CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 310-324
p-ISSN: 2086-0382; e-ISSN: 2477-3344

ResNet-50 and ResNeXt-50 for Multiclass Classification of Chronic Wound Images under Gaussian Blur

Reynaldi Ikbar Surya Andhika*, Sugiyarto Surono, and Aris Thobirin

Department of Mathematics, Faculty of Applied Science and Technology, Ahmad Dahlan University, Yogyakarta, Indonesia.

Abstract

Chronic wound image classification is important for supporting the assessment of conditions such as diabetic foot ulcers (DFU) and pressure ulcers (PU). While convolutional neural network (CNN)–based approaches have shown promising results, most previous studies focus on binary classification and rarely evaluate robustness in multiclass chronic wound scenarios. This study investigates multiclass classification of chronic wound images, distinguishing DFU, PU, and Normal Skin, using ResNet-50 and ResNeXt-50 architectures. A total of 2,146 publicly available images were stratified at the image level into training (70%), validation (15%), and test (15%) sets. Both models were trained under an identical configuration using data augmentation and class-weighted loss. On clean test images, ResNet-50 and ResNeXt-50 achieved strong and comparable performance, with accuracies of 0.9877 and 0.9938 and macro-averaged F1-scores of 0.9866 and 0.9928, respectively. Robustness was evaluated by applying Gaussian blur at the inference stage to simulate image defocus. Under stronger blur ($\sigma = 2.0$), ResNeXt-50 maintained higher performance (accuracy 0.9723, macro-F1 0.9679) than ResNet-50 (accuracy 0.9200, macro-F1 0.9123). These results highlight the contribution of this study in evaluating robustness to blur in multiclass chronic wound image classification, while emphasizing that robustness is limited to resistance against image blur or defocus.

Keywords: Chronic Wound; Diabetic Foot Ulcer; Medical Image Analysis; Pressure Ulcer; Wound Image Classification.

1. Introduction

Deep Learning (DL) is a subfield of Artificial Intelligence (AI) that deals with machine learning methods, specifically those that focus on modeling neural networks [1]. The development of DL in recent years has made it a leading approach in medical image processing due to its ability to extract complex visual representations automatically and accurately. Recent studies have shown that DL can improve the performance of image-based diagnosis, reduce the variability of assessments between clinicians, and accelerate the process of analyzing visual data in clinical contexts that require fast and consistent decision-making [2–4]. This advantage makes DL a key technology in image-based diagnostic aiding systems.

Among the various DL architectures, the Convolutional Neural Network (CNN) is the most commonly used architecture for analyzing medical images [5–7]. CNNs operate through

*Corresponding author. E-mail: andhikaikbar@gmail.com

Submitted: January 09, 2026 Reviewed: January 15, 2026 Accepted: February 08, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.40323

convolutional layers, nonlinear activation functions, pooling, and fully connected layers, enabling hierarchical feature learning without manual feature engineering [2, 8]. Transfer learning and fine-tuning approaches further reinforce the effectiveness of CNNs, particularly in cases where the availability of annotated data is relatively limited, such as in the chronic wound domain. The ResNet-50 architecture is one of the CNN models that is widely used due to its residual design [9, 10], which allows the network to be trained deeper without performance degradation [11, 12]. ResNeXt-50 extends the ResNet-50 architecture by introducing the concept of cardinality through grouped convolutions, enabling multiple residual transformation paths to be executed in parallel to increase feature representation capacity without significantly increasing model complexity [13]. Recent studies report that the ResNeXt-50 variant has been successfully applied to medical image classification tasks, including the classification of white blood cells and chronic wounds, with competitive performance compared to other CNN architectures [14, 15].

Chronic wounds such as Diabetic Foot Ulcer (DFU) and Pressure Ulcer (PU) remain significant global health problems. DFU affects millions of people with diabetes each year and is associated with an increased risk of infection, lower extremity amputation, and mortality. Epidemiological reports indicate a DFU prevalence of 6–13% and a lifetime risk of nearly 30% [16, 17]. Meanwhile, PU predominantly occurs in patients with limited mobility, with reported prevalence rates of 0.4–38% in hospitals and 2.2–39% in long-term care facilities [18, 19]. In clinical practice, the assessment of these wounds still relies heavily on visual inspection, which is subjective and often leads to inter-observer variability. Recent research also indicates that artificial intelligence-based wound analysis systems can improve assessment consistency, reduce clinical workload, and accelerate diagnostic workflows in chronic wound management [20, 21].

In the last five years, most CNN-based chronic wound classification studies have focused on binary classification scenarios, such as distinguishing ulcer versus non-ulcer images or DFU versus non-DFU cases, and have generally reported promising results [11, 22]. However, studies that explicitly address the joint classification of DFU, PU, and Normal Skin within a single CNN model remain limited. Beyond the scarcity of multiclass studies, additional challenges arise from differences in dataset characteristics, including variations in image sources, ulcer appearance, and class imbalance, which can hinder consistent performance on more diverse test data, particularly Normal Skin images obtained outside the training distribution. Moreover, aspects of model robustness to image-quality degradation, such as motion blur, camera defocus, or acquisition device limitations, are still rarely investigated in the context of chronic wound classification. These degradation conditions are commonly encountered in real clinical environments and can substantially reduce the reliability of deep learning-based systems. Explainable AI techniques such as Grad-CAM have therefore been widely adopted to provide heatmap visualizations highlighting salient image regions, thereby improving model transparency [15, 23].

Despite recent advances, several research gaps remain unresolved. Most CNN-based chronic wound classification studies continue to focus on binary classification, which does not fully reflect real clinical settings where multiple wound types may coexist. Studies integrating DFU, PU, and Normal Skin classification into a unified multiclass framework are still limited. Furthermore, although high accuracy is often reported, model robustness to image-quality degradation, particularly blur caused by camera motion or defocus, is rarely analyzed in a systematic manner. In addition, model interpretability is frequently underexplored, despite its importance for transparency and trust in medical artificial intelligence applications.

Based on these gaps, this study develops and compares two CNN architectures, namely ResNet-50 and ResNeXt-50, for multiclass classification of chronic wound images distinguishing DFU, PU, and Normal Skin. In addition, robustness is evaluated specifically in terms of resistance to Gaussian blur applied at the inference stage. Finally, Grad-CAM is employed as a qualitative interpretability tool to examine whether the resulting predictions focus on clinically relevant lesion regions.

2. Methodology

In summary, the research method's flow is shown in Fig. 1. The process starts with the collection and merging of the DFU, PU, and Normal Skin datasets, followed by preprocessing and splitting into training, validation, and testing sets. In addition, two CNN-based models, ResNet-50 and ResNeXt-50, that had been pre-trained on ImageNet were refined for three-class classification tasks. After training, each model was evaluated on an internal test set derived from the merged dataset. Robustness was further assessed by applying controlled Gaussian blur to the test images at the inference stage. Finally, the Grad-CAM technique was used to visualize the image regions contributing to the model predictions.

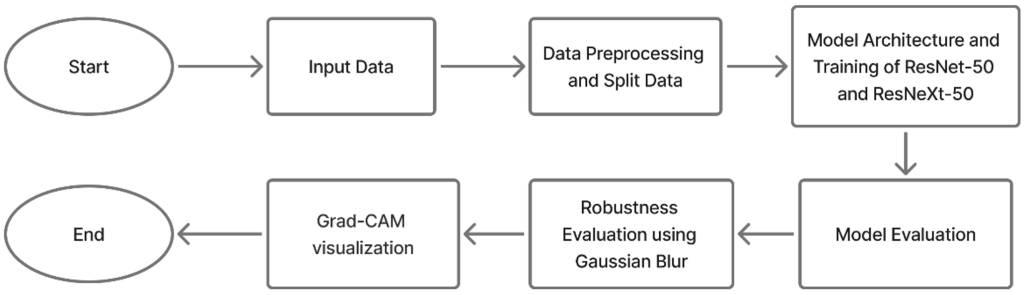

Fig. 1: Research Method

2.1. Input Data

The dataset used in this study consists of three image classes, namely DFU, PU, and Normal Skin. DFU and Normal Skin images were obtained from the DFU dataset developed by Alzubaidi et al., which is publicly available on the Kaggle platform [24–26]. Meanwhile, PU images were obtained from a set of PU images with severity annotations (stages 1–4); in this study, all the stages were combined into one PU class because the focus of the analysis was the differentiation between DFU, PU, and Normal Skin, not the classification of the severity of the wound [27].

No additional external dataset was used in this study. All robustness evaluations were conducted using the internal test set derived from the main dataset by applying controlled image-quality degradation. Specifically, Gaussian blur was introduced at the inference stage to simulate realistic variations in image quality that may occur under clinical imaging conditions.

2.2. Data Preprocessing and Split Data

Data pre-processing is done to prepare the dataset that has been obtained before proceeding to the next stage to be more structured and more accurate for data training. At this stage, all DFU, PU, and Normal Skin images are first resized to a fixed size of 224×224 pixels to match the input of the ResNet-50 and ResNeXt-50 architecture standards. Furthermore, the pixel intensity on each RGB channel was normalized using the mean values and standard deviations of ImageNet, so that the image distribution was more consistent with the pre-trained weights used on the backbone. For each pixel on channel $c \in \{R, G, B\}$, normalization is defined in Eq. (1).

$$x'_{i,c} = \frac{x_{i,c} - \mu_c}{\sigma_c} \tag{1}$$

where $x'_{i,c}$ denotes the normalized value of the i-th pixel in channel c, $x_{i,c}$ represents the original value of the i-th pixel in the same channel, μ_c is the mean value of channel c (R, G, or B), σ_c denotes the standard deviation of channel c, i refers to the pixel index in the image, and c indicates the color channel index.

After pre-processing, all images were organized into class-specific folders (DFU, PU, and Normal Skin) and then stratified into three subsets—approximately 70% for training, 15% for validation, and 15% for testing—such that the original class proportions were preserved in each

subset. Due to the absence of patient identifiers in the dataset, data splitting was performed at the image level. This may introduce a potential risk of data leakage if multiple images from the same patient appear in different subsets; this limitation is acknowledged in this study.

Training images are augmented during the training process using stochastic geometric and photometric transformations, including small rotations, horizontal flips, brightness–contrast adjustments, and random cropping and padding operations followed by resizing back to 224×224 pixels. Data augmentation is applied only to the training set, while validation and test images undergo only resizing and normalization. The distribution of images for each class in the training, validation, and test sets is presented in Table 1.

Table 1: Distribution of Dataset Images on Train, Validation, and Test Data

Class	Train	Val	Test	Total
DFU	358	76	78	512
PU	763	163	165	1091
Normal Skin	380	81	82	543
Total	1501	320	325	2146

2.3. Model Architecture and Training of ResNet-50 and ResNeXt-50

The two CNN architectures compared in this study are ResNet-50 and ResNeXt-50, both of which are pre-trained on ImageNet. Broadly speaking, ResNet-50 is a deep residual network consisting of 50 layers, composed of bottleneck blocks with convolutional sequences of 1×1, 3×3, and 1×1 filters complemented by shortcut connections that sum the inputs and outputs of each block. This residual mechanism helps to overcome the problem of performance degradation in deep networks by maintaining stable gradient flow during the training process [28].

ResNeXt-50 adopts the basic structure of ResNet-50 but replaces the standard residual blocks with aggregated residual transformations based on grouped convolutions. Within each block, the convolutional path is divided into several parallel transformations, with parameters partitioned into groups, and then rejoined before being summed with an identity path. This approach increases the capacity of feature representation through enhancement of cardinality (the number of parallel paths) without significantly increasing the number of parameters [29]. Thus, ResNeXt-50 is expected to extract more diverse texture and structural patterns of chronic wounds than ResNet-50 at comparable model complexity.

In this study, both backbones were used as the main feature extractors and modified at the final layer to produce three output classes (DFU, PU, and Normal Skin). The last fully connected (FC) layer converts the global average pooling feature vector into three logits, as defined in Eq. (2).

$$\mathbf{z} = \mathbf{W}\mathbf{H} + \mathbf{b} \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^d$ denotes the feature vector obtained from the global average pooling layer, $\mathbf{W} \in \mathbb{R}^{C \times d}$ represents the weight matrix of the final fully connected layer, and $\mathbf{b} \in \mathbb{R}^C$ is the bias vector associated with the same layer. In this study, $C = 3$.

The resulting vector $\mathbf{z} = [z_1, z_2, \ldots, z_C]$ contains the logit values, where each logit z_c represents the unnormalized score for class c. The predicted probability for each class is then computed using the softmax function, as defined in Eq. (3).

$$p_c = \text{softmax}(z_c) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)} \tag{3}$$

where p_c denotes the predicted probability for class c.

Training was performed using a fixed three-stage fine-tuning strategy for both models. All input images were resized to 224×224 pixels and processed in mini-batches of size 16. Data augmentation was applied only to the training set using fixed stochastic geometric and photometric transformations, including random affine transformations with rotation in the range $[-10°, +10°]$,

translation up to 5% of the image size, and scaling in the range $[0.9, 1.1]$; random horizontal flipping with probability $p = 0.5$; and random brightness adjustment in the range $[-10\%, +10\%]$. Validation and test images were not augmented and were only resized and normalized using ImageNet statistics.

The training procedure consisted of exactly three stages. (1) A warm-up stage in which only the fully connected classification layer was trained for 10 epochs while the backbone was frozen. (2) A first fine-tuning stage in which the top convolutional block (`layer4`) was unfrozen and jointly optimized with the classification head for 15 epochs. (3) A final fine-tuning stage in which all remaining backbone layers were unfrozen and optimized for 15 epochs.

Thus, the maximum number of training epochs was fixed at 40. Early stopping was applied based on validation accuracy with a patience of 7 epochs, and the checkpoint with the highest validation accuracy was selected for evaluation.

Optimization was performed using the Adam optimizer with fixed parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) and no weight decay. Stage-wise learning rates were set to 1×10^{-3} for the warm-up stage, 1×10^{-5} for the first fine-tuning stage, and 5×10^{-6} for the final fine-tuning stage. Learning-rate scheduling employed a *ReduceLROnPlateau* scheduler monitoring validation loss, with mode = "min", factor = 0.5, and patience = 3.

To ensure reproducibility, all experiments were conducted using a fixed random seed (SEED = 42) and implemented in PyTorch 2.1.0 with CUDA support and Python 3.10.

Due to class imbalance in the training dataset, class weights were introduced to penalize misclassification of minority classes more heavily. The class weight for class c is calculated based on inverse class frequency, as defined in Eq. (4).

$$w_c = \frac{N_{\text{train}}}{C \cdot n_c} \tag{4}$$

where w_c denotes the weight assigned to class c, n_c is the number of training samples belonging to class c, N_{train} is the total number of training samples, and C is the number of classes.

Using these class weights, the weighted cross-entropy loss is formulated as shown in Eq. (5).

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c \, y_{i,c} \log(p_{i,c}) \tag{5}$$

where N denotes the number of samples in a mini-batch, $y_{i,c} \in \{0, 1\}$ is the one-hot encoded ground-truth label of sample i for class c, and $p_{i,c}$ is the predicted probability of class c for sample i. This weighting strategy is commonly used in medical image classification tasks to mitigate class imbalance and improve balanced performance across all classes [30].

2.4. Model Evaluation

After the training process is completed, the best-performing ResNet-50 and ResNeXt-50 models, selected based on validation performance, are evaluated on the test set derived from the main dataset. The evaluation is conducted using a confusion matrix for three classes, namely DFU, PU, and Normal Skin. Standard classification metrics commonly used in medical image analysis are employed, including Accuracy, Precision, Recall, and F1-score for each class, to assess overall and class-wise performance.

Precision, Recall, and F1-score are computed based on the confusion matrix to evaluate the correctness and completeness of the model predictions for each class. These metrics are mathematically defined as shown in Eqs. (**??**). Overall classification accuracy is calculated to measure the proportion of correctly classified samples across all classes, as defined in Eq. (9). To address class imbalance, the macro-averaged F1-score is additionally employed by averaging the F1-scores of all classes equally, regardless of sample size, as formulated in Eq. (10) [31].

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{6}$$

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{7}$$

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{8}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}} \tag{9}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c \tag{10}$$

where TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively, computed in a one-vs-rest manner, and C denotes the total number of classes.

A comparison of confusion matrices, accuracy, and macro-averaged F1-score between ResNet-50 and ResNeXt-50 is presented in the Results section and serves as the basis for the comparative analysis of the two architectures. In medical image classification scenarios with class imbalance, previous studies have shown that the macro-averaged F1-score is more appropriate than accuracy alone, as it assigns equal importance to each class and prevents the performance of minority classes from being dominated by majority classes [32]. Therefore, in addition to reporting accuracy, this study uses class-wise F1-score and macro-averaged F1-score as the primary evaluation metrics.

2.5. Robustness Evaluation Using Gaussian Blur

To provide a visual overview of the robustness evaluation scenario used in this study, Fig. 2 shows an example of DFU and Normal Skin images before and after the application of Gaussian blur to the test data. This example is presented to visually illustrate the effects of image quality degradation, while quantitative evaluation is performed on all classes in a multiclass classification scenario, namely DFU, PU, and Normal Skin.

Fig. 2: Example of a clean test image and a test image with Gaussian blur ($\sigma = 2.0$) used to evaluate model robustness to image-quality degradation.

To evaluate the model's resistance to realistic image-quality degradation, additional testing was performed by applying Gaussian blur to the test data. Gaussian blur is used to simulate less-than-ideal clinical imaging conditions, such as camera defocus or mild motion disturbance during image acquisition, which are commonly encountered in real clinical practice. Several studies in medical image analysis have reported that image degradation such as blur can significantly affect the performance of deep learning models if not explicitly evaluated. Therefore, robustness

testing against Gaussian blur is an important aspect of assessing model reliability under degraded imaging conditions that may occur in clinical environments [33, 34].

Mathematically, Gaussian blur is expressed as a convolution operation between the input image and a Gaussian kernel, defined as shown in Eq. (11).

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{11}$$

where σ denotes the standard deviation controlling the blur intensity. In this study, Gaussian blur was applied at multiple levels ($\sigma = 1.0$, 1.5, and 2.0) at the inference stage without retraining the models. After applying Gaussian blur, model performance was re-evaluated using the same metrics as in the non-degraded test scenario, namely accuracy, precision, recall, and F1-score in the multiclass classification setting. Comparisons between unblurred and blurred test results were used to analyze changes in performance and predictive stability under image-quality degradation. Robustness in this study is specifically defined as resistance to image blur or defocus simulated using Gaussian blur and does not represent robustness to image degradation in general.

2.6. Grad-CAM Visualization

To assess the model's interpretability and ensure that the predictions are not only numerically strong but also clinically plausible, the Gradient-weighted Class Activation Mapping (Grad-CAM) technique is applied to the last convolutional layer of the model [15]. Grad-CAM generates a class-discriminative heatmap that highlights image regions contributing most strongly to the prediction of a particular class.

In this study, Grad-CAM visualizations are used in the Results section to illustrate that the models tend to focus on ulcer regions in DFU and PU images, while not emphasizing irrelevant regions in Normal Skin images. This observation supports the claim that the proposed classification system is not only accurate but also visually interpretable. Grad-CAM is employed strictly as a qualitative interpretability tool to provide visual insight into model attention, rather than as a quantitative localization or segmentation evaluation.

3. Results and Discussion

This section presents the experimental results and discussion of the proposed deep learning models for multiclass chronic wound image classification. The evaluation focuses on the performance of ResNet-50 and ResNeXt-50 models on the test dataset using standard classification metrics, including accuracy, class-wise F1-score, and macro-averaged F1-score. In addition, a robustness analysis is conducted by applying Gaussian blur with varying intensity levels to assess model stability under image-quality degradation. Visual interpretability is further analyzed using Grad-CAM to examine whether the models focus on clinically relevant regions when making predictions. The results are discussed in terms of classification performance, robustness, and comparative behavior between the two architectures.

3.1. Training and Validation Performance

The training and validation performance of the proposed models is evaluated to analyze their learning behavior, convergence characteristics, and generalization capability. In this subsection, the accuracy and loss curves of ResNet-50 and ResNeXt-50 are examined.

Fig. 3 presents the training and validation accuracy curves for both models. As shown in the figure, the accuracy values increase consistently during training and stabilize before reaching the maximum number of epochs due to the implementation of the early stopping criterion. This behavior indicates stable convergence for both models. Notably, ResNeXt-50 converges faster and achieves slightly higher validation accuracy than ResNet-50, suggesting superior generalization performance.

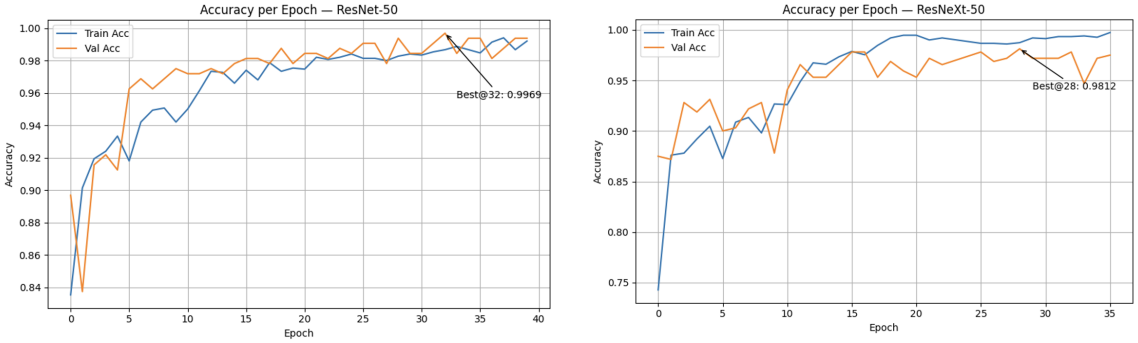

Fig. 3: Training and validation accuracy curves for ResNet-50 and ResNeXt-50.

To further investigate the optimization behavior of the models, the training and validation loss curves are analyzed, as illustrated in Fig. 4.

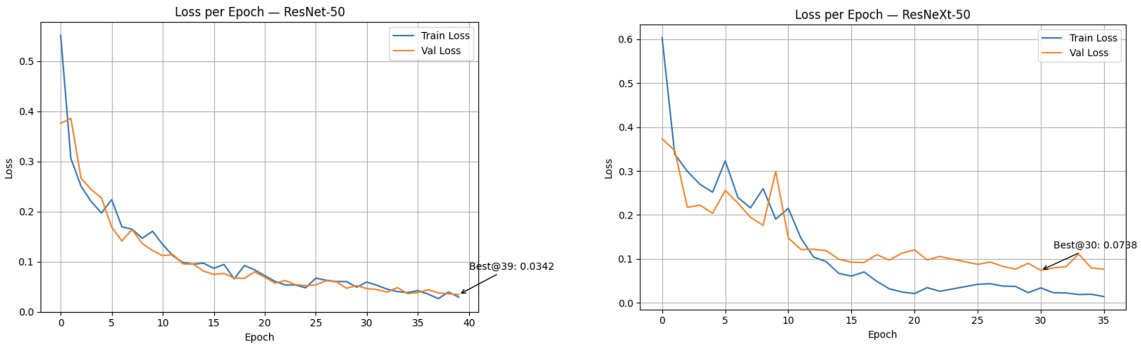

Fig. 4: Training and validation loss curves for ResNet-50 and ResNeXt-50.

The loss curves in Fig. 4 demonstrate a consistent decrease throughout the training process for both models, indicating effective optimization. The validation loss follows a similar decreasing trend and converges smoothly, implying that neither model experiences significant overfitting. In comparison, ResNeXt-50 attains a lower validation loss and stabilizes earlier than ResNet-50, which aligns with its faster convergence and improved generalization performance observed in the accuracy results.

3.2. Test Performance Evaluation

This subsection presents the classification performance of the proposed models on the test data using standard evaluation metrics.

3.2.1. ResNet-50 Performance

The quantitative evaluation results of the ResNet-50 model on the test data are summarized in Table 2.

Table 2: Results of ResNet-50 Model Evaluation on Test Data (overall accuracy: 0.9877).

Classes	Precision	Recall	F1-score	Support
DFU	0.9512	1.00	0.9750	78
Normal Skin	1.00	0.9878	0.9939	82
PU	1.00	0.9818	0.9908	165
Macro-averaged	0.9837	0.9899	0.9866	325
Weighted-averaged	0.9883	0.9877	0.9878	325

Based on the results presented in Table 2, the ResNet-50 model demonstrates strong classification performance on the test dataset, achieving an overall accuracy of 0.9877 and a macro-averaged

F1-score of 0.9866. These results indicate a high and well-balanced predictive capability across all classes.

To further analyze the classification behavior of the model, the corresponding confusion matrix is shown in Fig. 5.

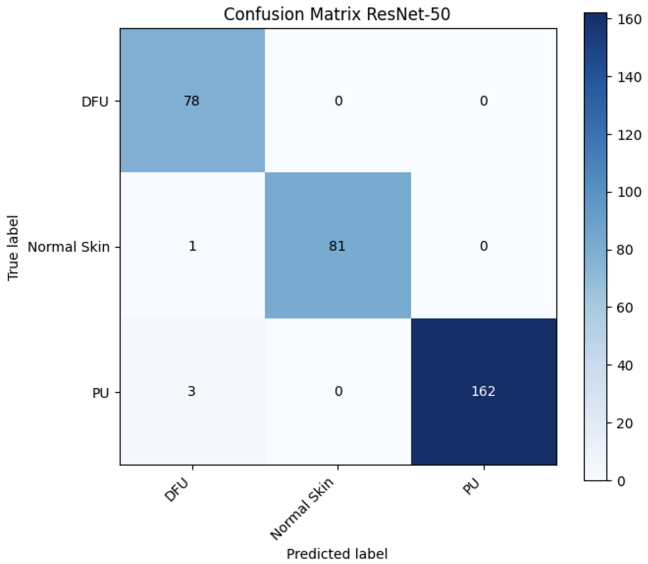

Fig. 5: Confusion matrix of ResNet-50 on the test dataset.

In the class-wise evaluation, the DFU class achieves a recall of 1.00, indicating that all DFU samples are correctly identified, although its precision is slightly lower at 0.9512 due to a small number of false positive predictions. The Normal Skin class shows excellent performance with perfect precision (1.00) and a recall of 0.9878, reflecting accurate identification of healthy skin images with minimal misclassification. The PU class also exhibits strong performance, achieving perfect precision (1.00) and a high recall of 0.9818.

As illustrated by the confusion matrix, most predictions lie along the diagonal, confirming the consistency of the classification results. The observed misclassifications primarily involve DFU predictions arising from other classes, indicating a slight tendency of the model to over-predict the DFU class rather than failing to detect DFU cases.

3.2.2. ResNeXt-50 Evaluation on Internal Test Data and Comparison with ResNet-50

The quantitative evaluation results of the ResNeXt-50 model on the test data are summarized in Table 3.

Table 3: Results of ResNeXt-50 Model Evaluation on Test Data (overall accuracy: 0.9938).

Classes	Precision	Recall	F1-score	Support
DFU	0.9873	1.00	0.9936	78
Normal Skin	0.9878	0.9878	0.9878	82
PU	1.00	0.9939	0.9970	165
Macro-averaged	0.9917	0.9939	0.9928	325
Weighted-averaged	0.9939	0.9938	0.9939	325

Based on the results presented in Table 3, the ResNeXt-50 model demonstrates excellent classification performance on the test data, achieving an overall accuracy of 0.9938 and a macro-averaged F1-score of 0.9928. These results indicate strong and balanced performance across all classes.

To further analyze the classification behavior of the model, the corresponding confusion matrix is shown in Fig. 6.

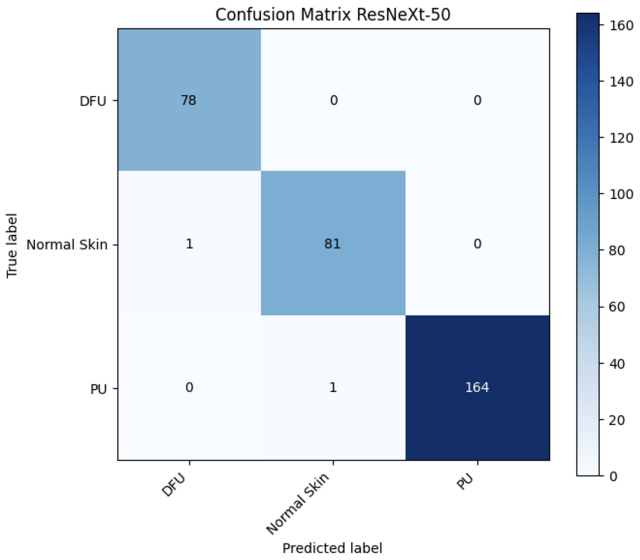

Fig. 6: Confusion matrix of ResNeXt-50 on the test dataset.

In the class-wise evaluation, the DFU class achieves a recall of 1.00 with a high precision of 0.9873, indicating that all DFU samples are correctly identified with only a small number of false positive predictions. The Normal Skin class shows stable performance, with precision, recall, and F1-score values of 0.9878, reflecting consistent discrimination of healthy skin images. The PU class exhibits the strongest performance, achieving perfect precision (1.00) and a recall of 0.9939, resulting in an F1-score of 0.9970.

The confusion matrix shown in Fig. 6 further supports these quantitative results. Most predictions are concentrated along the diagonal, confirming the high classification accuracy of the model. Only a very small number of misclassifications are observed, primarily involving occasional confusion between Normal Skin and ulcer classes. Overall, the confusion matrix indicates that ResNeXt-50 is able to effectively distinguish between DFU, PU, and Normal Skin categories.

3.2.3. Comparative Analysis

A comparative evaluation of ResNet-50 and ResNeXt-50 shows that both models achieve strong performance in multiclass classification of DFU, PU, and Normal Skin images. Both models demonstrate strong and comparable performance on clean test data, with ResNeXt-50 exhibiting slightly higher accuracy and macro-averaged F1-score. While ResNet-50 demonstrates reliable performance with high precision and recall values, particularly for the Normal Skin and PU classes, it exhibits a slightly lower precision for the DFU class, as reflected in the confusion matrix. In contrast, ResNeXt-50 achieves higher and more balanced class-wise performance, with perfect recall for the DFU class and fewer misclassifications overall.

The slightly higher performance of ResNeXt-50 on clean test data may be related to its architectural design; however, given the relatively small performance gap, both models can be considered to perform comparably under non-degraded conditions. As a result, ResNeXt-50 demonstrates an enhanced capability in capturing complex wound characteristics, leading to improved classification stability.

Overall, although both architectures perform competitively, ResNeXt-50 shows a consistent advantage over ResNet-50 in terms of accuracy, robustness of predictions, and balanced performance across classes.

3.2.4. Evaluation of Model Robustness Using Gaussian Blur

To evaluate the models' resistance to image-quality degradation, additional tests were conducted by applying Gaussian blur to the test data at several intensity levels (σ = 1.0, 1.5, and 2.0).

This evaluation was performed on both CNN architectures, ResNet-50 and ResNeXt-50, without retraining, to ensure a fair comparison.

Table 4: Robustness evaluation results using Gaussian blur.

Model	Gaussian Blur (σ)	Accuracy	Macro-averaged F1-score
ResNet-50	0 (Clean)	0.9877	0.9866
ResNet-50	1.0	0.9754	0.9731
ResNet-50	1.5	0.9446	0.9396
ResNet-50	2.0	0.9200	0.9123
ResNeXt-50	0 (Clean)	0.9938	0.9928
ResNeXt-50	1.0	0.9877	0.9866
ResNeXt-50	1.5	0.9754	0.9725
ResNeXt-50	2.0	0.9723	0.9679

Table 4 presents the robustness evaluation results of ResNet-50 and ResNeXt-50 under Gaussian blur at multiple intensity levels. All evaluations were conducted without retraining the models to ensure consistent and fair comparisons.

The results indicate that increasing Gaussian blur intensity leads to a gradual decline in performance for both models, as reflected in decreasing accuracy and macro-averaged F1-score values. Under clean image conditions ($\sigma = 0$), ResNeXt-50 achieves the highest performance with an accuracy of 0.9938 and a macro-averaged F1-score of 0.9928, slightly outperforming ResNet-50.

As the blur intensity increases, ResNet-50 experiences a more pronounced performance degradation. At $\sigma = 2.0$, its accuracy decreases to 0.9200 and its macro-averaged F1-score to 0.9123. In contrast, ResNeXt-50 demonstrates better resistance to image degradation, maintaining an accuracy of 0.9723 and a macro-averaged F1-score of 0.9679 at the same blur level.

At $\sigma = 1.0$, ResNeXt-50 attains accuracy and macro-averaged F1-score values that numerically match those obtained by ResNet-50 under clean image conditions. This numerical similarity arises from independent evaluation outcomes and should be interpreted as coincidental rather than indicative of identical model behavior. This observation suggests that low-intensity Gaussian blur does not immediately degrade classification performance and that both architectures remain stable under mild blur conditions.

Overall, these results indicate that ResNeXt-50 exhibits more stable feature representations under increasing Gaussian blur compared to ResNet-50, potentially due to its grouped convolution design and increased cardinality. The gradual degradation observed in both models further suggests that the proposed approach captures both fine-grained texture information and more global visual cues relevant to chronic wound assessment.

3.3. Visual Analysis of Model Prediction with Grad-CAM

Fig. 7 illustrates that the Grad-CAM heatmaps for the DFU and PU classes are strongly concentrated around ulcer regions, with high activation overlapping lesion areas in the affected skin. This indicates that the models tend to focus on visually relevant wound regions rather than background skin when generating predictions.

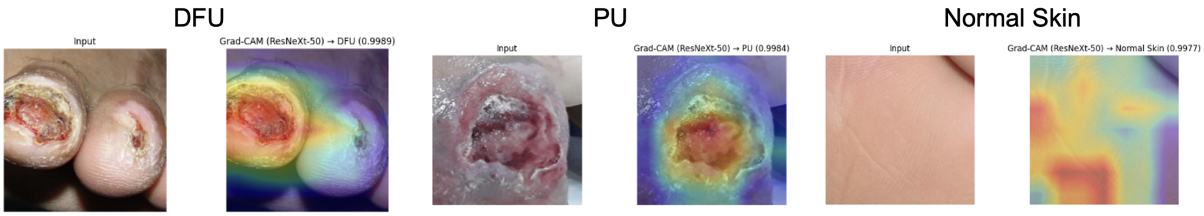

Fig. 7: Grad-CAM visualizations for DFU, PU, and Normal Skin images.

In contrast, Grad-CAM activation for Normal Skin images appears more diffuse across the

skin surface, without a distinct focus on a specific region, which is consistent with the absence of pathological lesions. These qualitative visual patterns are consistent with the quantitative evaluation results and provide insight into the models' decision-making processes. Overall, the use of Grad-CAM supports qualitative interpretability and may help increase confidence in the models' predictions; however, further external and prospective validation would be required before clinical deployment.

4. Conclusion

Overall, this study demonstrates that both ResNet-50 and ResNeXt-50, trained with data augmentation and class-weighted loss, achieve strong performance in multiclass classification of chronic wound images, including DFU, PU, and Normal Skin. Evaluation on clean test data shows that the two models exhibit strong and comparable accuracy and macro-averaged F1-scores, with balanced precision and recall across classes.

Robustness evaluation under Gaussian blur indicates that both models experience a gradual performance decline as image quality degrades. However, performance remains high under light to moderate blur levels ($\sigma = 1.0$ and 1.5), suggesting that the learned representations do not rely solely on fine-grained texture details but also capture clinically relevant global visual information. Compared to ResNet-50, ResNeXt-50 shows a smaller performance degradation under increasing blur intensity and maintains higher accuracy and macro-averaged F1-scores at stronger blur levels ($\sigma = 2.0$).

The observed robustness advantage of ResNeXt-50 under Gaussian blur conditions may be attributed to its architectural design, particularly the use of grouped convolution and increased cardinality, which can promote more diverse and stable feature representations under degraded imaging conditions. In addition, qualitative Grad-CAM analysis indicates that both models primarily focus on clinically relevant lesion regions in DFU and PU images, while showing diffuse activation patterns for Normal Skin images, supporting the qualitative interpretability of the predictions.

Despite these promising results, this study has several limitations that should be acknowledged. First, the dataset was compiled from multiple public sources, which may introduce domain-specific biases. Future work will address this limitation by incorporating larger and more diverse datasets collected from different clinical settings. Second, robustness evaluation was limited to Gaussian blur and did not consider other common image degradations such as noise or illumination variation. To overcome this limitation, future studies will evaluate model robustness under a wider range of realistic imaging degradations. Third, no external or prospective clinical validation was conducted. Future work will therefore focus on external and prospective validation to assess the generalizability and clinical reliability of the proposed approach. In addition, future research will explore lighter network architectures for deployment on resource-constrained devices and extend the classification task to include ulcer severity assessment.

In summary, the findings suggest that ResNeXt-50 offers improved robustness to Gaussian blur compared to ResNet-50 in multiclass DFU–PU–Normal Skin classification. While not yet intended for direct clinical deployment, the proposed approach shows potential as an explainable computer-aided screening support tool, subject to further validation and refinement.

CRediT Author Contributions

Reynaldi Ikbar Surya Andhika: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Visualization, Writing–Original Draft, Writing–Review & Rditing. **Sugiyarto Surono**: Supervision, Validation, Methodology Consultation, Review. **Aris Thobirin**: Supervision, Resources, Project Administration, Review.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, generative AI and AI-assisted tools (including ChatGPT and Grammarly) were used solely for language refinement and grammatical editing. No AI-generated content was used in the analysis, interpretation of results, or development of the scientific content. All analyses and conclusions were developed by the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and Acknowledgments

The authors would like to thank the Mathematics Study Program, Universitas Ahmad Dahlan, Yogyakarta, for the academic support and computing facilities provided during this research. The authors also gratefully acknowledge the researchers who released the DFU, PU, and Normal Skin image datasets to the public domain, which made the experimental evaluation in this work possible. Finally, the authors appreciate all colleagues and reviewers for their constructive comments and suggestions that helped improve the quality of this manuscript.

Data Availability

The datasets used in this study are publicly available. The DFU, PU, and Normal Skin image data were obtained from publicly accessible datasets as cited in the references. No proprietary or confidential data were used. All data preprocessing and experimental evaluations were conducted on these publicly available datasets.

References

[1] Sugiyarto Surono, D. K. E. Arofah, and Aris Thobirin. “Robust Convolutional Neural Network for Image Classification with Gaussian Noise”. In: *Frontiers in Artificial Intelligence and Applications* 378 (2023), pp. 67–76. DOI: 10.3233/FAIA231007.

[2] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono. “Deep Convolutional Neural Networks in Medical Image Analysis: A Review”. In: *Information* 16.3 (2025). DOI: 10.3390/info16030195.

[3] G. Latif, J. Alghazo, M. A. Khan, G. Ben Brahim, K. Fawagreh, and N. Mohammad. “Deep Convolutional Neural Network (CNN) Model Optimization Techniques—Review for Medical Imaging”. In: *AIMS Mathematics* 9.8 (2024), pp. 20539–20571. DOI: 10.3934/math.2024998.

[4] K. Song, J. Feng, and D. Chen. “A Survey on Deep Learning in Medical Ultrasound Imaging”. In: *Frontiers in Physics* 12 (2024). DOI: 10.3389/fphy.2024.1398393.

[5] X. Liu et al. “Advances in Deep Learning-Based Medical Image Analysis”. In: *Health Data Science* (2021), p. 8786793. DOI: 10.34133/2021/8786793.

[6] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang. “Pooling in Convolutional Neural Networks for Medical Image Analysis: A Survey and an Empirical Study”. In: *Neural Computing and Applications* 34.7 (2022), pp. 5321–5347. DOI: 10.1007/s00521-022-06953-8.

[7] M. Li, Y. Jiang, Y. Zhang, and H. Zhu. “Medical Image Analysis Using Deep Learning Algorithms”. In: *Frontiers in Public Health* 11 (2023), p. 1273253. DOI: 10.3389/fpubh.2023.1273253.

[8] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. "A Survey of the Recent Architectures of Deep Convolutional Neural Networks". In: *Artificial Intelligence Review* 53.8 (2020), pp. 5455–5516. DOI: 10.1007/s10462-020-09825-6.

[9] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam. "An Enhanced Technique of Skin Cancer Classification Using Deep Convolutional Neural Network with Transfer Learning Models". In: *Machine Learning with Applications* 5 (2021), p. 100036. DOI: 10.1016/j.mlwa.2021.100036.

[10] A. T. Ibrahim, M. Abdullahi, A. F. D. Kana, M. T. Mohammed, and I. H. Hassan. "Categorical Classification of Skin Cancer Using a Weighted Ensemble of Transfer Learning with Test Time Augmentation". In: *Data Science and Management* 8.2 (2025), pp. 174–184. DOI: 10.1016/j.dsm.2024.10.002.

[11] M. Harahap, S. K. Anjelli, W. A. M. Sinaga, R. Alward, J. F. W. Manawan, and A. M. Husein. "Classification of Diabetic Foot Ulcer Using Convolutional Neural Network (CNN) in Diabetic Patients". In: *Journal of INFOTEL* 14.3 (2022), pp. 196–202. DOI: 10.20895/infotel.v14i3.796.

[12] C. Wang, Z. Yu, Z. Long, H. Zhao, and Z. Wang. "A Few-Shot Diabetes Foot Ulcer Image Classification Method Based on Deep ResNet and Transfer Learning". In: *Scientific Reports* 14.1 (2024). DOI: 10.1038/s41598-024-80691-w.

[13] Y. Zhang. "3D Reconstruction of Monocular Images Based on ResNeXt Neural Network". In: *Proceedings of Atlantis Press*. 2024, pp. 816–829. DOI: 10.2991/978-94-6463-540-9_82.

[14] A. Priya and P. S. Bharathi. "SE-ResNeXt-50-CNN: A Deep Learning Model for Lung Cancer Classification". In: *Applied Soft Computing* 171 (2025), p. 112696. DOI: 10.1016/j.asoc.2025.112696.

[15] H. Neuwieser et al. "Interpreting Venous and Arterial Ulcer Images through the Grad-CAM Lens". In: *Diagnostics* 15.17 (2025). DOI: 10.3390/diagnostics15172184.

[16] J. M. Raja, M. A. Maturana, S. Kayali, A. Khouzam, and N. Efeovbokhan. "Diabetic Foot Ulcer: A Comprehensive Review of Pathophysiology and Management Modalities". In: *World Journal of Clinical Cases* 11.8 (2023), pp. 1684–1693. DOI: 10.12998/wjcc.v11.i8.1684.

[17] K. McDermott, M. Fang, A. J. M. Boulton, E. Selvin, and C. W. Hicks. "Etiology, Epidemiology, and Disparities in the Burden of Diabetic Foot Ulcers". In: *Diabetes Care* 46.1 (2022), pp. 209–221. DOI: 10.2337/dci22-0043.

[18] S. Vieira, A. Mostardinha, and P. Alves. "Unveiling the Burden: A Six-Year Retrospective Analysis of Pressure Ulcer Epidemiology in an ICU". In: *Nursing Reports* 14.4 (2024), pp. 3291–3309. DOI: 10.3390/nursrep14040239.

[19] S. Zhang, G. Wei, L. Han, W. Zhong, Z. Lu, and Z. Niu. "Global, Regional and National Burden of Decubitus Ulcers from 1990 to 2021". In: *Frontiers in Public Health* 13 (2025). DOI: 10.3389/fpubh.2025.1494229.

[20] Y. Patel et al. "Integrated Image and Location Analysis for Wound Classification". In: *Scientific Reports* 14.1 (2024), p. 7043. DOI: 10.1038/s41598-024-56626-w.

[21] G. Zhu, Z. Fu, and G. Guo. "Applications and Prospects of Artificial Intelligence in Wound Healing". In: *Regeneration, Repair and Rehabilitation* 1.4 (2025), pp. 12–19. DOI: 10.1016/j.rerere.2025.08.001.

[22] P. S. Rathore, A. Kumar, A. Nandal, A. Dhaka, and A. K. Sharma. "A Feature Explainability-Based Deep Learning Technique for Diabetic Foot Ulcer Identification". In: *Scientific Reports* 15.1 (2025). DOI: 10.1038/s41598-025-90780-z.

[23] E. Tjoa and C. Guan. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4793–4813. DOI: 10.1109/TNNLS.2020.3027314.

[24] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, and J. Zhang. "DFU_QUTNet: Diabetic Foot Ulcer Classification Using Novel Deep Convolutional Neural Network". In: *Multimedia Tools and Applications* 79.21 (2020), pp. 15655–15677. DOI: 10.1007/s11042-019-07820-w.

[25] L. Alzubaidi et al. "Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study". In: *Applied Sciences* 10.13 (2020). DOI: 10.3390/app10134523.

[26] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, and Y. Duan. "Robust Application of New Deep Learning Tools: An Experimental Study in Medical Imaging". In: *Multimedia Tools and Applications* 81.10 (2022), pp. 13289–13317. DOI: 10.1007/s11042-021-10942-9.

[27] B. Ay, B. Tasar, Z. Utlu, K. Ay, and G. Aydin. "Deep Transfer Learning-Based Visual Classification of Pressure Injuries Stages". In: *Neural Computing and Applications* 34.18 (2022), pp. 16157–16168. DOI: 10.1007/s00521-022-07274-6.

[28] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. "Aggregated Residual Transformations for Deep Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.

[30] Z. Song, Z. Shi, X. Yan, B. Zhang, S. Song, and C. Tang. "An Improved Weighted Cross-Entropy-Based Convolutional Neural Network for Auxiliary Diagnosis of Pneumonia". In: *Electronics* 13.15 (2024). DOI: 10.3390/electronics13152929.

[31] N. A. Sovia, N. W. S. Wardhani, and E. Sumarminingsih. "Enhancing Image Classification of Cabbage Plant Diseases Using a Hybrid Convolutional Neural Network and XGBoost Model". In: *Cauchy: Jurnal Matematika Murni dan Aplikasi* 10.1 (2025), pp. 278–289. DOI: 10.18860/cauchy.v10i1.30866.

[32] N. N. K. Krisnawijaya, C. Catal, B. Tekinerdogan, R. van der Tol, H. Hogeveen, and Y. Herdiyeni. "A Machine Learning Approach to Identifying Foot and Mouth Disease Incidence in Dairy Farms with Suboptimal Veterinary Infrastructure". In: *Smart Agricultural Technology* 12 (2025), p. 101261. DOI: 10.1016/j.atech.2025.101261.

[33] S. Gehlot, A. Gupta, and R. Gupta. "A CNN-Based Unified Framework Utilizing Projection Loss in Unison with Label Noise Handling for Multiple Myeloma Cancer Diagnosis". In: *Medical Image Analysis* 72 (2021), p. 102099. DOI: 10.1016/j.media.2021.102099.

[34] P. Celard, E. L. Iglesias, J. M. Sorribes-Fdez, R. Romero, A. S. Vieira, and L. Borrajo. "A Survey on Deep Learning Applied to Medical Images: From Simple Artificial Neural Networks to Generative Models". In: *Neural Computing and Applications* 35.3 (2023), pp. 2291–2323. DOI: 10.1007/s00521-022-07953-4.