 [21], which in our previous research [18] focusing on mammography patches was able to achieve the best results compared to other architectures such as VGG [22], Inception-V3 [23], DenseNet [24] and EfficientNet [25]. Although the other architectures produced comparable results, ResNet101 performed slightly better. While smaller, lightweight architectures (such as EfficientNet or MobileNet) might offer advantages in terms of faster inference time and deployment feasibility, our selection prioritized maximizing classification accuracy by utilizing the empirically superior ResNet101 backbone in conjunction with high-resolution patches (674 x 674 pixels). Transfer learning was employed using pre-trained weights from the ImageNet dataset [26]. Using pre-trained weights generally leads to faster convergence and improved performance with limited data [27]. The final classification layer was replaced with two freshly initialized neurons. During each training epoch, all layers' weights were unlocked for training. Due to the memory demands on the hardware, a mini-batch size of 8 was used in each experiment.

Augmentation of patches in the training set in the form of 90-, 180-, and 270-degree rotations of patches with microcalcification clusters was used in each experiment. Additional weighting was needed for the cross-entropy loss function to better balance the classes:

- 0.359 for the class without suspicious clusters
- 0.641 for the class with suspicious clusters

Each experiment was carried out for a maximum of 40 epochs, tracking the best validation accuracy achieved.

A. Summary of the experiments

The first experiment focused on finding the optimal learning rate for the Adam optimizer. The models performed best with learning rates between 1e-4 and 1e-6, as shown in Table III-A. The best model achieved 97.3% accuracy with a learning rate of 5e-6.

TABLE II. RESULTS OF EXPERIMENTS WITH DIFFERENT VALUES OF LEARNING RATE

Learning Rate	Val. Acc.	Train Acc.
1e-4	97.2%	97.3%
5e-5	97.2%	98.5%
1e-5	97.2%	98.9%
5e-6	97.3%	99.3%
1e-6	97.0%	98.2%
AVG	97.2%	98.4%

Next, regularization was applied using the weight decay hyperparameter to prevent over fitting by penalizing large weights (smaller weights result in more stable training). The goal of weight decay is to encourage the model to find simpler and more robust solutions. Weight decay works similarly to L2 regularization when used with the Adam optimizer. The learning rate of 5e-6 from the previous experiment was used. The results are in Table III-A. Most values improved accuracy, except for a value of 1e-1, which worsened accuracy, likely due to overly aggressive weight decay. The best value was 1e-3, achieving 97.6% validation accuracy and the best sensitivity (the proportion of correct predictions of suspicious clusters out of all suspicious clusters).

TABLE III. RESULTS OF EXPERIMENTS WITH DIFFERENT VALUES OF WEIGHT DECAY

Weight Decay	Val. Acc.	Train Acc.
1e-1	97.1%	100.0%
1e-2	97.4%	99.5%
1e-3	97.6%	99.0%
1e-4	97.6%	99.9%
1e-5	97.6%	100.0%
AVG	97.5%	99.7%

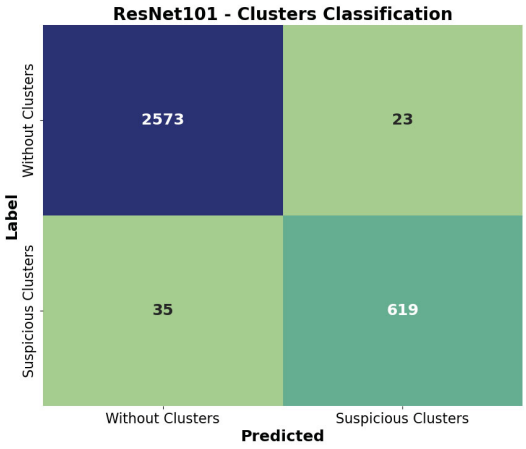

Fig. 4. Confusion matrix of the best model

The final experiment focused on better data augmentation performed directly during training. Each epoch applied random augmentation operations to each patch. Using PyTorch tensors, each patch had a 50% chance of a horizontal flip and the same chance for a vertical flip. Each patch could also be rotated

by 0 to 89 degrees. This additional augmentation, combined with the best-discovered hyperparameters, produced the best model, achieving the following metrics on the validation set: 98.2% accuracy, F1 score of 0.955, MCC of 0.944, specificity of 99.1%, and sensitivity of 94.6%. The confusion matrix is shown in Figure 4.

Grad-CAM interpretation confirmed that the model made decisions based on significant radiological features. Several images will be presented in pairs. The left image of the pair will be the original patch used as input for the model, supplemented by a red circle indicating the area of interest. The right image of the pair will be a Grad-CAM interpretation of the left patch for the predicted class. One figure will contain multiple pairs. The remaining images will be made up of triplets of images. The left image will be input patch. The middle image will show what contributed to the class with suspicious clusters. The right image will show the contribution to the class without suspicious clusters.

B. Grad-CAM Interpretation

From visual inspection, it was clear that all correct predictions for patches with suspicious clusters were based on actual suspicious microcalcifications. The model could accurately locate visually challenging calcifications (Figure 5), handle extensive clusters covering a large part of the patch (left pair in Figure 6), and correctly detect multiple significant clusters in one patch (right pair in Figure 6).

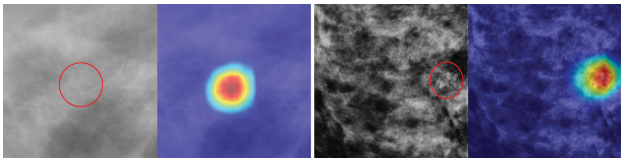

Fig. 5. Left pair: hard-to-distinguish cluster from CBIS-DDSM. Right pair: hard-to-distinguish cluster from OMI-DB.

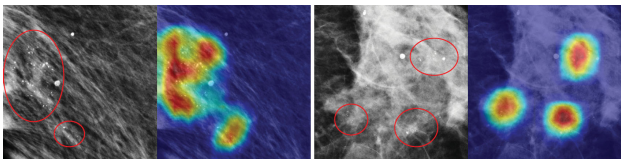

Fig. 6. Left pair: cluster covering large area of the patch. Right pair: correct localization of multiple clusters

Correct predictions: Correct predictions for patches without suspicious clusters were harder to interpret. In some cases, the network made decisions based on the presence of other abnormalities, such as macrocalcifications (left pair in Figure 7) or masses (right pair in Figure 7). In other cases, obvious abnormalities were deemed irrelevant, and the decision was based on a larger area of the patch without a localized focus (left pair in Figure 8). For patches of healthy tissue, the decision was based on a larger area of the patch (right pair in Figure 8). However, it was not possible to define exactly what the model was looking at. If macrocalcifications were present in a patch with a suspicious cluster, the model correctly classified it as containing a suspicious cluster (Figure 9). Even a larger number of macrocalcifications did not pose a problem (Figure 10).

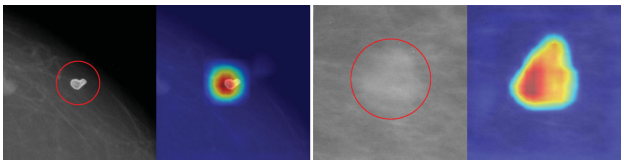

Fig. 7. Left pair: decision based on macrocalcification. Right pair: decision based on mass

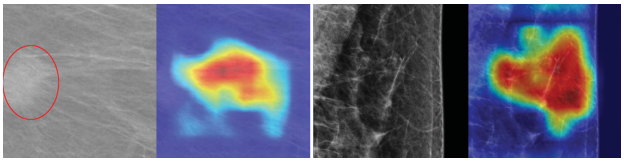

Fig. 8. Left pair: decision based on large area instead of focusing on the mass. Right pair: patch with healthy tissue and decision based on large area.

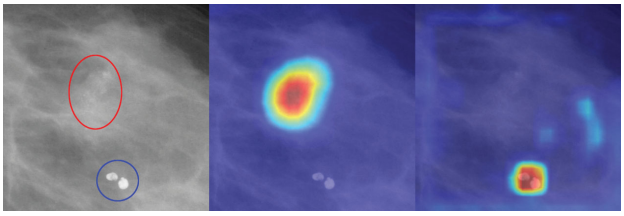

Fig. 9. Input patch with red circle for important cluster and blue circle for macrocalcification.

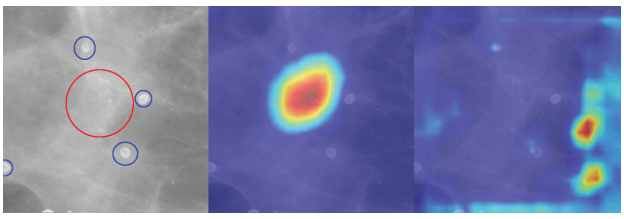

Fig. 10. Input patch with red circle for important cluster and blue circles for multiple macrocalcifications.

Incorrect predictions: Incorrect predictions for patches with clusters into the class without clusters are more difficult to explain. One issue for misclassification is a small number (up to three) of microcalcifications in a patch. It is objectively questionable whether such a patch is rightfully labeled as containing a cluster of microcalcifications. This issue is typical in the CBIS-DDSM database, but may also be found in OMI-DB (Figure 11). Another issue found on four patches is the presence of rare types of calcifications (Figure 12). Solving this problem is more challenging since databases do not offer a large number of such rare types of calcifications. The last issue is linked to patches with visually difficult-to-distinguish

microcalcifications (Figure 13). Even for a trained radiologist, it is challenging to locate such cases.

Nevertheless, the robustness of the model is demonstrated by its ability to correctly locate suspicious microcalcifications in most cases despite misclassification, see Figure 13. For CBIS-DDSM patches, prediction difficulty might be due to poorer detail capture from SFM technology or subsequent mammogram digitization.

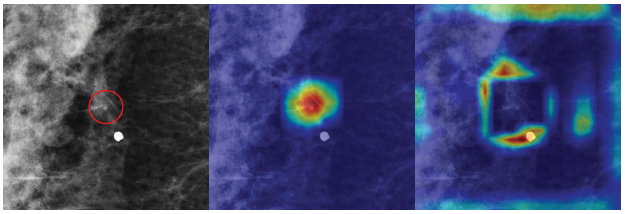

Fig. 11. Example of OMI-DB patch that had an incorrect prediction to a class with no clusters.

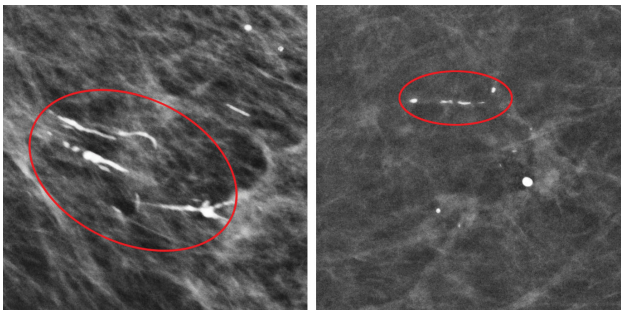

Fig. 12. Two examples of calcification types with very little representation in the databases used.

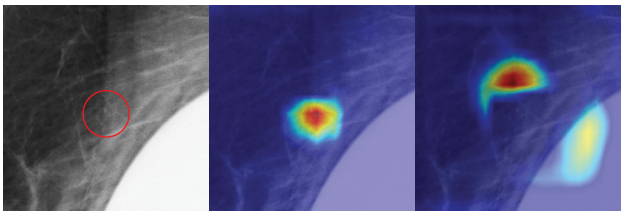

Fig. 13. A very difficult to distinguish cluster of microcalcifications classified to class without clusters. Despite the models incorrect prediction, the Grad-CAM method in the middle image correctly localized the cluster, proving the robustness of the method.

Incorrect predictions of the class without clusters into the class with clusters contain predictions based on macrocalcifications (Figure 14), predictions due to noise (Figure 15), and predictions based on small, uninteresting groups of microcalcifications (Figure 16). Better cleaning of the training set, particularly removing CBIS-DDSM patches containing small benign microcalcification groups, could resolve this issue.

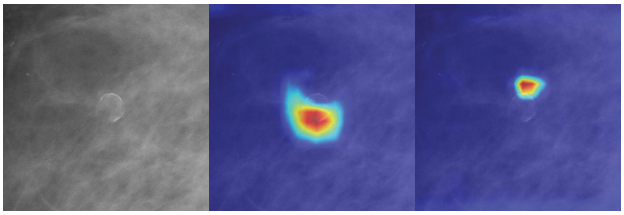

Fig. 14. A macrocalcification with indistinct margins resembling multiple microcalcifications was incorrectly categorized as a cluster of microcalcifications.

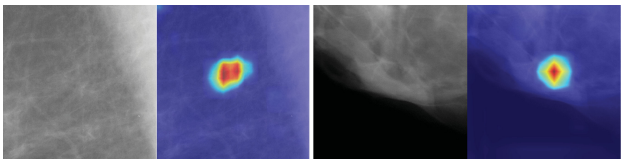

Fig. 15. Example of two incorrect predictions of patches without clusters. The model probably made a decision based on noise, but it is not possible for the human eye to detect anything resembling a cluster of microcalcifications.

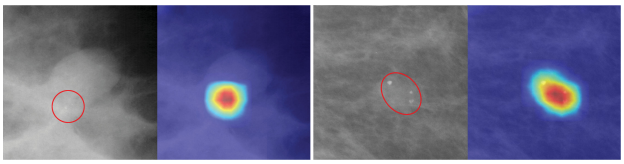

Fig. 16. Two small unimportant groups of microcalcifications categorized as a suspicious cluster. In the left pair, two calcifications are part of the mass.

IV. Discussion

The design choices implemented in this study reflect the intended clinical role of the model as a supporting tool within complex diagnostic workflows. The system is specifically aimed at functioning as a medical decision support system, assisting radiologists who have already identified and localized a suspicious abnormality. In this scenario, the model provides additional, interpretable evidence to help distinguish between a suspicious cluster requiring invasive follow-up and a benign formation.

This study is subject to several limitations. The model's final performance metrics were reported only on the validation set. The absence of an independent, held-out test set is an evaluation limitation that limits the immediate certainty of the model's generalizability.

Although the training incorporated diverse data from two databases utilizing different technologies—screen-film mammography (CBIS-DDSM) and Full-Field Digital Mammography (OMI-DB)—its generalizability to entirely different external datasets or new clinical environments remains uncertain. External validation and prospective clinical trials will be essential to confirm robustness across varied acquisition protocols and annotation standards.

The dataset exhibits class imbalance, with "Without Clusters" patches (21,786 total) predominating over "Suspicious Clusters" patches (11,406 total). This was partially addressed

using a weighted loss function (with a weight of 0.641 for the suspicious cluster class). Future work should explore more balanced sampling or augmentation strategies.

The creation of the high-quality dataset required manual review of nearly 9,700 patches, including removing approximately 550 ambiguous findings and making decisions on small microcalcification groups. This reliance on human review, while necessary for clinical relevance, introduces a degree of subjectivity into the preprocessing that could affect reproducibility.

The necessity of using high-resolution patches (674 x 674 pixels) to preserve microcalcification detail combined with the choice of the large ResNet101 architecture to maximize accuracy demanded significant GPU resources, resulting in a mini-batch size of 8 during training. This design choice implies high computational costs. The study does not provide discussion or quantification of the inference speed or deployment feasibility in high-throughput clinical workflows. Although ResNet101 was empirically superior, future work should compare performance with smaller architectures (e.g., EfficientNet or MobileNet) to explore the trade-off between accuracy and deployment practicability.

False positives remain a significant challenge for clinical deployment. Incorrect positive predictions often stemmed from benign findings such as macrocalcifications with indistinct margins, image noise, or small, uninteresting groups of microcalcifications. While these cases are not indicative of malignancy, they can lead to unnecessary patient follow-ups and increased radiologist workload.

To mitigate these limitations, future work should focus on reducing false positives through contextual modeling and incorporating radiologist feedback mechanisms to enhance specificity, exploring alternative sampling techniques to further reduce the effect of class imbalance, extending the methodology to clinically more relevant multi-class classification (benign versus malignant) and quantifying the inference time of the current model and investigating smaller, more efficient architectures suitable for practical clinical deployment.

V. Conclusions

The study first analyzed the CBIS-DDSM and OMI-DB mammography databases, highlighting their advantages and disadvantages.

- CBIS-DDSM: The main drawbacks include the older SFM technology used to capture the images and the presence of many less significant findings with benign microcalcifications forming small groups.
- OMI-DB: The primary disadvantages are the more complex data processing and the challenging process of gaining access to the database under specific conditions.

A detailed data preprocessing process was described, along with solutions for various issues encountered in the databases, such as masks with incorrect resolutions and inverted mammograms. The reasons for avoiding lower-resolution patches and the unsuitability of using convolutional networks for detection due to insufficient image annotations were also explained. Emphasis was placed on proper cleaning of the dataset.

Important insights were provided for creating the class without clusters. When generating random patches without clusters, manual evaluation is necessary since many breasts might contain groups/clusters of microcalcifications. Additionally, for breasts with malignant clusters, there's an increased risk of other suspicious clusters, so it's recommended to generate healthy patches from patients without any recorded abnormalities.

The experimental part focused on binary classification into classes with and without suspicious clusters. Various learning rate settings were explored, followed by the application of regularization using weight decay and additional data augmentation with random flips and rotations changing each epoch.

The best model achieved a very high validation accuracy of 98.2%. Model interpretation confirmed that decisions for the class with suspicious clusters were based on important radiological features. A significant advantage of the model was its ability to handle visually challenging microcalcifications. This capability could be useful in practice, as many of these clusters might not be detectable without sufficient image magnification (high-resolution monitors are used during examinations for this reason, but it can still be challenging to spot some clusters). The model has been made publicly available, and details are provided as Supplementary material.

Even in incorrect predictions, the model was able to locate the correct positions of clusters, even if the final decision was wrong. The main reason for incorrect predictions was the presence of small groups of microcalcifications in the patches. These cases were more typical for the CBIS-DDSM database. This issue could be mitigated by more thorough cleaning of the dataset.

Supplementary Materials

The supplementary data in the form of a published AI model is available at the GitHub repository https://github.com/icimrak/Microcalc-Detect accessed on 31 May 2024.

Acknowledgment

This research was partially funded (80%) by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the contract No. VEGA 1/0525/23.

This research was partially funded (20%) by project TEF-Health. The project TEF-Health has received funding from the European Union's Digital Europe programme under grant agreement No. 101100700.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, p. 209—249, May 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdfdirect/10.3322/caac.21660

[2] "Survival rates for breast cancer," https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html, mar 2022, accessed: 2022-05-31.

[3] D. Allred, "Ductal carcinoma in situ: Terminology, classification, and natural history," *Journal of the National Cancer Institute. Monographs*, vol. 2010, pp. 134–8, 10 2010.

[4] L. J. Grimm, H. Rahbar, M. Abdelmalak, A. H. Hall, and M. D. Ryser, "Ductal carcinoma in situ: State-of-the-art review," *Radiology*, vol. 302, no. 2, pp. 246–255, 2022, pMID: 34931856. [Online]. Available: https://doi.org/10.1148/radiol.211839

[5] H. Geijer and M. Geijer, "Added value of double reading in diagnostic radiology,a systematic review," *Insights Imaging*, vol. 9, p. 287–301, jun 2018.

[6] J. Chhatwal, O. Alagoz, and E. S. Burnside, "Optimal breast biopsy decision-making based on mammographic features and demographic factors," *Operations research*, vol. 58, no. 6, pp. 1577–1591, 2010.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[8] F. Pesapane, C. Trentin, F. Ferrari, G. Signorelli, P. Tantrige, M. Montesano, C. Cicala, R. Virgoli, S. D'Acquisto, L. Nicosia, D. Origgi, and E. Cassano, "Deep learning performance for detection and classification of microcalcifications on mammography," *European radiology experimental*, vol. 7, p. 69, 11 2023.

[9] R. Luna-Lozoya, H. Ochoa-Domínguez, J. Sossa-Azuela, V. Cruz-Sánchez, and O. Vergara, "Lightweight cnn for detecting microcalcifications clusters in digital mammograms," *Computacion y Sistemas*, vol. 28, pp. 245–256, 03 2024.

[10] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, and J. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Academic radiology*, vol. 19, pp. 236–48, nov 2011.

[11] M. Sakaida, T. Yoshimura, M. Tang, S. Ichikawa, and H. Sugimori, "Development of a mammography calcification detection algorithm using deep learning with resolution-preserved image patch division," *Algorithms*, vol. 16, no. 10, 2023. [Online]. Available: https://www.mdpi.com/1999-4893/16/10/483

[12] R. Khaled, M. Helal, O. Alfarghaly, O. Mokhtar, A. Elkorany, H. Kassas, and A. Fahmy, "Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research," *Scientific Data*, vol. 9, p. 122, 03 2022.

[13] R. Lee, F. Gimenez, A. Hoogi, K. Miyake, M. Gorovoy, and D. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, pp. 170–177, dec 2017.

[14] M. D. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. S. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "Optimam mammography image database: A large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence*, vol. 3, no. 1, p. e200103, 2021, pMID: 33937853. [Online]. Available: https://doi.org/10.1148/ryai.2020200103

[15] "Omidb documentation," https://scicomcore.bitbucket.io/omidb/index.html, accessed: 2022-05-31.

[16] A. Mračko, I. Cimrák, L. Vanovčanová, and V. Lehotská, "Deep learning in breast calcifications classification: Analysis of cross-database knowledge transferability," in *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2024.

[17] K. J. Geras, S. Wolfson, S. G. Kim, L. Moy, and K. Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *ArXiv*, vol. abs/1703.07047, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:14195715

[18] A. Mračko, I. Cimrák, and L. Vanovcanová, "Enhancing breast microcalcification classification: From binary to three-class classifier," 04 2024, pp. 473–481.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[20] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[24] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[25] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[27] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. PP, pp. 1–34, 07 2020.