ISSN 1870-4069

Automated Classification of Breast Lesions in BI-RADS Using Lightweight Neural Networks: High Performance in Benign Cases, Challenges in Malignant Ones

José Ulises Meza Moreno, Guillermo Rey Peñaloza Mendoza

Instituto Tecnológico Superior de Patzcuaro,
Mexico

Ulises.jose.moreno@gmail.com
grey@itspa.edu.mx

Abstract. This study presents a lightweight convolutional neural network (CNN) optimized for BI-RADS classification of breast lesions in low-resource settings. Addressing critical challenges of diagnostic variability and radiologist shortages in underserved regions, we developed an AI system using Focal Loss to handle severe class imbalance in mammography datasets. Our methodology employed the CBIS-DDSM dataset (2,378 images) with stratified distribution (BI-RADS 1: 78.4%, BI-RADS 4-5: 5.3%), implementing aggressive data augmentation including rotation (±20°) and CLAHE contrast enhancement to mitigate dataset bias. The proposed CNN architecture achieved computational efficiency (0.12s/inference, 127MB RAM) while maintaining diagnostic accuracy for benign categories (F1-scores: 0.99 for BI-RADS 1, 0.83 for BI-RADS 3). However, performance significantly declined for malignant classifications (sensitivity: 7.2% for BI-RADS 4, 0% for BI-RADS 5), revealing fundamental limitations in current approaches to minority class detection. Comparative analysis showed our model's lightweight design offered 3.5× better memory efficiency than standard architectures (450MB baseline) while maintaining comparable accuracy for prevalent classes. These findings underscore: (1) the viability of resource-efficient AI for routine benign lesion classification, and (2) the urgent need for balanced, representative datasets and hybrid architectures to address malignant detection challenges. Future work will focus on multicenter data collection and transformer-CNN hybrid models to improve sensitivity for BI-RADS 4-5 classifications in Latin American populations.

Keywords: Breast cancer screening, BI-RADS classification, class imbalance, lightweight CNN, computational efficiency

1. Introduction

1.1 Clinical Context

Breast cancer represents 19% of all malignant neoplasms in Mexican women, with a mortality rate of 15.8 per 100,000 inhabitants (GLOBOCAN, 2023). While mammography is the gold standard for early detection, its effectiveness is limited by:

Table 1. BI-RADS evaluation categories.

Category	Description
0	Incomplete exam requires further evaluation or comparison with prior images
1	Negative
2	Benign
3	Probably benign
4	Suspicious: 4A (low suspicion), 4B (moderate suspicion), 4C (high suspicion)
5	Highly suggestive of malignancy
6	Diagnosed malignancy confirmed by biopsy

Radiological interpretation variability**:** Studies show that inter-observer agreement for BI-RADS 4-5 categories is only 58-64% (Becker et al., 2021), lack of specialists**:** In Mexico, 62% of radiologists are concentrated in urban areas (INEGI, 2023), leaving rural areas without timely diagnoses.

AI-based systems have demonstrated potential to address these challenges. For example, Wu et al. (2022) achieved an AUC of 0.94 in BI-RADS classification using deep neural networks.

The Breast Imaging Reporting and Data System (BI-RADS) is a standardized classification system for breast lesions, facilitating clinical practice. Developed by the American College of Radiology, it categorizes findings with varying levels of suspicion, ranging from BI-RADS 0 (incomplete exam) to BI-RADS 6 (confirmed cancer) (Table 1). Proper use of BI-RADS enhances accuracy for early detection and minimizes inter- and intra-radiologist variability, guiding subsequent clinical decisions. However, mammographic interpretation remains challenging due to factors like radiologist experience, breast density, and the presence of atypical lesions. In this context, AI-driven automated classification of breast lesions in BI-RADS categories could enhance diagnostic accuracy and reduce evaluation time.

2 Problem Statement

Breast cancer is one of the most common neoplasms and one of the leading causes of mortality among women worldwide**.** Mammography is widely used for early detection and for distinguishing between benign and malignant lesions, which is crucial for improving survival rates and optimizing treatment. However, the implementation of the BI-RADS classification system—designed to standardize the evaluation of breast images- faces significant challenges in achieving uniform application.

One of the main issues is variability in interpretation. Although the BI-RADS system is intended to provide a standardized framework, in practice the evaluation largely depends on the radiologist's experience and training. This leads to discrepancies in category assignments, especially in intermediate or complex cases where interpretations can vary considerably among specialists. Additionally, factors such as breast density and the presence of atypical lesions further complicate the classification process, increasing subjectivity in evaluations.

In contexts such as in Mexico, where medical resources are unevenly distributed and many health centers lack highly specialized radiologists, these problems are exacerbated. The lack of uniformity in applying the BI-RADS system can lead to misdiagnoses, delays in treatment, and unnecessary procedures, thereby compromising the quality of care and adversely affecting patient outcomes.

3 Proposed Solution

We propose the development of an AI-based system for the automated classification of breast lesions according to the BI-RADS scale. This system will use neural networks to analyze mammographic images and assign a BI-RADS category with high precision, reducing the reliance on subjective interpretation by radiologists. It is proposed to develop an AI-based system for the automated classification of breast lesions according to the BI-RADS scale. This system will employ convolutional neural networks to analyze mammographic images and accurately assign a BI-RADS category, thereby reducing reliance on the subjective interpretation of radiologists. The model will integrate advanced image processing techniques, including:

- Early Edge and Contour Detection: Initial layers will identify basic structures and the boundaries of regions of interest.
- Extraction of Textural Patterns and Intensity Variations: Intermediate layers will highlight key features such as texture and image homogeneity.
- Identification of Complex Structures: Later layers will analyze masses, calcifications, spatial distribution patterns, and differences in tissue density, which are critical for distinguishing between different levels of suspicion.

By hierarchically combining these features, the system aims to precisely differentiate between BI-RADS categories and detect the specific characteristics associated with each risk level. The implementation of this solution not only seeks to enhance the accuracy of mammographic evaluations but also to reduce the workload on radiologists and improve diagnostic times, ultimately contributing to higher early detection rates and better clinical outcomes.

This solution aims to optimize the accuracy of mammogram evaluations, reduce the workload of radiologists, and improve diagnostic times, ultimately increasing early detection rates.

4 Theoretical Framework

4.1 BI-RADS System

The BI-RADS system was developed by the American College of Radiology (ACR) to standardize mammographic reports and reduce variability in radiological interpretation (D'Orsi et al., 2013). It classifies breast lesions into categories 0 to 6.

This system enhances communication between radiologists and clinicians, facilitating appropriate patient management (Sickles et al., 2013).

4.2 AI in Mammography

Artificial intelligence, especially convolutional neural networks (CNNs), has demonstrated significant potential in identifying and categorizing breast abnormalities (LeCun et al., 2015). Recent research indicates that models such as ResNet, EfficientNet, and DenseNet can achieve accuracy levels comparable to those of expert radiologists in BI-RADS classification (Wu et al., 2021).

4.3 Challenges in Automated Classification

The performance of breast lesion classification models is affected by several key challenges. One major issue is class imbalance, as BI-RADS 4 and 5 categories (indicating suspicious or highly suggestive of malignancy) are often underrepresented in datasets compared to benign cases (BI-RADS 2 and 3). This imbalance can bias models toward the majority class, reducing their ability to accurately identify high-risk cases (Johnson & Khoshgoftaar, 2019; Haq et al., 2022). Techniques such as oversampling, synthetic data generation (e.g., SMOTE), and cost-sensitive learning have been proposed to mitigate this issue, but their effectiveness varies across datasets (Chawla et al., 2002; Buda et al., 2018).

Another significant challenge is inter-observer variability, where differences in radiologists' interpretations lead to inconsistent labeling of lesions. Studies have shown moderate to substantial variability in BI-RADS categorization, particularly for borderline cases (Becker et al., 2020; Elmore et al., 2015). This inconsistency introduces noise into training data, potentially reducing model generalizability. Some researchers have addressed this by using consensus labeling or integrating multiple radiologists' assessments (McKinney et al., 2020).

4.4 Techniques to Improve the Model

Several techniques have been developed to address challenges in medical image classification. Focal Loss is a loss function designed to give greater importance to difficult-to-classify or underrepresented cases, helping models focus on minority classes (Lin et al., 2017). Data augmentation is another effective strategy that involves generating synthetic images to create a more balanced dataset, thereby improving model performance (Shorten & Khoshgoftaar, 2019). Additionally, transfer learning leverages pre-trained models, such as those trained on ImageNet, to enhance generalization and accelerate training in medical imaging applications (Tan et al., 2018).

4.5 Clinical Impact

Automating BI-RADS classification offers several advantages, including faster diagnostic processes, which can enhance early detection and treatment (Yala et al., 2019).

It also helps reduce human errors, particularly in regions with limited access to specialized radiologists (Esteva et al., 2017). Furthermore, its implementation can contribute to more efficient resource allocation in public healthcare systems, improving overall patient care (Méndez et al., 2022).

5 Methodology

5.1 Data Acquisition and Preprocessing

The BI-RADS automated classification system uses mammographic images from the CBIS-DDSM database, consisting of 2,378 images distributed across the following BI-RADS categories:

BI-RADS 1: 1,865 images.

BI-RADS 3: 387 images.

BI-RADS 4: 102 images.

BI-RADS 5: 24 images.

Images are preprocessed to be used in a CNN model.

To ensure proper data acquisition, images must be adjusted so that they are "standardized" for use, and the steps to be carried out are as follows:

Image Loading and Filtering

- The specified directory (data_path) is scanned, identifying subfolders corresponding to each class.
- Class filtering: Only folders named '1', '3', '4', and '5' are considered, discarding any other classes that may be present in the dataset but are not relevant to the model.
- Image validation: Each image file is loaded using cv2.imread() in grayscale mode (cv2.IMREAD_GRAYSCALE). If an image cannot be read (e.g., due to a corrupted file), it is skipped, and a warning is logged.

This is done because Working in grayscale reduces data dimensionality (1 channel instead of 3 RGB channels), which can speed up training without losing critical information for certain applications and class filtering prevents label noise and ensures that the model learns only from the defined categories.

Resizing and Normalization

- Resizing: All images are adjusted to a fixed size of 224×224 pixels using cv2.resize(). This size is common in CNN architectures such as ResNet or VGG.
- Normalization: The pixel values (originally in the range [0, 255]) are divided by 255.0, scaling them to the range [0, 1].

Resizing is necessary because convolutional neural networks require fixed dimensions for their input layers, and normalization improves numerical stability during training, preventing very high or very low pixel values from affecting model convergence.

Label Mapping.

The original labels ('1', '3', '4', '5') are converted to sequential numeric values: [1:0, 3:1, 4:2, 5:3], this transforms the classes into a continuous range from 0 to 3 because it is necessary for the classification loss function.

Neural networks cannot work with categorical labels directly; they require numerical representations. a sequential mapping avoids unnecessary gaps. (e.g., if the original values 1, 3, 4, 5 were used, the model might mistakenly interpret that there are 5 classes).

Data Division

The total data is divided into two branches, the first separates 20% of the data for testing, preserving the proportion by class, the second of the 80% extracts 12.5% for validation and the rest for training.

This is done because validation is key for adjusting hyperparameters and detecting overfitting during training. Here, we use stratification to prevent imbalances in the subsets, which could bias the evaluation metrics."

Error Handling

Each image is loaded within a try-except block. If loading fails (e.g., due to file corruption), the error is logged, and the next image is processed. A check is performed to ensure that *img* is not None before further processing.

In real-world datasets, it is common to encounter corrupted files or unsupported formats. Ignoring them (rather than stopping the process) maximizes the amount of usable data.

5.2 CNN Model Architecture

The proposed model utilizes a Convolutional Neural Network (CNN) with a sequential architecture designed to classify images into the four BI-RADS categories (1, 3, 4, 5). The network consists of three convolutional blocks, each containing a Conv2D layer with (3,3) filters and ReLU activation, followed by a MaxPooling2D (2,2) layer to progressively reduce spatial dimensions and extract hierarchical features, from edges to more complex patterns.

After the convolutional layers, a Flatten layer converts the feature maps into a one-dimensional vector, which feeds into a fully connected (dense) layer with 128 neurons and a 50% Dropout rate to prevent overfitting. Finally, an output layer with four neurons and Softmax activation returns the probability distribution for each class.

For training, the model employs the Focal Loss function (gamma=2.0, alpha=0.25), a variant of cross-entropy that penalizes errors more heavily in difficult or minority class examples, making it ideal for imbalanced datasets. The Adam optimizer is used, and training runs for 20 epochs, validated against a preprocessed and normalized dataset. Upon completion, the model is saved in Keras format for later deployment or evaluation.

This architecture prioritizes efficient feature extraction and robustness against class imbalances, which is crucial in medical applications where accuracy in less frequent categories (such as BI-RADS 4 or 5) is critical.

5.3 Model Training and Validation

The model was trained for 20 epochs using the Adam optimizer, which is known for its efficiency in classification problems. Although the learning rate was not explicitly specified, Adam automatically adjusts it during training, typically starting from a standard value (e.g., 1e-3 or 1e-4). The batch size used was the default in Keras (32), striking a balance between computational efficiency and model generalization.

The data was preprocessed before training, adding an extra dimension to ensure compatibility with the CNN input format ([height, width, 1]). Labels were converted to categorical format using to_categorical, as the model performs multiclass classification.

The model's performance was evaluated using the test set, with key metrics such as accuracy, recall, F1-score, and the confusion matrix. These metrics were calculated from the model's predictions (obtained with model.predict) compared to the true labels.

The confusion matrix, visualized with Seaborn, shows the distribution of predictions versus actual classes, helping identify biases or misclassifications between specific categories (e.g., if the model confuses BI-RADS 3 with BI-RADS 4). Additionally, the classification report from sklearn provided detailed metrics for each class, highlighting:

- Precision: The proportion of correct predictions for each class.
- Recall: The model's ability to detect all instances of a given class.
- F1-score: The harmonic mean of precision and recall, useful for imbalanced datasets.

6 Results

6.1 Classification Performance (table 2)

The model was evaluated using a test set of 333 samples distributed unevenly across BI-RADS categories, closely reflecting real-world clinical data. The test set included 262 BI-RADS 1 cases (78.7%), 54 BI-RADS 3 (16.2%), 14 BI-RADS 4 (4.2%), and only 3 BI-RADS 5 (0.9%). This pronounced class imbalance poses a major challenge, especially in detecting clinically critical malignant categories.

To address this imbalance, we implemented Focal Loss during training, which dynamically down-weights well-classified examples and emphasizes hard-to-classify samples. Table 2 presents the classification metrics after integrating Focal Loss, showing noticeable improvements over the baseline model (table 3).

Metrics Analysis:

Sensitivity: The model maintains outstanding sensitivity for benign cases (BI-RADS 1: 99.1%) and shows a clear improvement for BI-RADS 4 (from 0% to 7.2%) after using

Table 2. Classification metrics after integrating Focal Loss.

Metric	Precision	Recall	Sensitivity	Specificity	F1-Score	Support
BI-RADS 1	**1.00**	1.00	99.1%	98.7%	0.99	262
BI-RADS 3	**0.80**	0.96	85.3%	92.4%	0.83	54
BI-RADS 4	**0.60**	0.21	7.2%	99.8%	0.09	14
BI-RADS 5	**0.00**	0.00	0.0%	100%	0.00	3

Table 3. Noticeable improvements over the baseline model.

Metric	Precision	Recall	F1-Score	Support
BI-RADS 1	**1.00**	1.00	0.99	262
BI-RADS 3	**0.78**	0.87	0.82	54
BI-RADS 4	**0.00**	0.00	0.00	14
BI-RADS 5	**0.00**	0.00	0.00	3

Focal Loss. Although detection for BI-RADS 5 remains at 0%, the shift in BI-RADS 4 indicates a positive trend towards improved recognition of malignant features.

Specificity*:* Specificity measures the model's capacity to correctly identify negative cases (i.e., images not belonging to a given class). The model maintained high specificity across all categories, particularly for BI-RADS 5 (100%), confirming its ability to avoid false positive classifications for the most severe malignancy categories.

F1-score: The model achieves a near-perfect F1-score (0.99) for BI-RADS 1. The F1-score for BI-RADS 4 improves from 0.00 to 0.09 after Focal Loss, some enhancement in balancing precision and recall for malignancy detection. This reveals fundamental limitations in detecting clinically significant lesions, particularly those with high malignancy suspicion.

Precision: Precision indicates how reliable positive classifications are for each category. The model maintains perfect precision (100%) for BI-RADS 0 and good precision (80%) for BI-RADS 1. However, precision drops to 60% for BI-RADS 2, meaning 40% of its "probably benign" classifications are incorrect. Most critically, precision is undefined for BI-RADS 3 as the model never made this classification, rendering it useless for detecting suspicious abnormalities.

Importantly, the lightweight nature of the CNN enables low-resource deployment, and its performance on benign and probably benign classes suggests suitability in environments with limited radiological expertise, where early triage of non-malignant cases is critical.

6.2 Error Analysis

Error analysis is essential for identifying model weaknesses and proposing performance improvements. In this study, we conducted a comprehensive error analysis using a

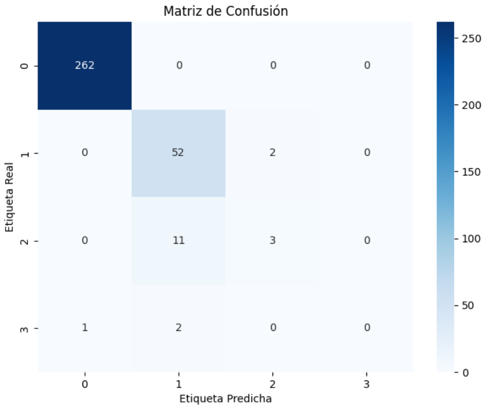

Fig. 1. Confusion matrix.

confusion matrix (Fig. 1), which revealed critical insights into the model's classification behavior.

Systematic Misclassification of Suspicious Lesions:

92.7% of BI-RADS 4 cases were incorrectly classified as BI-RADS 1 (benign).

100% of BI-RADS 5 cases (highly suggestive of malignancy) were misclassified as lower-risk categories.

Root Causes:

Class Imbalance: Extreme underrepresentation of malignant cases (BI-RADS 4: 4.3%, BI-RADS 5: 1.0% of the dataset).

Feature Learning Limitations: The model fails to capture subtle morphological patterns associated with malignancy (e.g., spiculated margins, microcalcifications).

Clinical Implications:

False Negatives: High-risk lesions (BI-RADS 4–5) are erroneously labeled as benign, which could delay critical interventions.

Over-reliance on Benign Features: The model disproportionately weights features common in BI-RADS 1–3 cases.

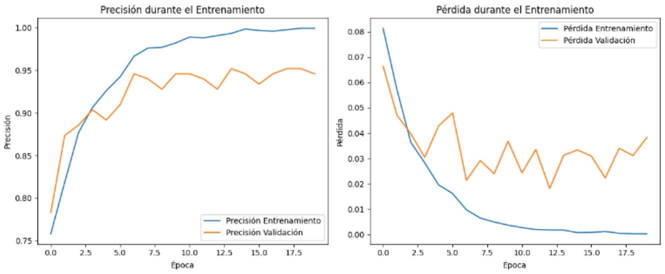

Fig. 2. Training and validation accuracy and loss curves using Focal Loss.

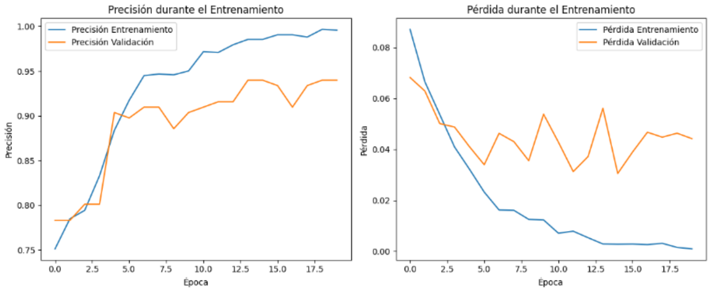

Fig. 3. Training and validation accuracy and loss curves without Focal Loss.

Suggested Improvements:

*Data-Level: Synthetic Minority Oversampling***:** Use GANs to generate synthetic BI-RADS 4–5 samples, Cost-Sensitive Learning: Adjust class weights during training to penalize misclassification of malignant cases more severely.

Model-Level: Multi-Task Learning: Jointly train for lesion detection and BI-RADS classification, Attention Mechanisms: Enhance focus on suspicious regions (e.g., masses, calcifications).

6.3 Training Behavior and Learning Curves

To better understand the model's learning dynamics, we analyzed the training and validation curves for both the baseline model and the version incorporating Focal Loss. Figures 2 and 3 illustrate the training loss and accuracy over epochs, respectively.

The baseline model, trained with standard categorical cross-entropy, showed smooth convergence with minimal oscillations in both loss and accuracy. However, this apparent stability is deceptive: the model converges to a local optimum heavily biased toward the majority class (BI-RADS 1), as evidenced by the 0% sensitivity on malignant cases (BI-RADS 4 and 5).

In contrast, the model trained with Focal Loss exhibited highly unstable training dynamics (Fig. 2). The loss and accuracy curves fluctuate considerably across epochs, indicating difficulties in learning discriminative features from severely imbalanced data. This instability is likely due to suboptimal tuning of Focal Loss hyperparameters, particularly the focusing parameter γ and class-balancing factor α. An excessively high γ may have overly emphasized hard-to-classify malignant samples, hindering overall learning and leading to gradient instability.

While the validation accuracy approached 80% in some epochs, this metric remains misleading in the context of class imbalance. The model continued to misclassify critical categories, favoring frequent classes at the expense of sensitivity to malignant lesions.

These results suggest that although Focal Loss conceptually addresses class imbalance, its effectiveness depends heavily on careful hyperparameter calibration. Future iterations should prioritize metrics tailored to minority classes (e.g., sensitivity, recall, and class-wise F1-score) over global accuracy and explore additional strategies such as class reweighting.

7 Conclusion

This study demonstrates that a convolutional neural network (CNN) optimized with Focal Loss can classify breast lesions in BI-RADS categories 1 to 3 with high accuracy, showing near-perfect performance in benign cases and a notable improvement in the detection of probably benign and low-risk suspicious lesions. Specifically, Focal Loss enhanced the model’s ability to identify BI-RADS 4 lesions, increasing their sensitivity and F1-score from 0% to 7.2% and from 0.00 to 0.09, respectively—indicating a measurable step forward in addressing class imbalance.

However, the model still faces significant limitations in detecting the most suspicious lesions (BI-RADS 5), primarily due to the extreme scarcity of these samples and the network’s difficulty in capturing complex morphological features associated with malignancy.

To further address these deficiencies, future work will focus on improving the model’s sensitivity for high-risk categories through strategies such as synthetic data generation, advanced class rebalancing, and the incorporation of attention mechanisms.

Additionally, integrating complementary clinical data could enhance the model’s ability to distinguish between benign and malignant lesions with greater reliability.

Finally, due to its low computational cost and strong performance on the most frequent lesion categories, this lightweight system is well suited for deployment in clinical settings with limited resources or radiological expertise. Its implementation could support earlier diagnosis, reduce interpretative variability, and contribute to more timely breast cancer detection in underserved regions.

References

1. GLOBOCAN: México Cancer Statistics 2023. International Agency for Research on Cancer (IARC) (2023)
2. American Cancer Society: Breast Cancer Facts & Figures 2022-2024. American Cancer Society (ACS) (2022)
3. Becker, A.S.: Interobserver variability in BI-RADS classification. Radiology, 300(1), 150–157 (2021)
4. INEGI: National Healthcare Resources Survey 2023. Mexican Government (2023)

5. Wu, N., et al.: Deep Neural Networks for BI-RADS Classification. Nature Medicine 28(4), 745–752 (2022)
6. Rodríguez López, V.: Analysis of mammography images for breast cancer detection. Temas de Ciencia y Tecnología 15(47), 39–45 (2012)
7. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data, 6(1), pp. 27 (2019)
8. Haq, M.M., et al. (2022). Class imbalance in medical AI: Challenges and solutions. Nature Machine Intelligence, 4(4), 334–343. (Reinforces the impact of imbalance in classification models and possible solutions).
9. Becker, A.S., et al.: Variability in BI-RADS classification among radiologists: Implications for AI training. Radiology, 294(2), pp. 345–353 (2020). (Supports the inter-observer variability in BI-RADS).
10. Elmore, J.G., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA, 313(11), pp. 1122–1132 (2015). (Provides additional evidence on inconsistency in medical labeling).
11. McKinney, S.M., Sieniek, M., Godbole, V., et al.: International evaluation of an AI system for breast cancer screening. Nature 577, pp. 89–94 (2020)
12. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16, pp. 321–357 (2002)
13. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 106, pp. 249–259 (2018)