Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 6 (2025) 1428 - 1438 e-ISSN: 2580-0760

Comparative Analysis of ResNet-Based Wagner-Scale Classification for Imbalanced DFU Data

Aditya Wahyu Ramadhan[1], Pulung Nurtantio Andono[2], Moch.Arief Soeleman[3]
[1,2,3]Master's Programme in Information Technology, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

[1]adityawr07@gmail.com, [2] pulung@dsn.dinus.ac.id, [3]arief2802@dsn.dinus.ac.id

Abstract

Diabetic foot ulcers (DFUs) constitute a serious complication associated with diabetes mellitus and significantly increase the likelihood of lower-extremity amputation if timely treatment is not provided. Conventional evaluation approaches that rely heavily on clinicians' visual judgment tend to be subjective and may result in inconsistent assessments. To overcome these limitations, this research proposes a multiclass DFU classification system grounded in the Wagner grading scheme (Grades 0–5), employing a ResNet-50 deep learning architecture enhanced through transfer learning. The study utilizes a dataset of 1,415 clinical wound images that were carefully labeled and verified by medical experts. However, the dataset exhibits a pronounced class imbalance, with 543 samples categorized as Grade 0, 110 as Grade 1, 252 as Grade 2, 145 as Grade 3, 293 as Grade 4, and only 72 samples representing Grade 5. To address this issue, random oversampling (ROS) was applied in conjunction with common preprocessing techniques, including image normalization and data augmentation, to increase the diversity of training data. The experimental evaluation indicates that the proposed approach achieves robust classification performance across multiple metrics, namely accuracy, precision, recall, and F1-score. The model records a precision value of 0.96, a recall of 0.95, and an F1-score of 0.95, reflecting consistent and reliable performance across all Wagner grades. The optimal model configuration, which integrates ResNet-50 with ROS, significantly improves prediction accuracy for underrepresented classes, particularly Grades 1 and 5. Moreover, the classifier demonstrates stable recognition of rare categories without signs of overfitting. The incorporation of the Adam optimization algorithm alongside data balancing strategies further strengthens the model's generalization capability. Overall, the results suggest that the proposed framework not only enables accurate automated classification of DFUs but also offers considerable potential to assist objective clinical decision-making and accelerate diagnostic processes, especially in resource-constrained healthcare settings.

Keywords: class imbalance; deep learning; diabetic foot ulcer; random oversampling; ResNet-50; Wagner Grades; medical image classification

How to Cite: A. W. Ramadhan, Pulung Nurtantio Andono, and M. Arief Soeleman, "Comparative Analysis of ResNet-Based Wagner-Scale Classification for Imbalanced DFU Data", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 6, pp. 1428 - 1438, Dec. 2025.
*Permalink/*DOI: https://doi.org/10.29207/resti.v9i6.7016

Received: July 28, 2025
Accepted: December 16, 2025
Available Online: December 30, 2025

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Diabetes mellitus is recognized as a long-standing metabolic disorder that stems from abnormalities in how insulin is produced or how effectively it functions in the body, leading to excessive accumulation of glucose in the blood. Indonesia is currently the world's fifth-largest contributor to diabetes cases, affecting approximately 19.5 million individuals, as reported by the IDF in 2021 [1].Diabetic foot ulcers (DFU) are among the most serio us consequences of diabetes, a condition that may progress to serious infection, sepsis, and eventual lower-limb amputation when adequate treatment is not provided. Globally, DFU represents the leading cause of limb loss unrelated to traumatic injury, with approximately 80% of amputation cases originating from untreated DFU wounds [2]. In Indonesia, the complication rate of DFU is relatively high. A study at Dr. Soetomo General Hospital reported that 68.26% of DFU patients developed sepsis, with a mortality rate of 40.93% [3]. Meanwhile, at Dr. Kariadi Central General Hospital, 36.3% of patients underwent amputation, a figure that increased significantly amid the COVID-19 outbreak [4]. According to Wagner's research [5], [6]. DFU is classified into 6 classes (grade 0 - grade 6), but this approach is subjective, reliant on clinical experience, and prone to inconsistencies. Therefore, technology-based solutions are needed to help the

process of diagnosing and classifying wounds automatically, quickly, and objectively [7], [8]. Conventional machine learning approaches for classifying DFU—such as SVM and KNN—typically focus on binary tasks, distinguishing categories like ischemic vs. non-ischemic or infectious vs. non-infectious, and these methods have reported strong accuracy levels [9]. However, this method has some drawbacks: feature extraction is done manually, processing a large volume of DFU images is time-consuming, and its practical use remains restricted because the classification is still relatively coarse and limited to binary categories [10].

Technological advances such as digital documentation and telemedicine have increased the accuracy of wound classification by up to 35% compared to conventional methods[11]. Image analysis techniques (semantic segmentation) are also used to objectively obtain all information about the wound through every pixel generated by the wound image [12].

The Convolutional Neural Network (CNN) method is a model that is often used in several studies [13] - [16] to automatically detect diabetic foot disease through the classification of foot images. Among these, Thotad et al. [16] utilized a clinical dataset of diabetic foot ulcer images collected from hospital patients, while other studies focused on CNN architectures and their generalization ability in image recognition tasks using public image datasets. The EfficientNet model, which is a development of CNN, is also applied by combining techniques such as residual connections, dropouts, global average pooling, and data augmentation so that it is able to enhance the model's ability to differentiate between healthy feet and diabetic feet [17]. In that study, Almufadi et al. [17] used the Diabetic Foot Ulcer Challenge (DFUC) dataset, which contains ulcer and non-ulcer foot images collected from multiple clinical sources.Deep learning approaches are also applied to detect DFU through several pre-trained models such as AlexNet [18], VGG16/19 [19] - [21], GoogleLeNet [9], ResNet50/101 [19], [20], [22], MobileNetV2 [15], SqueezeNet [23], and DenseNet201 [24]. Most of these models were evaluated on general medical imaging datasets (e.g., histopathology, chest X-ray, or skin lesion datasets) rather than DFU-specific datasets. However, based on several comparative studies, the ResNet model consistently shows the best performance because it can learn complex hierarchical features [19], [25] - [28], overcome accuracy degradation [10], [28] - [30], and mitigate gradient vanishing during training [28], [31].

The limited number of DFU images makes [10] using a few-shot learning approach by leveraging the ResNet-50 architecture and transfer learning. This model aims to classify images of diabetic foot ulcers into three levels of zero, mild, and severe severity automatically and efficiently based on the Wagner scale. In its implementation, the DFU data, which initially only amounted to 1415 images, was multiplied to 3000 through data augmentation using geometric transformations and the addition of random noise. The pre-trained ResNet-50 model is then fine-tuned for retraining. The results of the experiments show that this approach performs well when recognizing wound imagery, making it an effective solution, especially for clinical devices with limited computing resources.

Although the DFU classification approach based on three grades (zero, lightweight, and heavy) performs well, this simplification tends to ignore the clinical complexity inherent in the six-level Wagner Scale classification (grades 0–5) [6]. Each level on that scale has a significant clinical difference and has a direct impact on determining the right therapy. Classification by six classes allows the system to provide a more detailed and medically relevant diagnosis. Although this approach is more complex and faces challenges such as data imbalance and visual similarity between classes, advances in deep learning methods such as class Imbalance, focal loss, and attention mechanism can overcome these barriers. Thus, the classification of six classes based on the Wagner Scale is considered more representative to support accurate clinical decision-making than the generalized approach of the three-class classification.

A range of imbalance-handling techniques has been introduced to mitigate issues arising from unequal class distributions [32], [33], reduce algorithmic bias towards the majority class, as well as increase the sensitivity of the minority class [34]. These methods generally fall into three groups: data-level strategies that focus on resampling, algorithm-level strategies that modify the classification models themselves, and hybrid strategies that integrate elements of both approaches [35]. In a data-level approach, each resampling method works in a different way to address class imbalances [36] Oversampling works by expanding the minority-class samples, typically through methods like the Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE, Random Oversampling (ROS), and SVM-SMOTE [37].

2. Methods

This study uses an experimental design and a quantitative methodology to investigate and evaluate several pre-processing methods for ResNet-based diabetic ulcer image classification using the Wagner Scale [10], [38] as shown in Figure 1.

This method enables an objective evaluation of how clinical data imbalance influences the results, as well as the effectiveness of the pre-processing strategies implemented.

ResNet-50 was chosen because it is one of the stable deep convolutional neural network architectures and has a strong ability to capture hierarchical representations of features [39], [40]. This architecture consists of 50 layers with a residual learning mechanism, specifically, a shortcut connection that enables gradients to pass directly to earlier layers

without diminishing, hence assisting in reducing the vanishing-gradient problem that deep learning frequently faces [40], [41].

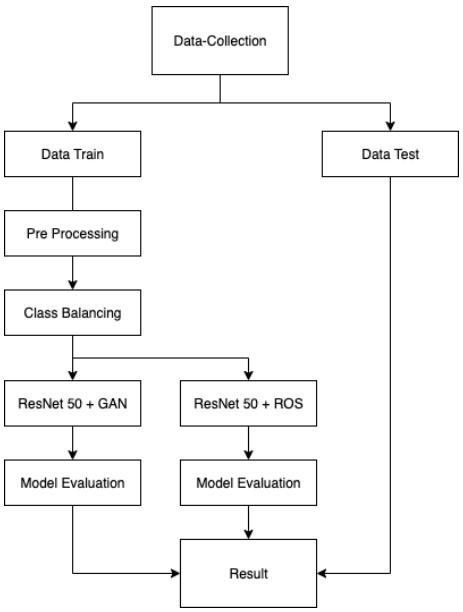


Figure 1. Research Workflow

2.1 Data Collection

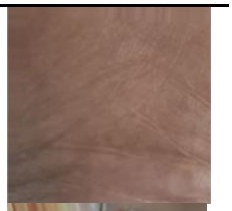
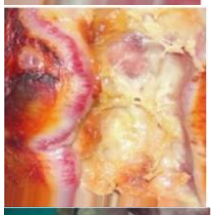
The dataset employed in this study comprised 1,415 images of diabetic foot ulcers, which have been classified based on the Wagner scale into six classes (Grade 0 to Grade 5). The dataset was obtained from a private hospital in Semarang, Indonesia, collected from patient records under ethical approval. Due to patient privacy and data protection regulations, this dataset is not publicly available. This dataset was collected from open dataset sources and the results of collaboration with health service institutions. Each image has gone through a label validation process by competent medical personnel.

In Table 1, image data is collected from 1,415 images of diabetic ulcer wounds and has 6 different classes. Grade 0 consists of 543 images, this grade is marked as no open ulcers but there is a deformity of the leg. Grade 1 consists of 110 images characterized by superficial ulcers that hit the surface layer of skin. Grade 2 consists of 252 images marked as deeper ulcers until they reach tendons and ligaments. Grade 3 consists of 145 images, this grade is an ulcer that is accompanied by an infection (abscess) or osteomyelitis. Grade 4 consists of 293 marked limited local gangrene, usually in just one toe stool. While grade 5 consisted of 72 images characterized by total gangrene of the leg and required amputation.

2.2 Preprocessing

The preprocessing stage aims to improve image quality, allowing models to learn from cleaner and more informative data [42]. In this study, preprocessing begins with the process of normalizing the image and changing the size of the original image to 224 x 224 pixels. Data Normalization is a preprocessing technique used to adjust the scale of data, so that features contribute to model performance [43], [44] The dataset is normalized by converting the pixel value to a range [0,1] This normalization step is performed by scaling the original pixel intensities through division by 255, ensuring that all input values fall within a uniform range and thereby enhancing the model training process's stability.

Table 1. Dataset

Label Name	Private Data	Photo
Grade 0	543	
Grade 1	110	
Grade 2	252	
Grade 3	145	
Grade 4	293	
Grade 5	72	

At this stage, data augmentation is also carried out which aims to create new data to increase the diversity and adequacy of training data so that features are more comprehensive [45], [46]. The augmentation methods applied include zooming the images by up to 20% to improve robustness against resolution changes, horizontal shifts of as much as 10% of the image's width and vertical shifts of up to 10% of its height, thereby simulating minor positional variations of the ulcer.

2.3 Class Balancing

To handle the class imbalance present in the DFU dataset, Random Oversampling (ROS) is applied as the subsequent step. This approach aims to equalize the sample distribution across Wagner scale categories by randomly replicating instances from underrepresented classes within the training data [47]. The oversampling procedure is iteratively performed until the number of samples in each minority class matches that of the majority classes, as formulated in Equation 1.

$$N_{oversample} = N_{max} - N_{minor} \quad (1)$$

$N_{\text{oversample}}$ is the number of additional samples that need to be doubled for the minority class; N_{max} are the number of samples in the majority class ; and N_{minor} are the number of samples in minority classes

Finding the difference between the dominant and minority classes is the first step. Minority-class cases are then randomly chosen and repeatedly added to the training set after the computed difference is utilized to estimate how many samples must be reproduced [48].

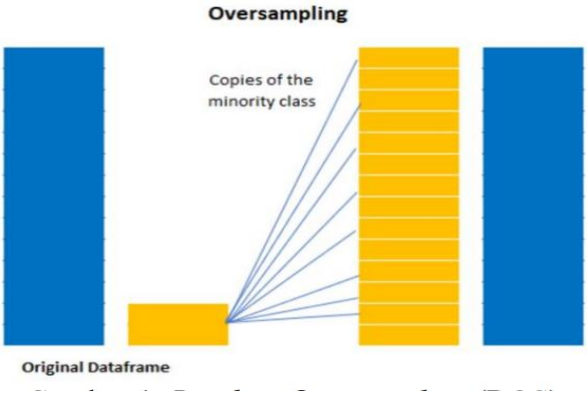


Figure 2. Random Oversampling (ROS)

2.4. Residual Network (ResNet)

Residual Network is a variant of CNN introduced by Kaiming He in 2016 that aims to address CNN performance issues on Deep learing networks. This architecture consists of 50 layers with a residual learning mechanism, namely a shortcut connection that allows the gradient to flow directly to the previous layer without degradation, thus overcoming the problem of vanishing gradient in deep learning [22], [29]. The design of this architecture aims to address challenges commonly encountered during deep learning model training, where processing time is substantial and performance tends to decline as the network depth increases [49]. ResNet addresses this complexity by incorporating skip, or shortcut, connections into the network [22]. Unlike conventional architectures that tend to experience a decrease in accuracy on very deep networks, ResNet is able to maintain performance even though the number of layers increases significantly [50] In addition, computational calculations are lighter and the ability to train the network is better [51]. The ResNet architecture is realized by inserting skip connections across two to three layers that include ReLU activation and batch normalization, as illustrated in Figure 3.

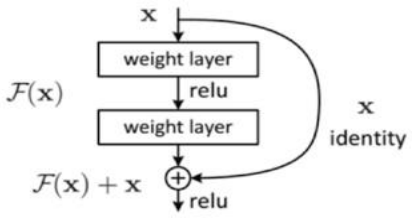


Figure 3. ResNet-50 Architecture

In research [30] these results demonstrate that the ResNet model is quite successful at extracting visual information, outperforming alternative designs in picture classification tasks. By introducing residual learning, He et al. made it possible to apply it again across several layers. Equation 2 defines residual blocks on ResNet [41].

$$y = F\ (x, \{W_I\}) + x \quad (2)$$

Equation 2 is a function of the transformation F to the input x using the parameter {,W-I.}. The result of the function is then summed with the original input x, resulting in the final output y [41]. This summation serves as a residual connection, which helps to overcome the problem of degradation in the deep network, and simplifies the training process by retaining information from the initial input [41]. The next steps to perform a Bottleneck Residual Block as shown in Equation 3.

$$F(x) = W_3.\sigma(W_2.\sigma(W_1.x)) \quad (3)$$

Equation 3 starts by matching the input x with the matrix weight W_1, then the result is passed to the activation function σ (such as ReLU). The output of such activation is then multiplied by the weight W_2, followed by activation σ finally the result is multiplied by the weight W_3 to produce the final output of the transform function $F(x)$ [41].

An overview of the ResNet50 model's hyperparameters, which were chosen as the experiment's training option, is shown in Table 2. The Adam optimizer is used to optimize the model, and the loss function is categorical cross-entropy, which makes it suitable for multi-class classification applications.

In addition to the basic configuration, this study also compares two derivative architectural approaches, namely ResNet-50 + GAN and ResNet-50 + Random Over-Sampling (ROS) as strategies for addressing data imbalance

The Generative Adversarial Network's (GAN) generative power and ResNet-50's feature extraction capabilities are combined in the ResNet-50 + GAN architecture. This approach allows the creation of synthetic imagery that resembles the original data, so that the amount of data in the minority class can be

significantly increased and the model obtains a more diverse variety of training data. Study by[52] states that the combination of ResNet-50 and DC-GAN in the classification of medical images can increase the F1-score by 5.8% compared to pure ResNet-50 [52]. Ribas et al [53] In its survey study, it was stated that the use of Generative Adversarial Network (GAN) in ResNet-based classification architecture was generally able to improve model performance in minority classes, especially in datasets with extreme imbalances. However, they also emphasized that this approach has the drawback of training instability and requires longer training times, especially on datasets with limited sizes [53].

Table 2. Model training process

Parameter	Value	Classic Layers	Resnet50 Layers
Optimizer	Adam	Input Image (200 x 200 x 1)	Input Image (224 x 224 x 1)
Max Epoch	100	Convolutional (3 x 3 x 8)	Convolutional (7 x 7 x 64)
		Batch Normalization	Batch Normalization
		ReLU	ReLU
Validation Frequency	30	Max Pooling	Max Pooling
InitialLearnRate	0.0001	Convolutional (3 x 3 x 16)	Residual Block (64, 64)
		Batch Normalization	(64, 64)
		ReLU	(64, 64)
			Residual Block
MiniBatchSize	64	Convolutional (3 x 3 x 32)	Residual Block (128, 128)
		Batch Normalization	(128, 128)
		ReLU	(128, 128)
			Residual Block
			Residual Block (256, 256)
			(256, 256)
			(256, 256)
			Residual Block
Execution by GPU		Max Pooling	Residual Block (512, 512)
			(512, 512)
			Average Pooling Layer (1 x1)
			Fully Connected Layer (3)
			Softmax
			Classified

Random Over-Sampling (ROS) is one of the data-level balancing techniques widely used to address classroom imbalances prior to convolutional neural network (CNN) training. Senturk & Budak in his research [54] explained that ROS has a major advantage in the form of simplicity of implementation, because it is carried out only in the pre-processing stage without changing the CNN architecture. In addition, ROS is able to increase sensitivity to minority classes and improve overall prediction results when the dataset used has an unbalanced distribution. The same thing was also discovered by Gupta et al. On the study [55] reported that the application of ROS to CNN improved the model's ability to recognize minority classes, resulting in higher recall and F1-score rates than training without oversampling.

Although Generative Adversarial Networks (GAN) are frequently regarded as a more advanced approach for handling class imbalance, in this case Random Over-Sampling (ROS) provides several practical advantages. Compared to GAN, ROS is characterized by its ease of implementation, lower computational cost, and higher training stability, as no adversarial training between a generator and discriminator is needed. We formally address this in the case of GANs by definition, a Generator and Discriminator pair, mathematically. The generator generates fake samples G(z) that are similar to data x, and the discriminator estimates and differentiates real data from the generated ones. The train of the training is cast into min-max optimization, as demonstrated in Equation 4.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (4)$$

On the other hand, GAN-based approaches are capable of producing diverse and realistic synthetic samples, which makes them suitable for extremely imbalanced datasets with very limited minority class representations. Nevertheless, GANs usually demand careful tuning of training parameters and are prone to instability issues, such as mode collapse, which may negatively affect model performance. In this study, the available dataset still contained a sufficient number of samples and the degree of class imbalance was relatively moderate; therefore, the ROS-based approach yielded more stable and accurate results compared to the GAN-based method, which tends to be more effective only in scenarios where the minority class is extremely underrepresented and additional synthetic variability is necessary.

2.5. Evaluation Model

The evaluation of the proposed model was conducted using a confusion matrix that compares the predicted class outcomes with the corresponding ground-truth labels [15]. Performance assessment was carried out using accuracy, precision, recall, and F1-score metrics, all of which were derived from the confusion matrix presented in Table 3.

Table 3. Confusion Matrix

Class	Actual Yes	Actual No
Predict Yes	True Positive (TP)	False Positive (FP)
Predict No	False Negative (FN)	True Negative (TN)

A number of assessment metrics are used to evaluate the model's performance[56]. The right percentage of the prediction is used to compute accuracy. While recall measures how well the model detects all real positive

instances, precision shows the percentage of positive predictions that are truly accurate. By combining recall and precision into a single, cohesive metric, the F1-score offers a fair evaluation.

To overcome this, Random Oversampling is carried out by doubling the sample from the minority class until the number is equal to the class with the largest number. This process is calculated using Equations 5 to 7.

$$Akurasi = \frac{TP+FN}{TP+TN+FP+FN} \quad (5)$$

$$Presisi = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

3. Results and Discussions

3.1 Result

This research is directed to build and test an effective classification model in the context of the problem, namely medical imaging which is specifically focused on the detection and recognition of diabetic ulcer lesions. The model used adopts the ResNet-50 deep learning architecture with a transfer learning approach, which allows the utilization of the weight of the initial training results (pretrained) to improve the efficiency and accuracy of training on a limited dataset. The main goal of this model is to automatically identify and classify diabetic ulcer lesions into six severity levels according to the Wagner scale, namely Grade 0 to Grade 5. This classification is expected to assist medical personnel in assessing the severity of injuries objectively and quickly, thereby accelerating clinical decision-making and supporting more optimal management of diabetes patient care.

The findings presented in Figure 4 show that the image data used in this study consisted of 1,415 images classified into six different classes based on the severity of the injury.

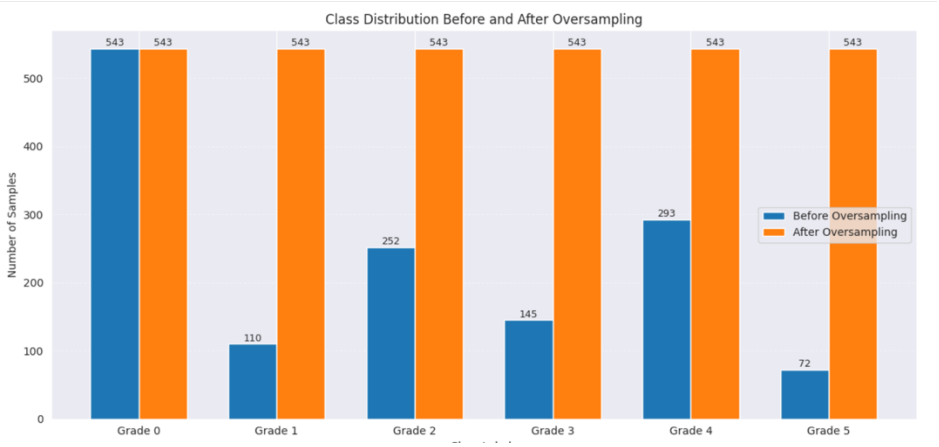


Figure 4. Implementation of Random OverSampling (ROS)

The distribution of data shows a significant imbalance. Grade 0 is the majority class with 543 images, marked by the absence of open ulcers but there is a deformity of the legs. Meanwhile, other classes such as Grade 1 (110 images), Grade 3 (145 images), and Grade 5 (72 images) had a much smaller number of images, indicating that the data falls into the category of imbalanced datasets. This inequality can negatively impact the performance of classification models, especially in detecting minority classes such as Grade 1 and Grade 5. The model tends to be biased towards the majority class and ignores classes with small amounts of data [33], [35]. Therefore, two balancing techniques are applied: SMOTE and Random Over Sampling (ROS) [32].

3.2 Baseline ResNet Model Result

The model was trained for 100 epochs without using EarlyStopping, utilizing the ReduceLROnPlateau method to automatically set the learning rate. The training process showed good convergence, characterized by a decrease in losses and an increase in accuracy in validation data.

In Figure 5, the training curve for ResNetNet-50 without a data balancing method shows that the loss decreased drastically during the first 10 epochs, and then stabilized. However, in the validation accuracy graph (right), there is an inequality between training accuracy and validation. This reflects the potential for overfitting because models tend to learn better against the majority class, but have difficulty generalizing to minority data. Validation accuracy stagnated in the 60–70% range after the 40th epoch, suggesting that the model was not able to recognize patterns in minority classes consistently.

Figure 6 illustrates the training of the ResNet-50 model with data balancing using the Generative Adversarial Network (GAN). Although the model can learn from the GAN's synthetic data, the learning process is not yet completely stable, as evidenced by the validation loss's large fluctuations and the training loss curve's relative

stability. However, the validation accuracy shows a trend of slowly but consistently increasing, reaching a value of around 85%. This shows that the GAN approach is able to improve the performance of the model in recognizing minority data, but the complexity of the GAN model requires further tuning and stabilization.

Figure 7 Training and validation loss curves (left) and corresponding accuracy curves (right). The loss curve shows that the model experienced a consistent decrease in error to a steady decline at around the 30th epoch. The accuracy curve shows a rapid improvement, with validation values reaching an average of 92.7%, reflecting good generalizations.



Figure 5. ResNet-50 curve



Figure 6. ResNet-50 curve Using GAN



Figure 7. ResNet-50 curve using ROS

There is no indication of significant overfitting, as the loss and accuracy of the validation data tend to be parallel to the training data. The classification evaluation is shown in Table 4.

The classification results are mostly correct, with the dominance of correct predictions in the main diagonal. Misclassification mainly occurs between Grade 2 and Grade 3, which is most likely due to the similarity of the visual characteristics of the wound in those classes. The fixed model is able to recognize minority classes such as Grade 1 and Grade 5 with fairly good accuracy as shown in Table 5.

The diabetic ulcer imagery dataset used has an unbalanced class distribution. To overcome this, a

combination of techniques is used, namely Data augmentation including rotation, flipping, zoom, and lighting. Furthermore, Oversampling is carried out to balance the number of minority class samples.

Table 4 Performance Comparison of ResNet-50 Model with and without Imbalance Handling Techniques

	Accuracy	Precision	Recall	F1-Score
Resnet 50	0.71	0.71	0.69	0.66
Resnet 50 + GAN	0.85	0.85	0.85	0.84
Resnet 50 + ROS	0.96	0.96	0.95	0.95

Table 5 Performance Comparison of ResNet-50, ResNet-50 + GAN, and ResNet-50 + ROS by Grade

Grade	Resnet 50	Resnet 50 + GAN	Resnet 50 + ROS
0	0.97	0.86	1.00
1	0.40	0.94	0.98
2	0.48	0.80	0.90
3	0.45	0.94	0.98
4	0.60	0.70	0.98
5	1.00	0.86	0.98

No substantial signs of overfitting are observed, as the validation loss and accuracy trends closely follow those of the training phase. The comprehensive classification outcomes are presented in Table 4, demonstrating that the incorporation of class imbalance mitigation techniques significantly enhances the performance of the ResNet-50 model across Precision, Recall, and F1-score metrics. Specifically, the standard ResNet-50 model achieves an F1-score of 0.66, while the integration of GAN-based augmentation and Random Oversampling improves the F1-score to 0.84 and 0.95, respectively.

To better understand the source of this improvement, Table 5 presents a detailed breakdown of the model performance per class. As shown in Table 5, the fixed models (GAN and ROS) produce substantially higher scores across minority classes such as Grade 1 and Grade 5, where the baseline ResNet-50 only achieved 0.40 and 1.00, while the ROS-based model reached 0.98 for both grades. These results support the aggregated performance values reported in Table 4 and demonstrate that the improvement is not limited to the majority class (e.g., Grade 0), but also occurs across minority classes. Misclassification mainly occurs between Grade 2 and Grade 3, most likely due to the similarity of visual characteristics and overlapping feature distributions in these classes. The diabetic ulcer dataset used in this study is highly imbalanced. To address this issue, a combination of data augmentation (rotation, flipping, zooming, and lighting variations) and oversampling has been applied. The evaluation shows that this approach is effective, since there is no noticeable decline in the performance of the minority classes. In fact, the highest F1-Score values were obtained in Grade 0 (1.00) and Grade 5 (0.98), while the lowest performance appeared in Grade 2 (0.90) due to visual similarities with Grade 3.

3.3 Model Optimization and Data Balancing Result

This research combines model optimization strategies with data balancing methods to improve overall classification performance. The integration of these techniques not only speeds up the training process but also enhances the model's generalization capability when applied to previously unseen data. In the optimization stage, the ResNet-50 model is trained using the Adam (Adaptive Moment Estimation) optimizer, which is commonly recognized as a highly effective optimization algorithm in deep learning applications [30], [57], [58]. By utilizing estimations of both the first-order moment (mean of the gradients) and the second-order moment (variance of the gradients), the Adam optimizer is able to automatically adjust the learning rate. This mechanism facilitates faster convergence while minimizing the likelihood of the optimization process becoming trapped in local minima [30], [57], [58]. As a result, the model can attain high levels of accuracy within a relatively short and stable training duration.

The distribution of classes in the dataset shows a considerable imbalance, where majority classes such as Grade 0 have a much larger number of images than minority classes such as Grade 1 and Grade 5. This imbalance can introduce classification bias. To address it, random oversampling is performed by repeatedly replicating minority-class samples until their quantity equals that of the majority class. Figure 2 shows the data distribution both prior to and following the oversampling procedure.

The augmentation technique is carried out by producing new variations of the original image using transformations such as rotation (20°), zoom (20%), shear (10%), shifting (horizontal and vertical), and flipping. This aims to increase the diversity of training data and help the model be more resilient to visual variation in the real world.

After augmentation, the distribution of data still shows inequality between classes. Therefore, random oversampling is carried out, which is the random addition of copies of data from the minority class until the number is equivalent to the majority class [55]. Figure 2 shows the initial unequal distribution of data, and after oversampling shows an even distribution of class data. The implementation of this strategy showed a significant improvement in classification performance, especially in classes with a small amount of initial data. The F1-score for Grade 1 and Grade 5 reached 0.97 and 0.99 respectively, close to the performance of the majority class. This suggests that this approach has succeeded in reducing classification bias and improving minority class representation, which is especially important in the context of uneven medical data.

The findings of this study indicate that the ResNet-50 model, when combined with transfer learning, performs highly effectively in multi-class classification tasks

involving limited medical imaging datasets. The high and stable F1-score across the class shows that the model can generalize well, even without regulatory techniques such as early stopping. Compared to previous studies that used standard CNN for the classification of diabetic ulcer lesions, this approach provides a significant improvement in accuracy and resistance to overfitting. Some previous studies reported an accuracy of around 85–88%, while the model in this study achieved an average accuracy of 97%. One of the main strengths of this model is the ability to classify minority classes, such as Grade 1 and Grade 5, thanks to oversampling and data augmentation strategies. This is important because in clinical practice, it is often the minority class that needs special attention the most.

However, there are some limitations that need to be noted. One of them is the visual similarity between classes, which causes ambiguity in classification. A potential solution to this challenge is to integrate additional clinical features (e.g. wound size or healing time) or to use model interpretability techniques such as Grad-CAM.

The implication of this study is that AI-based systems like this can be a reliable diagnostic tool in clinical settings, particularly in areas with limited medical resources. With further integration into clinical applications, these models can accelerate the wound triage process and improve therapeutic decision-making.

4. Conclusions

This research successfully developed a diabetic foot ulcer wound image classification model based on the Wagner grading system by utilizing the ResNet-50 architecture combined with a transfer learning strategy. The proposed model is capable of identifying and categorizing wound images into six severity levels, ranging from Grade 0 to Grade 5.

Based on the training and evaluation outcomes, the model exhibited excellent performance, achieving an average validation accuracy of 97%. Additional evaluation using precision, recall, and F1-score metrics produced a weighted F1-score of 0.97, demonstrating the model's reliability and consistency in multi-class classification tasks. The highest F1-scores were observed for Grade 0 (1.00) and Grade 5 (0.99). In contrast, the lowest precision value was recorded for Grade 2 (0.90), which may be attributed to visual similarities between Grade 2 and Grade 3 wound characteristics.

The application of data augmentation and oversampling techniques has proven to be effective in overcoming class distribution imbalances, so that the model is able to classify minority classes such as Grade 1 and Grade 5 optimally. In addition, the model showed stable convergence during training, without overfitting, although it did not use the early stopping.

Overall, the findings of this study indicate that the ResNet-50 architecture demonstrates strong potential for integration into artificial intelligence–based systems to support automated diagnosis of diabetic foot ulcer wounds.. Further research can be developed by adding non-visual features, expanding the variation of the dataset across institutions, and applying model interpretability techniques such as Grad-CAM to increase clinical confidence in the developed system.

Acknowledgements

The author would like to thank Universitas Dian Nuswantoro (UDINUS) and all lecturers in the Master of Informatics Engineering Study Program for the guidance, direction, and support provided throughout this research process. The author would like to express sincere appreciation to the Beasiswa Unggulan program under the Ministry of Research, Technology, and Higher Education (Kemenristek Dikti) for the financial support that enabled the completion of this study. The author also extends gratitude to all parties who contributed, either directly or indirectly, to the successful execution of this research.

References

[1] International Diabetes Federation, "IDF Diabetes Atlas," 2021, *Brussels, Belgium*.

[2] H. Zhang, C. Huang, J. Bai, and J. Wang, "Effect of diabetic foot ulcers and other risk factors on the prevalence of lower extremity amputation: A meta-analysis.," *Int Wound J*, vol. 20, no. 8, pp. 3035–3047, Oct. 2023, doi: 10.1111/iwj.14179.

[3] A. S. Hariftyani, H. Novida, and M. Edward, "Profile of Diabetic Foot Ulcer Patients at Tertiary Care Hospital in Surabaya, Indonesia," *Jurnal Berkala Epidemiologi*, vol. 9, no. 3, p. 293, Sep. 2021, doi: 10.20473/jbe.V9I32021.293-302.

[4] E. Yunir *et al.*, "Characteristics of Diabetic Foot Ulcer Patients Pre- and During COVID-19 Pandemic: Lessons Learnt From a National Referral Hospital in Indonesia," *J Prim Care Community Health*, vol. 13, Jan. 2022, doi: 10.1177/21501319221089767.

[5] L. A. Lavery, D. G. Armstrong, and L. B. Harkless, "Classification of diabetic foot wounds.," *J Foot Ankle Surg*, vol. 35, no. 6, pp. 528–31, 1996, doi: 10.1016/s1067-2516(96)80125-6.

[6] P. Shah *et al.*, "Wagner's Classification as a Tool for Treating Diabetic Foot Ulcers: Our Observations at a Suburban Teaching Hospital," *Cureus*, Jan. 2022, doi: 10.7759/cureus.21501.

[7] C. Lauri, A. Leone, M. Cavallini, A. Signore, L. Giurato, and L. Uccioli, "Diabetic Foot Infections: The Diagnostic Challenges.," *J Clin Med*, vol. 9, no. 6, Jun. 2020, doi: 10.3390/jcm9061779.

[8] P. Sivakamasundari and G. Niranjana, "An automatic detection and classification of diabetic foot ulcers using chebyshev chaotic ladybug beetle optimized extended swin transformer-inceptionV3 model," *Biomed Signal Process Control*, vol. 110, p. 108268, Dec. 2025, doi: 10.1016/j.bspc.2025.108268.

[9] J. Yogapriya, V. Chandran, M. G. Sumithra, B. Elakkiya, A. Shamila Ebenezer, and C. Suresh Gnana Dhas, "Automated Detection of Infection in Diabetic Foot Ulcer Images Using Convolutional Neural Network," *J Healthc Eng*, vol. 2022, pp. 1–12, Apr. 2022, doi: 10.1155/2022/2349849.

[10] 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," *Sci Rep*, vol. 14, no. 1, p. 29877, Dec. 2024, doi: 10.1038/s41598-024-80691-w.

[11] C. Shi, J. C. Dumville, H. Juwale, C. Moran, and R. Atkinson, "Evidence assessing the development, evaluation and implementation of digital health technologies in wound care: A rapid scoping review," *J Tissue Viability*, vol. 31, no. 4, pp. 567–574, Nov. 2022, doi: 10.1016/j.jtv.2022.09.006.

[12] G. Scebba *et al.*, "Detect-and-segment: A deep learning approach to automate wound image segmentation," *Inform Med Unlocked*, vol. 29, Jan. 2022, doi: 10.1016/j.imu.2022.100884.

[13] G. Rangel, J. C. Cuevas-Tello, J. Nunez-Varela, C. Puente, and A. G. Silva-Trujillo, "A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks," *J Sens*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/2797320.

[14] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis.," *Multimed Syst*, vol. 28, no. 6, pp. 2335–2355, 2022, doi: 10.1007/s00530-022-00960-4.

[15] Achmad Naila Muna Ramadhani, Galuh Wilujeng Saraswati, Rama Tri Agung, and Heru Agus Santoso, "Performance Comparison of Convolutional Neural Network and MobileNetV2 for Chili Diseases Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 4, pp. 940–946, Aug. 2023, doi: 10.29207/resti.v7i4.5028.

[16] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Diabetic foot ulcer detection using deep learning approaches," *Sensors International*, vol. 4, p. 100210, 2023, doi: 10.1016/j.sintl.2022.100210.

[17] N. F. Almufadi, H. F. Alhasson, and S. S. Alharbi, "E-DFu-Net: An efficient deep convolutional neural network models for Diabetic Foot Ulcer classification," *Biomolecules and Biomedicine*, vol. 25, no. 2, pp. 445–460, Jan. 2025, doi: 10.17305/bb.2024.11117.

[18] N. W. Asnake, A. M. Ayalew, and A. A. Engda, "Detection of oral squamous cell carcinoma cancer using AlexNet on histopathological images," *Discover Applied Sciences*, vol. 7, no. 3, p. 155, Feb. 2025, doi: 10.1007/s42452-025-06514-3.

[19] D. Theckedath and R. R. Sedamkar, "Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks," *SN Comput Sci*, vol. 1, no. 2, p. 79, Mar. 2020, doi: 10.1007/s42979-020-0114-9.

[20] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.

[21] T. Fatima and H. Soliman, "Application of VGG16 Transfer Learning for Breast Cancer Detection," *Information*, vol. 16, no. 3, p. 227, Mar. 2025, doi: 10.3390/info16030227.

[22] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah, I. N. Yulita, and M. H. Asnawi, "A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image," *Applied Sciences*, vol. 13, no. 24, p. 13111, Dec. 2023, doi: 10.3390/app132413111.

[23] D. W. Ayu and G. A. Pradipta, "SqueezeNet Feature Extraction dan Gradient Boosting untuk Klasifikasi Penyakit Monkeypox pada Citra Kulit," *Jurnal Sistem dan Informatika (JSI)*, vol. 18, no. 2, pp. 177–183, May 2024, doi: 10.30864/jsi.v18i2.612.

[24] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J Biomol Struct Dyn*, vol. 39, no. 15, pp. 5682–5689, Oct. 2021, doi: 10.1080/07391102.2020.1788642.

[25] L. Wen, Z. Xiao, X. Xu, and B. Liu, "Disaster Recognition and Classification Based on Improved ResNet-50 Neural Network," *Applied Sciences*, vol. 15, no. 9, p. 5143, May 2025, doi: 10.3390/app15095143.

[26] M. N. Musa, "MRI-Based Brain Tumor Classification using ResNet-50 and Optimized Softmax Regression," *JURNAL INFOTEL*, vol. 16, no. 3, Sep. 2024, doi: 10.20895/infotel.v16i3.1175.

[27] L. Zhang, Y. Bian, P. Jiang, and F. Zhang, "A Transfer Residual Neural Network Based on ResNet-50 for Detection of Steel Surface Defects," *Applied Sciences*, vol. 13, no. 9, p. 5260, Apr. 2023, doi: 10.3390/app13095260.

[28] T. A. Sugiarto, M. A. Soeleman, and P. Pujiono, "Enhancing Augmentation-Based ResNet50 for Car Brand Classification," *Journal of Applied Intelligent System*, vol. 8, no. 3, pp. 400–412, 2023.

[29] S. R. Klomp, R. G. J. Wijnhoven, and P. H. N. de With, "Performance-Efficiency Comparisons of Channel Attention Modules for ResNets," *Neural Process Lett*, vol. 55, no. 5, pp. 6797–6813, Oct. 2023, doi: 10.1007/s11063-023-11161-z.

[30] C. Raptis, E. Karavasilis, G. Anastasopoulos, and A. Adamopoulos, "Comparative Analysis of Conventional CNN v's ImageNet Pretrained ResNet in Medical Image Classification," *Information*, vol. 15, no. 12, p. 806, Dec. 2024, doi: 10.3390/info15120806.

[31] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau, "Stable ResNet," Mar. 2021.

[32] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, p. 54, Jan. 2023, doi: 10.3390/info14010054.

[33] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif Intell Rev*, vol. 57, no. 6, p. 137, May 2024, doi: 10.1007/s10462-024-10759-6.

[34] C. de Alvis and S. Seneviratne, "A Survey of Deep Long-Tail Classification Advancements," Apr. 2024.

[35] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," *J Big Data*, vol. 12, no. 1, p. 71, Mar. 2025, doi: 10.1186/s40537-025-01119-4.

[36] A. Riyandi, Mahazam Afrad, M Yoka Fathoni, and Yogo Dwi Prasetyo, "Obesity Status Prediction Through Artificial Intelligence and Balanced Label Distribution Using SMOTE," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 3, pp. 519–524, Jun. 2025, doi: 10.29207/resti.v9i3.6204.

[37] "A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 3, pp. 345–361, Jun. 2023, doi: 10.22266/ijies2023.0630.28.

[38] G.-X. Zhou *et al.*, "Construction and validation of a deep learning-based diagnostic model for segmentation and classification of diabetic foot," *Front Endocrinol (Lausanne)*, vol. 16, Apr. 2025, doi: 10.3389/fendo.2025.1543192.

[39] C.-L. Lin and K.-C. Wu, "Development of revised ResNet-50 for diabetic retinopathy detection," *BMC Bioinformatics*, vol. 24, no. 1, p. 157, Apr. 2023, doi: 10.1186/s12859-023-05293-1.

[40] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-Maadeed, "Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches," Jun. 2022, doi: 10.1016/j.cosrev.2024.100645.

[41] B. Mahaur, K. K. Mishra, and N. Singh, "Improved Residual Network based on norm-preservation for visual recognition," *Neural Networks*, vol. 157, pp. 305–322, Jan. 2023, doi: 10.1016/j.neunet.2022.10.023.

[42] Aulia Afifah, E. R. Arumi, M. Maimunah, and Setiya Nugroho, "Comparative Evaluation of Preprocessing Methods for MobileNetV1 and V2 in Waste Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 3, pp. 444–452, May 2025, doi: 10.29207/resti.v9i3.6211.

[43] I. Corley, C. Robinson, R. Dodhia, J. M. L. Ferres, and P. Najafirad, "Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters," May 2023.

[44] X. Pei *et al.*, "Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences," *Mater Des*, vol. 232, p. 112086, Aug. 2023, doi: 10.1016/j.matdes.2023.112086.

[45] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature

review.," *Comput Biol Med*, vol. 152, p. 106391, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106391.

[46] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review," *J Imaging*, vol. 9, no. 4, p. 81, Apr. 2023, doi: 10.3390/jimaging9040081.

[47] G. Eom and H. Byeon, "Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique," *Mathematics*, vol. 11, no. 16, p. 3605, Aug. 2023, doi: 10.3390/math11163605.

[48] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front Digit Health*, vol. 6, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.

[49] I. Bello *et al.*, "Revisiting ResNets: Improved Training and Scaling Strategies," Mar. 2021.

[50] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," *Procedia Comput Sci*, vol. 179, pp. 423–431, 2021, doi: 10.1016/j.procs.2021.01.025.

[51] A. Thakur, H. Chauhan, and N. Gupta, "Efficient ResNets: Residual Network Design," Jun. 2023.

[52] L. Ma, R. Shuai, X. Ran, W. Liu, and C. Ye, "Combining DC-GAN with ResNet for blood cell image classification," *Med Biol Eng Comput*, vol. 58, no. 6, pp. 1251–1264, Jun. 2020, doi: 10.1007/s11517-020-02163-3.

[53] L. C. Ribas, W. Casaca, and R. T. Fares, "Conditional Generative Adversarial Networks and Deep Learning Data Augmentation: A Multi-Perspective Data-Driven Survey Across Multiple Application Fields and Classification Architectures," *AI*, vol. 6, no. 2, p. 32, Feb. 2025, doi: 10.3390/ai6020032.

[54] W. M. Thuzar and M. Moe Htay, "Ensemble Learning for Breast Cancer Image Classification Using Pre-trained CNNs and Random Oversampling Method," in *2024 5th International Conference on Advanced Information Technologies (ICAIT)*, 2024, pp. 1–6. doi: 10.1109/ICAIT65209.2024.10754947.

[55] E. Y. Puspaningrum, R. R. Wahid, R. P. Amaliyah, and C. Nisa', "Alzheimer's Disease Stage Classification using Deep Convolutional Neural Networks on Oversampled Imbalance Data," in *2020 6th Information Technology International Seminar (ITIS)*, 2020, pp. 57–62. doi: 10.1109/ITIS50118.2020.9321061.

[56] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020.

[57] M. Uppal *et al.*, "Enhancing accuracy in brain stroke detection: Multi-layer perceptron with Adadelta, RMSProp and AdaMax optimizers," *Front Bioeng Biotechnol*, vol. 11, Sep. 2023, doi: 10.3389/fbioe.2023.1257591.

[58] S. Shedriko and M. Firdaus, "Perbandingan Optimizer Adagrad, Adadelta dan Adam dalam Klasifikasi Gambar Menggunakan Deep Learning," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 8, no. 1, p. 103, Aug. 2023, doi: 10.30998/string.v8i1.16564.