Enhancing Breast Microcalcification Classification: From Binary to Three-Class Classifier

Adam Mračko, Ivan Cimrák
University of Žilina,
Žilina, Slovakia
adam.mracko, ivan.cimrak@fri.uniza.sk

Lucia Vanovčanová
St. Elizabeth Cancer Institute and Comenius University,
Bratislava, Slovakia
lucia.vanovcanova@ousa.sk

***Abstract*—This research explores the optimization of convolutional architectures for breast microcalcification classification and investigates the transition from binary to three-class classifiers with emphasis on the interpretation results of the Grad-CAM method. The study begins by identifying ResNet101 as the most suitable architecture, achieving competitive results across various models. Subsequent experiments reveal the detrimental impact of reducing image size from 674×674 to the standard 224×224 pixels, attributing decreased model accuracy to the loss of crucial details in already small microcalcifications. Building on these findings, the study introduces a three-class classifier to address limitations observed in binary classification. While the best binary classifier achieves 74,7% accuracy and an MCC of 0,458, interpretation highlights intuitive decision-making based on significant features, albeit with identified shortcomings such as several non-intuitive classification and challenges posed by artifacts and macrocalcifications. Transitioning to a three-class model significantly improves interpretability and model credibility, yielding a 91,7% accuracy and an MCC of 0,767. However, this expansion uncovers new challenges, including misclassification of vascular calcifications and issues with breast implants, emphasizing the complexity of incorporating additional classes.**

I. Introduction

Breast cancer is the most common type of cancer among women [1], emphasizing the importance of early detection and treatment for better patient prognoses. To address this, many countries have implemented mammography screening to detect cancer before any symptoms appear. Common abnormalities identifiable through mammography include masses, calcifications (macro- and micro-), architectural distortions, and asymmetries. This study focuses on clusters of microcalcifications as mammography excels in their detection. Suspicious clusters of microcalcifications often lead to the diagnosis of ductal carcinoma in situ (DCIS), a pre-invasive type of breast cancer that can progress to invasive cancer. DCIS accounts for approximately 20-30% [2] of all breast cancer types, with mammography diagnosing around 80-90% of DCIS cases [3].

Accurate diagnosis of microcalcifications is challenging due to variations in shape, density, size, number, and distribution (either diffuse or clustered). This complexity results in a high number of false positives, with only 15-45% of cases confirmed as malignant after tissue biopsy [4]. Each mammography examination undergoes double reading, where two radiologists must independently agree on the assessment.

Introducing artificial intelligence models could potentially improve and expedite the diagnosis of suspicious abnormalities. If highly accurate, these models could replace the second radiologist in double reading. Currently, convolutional neural networks (CNNs) are the most suitable models for image classification tasks. This study will utilize two large well-known databases with images obtained through different technologies.

The main objective of the study is to compare binary and three-class classifiers. The binary classifier aims to correctly classify clusters of microcalcifications into benign or malignant classes, while the three-class classifier includes a background class containing other findings and healthy tissue. A significant focus will be on explaining the models using the Grad-CAM interpretation method. Interpreting medical models is essential as standalone black-box CNN models lack credibility (we cannot determine the basis of their decisions).

II. Mammography data

For experimental purposes, mammography images were obtained from the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [5] and the Optimam database (OMI-DB) [6].

CBIS-DDSM is publicly available without the need for registration. The images were obtained by digitizing mammograms created using screen-film technology (indirect digital mammography). The DICOM format was used for image data, which is the current standard in medicine for storing and working with visual data obtained from various modalities (mammography, magnetic resonance, computed tomography, etc.). The database provides distributions for training and validation sets, which were applied in the study. In addition to the calcifications (macrocalcifications and microcalcifications), the database also includes mass findings. Binary segmentation masks are used to describe the position and size of the findings. An important aspect of the database is the presence of pathological results associated with the findings (information about malignancy or benignity).

OMI-DB is a newer and more comprehensive database. It is available to groups affiliated with an organization (commercial, non-profit, or academic) upon a submitted scientific project evaluated by database experts. Access is subsequently granted to only a subset of the database depending on an individual agreement. Data is collected from several institutions

across the United Kingdom, however, they do not indicate whether new cases have been collected since 2021. The images come from direct digital mammography known as Full-Field Digital Mammography in DICOM format. This is a newer technology approved for screening, generally capable of capturing smaller details in the breast more sharply. The database contains individual as well as combinations of abnormalities such as clusters of microcalcifications, masses, architectural distortions, and focal asymmetries. Custom distributions were created for training and validation data. Only findings containing clusters of microcalcifications without combinations with other abnormalities were used. The position and size of the findings are marked using rectangular bounding boxes. Histopathology results from tissue biopsy are included.

In our previous study [7] we studied cross-database transferrability of the CNN models and we concluded that the combination of these databases, with images acquired through different technologies, not only improved accuracy of the CNN models but also enhanced the interpretation results of the models.

A. Data preprocessing

For the main study, two datasets consisting of patches extracted from mammograms were created from the databases. The first dataset focused on binary classification of patches containing clusters of microcalcifications into malignant or benign classes. The second dataset included a third class (serving as background), consisting of healthy tissue, tissue containing masses, and tissue containing macrocalcifications.

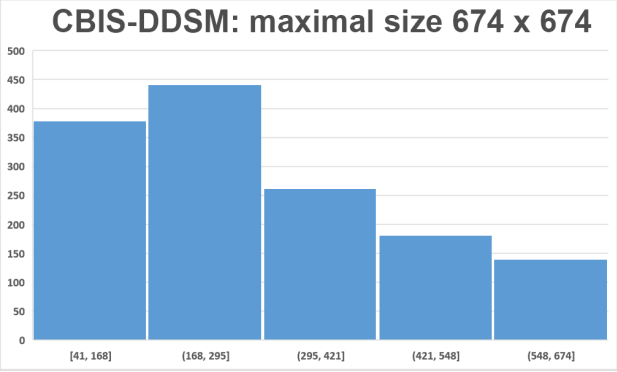

Fig. 1. Size distribution of calcification findings, not higher than 674x674 pixels, in the CBIS-DDSM.

The patches had a size of 674×674 pixels, with larger size findings not being used. For smaller sizes findings (Fig. 1 and 2), the surrounding area from the mammogram was added to keep the size of a patch 674×674. When possible, the findings were centered within the patch frame. For findings located at the edges of the mammogram, the patch frame was shifted towards the center. Before patch extraction, the mammograms were normalized to values ranging from 0 to 1. Models trained on patches can later be transformed and fine-tuned to classify the entire image, for example, using the end-to-end approach published in [8].

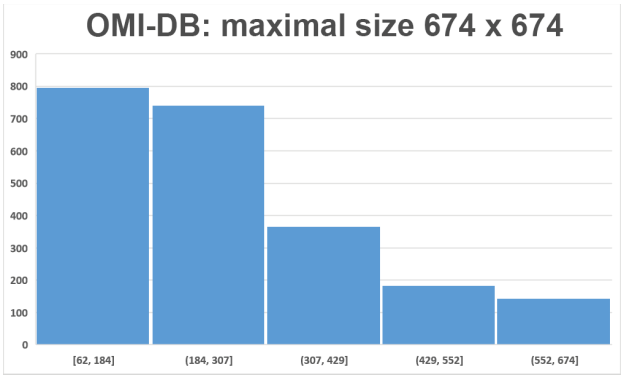

Fig. 2. Size distribution of calcification findings, not higher than 674x674 pixels, in the OMI-DB.

The CBIS-DDSM database contained several masks with varying sizes compared to the corresponding mammogram. Such masks were scaled to match the size of the mammogram. Approximately 30 subsequent adjustments were made in the dataset. If a mammogram contained multiple masks close to each other, they were unified. Minor shifts were made on some masks if the segmentation was adjacent to a finding. Some findings were removed if no calcifications could be localized using the mask.

The OMI-DB database contained several inverted images (a black background is expected from the mammogram; however, radiologists sometimes use inverted images for better visibility of certain observed abnormalities), which were corrected using inversion. Images with unexpected gray backgrounds associated with poor image quality and sharpness were discovered and retained in the dataset.

The purpose of the third class was to exclude any clusters of suspicious microcalcifications. Patches of healthy tissue were sourced from the OMI-DB database, representing patients with no malignant or benign histopathological records across all examinations. One patch was generated from each image, with at least 70% of the patch covering breast tissue. Patches with mass findings (malignant and benign) were sourced from both databases (with the same filter applied as with microcalcifications, focusing solely on mass findings without any combination with other abnormalities). Subsequently, 14,007 patches were manually checked, removing those containing clusters of microcalcifications. Patches with individual microcalcifications (which did not form a cluster) were retained. Additionally, macrocalcifications from the CBIS-DDSM database were added to the class (OMI-DB did not contain such findings), as they are benign abnormalities that are relatively easy to classify correctly.

After data preprocessing, the first dataset comprised 3350 samples (2691 training and 659 testing). The second dataset, including the background class, contained an additional 12994 samples (total of 16344 – 13073 training and 3271 testing), with a more detailed data distribution as shown in Table II-A.

TABLE I. DISTRIBUTION OF TRAINING AND VALIDATION DATA IN THE CREATED DATASETS OF CBIS-DDSM AND OMI-DB DATABASES

	Train		Test	
	CBIS-DDSM	OMI-DB	CBIS-DDSM	OMI-DB
Benign	603	457	138	116
Malign	309	1322	73	332
Background	1280	9102	337	2275
Sum	2192	10881	548	2723

III. PRELIMINARY EXPERIMENTS

Two preliminary experiments were initially conducted, and the insights gained from these experiments were utilized in the main experiment. The first experiment focused on testing various architectures of convolutional neural networks, while the second experiment aimed to reduce the size of input patches. For these experiments, the binary dataset was used with a slight modification, including macrocalcifications in the benign class.

A. Convolutional Architectures

The architecture of a neural network is among the most crucial elements of the model. Along with the correct setting of the learning rate and high-quality training data, it can significantly influence the overall accuracy of the model. Currently, several research groups are engaged in the development of convolutional architectures, with their performance commonly evaluated on the ImageNet dataset. The most frequently used version is ImageNet-1k, containing a total of 1000 output classes (different categories such as animals, objects, plants, etc.).

The advantage of architectures trained on the ImageNet dataset is the possibility of using transfer learning. In transfer learning, the model's weights can be transferred (instead of freshly initialized weights) and fine-tuned on a different domain. These transferred weights facilitate faster training because the model does not have to learn common features from scratch. In addition to faster training, potentially better results can also be achieved.

The experiment will explore several older and newer well-known architectures such as VGG, Inception-V3, ResNet, DenseNet, and EfficientNet.

1) Description of the architectures:

- Year 2014 – VGG (Visual Geometry Group) [9]: One of the first deeper architectures (the deepest version being 19 layers), introduced the idea of reducing the number of trainable parameters by using multiple 3x3 filters (kernels). For example, a 5x5 filter (25 parameters) can be replaced by two 3x3 filters (18 parameters), while the output feature map will have the same size. A major advantage of this architecture is its simplicity; it does not use any advanced methods and is therefore suitable for implementing various techniques (e.g., interpretation methods).
- Year 2016 – Inception-V3 [10]: Successor to the GoogLeNet architecture (Inception-V1). The main idea is to use Inception modules, which contain different-sized filters next to each other in a single layer (in parallel). It also uses 1x1 filters to reduce the number of feature maps.
- Year 2016 – ResNet (Residual Network) [11]: Created a breakthrough by introducing skip connections, allowing the creation of very deep networks (the deepest version being 152 layers) at the cost of increased model complexity. It consists of residual blocks, and the principle involves sending the identity (via skip connection) to a lower layer. The introduction of skip connections helped address the vanishing/exploding gradient problem [12].
- Year 2017 - DenseNet (Densely Connected Convolutional Network) [13]: Introduced dense connectivity, where each layer receives the identity from all preceding layers in a single dense block (in the ResNet architecture, the identity was only sent to the nearest lower layer).
- Year 2019 – EfficientNet [14]: Previous architectures typically expanded in only one way (e.g., in depth). This architecture introduced the compound scaling method, which simultaneously adjusts the width, depth, and resolution of the model using a compound coefficient.

In addition to the year of creation, the architectures are also ranked according to their success on the ImageNet dataset. Generally, newer architectures achieve better results compared to older ones, partly due to inspiration gained from the methods of their predecessors.

2) Exploration of a suitable architecture: Most architectures consist of multiple versions (usually identified by a number), differing in the number of trainable parameters. Implementations provided in the PyTorch [15] framework were used for architectures and their pretrained weights. The following learning rate values were tested: 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7 using the Adam optimizer. Early stopping was also applied. All versions that could fit into the GPU memory (RTX 4080 16GB) with a mini-batch size of 8 were tested.

- VGG: all versions (11, 13, 16, 19) + versions with batch normalization
- Inception-V3: only one version
- ResNet: all versions (18, 34, 50, 101, 152)
- DenseNet: only version 121
- EfficientNet: versions B0 to B2

TABLE II. TOP 10 MODELS - ALL ARCHITECTURES

Model	LR	Val. Acc.	Train Acc.
ResNet101	1e-6	77,2%	81,3%
DenseNet-121	1e-6	76,5%	85,5%
DenseNet-121	1e-5	76,3%	83,1%
EfficientNet-B1	1e-5	75,9%	78,7%
ResNet50	1e-6	75,9%	79,9%
ResNet152	1e-6	75,7%	90,9%
VGG-16-BN	1e-6	75,7%	83,4%
VGG-19-BN	1e-6	75,7%	81,5%
VGG-13	1e-6	75,6%	83,4%
VGG-19-BN	1e-5	75,6%	83,7%

The top 10 models are displayed in Table II. Unexpectedly, architectures performed very comparably, with the best architecture being ResNet101 with an accuracy of 77,2%. It is worth noting that due to the small amount of data, the resulting accuracy may vary between two identical runs by approximately +-1,5%. It was found that the most suitable learning rate across architectures was 1e-6. All architectures except Inception-V3 made it to the top 10, with Inception-V3 achieving 12th position with an accuracy of 75,4% and a learning rate of 1e-4.

B. Reducing the size of patches

Using such high size as 674×674 pixels for input images is not standard in convolutional neural networks. The disadvantage is a significant slowdown in model training and significantly greater GPU memory requirements. Our hypothesis is that reducing already small microcalcifications could result in the loss of essential details in the image, which would affect the model's accuracy. The most commonly used size is 224×224 pixels, which was typical for architectures pretrained on the ImageNet dataset.

TABLE III. VALIDATION ACCURACY COMPARISON OF PATCHES WITH DIFFERENT SIZE

	Size 674x674		Size 224x224	
LR	Avg. Acc.	Best Acc.	Avg. Acc.	Best Acc.
1e-6	75,4%	76,0%	-2,79%	-3,03%
1e-5	74,9%	75,4%	-2,00%	-1,72%
1e-3	72,5%	72,8%	-0,97%	-0,82%
1e-4	72,3%	72,8%	-0,28%	-0,41%

For better statistical sampling, training was run three times for the same hyperparameter settings. The ResNet50 architecture was used. The results can be observed in Table III. For the best learning rate of 1e-6, there was a decrease in accuracy on downsized patches by up to 2,79%, with an average decrease of 1,51% across different learning rates.

A similar experiment was conducted in the past, with the difference being that only the CBIS-DDSM database was used. In this case, there was an even more significant decrease in accuracy on downsized patches, with an average decrease of 6,0% in accuracy across different learning rates. This more pronounced decrease could be explained in two ways. First, there was a smaller amount of training/testing data in the dataset. Second, there was poorer image sharpness capturing finer details due to the older screen-film technology.

It was confirmed that downsizing patches is not suitable as it leads to the loss of crucial details. Other studies, such as [16], also addressed the reasons for not downsizing mammography images.

IV. COMPARISON OF BINARY AND THREE-CLASS CLASSIFIER

The study for hyperparameter tuning leveraged insights gained from previous experiments. The main goal of comparing the binary and three-class classifiers is to determine the limitations and advantages of each approach. Emphasis will be placed on explaining the models using the interpretational method Grad-CAM [17]. It will be necessary to confirm that the models indeed make decisions based on important features, in our case, confirming that the models decide on the benign and malignant classes based on clusters of microcalcifications. In the medical field, the use of these methods is necessary, as the model itself functioning as a black box is not sufficiently credible and transparent.

The decision to remove/relocate macrocalcifications to the background class was driven by two ideas. Macrocalcifications are benign abnormalities that are very easy to visually classify (for radiologists as well as models). Compared to microcalcifications, they are too distinct (especially in size), and therefore, it does not make sense to categorize them in the same class. Their presence is unique to the CBIS-DDSM database, as other databases do not consider this abnormality significant. The second reason was identified when using the Grad-CAM method on models from convolutional architecture experiments. It was found that in the case of a malignant finding, the presence of macrocalcifications could lead the model to classify it as benign. This situation occurred in some patches due to preprocessing, where a finding of smaller size than 674×674 pixels was supplemented with surrounding areas that could contain macrocalcifications. Removing macrocalcifications from the benign class will lead to a partial decrease in accuracy since macrocalcifications are easy to classify and their presence in the test set increases accuracy. On the other hand, the resulting accuracy will be more credible.

A. Binary Classification

Due to the unbalanced classes, the following weights were used for loss computation:

- Benign class: 0,606
- Malignant class: 0,394

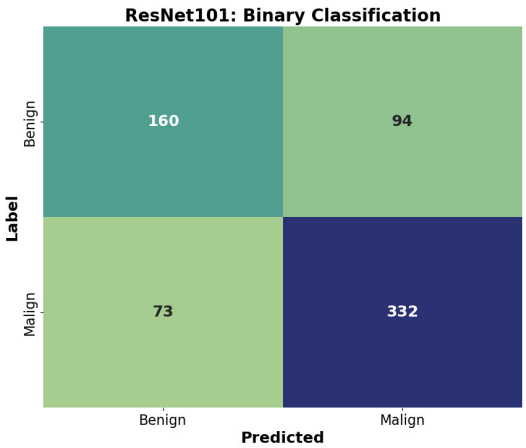

Fig. 3. Confusion matrix for the binary classification.

The best model achieved an accuracy of 74,7%, with an average accuracy of 74,3% across three runs. As expected,

there was a slight decrease in accuracy after removing macrocalcifications (compared to experiments with architectures). The confusion matrix of the best model is displayed in Fig. 3, with the following additional metrics: sensitivity – 0,820, specificity – 0,630, MCC – 0,458.

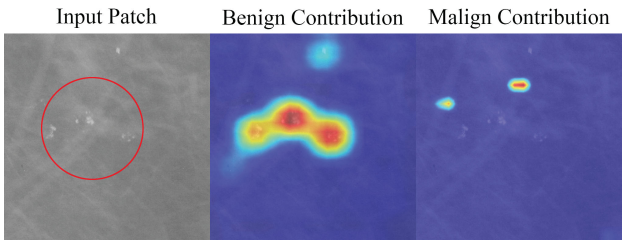

Fig. 4. Correct prediction of benign cluster with correct Grad-CAM high-lighting

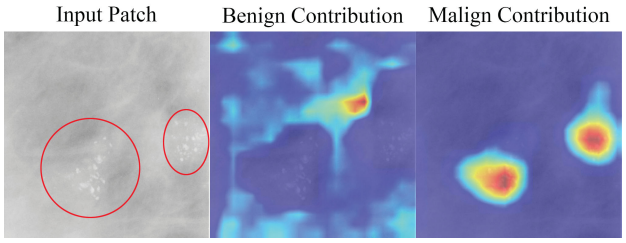

Fig. 5. Correct prediction of malign cluster with correct Grad-CAM high-lighting

Using Grad-CAM interpretation, it's possible to highlight areas on the input image that contributed most to the model's decision for a particular class. Examples of correct predictions for both classes are shown in Fig. 4 and 5. In both cases, the finding was correctly identified and placed in the correct class. Similar behavior was observed for most correct predictions, confirming that the model can make decisions based on important areas in the tissue with microcalcifications. Areas in the incorrect class can be disregarded due to their weak contribution. The predicted class is an important indicator, as its contribution to the area influenced the final classification decision.

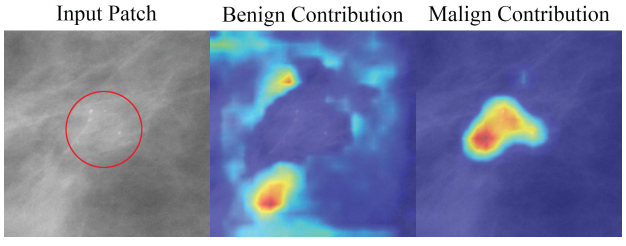

Fig. 6. Correct prediction of benign patch with incorrect Grad-CAM high-lighting the surrounding area of a finding and not the finding itself

However, for some benign patches, there were cases where the interpretation was not intuitive. An example is shown in Fig. 6, where a correct benign prediction was made, but the model did not base its decision on the area with microcalcifications but on the surrounding area. Conversely, the malignant class managed to detect the correct cluster area.

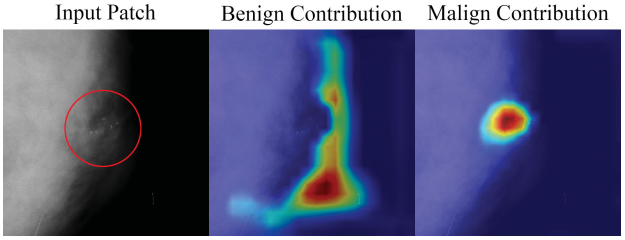

Fig. 7. Incorrect prediction of malign patch in dependence on the black background of the mammogram

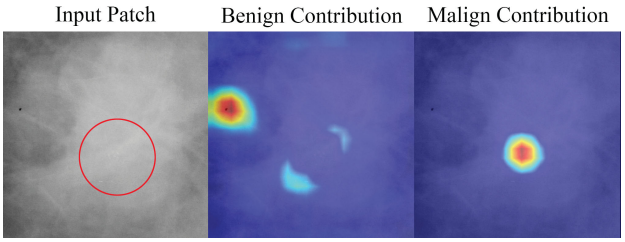

Fig. 8. Incorrect prediction of malign patch in dependence on the small artifact

Several limitations were discovered with incorrect predictions. Fig. 7 illustrates a case where a malignant cluster was identified, but the patch was placed in the wrong class based on the black background of the mammogram without abnormalities. Fig. 8 shows a black artifact that caused similar behavior.

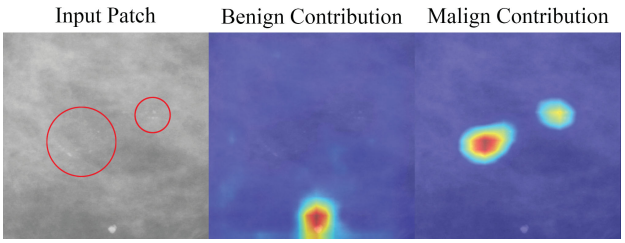

Fig. 9. Incorrect prediction of malign patch in dependence on the uninteresting benign abnormality

Despite removing macrocalcifications from the binary set, the problem of a random macrocalcification biasing the classification into the benign class was not completely eliminated (Fig. 9).

Breast implants also caused classification problems because their strong white color made it difficult for the model to determine where to look. White color is associated with abnormalities, and larger white areas may indicate the presence of masses. Fig. 10 shows a case with very fine malignant calcifications that the implant prevented from detecting.

The aforementioned incorrect predictions share one common characteristic: they were all mispredictions into the

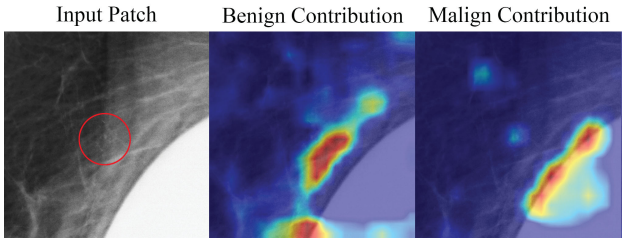

Fig. 10. Incorrect prediction of malign patch, the model failed to correctly identify the correct area due to the breast implants present

benign class. From this, it can be concluded that the model currently uses the benign class not only for predictions dependent on microcalcification clusters but also for other factors. This behavior was the main inspiration for creating the three-class classifier.

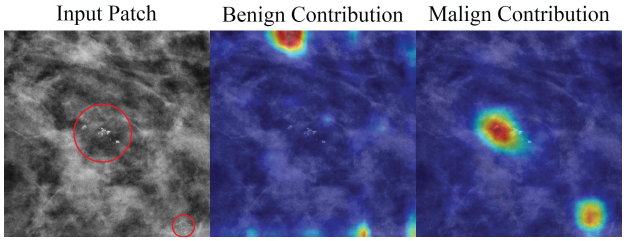

Fig. 11. Incorrect prediction of benign patch, the correct region was found only by the malignant class

The last expected type of incorrect predictions were patches where a suspicious cluster was detected and placed in the wrong class (Fig. 11). After consulting with specialists, it was confirmed that most of these cases would indeed pose a significant challenge for correct classification. In some cases, correct visual classification may not even be possible, and a tissue biopsy would have to be performed. This limitation could be partially addressed by having a larger amount of quality data, which is currently lacking from accessible sources.

B. Three-class classification

The goal of adding a third class was to help the model make decisions about benign and malignant classes based solely on clusters of microcalcifications. The benign class should no longer function multifunctionally, as observed in binary classification. Such behavior would significantly improve the model's credibility.

Similar to the previous case, different weights were used for the classes:

- Benign class: 0,571
- Malignant class: 0,371
- Background class: 0,058

The prediction results on the validation set are displayed in Fig. 12. The metric values were as follows: accuracy – 91,7%, macro-recall - 0,780, macro-precision - 0,754, MCC - 0,767. There was a slight decrease in accuracy for significant

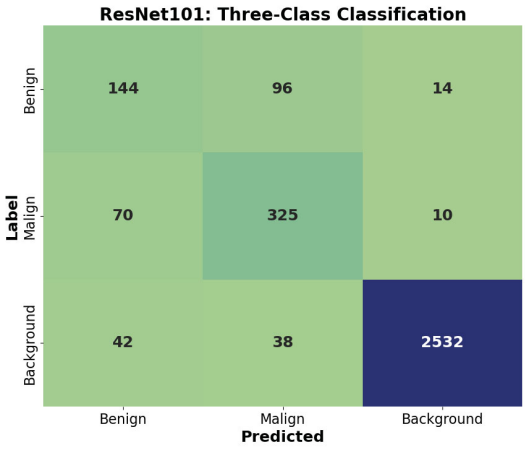

Fig. 12. Confusion matrix for the three-class classification

malignant and benign predictions. The model achieved high accuracy in deciding the background class.

Predictions of the benign class into the background, and vice versa, cannot be considered entirely wrong. Most of these interchangeable predictions contained isolated or small clusters of microcalcifications (more typical for the CBIS-DDSM database). Isolated calcifications are a common abnormality found in almost every breast, and manually filtering them out would be challenging, so they were left in the background class. This fact was not taken into account in the metric calculation.

The Grad-CAM method revealed a significant improvement in interpretation results. The goal was achieved where the model decides on both the benign and malignant classes based on the same cluster, as observed in Fig. 13. Moreover, in many cases, noise reduction was achieved, where the model attributed significance even to areas without abnormalities.

Many issues from binary classification were addressed:

- Non-intuitive classification based on the surrounding area visible in Fig. 6 for binary classification was eliminated in three-class classification, see Fig. 14.
- The issue with black background affecting predictions from Fig. 7 was diminished in Fig. 15.
- Small black artifact, depicted in Fig. 8, no longer pose a problem, as evidenced by the improvements shown in Fig. 16.
- The significance of uninteresting abnormalities was reduced as may be seen in Fig. 17 when compared with binary classification in Fig. 9.
- Fig. 18 illustrates how breast implants no longer hinder the localization of suspicious clusters, as it did in Fig. 10.
- Even in cases where the model made incorrect predictions, as in Fig. 11, now it decided based on a suspicious cluster/clusters in both classes visible in Fig. 19.

However, the three-class classification also revealed new challenges. The model struggled with malignant findings (cat-

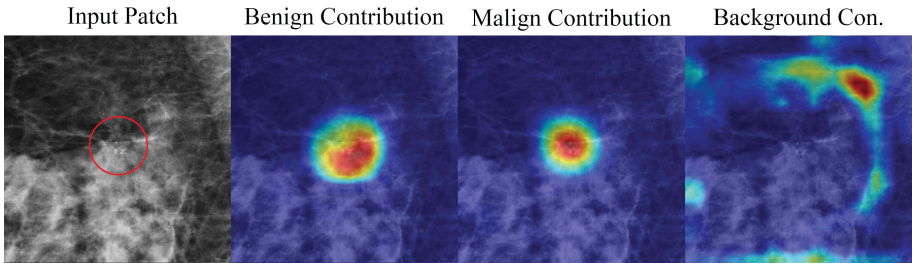

Fig. 13. Correct prediction of malign cluster, both malignant and benign classes correctly located the suspicious cluster

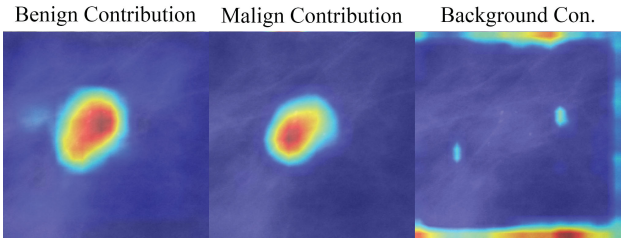

Fig. 14. Improving the classifier, the model no longer makes decisions based on the surroundings in the benign class

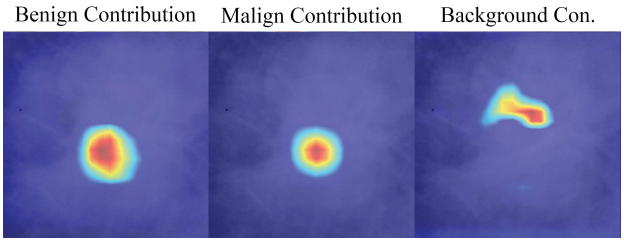

Fig. 16. A small black artifact no longer affects the classification. The malignant cluster has been classified as benign

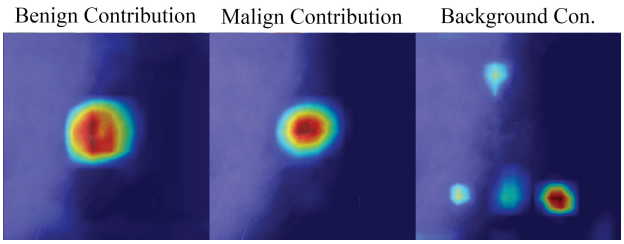

Fig. 15. The black area in the patch no longer influences the decision in the benign class. The model, however, still made an incorrect prediction of the malignant cluster into the benign class

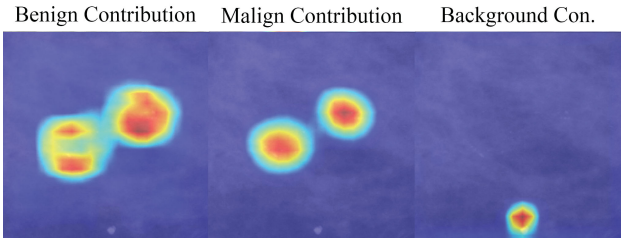

Fig. 17. The benign abnormality was localized in the background class. A correct prediction was made

egorized as background) in images with high fibroglandular tissue density and images with poorer sharpness (or their combination), where detecting calcifications was also very challenging for the human eye. The significant advantage was that the model managed to detect suspicious calcifications even in such images, but the resulting prediction was incorrect.

Another limitation was revealed in incorrect predictions of the background class into the malignant class. It turned out that the model had difficulty correctly classifying a specific type of calcifications found in vessels, known as vascular calcifications. These calcifications are straightforward for specialists to classify, but their properties may resemble malignant clusters (Fig. 20). Due to easy human classification, these findings are not included in any databases as benign findings, as they do not undergo biopsy. This limitation also highlights the situation where models trained on these databases are biased toward data containing suspicious clusters (as all findings underwent biopsy – specialists were not convinced of their benign nature). Therefore, models did not encounter clusters of calcifications during training where radiologists are quite sure of the benign nature of the finding. Using such findings in models could lead to unexpected results.

It was also discovered that breast implants still pose problems (despite resolving the issue in Fig. 18), in cases where no cluster of calcifications is present in the patch. Due to implants, patches belonging to the background were classified as malignant predictions. Interpretation revealed that some clusters were not successfully removed in manual preprocessing.

V. Conclusion

The first part of the work focused on finding the most suitable convolutional architecture. The architecture that achieved the best results was ResNet101, although other architectures showed very comparable results. It was also revealed that across different architectures, the best results were achieved with a learning rate of 1e-6.

The second part aimed to reduce the size of patches from 674×674 pixels to the standard size of 224×224 pixels. Experiments demonstrated that reducing size led to a decrease in model accuracy, which could be attributed to the loss of significant details in already small microcalcifications.

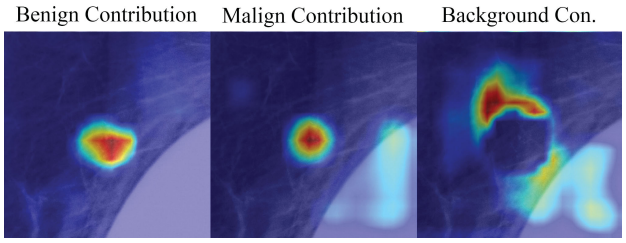

Fig. 18. The model was able to localize visually very challenging microcalcifications, the prediction was correct, and the implant no longer affected the classification significantly

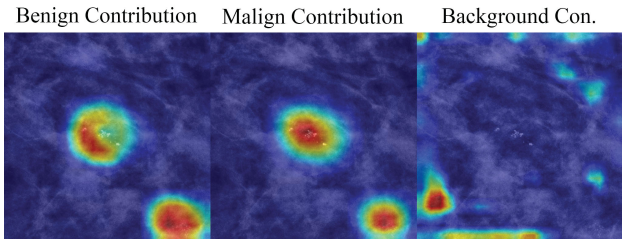

Fig. 19. Both classes correctly localized both clusters of microcalcifications. The resulting prediction was unchanged

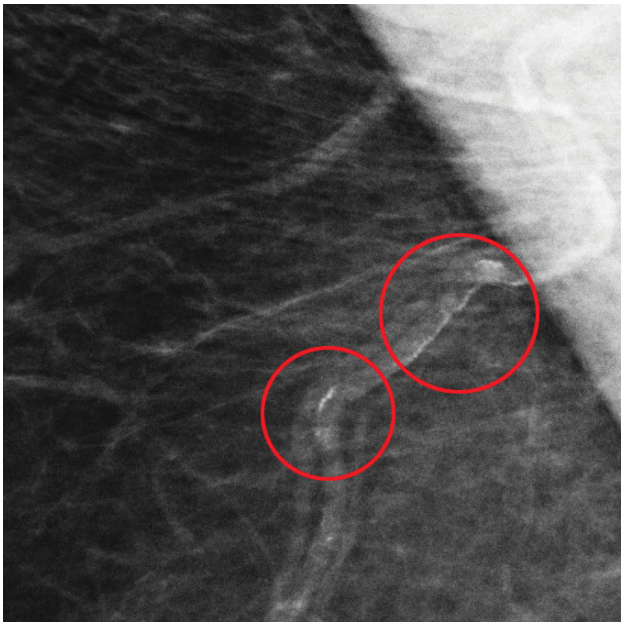

Fig. 20. Example of vascular calcifications

The final part utilized these findings and focused on comparing binary and three-class classifiers. The best binary classifier achieved an accuracy of 74,7% (MCC - 0,458). Interpretation confirmed that the model indeed makes decisions based on important features. However, limitations were also discovered, where the model made non-intuitive decisions based on the area around the finding, problems caused by breast implants, and introducing the model into incorrect predictions based on other abnormalities/artifacts. It was possible to define that the benign class decides on benignity not only based on benign clusters but also other features.

This led to the creation of a three-class classifier, which managed to eliminate the described limitations and guide the model to make decisions based on the same cluster of microcalcifications for both malignant and benign classes. The advantage of this model was that it correctly identified even hard-to-see calcifications, and interpretation demonstrated noise reduction. The best model achieved an accuracy of 91,7%, but the more appropriate metric this time was MCC – 0,767 due to a large number of patches belonging to the background. When comparing important predictions for malignant and benign classes, a slight deterioration in results was observed. However, adding a new class also revealed new limitations. It was found that the model had difficulty with vascular calcifications (included in the background class – as they are not very interesting findings for radiologists), which were incorrectly classified as malignant. Problems were also caused by patches without abnormalities with breast implants. Generally, the use of three classes significantly improved interpretational results and thereby the credibility of the model. However, creating a third class is not straightforward because when creating patches from a random location in the breast, it is possible to include unwanted abnormalities (in our case, clusters of microcalcifications) in the patch.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the contract No. VEGA 1/0525/23.

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] 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.

[3] 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

[4] 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.

[5] 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.

[6] 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

[7] 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.

[8] L. Shen, L. Margolies, J. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," Scientific Reports, vol. 9, pp. 1–12, 08 2019.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.
[11] 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
[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Journal of Machine Learning Research - Proceedings Track, vol. 9, pp. 249–256, 01 2010.
[13] 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
[14] 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
[15] 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
[16] 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
[17] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.