Explainable Deep Learning for Diabetic Foot Ulcer Diagnosis: A Comparative CNN Framework

Rania Kadhim, Ahmad Lateef*, Mohammed Kamil

College of Science, Mustansiriyah University, Baghdad, Iraq

Received 01 June 2026; revised 08 August 2026; accepted 10 August 2026

DOI: https://doi.org/10.46604/aiti.2026.16496

Abstract

This study aims to develop and evaluate an automated deep learning framework for diabetic foot ulcer (DFU) classification to reduce reliance on manual visual inspection. Five transfer-learning architectures (EfficientNetB0, ResNet50, VGG16, InceptionV3, and MobileNetV2) are evaluated on 1,055 localized, augmented foot images using stratified 5-fold cross-validation. The ResNet50 model achieves excellent diagnostic results, with a mean F1-score of 0.997 ± 0.004 and a mean ROC-AUC of 1.000 ± 0.00001 in all the folds. These results demonstrate the potential of the proposed explainable deep learning framework for automated DFU classification. However, the lack of patient-level information in the dataset used may lead to data leakage due to image-level splitting. Therefore, validation of this explainable model on external clinical cohorts is highly recommended for reliable and safe use.

Keywords: deep learning, convolutional neural networks, transfer learning, explainable AI, medical image analysis

1. Introduction

Diabetes mellitus is a highly prevalent metabolic syndrome that has become a major global health concern. At present, the number of people aged 20–79 years living with this disorder exceeds 400 million, while epidemiological predictions report a striking rise to 780 million by 2046 [1]. Diabetes can also cause a wide range of other complications in the body, such as cardiovascular disease, nephropathy, and retinopathy, as well as glycemic imbalance [2]. Among the above complications, diabetic foot ulcers (DFUs), which are chronic and complex wounds that typically develop on the foot, are particularly damaging consequences of diabetes. Due to a lack of regular observation and effective treatment, DFUs evolve and often require complicated surgeries, ending in limb amputations [3]. According to epidemiological data, about 25% of people suffering from diabetes will get a DFU during their lifetime [4]. These ulcers cause more than a million amputations around the world each year. This burden falls disproportionately on the shoulders of developing countries due to poor access to podiatry preventative healthcare services [5]. The development of such ulcers is due to chronic hyperglycemia, which contributes to microvascular damage and peripheral neuropathy [6]. Although there have been significant advances in the treatment strategies available, it can be a challenge to prevent the progression of DFUs in this long-term condition.

Diagnostic delays are partly attributable to the complex and variable visual appearance of ulcer tissues during the evolution process, which makes the manual assessment highly subjective and liable to clinical discrepancy [7]. With the success of computer vision in medical imaging, the use of automated diagnostic systems for DFU screening is an important and essential step to standardizing patient care. The high number of patients who are not diagnosed or diagnosed too late, due to limited access to specialists and low health literacy in resource-limited countries, also contributes to this pressure.

* Corresponding author. E-mail address: a97s21@uomustansiriyah.edu.iq

In recent studies, more AI-based methods have been used to automate DFU detection. For instance, Cruz Vega et al. [8] applied the thermal imaging method with CNNs, and Khandakar et al. [9] adopted hand-crafted features with traditional machine learning methods. Munadi et al. [10] investigated the integration of thermal images with clinical data using the concept of decision fusion. For standard optical imaging, Goyal et al. proposed DFUNet for RGB images with a 93.9% F1 score [11], and Ahsan et al. [12] showed the efficiency of ResNet models for DFUs using the DFUC2020 dataset. Despite these recent advances, several methodological gaps remain in the existing literature: (i) the use of thermal or special imaging devices, which make high-performance models difficult to use in normal primary care settings, (ii) the low quality of feature extraction, which is not able to determine the texture characteristics of ulcers corresponding to the pathological conditions, and (iii) lack of systematic comparisons between the heavy models of contemporary times and the accurate deep CNNs based on standard RGB imaging.

To bridge these important gaps, this study introduces a systematic automated framework for DFU analysis based on five convolutional neural networks (CNNs): EfficientNetB0, ResNet50, VGG16, InceptionV3, and MobileNetV2. This framework uses a transfer learning technique with a customized classification head trained and tested on a wide range of images (1,055 images localized in RGB cameras). Although deep learning methods have been increasingly applied to DFU classification, most existing studies have primarily reported performance metrics such as F1-score without assessing statistical significance or model interpretability. The key strength of this work is thus not so much in a new architecture, but in providing a statistically valid and clinically meaningful assessment system. That is, the core impacts are:

(1) Rigorous statistical validation: Rather than sticking to traditional single-train-validation splits, this study follows a stratified 5-fold cross-validation approach. All clinical evaluation parameters (sensitivity, specificity, PPV, NPV, F1-score, ROC-AUC, and PR-AUC) are presented with 95% confidence intervals (CIs).

(2) Statistical significance test for architectures: This study goes further than simple comparison of performance metrics using the paired statistical McNemar's test. It allows us to obtain a formally proven mathematical significance of the diagnostic superiority of the chosen architecture over other widely used benchmark architectures.

(3) Structured clinical explainability: To demystify the "black box" nature of deep learning, Grad-CAM visual analysis is employed in this study. Through explicit visualization of true positives, true negatives, false positives, and false negatives, this study provides clinicians with a morphologic picture of successful and unsuccessful diagnoses, such as slough, granulation tissue, and erythema.

The rest of this paper is organized as follows. In Section 2, a comprehensive review of the literature on the diagnosis of DFUs using artificial intelligence is carried out. The methodology followed is explained in Section 3 and involves the description of the data sets, the augmentation process, and modeling techniques. The outcomes of the experiments and their analysis are shown in Section 4, along with their clinical interpretation, while the conclusions of this study are described in Section 5.

2. Related Works

As can be seen from current studies in the field of medical science, epidemiological surveillance and pattern recognition techniques have become very popular tools in solving urgent problems of public health. Due to the increasing number of cases of DFUs in the world, it has become a topical issue for researchers and physicians to apply AI and ML algorithms to develop reliable diagnostic approaches [13]. Early identification of DFUs is a clinically urgent task since timely treatment provides a high chance of recovery and significantly reduces the chances of amputation. Therefore, AI technology offers a promising solution for developing fast, accurate, and inexpensive screening methods.

One of the early approaches to automated DFU classification involved analyzing thermal images. Cruz-Vega et al. [8] were pioneers in using CNNs to classify thermograms as healthy or ulcerated plantar surfaces. Khandakar et al. [9] also developed an automated diagnostic method based on thermal images, handcrafted features, and traditional classifiers. Munadi et al. [10] addressed this problem and created an approach based on a decision-fusion DL algorithm that used patients' data and thermal images. However, the use of thermograms involves several translation issues. Thermal cameras are costly and not readily available in low-resource settings such as primary care, especially in developing countries where optical imaging is the current standard [6]. Moreover, manually extracted features [9] may miss important, complex morphological features [6], and multimodal fusion approaches [10] may increase computational complexity.

To overcome the limitations of using specialised equipment, research has been increasingly focused on standard RGB optical imagery. In this area, Goyal et al. [11] compiled one of the most popular datasets in the field, called DFU Part A, and reported an F1-score of 93.9% and an area under the curve (AUC) of 96.1% using their proposed DFUNet. While following this optical imaging paradigm, Ahsan et al. [12] tested several deep learning architectures on the DFUC2020 dataset, showing that deep residual networks (ResNet) outperform other architectures on this dataset in every case. In a classical ML approach, Wang et al. [14] used a two-phase support vector machine (SVM) pipeline to examine the color and texture features of the standard images for ulcer border detection, which resulted in a 73.3% sensitivity and 94.6% specificity.

Remote DFU monitoring and integration of mobile health (mHealth) is another field that has gained momentum due to the ubiquity of smartphones. Niri et al. [15] introduced a tissue classification and segmentation system based on an iterative linear clustering algorithm with 92.68% accuracy and a Dice similarity coefficient (DSC) of 75.74% on a small dataset of 219 images. Amin et al. [16] experimented with a custom architectural design, where 16 layers CNN was used as a feature extractor for various terminal algorithms and found that Decision Tree and Softmax classifiers worked best in terms of predictive performance. Recently, Das et al. [7] designed ResKNet, a special type of residual network comprising separate functional models: Res4Net for ischemia detection (with an accuracy of 97.8% and F1-score of 97.8%) and Res7Net for diagnosis of infection (with an accuracy of 80.0%). In addition to pure image analysis, Hyun et al. [17] added customized sensor systems, combined with NeuralProphet forecasting and ML ensemble, to confirm DFU progression.

Despite these advances in technology, there is a need for cautious optimism to translate these models into reliable clinical tools. However, the clinical validity of such uncontrolled, smartphone-captured images for final clinical diagnosis remains a concern. Van Netten et al. [18] raised important concerns about the reliability of such images for definitive clinical diagnosis, highlighting the requirement for very strong feature extraction. Additionally, although high theoretical accuracy is shown in the literature, most of the above-mentioned studies are based on a single split of the data for testing purposes, and the results presented are not easily understood. These limitations highlight the need to leverage practical and accessible RGB imaging while at the same time placing a premium on strict cross-validation and clinical explainability, so that models are not just accurate in an abstract context but useful in clinical practice.

3. Materials and Methods

The implementation was developed using TensorFlow 2.10 and executed on a system equipped with an NVIDIA GeForce RTX 4050 Laptop GPU with 6 GB VRAM, 16 GB system RAM, and a 477 GB SSD. This section presents different architectures of deep convolutional neural networks for the proposed framework to diagnose DFUs from RGB images. The evaluated CNN architectures were EfficientNetB0, InceptionV3, MobileNetV2, ResNet50, and VGG16. These pretrained CNN architectures were used to deploy transfer learning techniques to extract visually discriminative and rich features. Finally, the visual patterns were classified using a softmax layer. The key elements of the proposed framework are discussed in the subsections.

3.1 Dataset description

The scarcity of public image databases for DFUs led to the utilization of a dataset originally collected and described in [19]. The primary dataset consists of macroscopic images captured ethically from patients at Al-Nasiriyah Hospital, Nasiriyah, Iraq, where all subjects provided informed consent for their anonymized images to be used in academic research. The initial dataset consisted of 754 high-resolution images obtained by a Samsung Galaxy Note 8 and an iPad under different illumination levels and clinical views. Images were selected if DFUs were visible or if healthy foot tissue was visible, and excluded when severe motion blur or lighting problems completely obscured the foot tissue morphology. Due to strict anonymization procedures in the public release, no information about patient identities and the actual number of unique patients is available.

A lot of pre-processing was performed on the first images to create a strong set of training data and prevent the possibility of learning some sort of irrelevant patterns from the surrounding background. As wide-field images might have clinical artifacts (e.g., rulers, gloves, medical tools, or sheets), these objects can be used by deep learning networks as shortcuts in their learning process. To avoid these biases and to increase the amount of data for training, the whole image was manually annotated and split into smaller non-overlapping patches. This splitting procedure enabled isolating the region of interest (ROI) – healthy or ulcerated skin tissue – excluding irrelevant background artifacts and letting the network learn only pathologies. The resulting patches were labeled with ‘Healthy’ and ‘Ulcer’ labels. Fig. 1 shows an example of the localized patches obtained from the initial images. Dataset characteristics are described in Table 1.

Fig. 1 Sample of patches

Table 1 Summary of the DFU dataset characteristics

Characteristic	Description / Value
Source Location	Al-Nasiriyah Hospital, Nasiriyah, Iraq [19]
Capture Devices	Samsung Galaxy Note 8; Apple iPad
Original Image Count	754 full-frame images
Preprocessing Strategy	Non-overlapping patch extraction to isolate tissue and remove background artifacts.
Total Patches (Images)	1,055
Ulcer Class Count	512 patches
Healthy Class Count	543 patches
Class Balance	49% Ulcer / 51% Healthy
Patch Resolution	224×224 pixels

3.2 Data partitioning

As opposed to using the single hold-out method, the performance of the model was rigorously validated using the 5-fold cross-validation procedure on the whole dataset (N = 1055). In each fold, the dataset was dynamically divided in such a way that 80% (n = 844) of the dataset acted as the training set, while the remaining 20% (n = 211) acted as the test set. Since the

size of the dataset used for training deep convolutional neural networks is fairly small, data augmentation was applied only to the training dataset in each fold.

3.3 Data augmentation

To reduce the likelihood of overfitting and enhance the ability of the model to generalize, the training set was increased by means of data augmentation performed on-the-fly. The augmentation strategy included changes at both the spatial and the pixel levels. Spatial transformations consisted of randomly rotating images by 15°, translating them in the plane by 10% horizontally and vertically, and flipping them. On the other hand, photometric variance was achieved through Gaussian noise ($\sigma = 0.09$) and increasing/decreasing the contrast of the whole image by ±10%. A full set of parameters for data augmentation is provided in Table 2.

3.4 Methodology

The reproducibility of the experiment is supported by the precise definition of all experimental settings, hardware, and hyperparameters. During training, the pre-trained weights were frozen from the ImageNet dataset, and only the classification layer was fine-tuned. This feature extraction approach helps prevent the rapid overfitting observed on medical datasets. Detailed information about the hyperparameter settings is provided in Table 2.

Table 2 Experimental setup, hyperparameter configuration, and the augmentation layer parameters

Parameter / Setup	Value / Description
Input Resolution	224 × 224 × 3 pixels
Backbone Weights	Pre-trained on ImageNet
Freezing Schedule	Base architectures completely frozen (feature extraction only)
Classification Head	GlobalAveragePooling2D → BatchNormalization → Dropout (Rate = 0.4) → Dense (Softmax)
Optimizer	Adam (initial learning rate = 0.001)
Learning Rate Schedule	ReduceLROnPlateau (Factor = 0.2, Patience = 3 epochs, Monitor = val_loss)
Batch Size	32
Maximum Epochs	50
Early Stopping	Patience = 6 epochs (Monitor = val_loss, Restores best weights)
Random Seed	42 (Applied to NumPy, TensorFlow, and K-Fold splits)
Hardware	NVIDIA GeForce RTX 4050 Laptop GPU (6 GB)
Software & Frameworks	Python 3.10, TensorFlow/Keras 2.10.0, Scikit-Learn
RandomRotation	0.15
RandomTranslation	height_factor = 0.1, width_factor = 0.1
RandomFlip	-
GaussianNoise	0.09
RandomContrast	0.1

A diverse suite of five convolutional neural network architectures was strategically selected to evaluate a wide spectrum of structural paradigms. In particular, EfficientNetB0 was selected for its favorable trade-off between the number of computations and predictive accuracy [20]; ResNet50 because it was the foundational architecture for deep residual learning [21]; VGG16, because it provides consistent and highly interpretable feature maps [22]; InceptionV3, to enable complex, multi-resolution feature extraction [23]; and MobileNetV2, to evaluate the viability of the algorithms in low-resource or mobile point-of-care settings [24]. To ensure a consistent comparison across architectures, the weights used for each network topology were initialized with those from ImageNet. The models were then fine-tuned on the DFU dataset using a set of identical hyperparameters to minimize configuration bias so that the analysis was conducted in an unambiguous manner.

After data preprocessing and data augmentation, the five selected pre-trained networks were subjected to a thorough fine-tuning process. All input images were resized to 224 × 224 pixels to make the dimensions consistent between the different evaluation pipelines. Their native fully connected layers had to be removed to make the transition from their original 1,000-class ImageNet configuration to a targeted binary classification framework. In particular, the top layers were removed to retain

only the feature extraction layers at the base and the initial weights of these base models were deliberately frozen. This transfer learning strategy leveraged well-backed visual representations to accelerate training, reduce computational requirements, and mitigate overfitting risks associated with the limited-size clinical dataset. A customized classification module was added to each network to replace the dense structures. This newly added head consisted of globally pooled layer, batch normalization, and a dropout regularisation layer with a rate of 0.4. The network output was then passed through a terminal dense layer with a softmax activation function to enable the learned features to be mapped to the two diagnostic classes. The optimization protocol was standardized across the different network topologies using Adam optimization with a fixed initial learning rate of 1×10^{-3} and a sparse categorical cross-entropy loss function. Fig. 2 provides a concise overview of this methodological pipeline through a comprehensive visual summary.

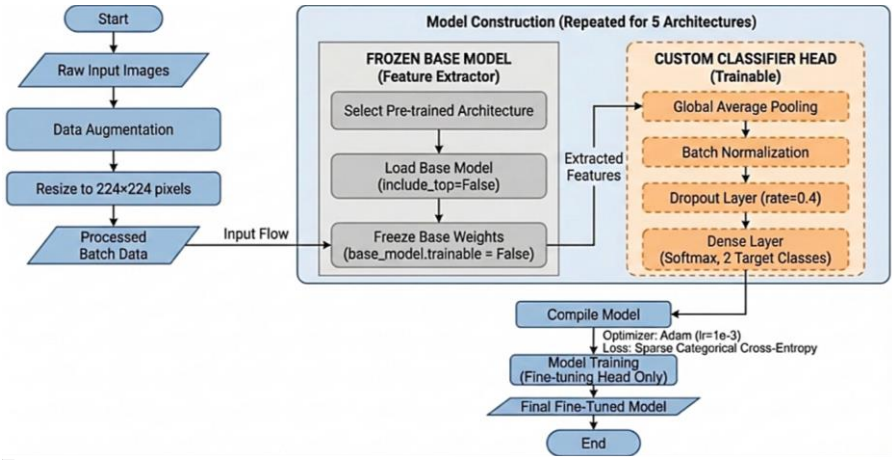


Fig. 2 Flowchart of the proposed system

4. Results and Discussion

Each of the architectures was trained for up to 50 epochs on an NVIDIA GeForce RTX 4050 Laptop GPU (6 GB). To prevent overfitting, an early stopping strategy was built into the training process and was configured to prematurely end the optimization sequence if the validation accuracy did not improve for 6 epochs. The performance of the models developed was carefully assessed using five basic metrics: accuracy, precision, recall, F1 score, and the AUC. The mathematical equations that describe the first four indices are expressed as follows [25]:

$$Accuracy = \frac{TP + TN}{Total\ Predictions} \tag{1}$$

$$Precision(PPV) = \frac{TP}{TP + FP} \tag{2}$$

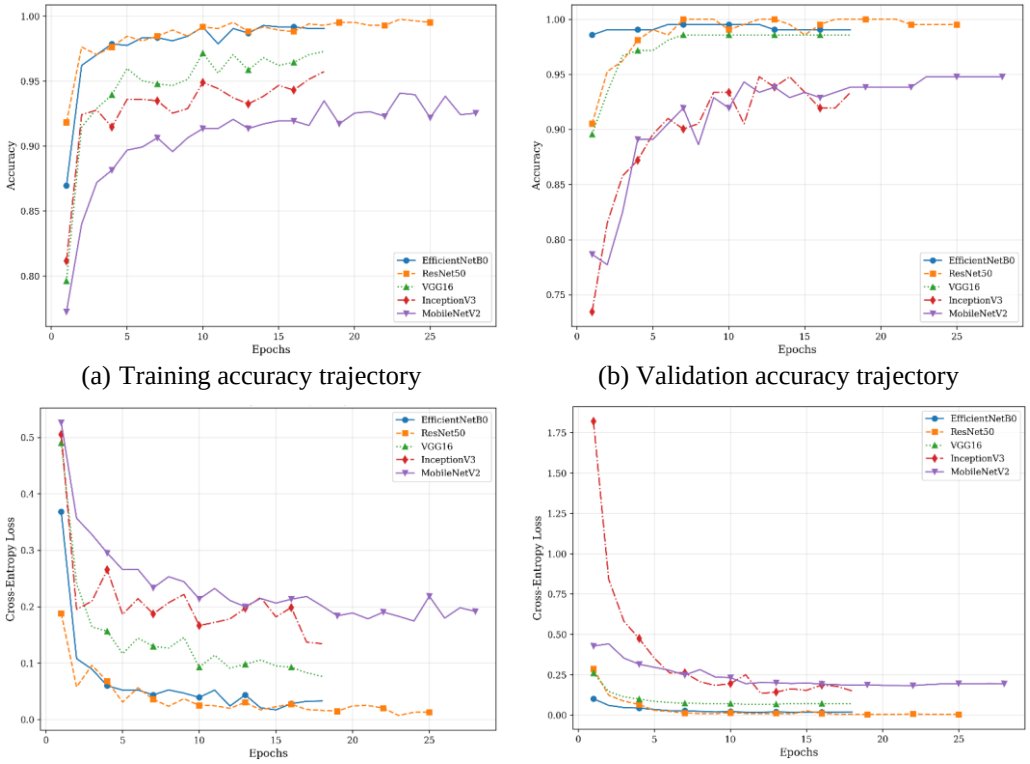
$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

In these formulations, true positives (TP) denote the count of diabetic ulcer images accurately classified by the model as infected. False positives (FP) represent samples where healthy foot images were misclassified as DFU. True negatives (TN) correspond to the normal foot images correctly classified as healthy. Finally, false negatives (FN) designate the critical clinical errors wherein genuine ulcerations were incorrectly dismissed as healthy tissue.

To empirically demonstrate that the models did not suffer from overfitting, the learning trajectories for all five architectures were recorded. Fig. 3 visualizes the training and validation accuracy, alongside the cross-entropy loss, across the training epochs. As illustrated, the implementation of an Early Stopping criterion (monitoring validation loss with a patience of 6

epochs) successfully terminated the training phase at the optimal point of convergence. The close tracking of the validation curves with the training curves, before early termination, confirms that the networks generalized well to the unseen validation folds without memorizing the training data. It can be observed that the ResNet50 model achieved the highest accuracy, reaching 99%, alongside a loss approaching zero over 25 epochs, while EfficientNetB0, InceptionV3, and VGG16 completed training after 18 epochs, with accuracies of 98%, 96%, and 97%, respectively. MobileNetV2 required the highest number of epochs, completing training after 29 epochs with an accuracy of 95%.



(a) Training accuracy trajectory (b) Validation accuracy trajectory

(c) Training loss trajectory (d) Validation loss trajectory

Fig. 3 Epoch-wise training trajectories of the deep learning models, averaged over the 5-fold cross-validation framework

The confusion matrices for the five evaluated models are presented in Fig. 4. It can be observed that the ResNet50 model achieved the highest accuracy, misclassifying only one infected case as healthy (FN) and two healthy cases as infected (FP). The remaining models exhibited varying degrees of accuracy; EfficientNetB0 demonstrated the second-highest performance among the developed models, yielding 11 misclassifications. Eight of these were infected cases that the model failed to detect, incorrectly classifying them as healthy. However, the remaining three models had inconsistent performance, and MobileNetV2 was the least accurate, with a total of 84 misclassifications, including 40 DFU cases incorrectly classified as healthy.

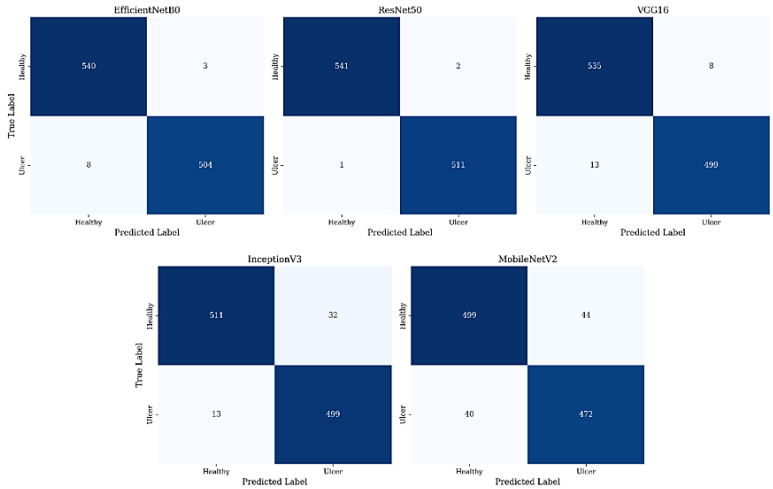
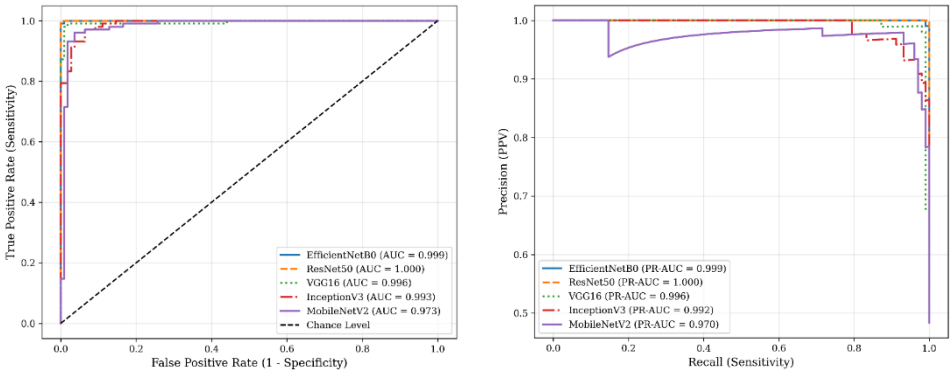


Fig. 4 Overall confusion matrix derived from the 5-fold cross-validation process

A thorough analysis of the quantitative performance metrics shows that the ResNet50 architecture outperforms other neural network architectures in terms of all performance metrics. For a visual comparison of these diagnostic classifiers, Fig. 5(a) provides ROC curves of the five models, while their respective PR curves are provided in Fig. 5(b).



(a) Receiver operating characteristic (ROC) (b) Precision-Recall curve

Fig. 5 ROC and PR curves comparison of different models

Based on the 5-fold cross-validation splitting, ResNet50 proved to be the best model due to its ability to obtain a mean F1-score of 0.997 ± 0.004 and an outstanding sensitivity of 0.998 ± 0.004, as presented in Table 3 below. This was closely followed by EfficientNetB0 (F1-score: 0.989 ± 0.009) and VGG16 (F1-score: 0.979 ± 0.010), while InceptionV3 and MobileNetV2 achieved F1-scores of 0.957 ± 0.010 and 0.918 ± 0.013, respectively. The successful performance of ResNet50 demonstrates the highly effective feature extraction ability of residual learning: The morphological textures of ulcerations are extracted effectively and successfully, capturing subtle and complex textures even when dealing with a massive number of parameters and vanishing gradients [26, 27]. More importantly, ResNet50's high negative predictive value (NPV: 0.998 ± 0.004) means that it produces almost no false negatives, which is particularly important for medical screening applications.

Table 3 Performance metrics of each model through the five-fold cross-validation

Model	Sensitivity±SD (95% CI)	Specificity±SD (95% CI)	PPV±SD (95% CI)	NPV±SD (95% CI)	F1-Score±SD (95% CI)
EfficientNetB0	0.984±0.018 (0.022)	0.994±0.005 (0.006)	0.994±0.005 (0.007)	0.986±0.016 (0.020)	0.989±0.009 (0.011)
ResNet50	0.998±0.004 (0.005)	0.996±0.005 (0.006)	0.996±0.005 (0.007)	0.998±0.004 (0.005)	0.997±0.004 (0.005)
VGG16	0.975±0.009 (0.011)	0.985±0.019 (0.024)	0.985±0.020 (0.025)	0.976±0.008 (0.010)	0.979±0.010 (0.013)
InceptionV3	0.975±0.011 (0.014)	0.941±0.017 (0.021)	0.940±0.016 (0.020)	0.975±0.011 (0.013)	0.957±0.010 (0.012)
MobileNetV2	0.922±0.032 (0.040)	0.919±0.021 (0.026)	0.915±0.019 (0.024)	0.927±0.028 (0.035)	0.918±0.013 (0.016)

EfficientNetB0 achieved second place in this cross-validation framework, but the metrics were still very competitive (Sensitivity: 0.984 ± 0.018). This is because its depth, width, and resolution of its input are optimally scaled while also being highly computationally efficient (only about 5.3 million parameters). Even with its base performance (F1: 0.918), which needs just 3.5 million parameters, MobileNetV2 shows potential for deployment in resource-constrained environments. With a significantly lower Sensitivity (0.922 ± 0.032), though, it also has a higher rate of false negatives. This is because the inherent representational compromises of depthwise separable convolutions can cause them to miss distinguishing between those tissue pathologies that fall near the boundary between two tissue classes, as well as the deeper residual architectures.

To rigorously validate the performance advantages of the proposed framework and ensure the results were not due to random sampling variance, McNemar's test was conducted and depicted in Table 4. The test evaluates whether the predictive disagreement between the two models is statistically significant based on their discordant prediction counts. The reference model chosen based on the results of the cross-validation is ResNet50, for which the significance level of $\alpha = 0.05$ was used.

Table 4 McNemar's test results for pairwise comparisons

ResNet50 vs	Chi-square	p-value
EfficientNetB0	6.125	0.0133
VGG16	16.056	0.0001
InceptionV3	38.205	0.00001
MobileNetV2	77.108	0.00001

The statistical results confirm that ResNet50's superior diagnostic performance is highly significant across all comparisons. Most notably, while EfficientNetB0 achieved highly competitive raw metrics (ranking second overall), the McNemar's test between ResNet50 and EfficientNetB0 yielded a χ^2 value of 6.125 and a p-value of 0.0133 ($p < 0.05$). This provides evidence that ResNet50's advantage—particularly its ability to minimize false negatives—is statistically significant, leading to a demonstrable improvement in clinical reliability.

Furthermore, the divergence in performance becomes even more pronounced when comparing ResNet50 to the remaining architectures. The comparisons against VGG16 ($\chi^2 = 16.056, p = 0.0001$), InceptionV3 ($\chi^2 = 38.205, p < 0.0001$), and MobileNetV2 ($\chi^2 = 77.108, p < 0.0001$) all yielded p-values well below the 0.001 threshold. This evidence clearly and consistently disproves the random prediction hypothesis, proving that ResNet50 has a strong statistical and structural capacity for feature extraction for DFU identification over the other networks considered in the experiment.

Visual interpretability was assessed to ensure the classification model is clinically credible, and the classification framework is based on real pathological markers, instead of background shortcuts or dermal pigmentation shortcuts. To identify spatial regions that make up each diagnostic decision, saliency maps were generated from the optimal ResNet50 model using the terminal convolutional layer (Fig. 6). In the case of correct classifications, the activation overlays in true positive cases are tightly localized above the ulcerated areas, with the activation concentrated on the morphological markers (yellow slough, necrotic tissue, epidermal erosion). In true negative cases, the activation maps are primarily focused on regions of healthy tissue rather than on natural skin contours, creases, or tonal variations. This means that the patch-based preprocessing was effective in limiting the network to different tissue properties in the local regions and not distracting it with visual context.

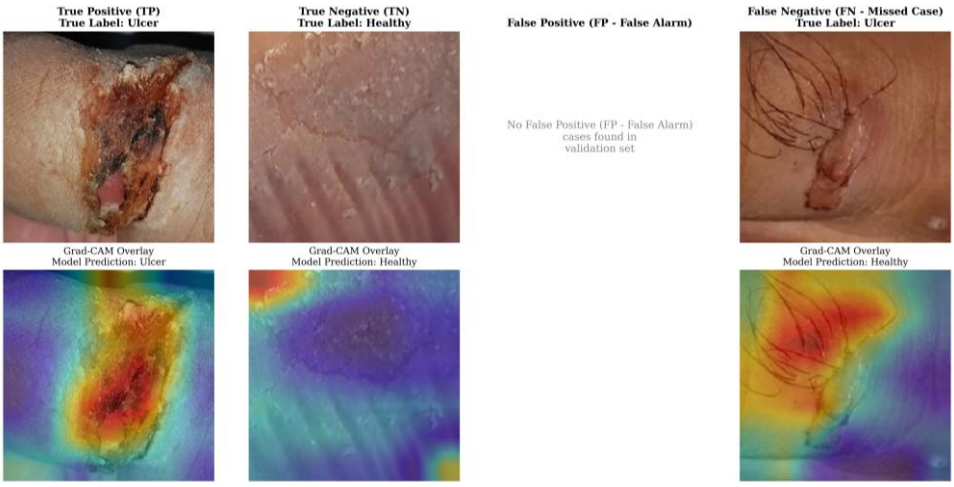


Fig. 6 Grad-CAM overlay for different validation set images

Case-mix analysis helps to clarify boundaries of model operation. The false positive errors mostly occurred in areas with severe hyperkeratosis (calluses) or localized erythema or deep shadow, which have similar surface textures to areas of early tissue breakdown. In contrast, false negative errors occurred mostly in superficial or partially obscured ulcers with poorly defined ulcer boundaries, particularly in cases with limited erythema or without clearly visible granulation or slough. Importantly, even if there were some misclassifications, the Grad-CAM heatmaps indicated that the model remained biologically plausible: it focused on structurally abnormal skin areas in comparison to irrelevant background noise, supporting the interpretability of the proposed classification framework.

The results of this experiment provide further insights into diagnostics in many important ways. Unlike methods that rely on specialized thermal imaging equipment [8-10], the proposed method uses only common RGB photography and provides an extremely practical approach for resource-limited primary healthcare centers. In addition to accessibility, the empirical performance of this approach is better than the F1 score of 93.9% reported by Goyal et al. on the DFUC2020 dataset. These findings are consistent with the results of Ahsan et al., which concluded that ResNet topologies are an extremely strong baseline model [12]. Finally, the sophisticated decision-fusion methods proposed by Munadi et al. could be unnecessary for this particular task, as the findings suggest that pixel-level spatial characteristics are, in fact, enough to make clinically exact classifications.

For medical screening, it is important to consider the overall accuracy of deep learning models to not miss important diagnostic issues [28]. Misclassifications have very asymmetric clinical costs in the context of diagnosis of DFU. A false positive (a false alarm in which healthy tissue is marked as an ulcer) is a relatively low penalty, as the patient is likely to have a second visual examination by a podiatrist or nurse. A false negative (a missed ulcer), on the other hand, is a major clinical failure. If left untreated and undiagnosed, a DFU can quickly worsen to deep tissue infection, osteomyelitis, gangrene, and finally to lower extremity amputation. A clinically viable model therefore has to carefully focus on the high sensitivity (recall) value and high negative predictive value (NPV) to ensure that true pathologies are not overlooked. The proposed framework shows excellent overall F1 and ROC-AUC scores, but the integrated confusion matrix and precision-recall curve show the models' individual abilities to prevent these clinically consequential false negatives. For real-world deployment, the classification boundaries of this proposed framework will be intentionally tuned to maximize Sensitivity – patient safety is the highest priority, not statistical accuracy.

The proposed framework is very high-performance, but there are several key restrictions to be recognised. One of the major restrictions is the lack of patient-level metadata in the publicly available dataset. All the data is anonymized, and there is no patient-level information or group metadata. Thus, 5-fold cross-validation was performed across images and not patients. In clinical image datasets, often there are several frames or slightly different angles from the same patient, the same foot, or the same imaging session. Since splitting patients at the level of the individual was not possible, it is possible that near-duplicate images were split between the training and validation folds. The possibility of this data leakage is a limitation of using an anonymized version of public datasets and may partly explain the very high performance across the different measures (e.g., ~99% accuracy and F1-score) reported in this study and in some other recent benchmark studies. Further studies are needed to confirm these architectural results on proprietary datasets with well-defined partitions for training and testing, where patient data is isolated.

Another shortcoming of the current study is that the data come from only one clinical center. The approach was to apply robust strategies to avoid overfitting, such as 5-fold stratified cross-validation, extensive data augmentation, and early stopping, but the models may experience reduced performance when applied to data from other centers due to variability in the camera hardware, lighting protocols, and patient demographics. The models were developed on localized and artifact-free tissue patches, which means that a direct evaluation on external public datasets (which mainly consist of uncropped images with clinical artifacts) may introduce a substantial domain shift. A key direction for future work is therefore the development of an automatic region-of-interest (ROI) extraction pipeline to address this gap, which will allow for systematic cross-dataset validation and domain adaptation testing on a variety of multi-center cohorts.

5. Conclusions

The current study has proposed an automated and interpretable deep learning-based approach for detecting DFUs. Five CNNs (EfficientNetB0, InceptionV3, MobileNetV2, ResNet50, and VGG16) were leveraged via transfer learning. To avoid any issues with generalization, the models were trained and tested on a dataset of 1,055 localized RGB images via the stratified

five-fold cross-validation method. Rather than relying solely on accuracy, the analysis among these models was performed considering rigorous clinical parameters such as Sensitivity, Specificity, PPV, NPV, and F1-score in addition to visual interpretability (Grad-CAM). The key results are listed below:

(1) Superior Clinical Efficacy of ResNet50: In the strict cross-validation scheme, ResNet50 proved to be the best clinically viable architecture with the highest mean F1-score (0.997 ± 0.004) and very high Sensitivity (0.998 ± 0.004). This near-perfect detection rate of true cases (high NPV of 0.998 ± 0.004) allows minimizing the risk of false negative cases.

(2) Architectural Trade-offs and Efficiency: While the deepest architecture of ResNet50 provided the highest performance due to the residual connections, EfficientNetB0 was capable of maintaining competitive diagnostic potential (mean F1-score: 0.989 ± 0.009) with a significant gain in computational efficiency thanks to the compound scaling.

(3) Viability of Standard Optical Imaging: The results suggest that standard RGB images can provide sufficient information for DFU classification. It also supports the effectiveness of modern CNNs' pixel-based features and may not be necessary for specialized hardware for initial screening of patients, such as thermal cameras.

(4) Clinical Risks of Lightweight Architectures: Lightweight architectures designed for mobile deployment should be evaluated carefully. Although MobileNetV2 managed to show a decent baseline mean F1-score (0.918 ± 0.013) with 3.5M parameters, its considerably lower Sensitivity (0.922 ± 0.032) means that there is a higher probability of false negative cases. Due to the asymmetrical nature of missed diagnoses, the deployment of such models may negatively affect the patient management process.

(5) Interpretability and Asymmetry in Error Cost: The use of Grad-CAM interpretation clearly shows that the best-performing models focus on true pathology and not the background noise. Additionally, the asymmetrical error distribution is a reflection of clinical practice for DFU screening, whereby false negatives have to be reduced at the expense of higher false positives because the clinical impact of missing a DFU is greater than the administrative cost of a false positive.

This evidence supports the feasibility of using AI for DFU screening with conventional digital images. Future research should focus on validation of these model designs in isolated patients from multi-centers and the automation of ROI extraction.

Conflicts of Interest

The authors declare no conflict of interest.

Statement of Ethical Approval

The dataset is approved by the Ethics Committee of Al-Nasiriyah Hospital as reported in the primary study [19] (approval number omitted in open-access metadata).

Statement of informed consent

For this type of study, informed consent is not required.

Data and Code Availability

The datasets generated and analyzed during the current study are available in the Kaggle repository, accessible at: https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu. The original primary data collection is described in Ref. [19].

The complete source code utilized for model training has been made publicly available on GitHub repository. The code repository can be accessed online at: https://github.com/ahd-sd/Explainable-Deep-Learning-for-Diabetic-Foot-Ulcer-Diagnosis-A-Comparative-CNN-Framework.

References

[1] S. K. Das, P. Roy, P. Singh, M. Diwakar, V. Singh, A. Maurya, et al., "Diabetic Foot Ulcer Identification: A Review," Diagnostics, vol. 13, no. 12, article no. 1998, 2023.

[2] P. L. Li, Q. F. Xiao, K. L. Yick, Q. L. Liu, and L. Y. Zhang, "A Novel Deep Learning Approach to Classify 3D Foot Types of Diabetic Patients," Scientific Reports, vol. 15, no. 1, article no. 13819, 2025.

[3] H. Xie, Z. Chen, G. Wu, P. Wei, T. Gong, S. Chen, et al., "Application of Metagenomic Next-Generation Sequencing (mNGS) to Describe the Microbial Characteristics of Diabetic Foot Ulcers at a Tertiary Medical Center in South China," BMC Endocrine Disorders, vol. 25, no. 1, article no. 18, 2025.

[4] F. Mostafavi, M. R. Amini, Y. Mehrabi, E. Nasli Esfahani, and S. S. H. Nazari, "Machine Learning Insights into Amputation Risk: Evaluating Wound Classification Systems in Diabetic Foot Ulcers," International Wound Journal, vol. 22, no. 6, article no. e70515, 2025.

[5] 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," Scientific Reports, vol. 15, no. 1, article no. 6758, 2025.

[6] F. Arnia, K. Saddami, R. Roslidar, R. Muharar, and K. Munadi, "Towards Accurate Diabetic Foot Ulcer Image Classification: Leveraging CNN Pre-Trained Features and Extreme Learning Machine," Smart Health, vol. 33, article no. 100502, 2024.

[7] S. K. Das, P. Roy, and A. K. Mishra, "Recognition of Ischaemia and Infection in Diabetic Foot Ulcer: A Deep Convolutional Neural Network Based Approach," International Journal of Imaging Systems and Technology, vol. 32, no. 1, pp. 192-208, 2022.

[8] I. Cruz-Vega, D. Hernandez-Contreras, H. Peregrina-Barreto, J. D. J. Rangel-Magdaleno, and J. M. Ramirez-Cortes, "Deep Learning Classification for Diabetic Foot Thermograms," Sensors, vol. 20, no. 6, article no. 1762, 2020.

[9] A. Khandakar, M. E. H. Chowdhury, M. B. Ibne Reaz, S. H. Md Ali, M. A. Hasan, S. Kiranyaz, et al., "A Machine Learning Model for Early Detection of Diabetic Foot Using Thermogram Images," Computers in Biology and Medicine, vol. 137, article no. 104838, 2021.

[10] K. Munadi, K. Saddami, M. Oktiana, R. Roslidar, K. Muchtar, M. Melinda, et al., "A Deep Learning Method for Early Detection of Diabetic Foot Using Decision Fusion and Thermal Images," Applied Sciences, vol. 12, no. 15, article no. 7524, 2022.

[11] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap, "DFUNET: Convolutional Neural Networks for Diabetic Foot Ulcer Classification," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 5, pp. 728-739, 2020.

[12] M. Ahsan, S. Naz, R. Ahmad, H. Ehsan, and A. Sikandar, "A Deep Learning Approach for Diabetic Foot Ulcer Classification and Recognition," Information, vol. 14, no. 1, article no. 36, 2023.

[13] A. M. Fahmy, "Assessing the Efficacy of AI Models in Medical Imaging Analysis for the Early Identification of Diabetic Foot Ulcers and Gangrene: A Review," Premier Journal of Science, vol. 12, article no. 100093, 2025.

[14] L. Wang, P. C. Pedersen, E. Agu, D. M. Strong, and B. Tulu, "Area Determination of Diabetic Foot Ulcer Images Using a Cascaded Two-Stage SVM-Based Classification," IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pp. 2098-2109, 2017.

[15] R. Niri, H. Douzi, Y. Lucas, and S. Treuillet, "A Superpixel-Wise Fully Convolutional Neural Network Approach for Diabetic Foot Ulcer Tissue Classification," Proceedings of Pattern Recognition. ICPR International Workshops and Challenges, pp. 308-320, 2021.

[16] J. Amin, M. Sharif, M. A. Anjum, H. U. Khan, M. S. A. Malik, and S. Kadry, "An Integrated Design for Classification and Localization of Diabetic Foot Ulcer Based on CNN and YOLOv2-DFU Models," IEEE Access, vol. 8, pp. 228586-228597, 2020.

[17] J. Hyun, Y. Lee, H. M. Son, S. H. Lee, V. Pham, J. U. Park, et al., "Synthetic Data Generation System for AI-Based Diabetic Foot Diagnosis," SN Computer Science, vol. 2, no. 5, article no. 345, 2021.

[18] J. J. van Netten, D. Clark, P. A. Lazzarini, M. Janda, and L. F. Reed, "The Validity and Reliability of Remote Diabetic Foot Ulcer Assessment Using Mobile Phone Images," Scientific Reports, vol. 7, no. 1, article no. 9480, 2017.

[19] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, "DFU_QUTNet: Diabetic Foot Ulcer Classification Using Novel Deep Convolutional Neural Network," Multimedia Tools and Applications, vol. 79, pp. 15655-15677, 2019.

[20] S. K. Das, S. Namasudra, and A. K. Sangaiah, "HCNNet: Hybrid Convolution Neural Network for Automatic Identification of Ischaemia in Diabetic Foot Ulcer Wounds," Multimedia Systems, vol. 30, no. 1, article no. 36, 2024.

[21] S. Alkhalefah, I. AlTuraiki, and N. Altwaijry, “Advancing Diabetic Foot Ulcer Care: AI and Generative AI Approaches for Classification, Prediction, Segmentation, and Detection,” Healthcare, vol. 13, no. 6, article no. 648, 2025.

[22] M. R. Naeemah and M. Y. Kamil, “SkinVGG-Net: A Modified and Fine-Tuned VGG19-Based Deep Learning Architecture for Skin Cancer Classification,” International Journal of Advances in Signal and Image Sciences, vol. 11, no. 1, pp. 169-179, 2025.

[23] A. K. C. Huong, J. M. Sukki, and X. T. K. Ngu, “MOBILEDFU: Classification of Diabetic Foot Ulcer Infection on the Edge,” Journal of Engineering Science and Technology, vol. 20, no. 2, pp. 577-589, 2025.

[24] S. Biswas, R. Mostafiz, B. K. Paul, K. M. M. Uddin, M. A. Hadi, F. Khanom, “DFU_XAI: A Deep Learning-Based Approach to Diabetic Foot Ulcer Detection Using Feature Explainability,” Biomedical Materials & Devices, vol. 2, no. 3, pp. 1225-1245, 2024.

[25] R. R. Kadhim and M. Y. Kamil, “Advancing Dermatological Image Classification: GLCM-Based Machine Learning Insights,” Advance Sustainable Science, Engineering and Technology (ASSET), vol. 7, no. 1, article no. 0250116, 2025.

[26] N. Almufadi and H. F. Alhasson, “Classification of Diabetic Foot Ulcers from Images Using Machine Learning Approach,” Diagnostics, vol. 14, no. 16, article no. 1807, 2024.

[27] N. K. I. S. Anggreni, H. Kristianto, D. Handayani, Y. Yueniwati, P. L. T. Irawan, R. Rosandi, et al., “Artificial Intelligence for Diabetic Foot Screening Based on Digital Image Analysis: A Systematic Review,” Journal of Diabetes Science and Technology, vol. 20, no. 4, pp. 1381-1388, 2026.

[28] T. Zhao, J. Yu, R. Deng, L. Ding, X. Li, and J. Mi, "Accuracy of Deep Learning for Risk Prediction and Screening of Diabetic Foot Ulcers: A Systematic Review and Meta-Analysis," Diabetes Research and Clinical Practice, vol. 236, article no. 113262, 2026.