IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 June 2025, accepted 17 June 2025, date of publication 25 June 2025, date of current version 9 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3583111

RESEARCH ARTICLE

Enhancing Diabetic Foot Ulcer Classification Through Fine-Tuned Multilevel CNN

ELINEIDE SILVA DOS SANTOS [1], **RODRIGO DE MELO SOUZA VERAS** [1], **FRANCISCO DAS CHAGAS TORRES DOS SANTOS** [1], **MÁRCIA ITO** [2], **ANDREA GOMES CAMPOS BIANCHI** [3], **KELSON RÔMULO TEIXEIRA AIRES** [1], **AND JOÃO MANUEL R. S. TAVARES** [4]

[1]Departamento de Computação, Universidade Federal do Piauí, Teresina 64049-550, Brazil
[2]Mestrado Profissional em Sistema Produtivo, CEETEPS, São Paulo 01124-060, Brazil
[3]Departamento de Computação, Universidade Federal de Ouro Preto, João Monlevade, Minas Gerais 35931-022, Brazil
[4]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, 4099-002 Porto, Portugal

Corresponding author: Rodrigo de Melo Souza Veras (rveras@ufpi.edu.br)

This work was supported in part by ''Fundação de Amparo à Pesquisa do Estado do Piauí'' (FAPEPI) and ''Conselho Nacional de Desenvolvimento Científico e Tecnológico'' (CNPq), Brazil; and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under Grant identifier ROR: 00x0ma614.

ABSTRACT Diabetic foot ulcers (DFUs) are especially problematic with their high rates of recurrence. Recognizing DFUs with infection and ischemia is vital for appropriate treatment and prevention of amputations. The proposed approach combines two convolutional neural networks (CNNs), fine-tuned through multilevel refinement, to classify DFU images into four categories: nonexistent, infection, ischemia, and both (infection/ischemia). Transfer learning was used and the adopted multilevel design includes fully connected layers with varying numbers of neurons and batch normalization. Data augmentation techniques were applied to a dataset comprising 8,242 images to deal with the class imbalance problem, mitigate the overfitting possibility, and enhance the approach's performance. Based on a five-fold cross-validation, the proposed approach obtained 95.91%, 93.28%, and 95.1% of accuracy, Kappa, and F1-score metrics, respectively, confirming its superiority compared to the current solutions for the same number of image classes. Furthermore, a cross-dataset validation method was employed to evaluate the model's generalization ability. Employing this methodology, the Kaggle DFU dataset obtained accuracy, Kappa, and F1-score values of 79.79%, 59.80%, and 79.02%. Therefore, the proposed approach appears promising for the precise and timely diagnosis of DFU images.

INDEX TERMS Deep learning, diabetic foot ulcer, image analysis, multilevel.

I. INTRODUCTION

Diabetes is a chronic condition that affects the function of the pancreas. In abnormal conditions, this organ either produces insufficient insulin (type 1 diabetes) or produces insulin in excess (diabetes mellitus). The IDF Diabetes Atlas Reports [1] indicate an urgent need for measures to address the rising number of cases of this disease, particularly those related to diabetes mellitus.

The associate editor coordinating the review of this manuscript and approving it for publication was Olarik Surinta.

A patient with diabetes mellitus can have various complications. For example, diabetic peripheral nephropathy can lead to symptoms such as numbness, pain, and tingling in the extremities, particularly in the hands or feet [2]. Another complication is Peripheral Arterial Disease, which results in stenosis or occlusion of the arteries responsible for supplying blood to the lower limbs [3], causing ulcers and wounds.

The healing process of ulcers takes time. In the case of diabetic foot ulcers (DFU), they have slow healing with a high chance of recurrence, increasing the risk of amputation [4]. Moreover, patients with DFU often experience psycho-social factors such as depression, anxiety, and loneliness [5]. For

FIGURE 1. Example images of diabetic foot ulcers (image from the DFUC 2021 dataset).

this reason, prompt treatment of the lesions is demanded. In severe cases, these lesions can lead to infections or ischemia. Fig. 1 demonstrates examples of lesions. Infected lesions may show abscesses, blisters, or purulence, while ischemia lesions often present hairlessness, skin peeling, muscle atrophy, or gangrene. Some ulcers exhibit both conditions. After successful treatment, they heal and resemble healthy skin. The four classes (infection, ischemia, both and none) were defined based on clinical relevance by reflecting the presence of infection, ischemia, or a combination of both.

Diagnosing and treating DFUs is challenging, often relying on manual or image-based inspection. Automated image processing solutions are increasingly needed to support specialists and have allowed the development of computer-aided diagnosis (CAD) systems to enhance e tasks often performed by the medical community, such as highlighting regions of clinical interest [6], analyzing the progression of pathology [7], and achieving effective image classification [8]. Building on this foundation, researchers have explored how using artificial intelligence tools can play a vital role in enhancing the efficiency of monitoring processes [9].

Modifications were made to seven commonly used convolutional neural networks (CNN) to develop the proposed multilevel architecture: DenseNet-201, EfficientNetB0, InceptionV3, MobileNetV2, ResNet-50, VGG-16, and VGG-19. These were studied under various conditions, starting with the standard network architecture, followed by modifications to their dense layers. The tests included using a single fully connected layer with 256, 512, or 1024 neurons or two fully connected layers with 512−256, 1024−256, or 1024−512 neuron configurations. The internal architectures of the VGG-16 and VGG-19 networks were then adjusted based on the results of these tests, incorporating dropout (DP) layers, batch normalization (BN) layers, and combinations of both.

A multilevel CNN classification-based approach is quite innovative, as it allows information integration at different levels of complexity throughout the classification process. This allows the model to capture both detailed and contextualized nuances within the images, by analyzing the simplest local features of more complex global patterns. The resulting classification is thus much more accurate and robust because the model can focus on all possible details in the images at any level of complexity and the relationships between different levels of information.

To validate the results, we used a cross-dataset method, that consisted of testing one dataset while using the remaining datasets for training. This method improves integrity and model generalization capability by ensuring that no test images are included in the training process.

In this context, the main contributions of this work are: (1) development of an approach for the classification of images of diabetic foot ulcers into four categories: nonexistent, infection, ischemia, and both (infection and ischemia); (2) analysis of 55 individual convolution neural network architectures; and (3) implementation of a classification system multilevel with promising defined parameters.

Despite these advances, the multilevel approach presents some limitations: (1) a higher computational cost, which increases the training and inference time owing to the combination of CNNs; however, this cost is offset by gains in accuracy and robustness, especially in critical clinical applications; (2) the complexity of the multilevel model requires larger datasets to avoid overfitting, but this requirement contributes to greater reliability in the face of variation in acquisition conditions; and (3) the study lacked patient-level clinical data (e.g., age, comorbidities, and sex), which limits the analysis of potential clinical correlations.

This article is organized as follows. A literature review is presented in Section II. Section III describes the materials used in the proposed approach. Next, Section IV discusses the experiments with image datasets, the evaluation metrics, and the results obtained. Section V provides the conclusion and insights for further research paths.

II. STATE OF THE ART

A review was conducted in this work to identify the current approaches for classifying diabetic foot ulcers. Goyal et al. [10] pioneered DFU classification using convolutional neural networks, achieving 92.5% of accuracy. Their approach uses convolution parallel blocks to concatenate characteristics of convolution operations and individual convolution layers inspired by the LeNet, AlexNet, and GoogLeNet architectures. This initial work laid the foundation for subsequent advancements in the field.

Building on Goyal et al. work, Das et al. [11], and Venkatesan, Sumithra, and Murugappan [12] improved the achieved classification accuracy by experimenting different kernel sizes with the same dataset (1,679 images).

Das et al. [11] achieved an accuracy of 96.4%. Meanwhile, Venkatesan et al. [12] achieved 100% accuracy by combining diverse kernel sizes with ReLU and PReLU activation functions to extract feature characteristics. In addition, the layers used were connected after a max pooling layer and local response normalization to avoid the computational complexity increase.

Another significant contributioncame from Alzubaidi et al. [13], who introduced DFU_QUTNet and, using a CNN combined with an SVM classifier, achieved a better classification accuracy of 95.4%. However, a major limitation of this approach is the use of a kernel size of 3 × 3 in the parallel convolution layers. Similarly, Yogapriya et al. [14] improved the training speed and accuracy of the used CNN using Rectified Linear Unit (ReLU) convolution layers.

Further advancements were observedseen by Al-Garaawi et al. [15] and Arumuga and Hepzibai [16], who incorporated Local Binary Pattern (LBP) texture features and employed advanced preprocessing techniques such as cascaded fuzzy filters and Non-parametric differential evolution (NPDE). These approaches enhanced the models' classification accuracy and robustness, especially when dealing with noise and segmentation challenges in large datasets.

The use of advanced CNN architectures continued with Lan et al. [17], who combined global and local features using the ResNet-34 model with a Convolutional Block Attention Module (CBAM), achieving an accuracy of 95.8%. Meanwhile, Liu et al. [18] applied modified EfficientNet networks to a larger dataset of 15,760 images, reaching an accuracy exceeding 97%. The Fast Convolutional Neural Network employed by Rubavathi [19] and Sathya [20] yielded accuracies of 92.9% and 99.3%, respectively, showing the continued evolution and refinement of CNN-based methods for DFU classification.

Innovations in feature extraction and classification techniques also played a crucial role. Xu et al. [21] used knowledge banks (CKBs) to enhance training performance. In contrast, Das et al. [22] combined multiple hybrid blocks of inception, residual, and dense modules with Squeeze-and-Excitation blocks, resulting in an accuracy of 98.8%. Ensemble methods further boosted the classification performance, as demonstrated by Rostami et al. [8], who combined AlexNet with a sliding window-based patch and a multi-layer perceptron network, achieving promising accuracies across binary and problems multiclass.

Recent advancements include the work of Ai et al. [23], [24], who introduced asymmetric convolution modules and the LeakyReLU activation function. Others, such as Santos et al. [25], developed a new approach based on known CNN as VGG-19 models, incorporating batch normalization and dense layers to enhance the output performance. Bansal and Vidyarthi [26] developed a dense neural network combined with data augmentation techniques, achieving an accuracy of 98.87% and an AUC of 98.13%.

Diverse scales exist in the classification of DFU; however, due to the parameters used, the Wagner scale is commonly adopted. The work of Gamage et al. [27] and Cao et al. [28] modified different CNN architectures for working with this scale. The approach proposed by Gamage et al. [27] performed best by combining the DenseNet-201 model with an ANN classifier, resulting in a difference of 19.2% in the F1-score metric relative to the approach suggested by Cao et al. [28].

The studies by Goyal et al. [29], Rostami et al. [8], and Santos et al. [30] achieved better accuracy performance when combining different classification techniques in an Ensemble scheme. In the study by Goyal et al. [29], a combination including Inceptionv3, ResNet-50, and Inceptionresnetv2 was employed along with an SVM classifier for prediction. Additionally, this work employed superpixels to extract color and texture information for training purposes. On the other hand, in the work of Santos et al. [30], an ensemble approach based on weighted majority voting was proposed. This ensemble comprised five modified convolution neural networks: VGG-16, VGG-19, Resnet-50, InceptionV3, and Densenet-201; besides, dropout and batch normalization were integrated into the final architectures of VGG-16 and VGG-19 before their combination with the other networks. Among them, Rostami et al. [8], achieved the highest accuracy in binary classification with 96.4%. In multiclass classification, the best result was achieved by Santos et al. [30] with 95%.

Biswas et al. [31] achieved a high accuracy of 99.05% using a hybrid model enhanced with explainable artificial intelligence (XAI) techniques, which improved the interpretability of predictions and increased the trustworthiness. Similarly, Thotad et al. [32] achieved an accuracy of 98.97% by applying EfficientNet. It is important to emphasize that both studies utilized the public Kaggle DFU dataset for training and testing, which may have made it difficult to evaluate the generalization capability of the models.

III. MATERIALS AND METHODS

This section outlines the experimental steps to develop the proposed approach for classifying images of foot ulcers in people with diabetes. It begins by describing the methodology used. Subsequently, it addresses critical components, including data augmentation for dataset expansion, cross-validation for performance evaluation, transfer learning using pre-trained models, dropout and batch normalization for training stability, multilevel, which integrates information across varying levels of complexity, and explainability to enhance CNN decision transparency.

A. METHODOLOGY

The methodological flowchart shown in Fig. 2 represents the adopted steps of the proposed approach development.

First, we selected pre-trained convolutional neural networks (CNNs) based on their usefulness in image classification, extracting their convolutional layers as the first

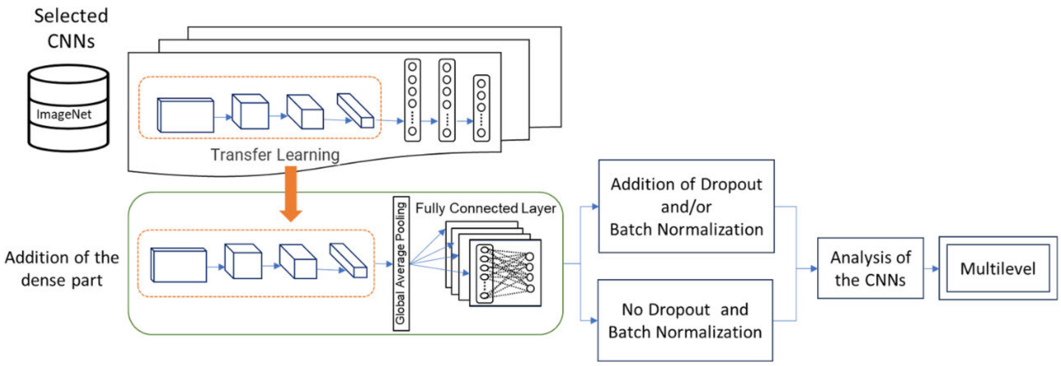


FIGURE 2. Methodological flowchart of the proposed approach development.

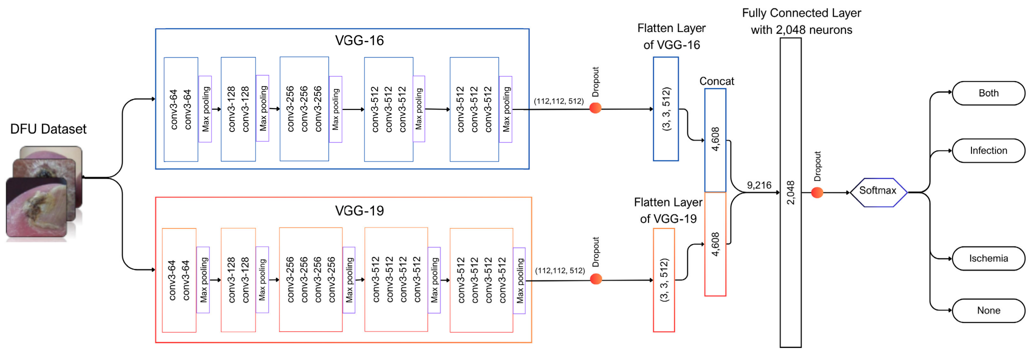


FIGURE 3. Proposed multilevel M2_V16V19_S_2048 architecture.

step. Next, we added a Global Average Pooling layer and fully connected layers. In CNNs based on the VGGNet architecture, the performance was enhanced by incorporating a dropout layer to promote more generalization. In order to guarantee quicker convergence during training and stabilize learning, we also incorporated batch normalization layers.

The networks were trained using the Deep Fine-Tuning (DFT) technique, which adjusts them to images from domains different from those in their original training datasets. This technique, replicated on diabetic foot ulcer (DFU) images, was validated by Santos et al. [25].

After analyzing the performance of each network individually, we combined the top performing architectures using a multilevel technique. This technique merges the outputs of various networks to create a more powerful model that captures different levels of abstraction. Fig. 3 illustrates the overall architecture of the proposed multilevel model. The M2_V16V19_S_2048 architecture concatenates the feature maps extracted from the VGG-16 and VGG-19 CNNs after undergoing an ablation process that includes the insertion of batch normalization layers. A fully connected layer was then used to synthesize the combined information and generate the final output. This approach demonstrated superior accuracy in classifying DFUs compared to existing methods in the literature.

Sections III-B and III-C describe the techniques used to develop and validate the proposed approach. Sections III-D, III-E, and III-F focus on the CNNs evaluated, while Section III-G explains how they were combined. Finally, Section III-H discusses the use of explainable artificial intelligence to enhance understanding of the models.

B. DATA AUGMENTATION

The limited data available in the training dataset is a significant challenge when using deep learning models. The class imbalance also worsens this problem when some classes have fewer samples than others. These problems, combined with CNNs high dimensionality, can result in overfitting, a situation where the model excessively adjusts to the training data, impairing its ability to generalize new information. As mentioned by Shorten and Khoshgoftaar [33], overfitting compromises the model's ability to make accurate predictions on unseen data.

An effective strategy to deal with this is data augmentation, which enlarges the training dataset by generating new images through various transformations applied to existing images. These transformations include rotation, translation, flipping, zooming, and shearing. The reflection padding typically fills the gaps created by rotation or translation, ensuring a smooth transition by replicating neighboring pixels without losing information.

Data augmentation increases the diversity of the training dataset, enabling the model to learn a broader range of features and patterns from the input of a given CNN [34]. Therefore, it helps reduce the risk of overfitting and enhances the model's generalization ability to unseen data. Consequently, it is an essential technique in scenarios with limited data.

This study implemented data augmentation randomly and strategically, resulting in an expanded dataset up to five times the original size. Additionally, all pixels in the input images were normalized (range between 0 and 1) before training. This normalization process stabilizes the model training and improves the learning convergence.

C. CROSS-VALIDATION

Cross-validation is commonly used to evaluate individual classifiers and multilevel models. In this process, the original dataset is randomly divided into k mutually exclusive sub-datasets, maintaining the same class proportion as in the original dataset.

During each iteration, one sub-dataset is used for testing, while the other $k - 1$ sub-datasets are for training. The model tests in k times, enabling a comprehensive evaluation of the classifier using the entire available dataset. The training sub-dataset adjusts the network weights, maximizing the model's accuracy. Meanwhile, the validation sub-dataset monitors learning during training, detecting potential issues such as overfitting. Thus, cross-validation provides a more reliable evaluation of a model's performance, ensuring that it can be generalized effectively.

This work divided the used image dataset into five folds ($k = 5$) conducted at the image level, because datasets do not provide patient-identifying information. Therefore, we reserved 20% of the instances to form the test sub-dataset, completely separate from the training and exclusively used for the final evaluation of the classifier under study. The remaining folds were divided into 70% training and 10% validation sub-datasets.

In addition to cross-validation, we conducted a cross-dataset evaluation to assess the generalization of the modelacross different data distributions. In this approach, the model is trained on one dataset and tested on a separate, independent dataset that is not observed during training [35]. Cross-dataset evaluation simulates real-world deployment scenarios in which models must perform well on data from different sources or institutions.

D. TRANSFER LEARNING

Transfer learning is widely adopted across diverse domains [36], is a strategy that uses pre-trained weights obtained, for example, from the ImageNet dataset [37], which is ample and diverse. This approach eliminates the need to start training the model from the beginning. It is particularly beneficial in deep learning scenarios where data is limited. This strategy frequently uses two principal techniques: shallow fine-tuning (SFT) and deep fine-tuning (DFT) [38]. SFT froze the initial layers of the model, and only the final layers are re-adjusted. However, in DFT, all layers are re-adjusted.

This work used this DFT technique based on previous studies, such as the one conducted by Santos et al. [25], which demonstrated the effectiveness of DFT for the specific problem of diabetic foot ulcers. The choice has proven to accelerate model training, especially in the current scenario where the availability of large volumes of data is limited. CNNs imported from ImageNet have a high capacity to extract relevant characteristics. However, the networks were pre-trained for a task distinct from the proposal. Consequently, changes in the internal layers, such as dropout and batch normalization, are needed to fit the problem under study better.

E. DROPOUT AND BATCH NORMALIZATION

To mitigate the overfitting problem in neural networks when data is scarce, Srivastava et al. [39] proposed dropout as a measure to reduce model complexity. This technique temporarily deactivates neurons during training while preserving information on connections before deactivation.

Furthermore, when using neural networks, one of the difficulties encountered is parameter tuning, as an inappropriate configuration can delay model convergence. Therefore, meticulous tuning is crucial. According to Wu et al. [40], batch normalization at layers' input can address this issue. This technique normalizes model inputs, enabling higher learning rates and reducing the time required for training convergence.

F. CNNS EVALUATED

In this study, after a comprehensive literature review, seven convolutional neural network architectures were selected: ResNet-50 [41], InceptionV3 [42], DenseNet-201 [43], MobileNetV2 [44], EfficientNetB0 [45], and the VGGNet [46] VGG-16 and VGG-19 versions. Table 1 allows a detailed comparison of these architectures, considering factors such as network topology depth, number of parameters, publication year, Top-1 accuracy, and size. The Top-1 accuracy explicitly describes the efficiency of the architectures on the ImageNet dataset, where the network's expected response matches the output exactly. It can be observed in Table 1 that all selected architectures had an accuracy above 70%, which demonstrates their generalization capability [47].

Another relevant point is the exclusive use of these architectures' convolutional part due to the transfer learning technique (DFT) technique application. In this context, it replaces the entire dense layer of the networks with fully connected layers, adapted for classifying diabetic foot ulcers.

Six variations were built in total: three included only the addition of a connected layer with 256, 512, or 1,024 neurons, while the remaining three consisted of two connected layers with configurations of 512−256, 1,024−256, or 1,024−512 neurons, respectively. Additionally, we integrated the Global

TABLE 1. Summary of the characteristics of the evaluated CNN architectures.

Architecture	Abbreviation	Topological Depth	Number of Parameters	Year	Top-1 Accuracy ImageNet	Size (MB)
VGG-16	V16	23	138,357,544	2014	71.3%	528
VGG-19	V19	26	143,667,240	2014	71.3%	549
ResNet-50	Res50	168	25,636,712	2015	74.9%	98
InceptionV3	InV3	159	23,851,784	2016	77.9%	92
DenseNet-201	Den201	201	20,242,984	2017	77.3%	80
MobileNetV2	MobV2	88	3,538,984	2018	71.3%	14
EfficientNetB0	EfB0	240	5,330,571	2019	77.1%	29

Average Pooling layer to reduce the dimensionality of the feature map and generate one-dimensional vectors, thus replacing the removed dense layer.

Each network's training was monitored for 500 epochs using the following parameters: an SGD optimizer with a momentum of 0.8, a learning rate set to 0.0001, and a batch size of 32. During this process, we saved the network model that achieved the best accuracy in the validation dataset (val_categorical_accuracy). The experiments were carried out using Python within the Anaconda environment. The tests ran on a computer with a GPU (Nvidia GeForce RTX 2060 16GB RAM) board.

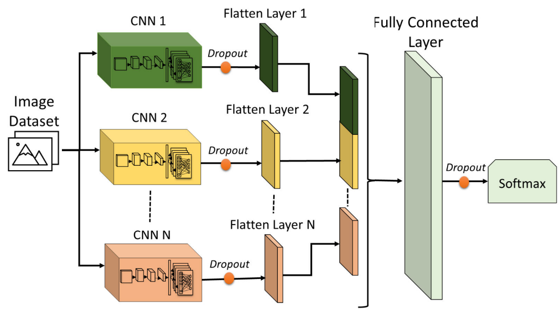


FIGURE 4. Generic scheme of a multilevel approach.

G. MULTILEVEL

A multilevel approach merges features extracted by different CNN architectures. Each CNN architecture extracts features such as edges, textures, shapes, or color patterns to form feature maps, which capture the image information at various levels of abstraction. Thus, a multilevel approach aims to extract feature maps from all CNNs involved and concatenate them into a one-dimensional vector, which is then connected to a fully connected layer [48]. This vector serves to encapsulate an all-around scope of visual information. Fig. 4 provides a schematic representation of this process.

For example, one CNN might effectively identify global structures in the input image, e.g., shape and outline, while another CNN might concentrate on more delicate elements, like texture variations and color changes. By merging these features, the multilevel ensures the model can understand the input image holistically. Next, it goes through a fully connected layer, which is more critical in refining the integrated features and synthesizing the information output that determines the final classification decision.

This multilevel approach is promising in classifying DFU. Ulcers present a broad array of visual characteristics, including notable divergences in color, texture, size, or location on the foot. Moreover, it improves the generalization of the model to new cases rather than becoming overly dependent on a narrow set of characteristics applied to different patient populations or imaging conditions.

The process of creating a multilevel must consider the following:

1) It is necessary to analyze the performance of each CNN to select the best ones to compose the multilevel;

TABLE 2. Defined criteria considered in creating the tested multilevel approach.

Number of Members	CNNs	Dropout	Number of Neurons			
4	V16 V19 Res50 Den201	yes no	256	512	1,024	2,048
3	V16 V19 Res50	yes no	256	512	1,024	2,048
3	V16 V19 Den201	yes no	256	512	1,024	2,048
3	V16 Res50 Den201	yes no	256	512	1,024	2,048
3	V19 Res50 Den201	yes no	256	512	1,024	2,048
2	V16 V19	yes no	256	512	1,024	2,048
2	V16 Res50	yes no	256	512	1,024	2,048
2	V16 Den201	yes no	256	512	1,024	2,048
2	V19 Res50	yes no	256	512	1,024	2,048
2	V19 Den201	yes no	256	512	1,024	2,048
2	Res50 Den201	yes no	256	512	1,024	2,048

2) It is crucial to evaluate the need for dropout usage, a technique that temporarily drops neurons during training to prevent overfitting;
3) It is essential to find the number of neurons in the fully connected layer. It plays a fundamental role in feature extraction and data classification.

To create multilevel configurations, we considered these requirements by combining 2, 3, and 4 members from different CNN architectures. We tested the dropout approach and configurations with and without dropout at distinct points in the model. According to Lyu and Ling [48], dropout can be added before feature extraction or after the connected layer. Regarding the formation of the fully connected layer, different numbers of neurons were considered and evaluated: from 256, 512, and 1,024 to 2,048.

Table 2 lists the criteria for creating the multilevel approach, offering a more comprehensive view of the different configurations tested. An example of one of the tested multilevel configurations is: "M4_V16V19Res50Den201_S_512", where M represents the model, 4 indicates the number of combined networks, $V16V19Res50Den201$ identifies the CNN architectures used, S refers to the use of dropout, and 512 represents the number of neurons.

H. EXPLAINABLE IN CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks are often considered "black-box" models because knowing all the neurons used in their decision-making is impossible. Although one can have a detailed understanding of their structure, it is impossible to decipher them fully. However, the literature presents studies that aim to explain the decisions made by a CNN through explainable artificial intelligence (XAI). This allows the production of indicators that facilitate human understanding, especially in the context of medical acceptance. Since CNNs can recognize and classify complex patterns, they possess many parameters. XAI investigates the internal mechanisms of these models through interpretability techniques that are not dependent on the used model.

Various techniques can be applied to imaging cases. One is the gradient-weighted class activation map (Grad-CAM) [49]. This technique, based on class activation mapping (CAM), generates heatmaps over the input image based on the gradients of the model and highlights specific regions of the image that most influence the classification of a particular class in the prediction layer of the model. Another technique used is the Local Interpretable Model-agnostic Explanations (LIME) [50], which generates disturbances in a specific sample to identify the image regions most relevant for the model's prediction. Initially, the image is segmented into superpixels, corresponding to neighboring pixels with similar characteristics, such as color and location. LIME then evaluates the impact of local information by selectively activating or disabling certain superpixels, allowing the visualization of their contribution to the model's decision.

IV. RESULTS AND DISCUSSION

The experiments characterize the image datasets and evaluation metrics used to develop and validate the approach in Sections IV-A and IV-B. We calculated the evaluation metrics using the test sub-datasets, with each sub-dataset generating its results, creating a confusion matrix. As a result, we computed the arithmetic mean and standard deviation from the five cross-validation (k-fold) runs. To enhance the interpretation of the presented tables, the Kappa coefficient was multiplied by 100. This section highlights the best results from analyzing individual (Sections IV-C and IV-D) and multilevel (Section IV-E) architectures. In Section IV-F, we compare the proposed method to current state-of-the-art approaches.

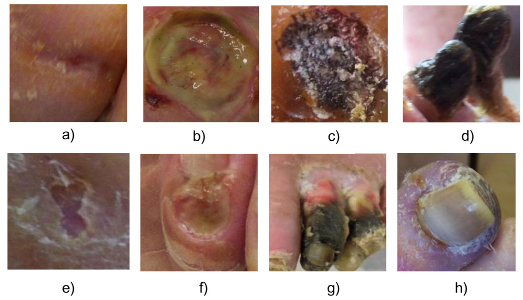


FIGURE 5. Examples of images in the DFUC dataset, (a, b, c, d) are of the 2020 set, and (e, f, g, h) are of the 2021 set. Images (a) and (e) are of the nonexistent class, (b) and (f) are of the infection class, (c) and (g) are of the ischemia class, and (d) and (h) are of the infection + ischemia class.

A. IMAGE DATASETS

Two publicly available datasets were used to evaluate the generalization capability of the proposed system. The first corresponds to the DFUC challenge dataset, which comprises images released in the 2020 [51] and 2021 [52] editions. Access to this dataset requires prior approval, which can be requested from an official website (https://dfu-challenge.github.io/). The second dataset used was the Kaggle DFU dataset, which contains labeled images acquired under different conditions from the DFUC dataset. This dataset is publicly available at https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu. Both datasets present variations in acquisition conditions, including lighting, resolution, and background, allowing evaluation of the robustness and generalization of the proposed model across domains.

The DFUC dataset was collected over five years from hospitals in Lancashire, UK, using three different cameras at a distance of 30 to 40 cm (Kodak DX4530, Nikon D3300, and Nikon COOLPIX P100) under uniform ambient lighting conditions with no flash to preserve colors. Images were annotated by two experts (a podiatrist and a medical) and validated with medical records [53]. In case of disagreement, the final annotation of both experts were used. Fig. 5 presents examples of the images used in each case. The images used were centered on the lesion area and had different dimensions, ranging from 34×31 pixels to $1{,}103 \times 1.127$ pixels, and were divided into four classes:

1) None: includes images of healthy skin, ulcers in the healing process, and ulcers without infection or ischemia (DFUC 2020 - 1,281; DFUC 2021 - 2,547; Total - 3,828);
2) Infection: contains only images of ulcers with infection (DFUC 2020 - 26; DFUC 2021 - 227; Total - 253);
3) Ischemia: contains only images of ulcers with ischemia (DFUC 2020 - 779; DFUC 2021 - 2,553; Total - 3,332);
4) Both: images of ulcers with both infection and ischemia simultaneously (DFUC 2020 - 209; DFUC 2021 - 620; Total - 829);.

a) b)

FIGURE 6. Examples from the Kaggle DFU dataset - Patches fold: (a) normal class (healthy skin) and (b) abnormal class (ulcerated region).

Four classes were defined based on their clinical relevance [54]. Initial lesions (ulcers without ischemia) can often be managed with wound dressings, glycemic control, antibiotics, and other conservative measures. By contrast, ulcers with ischemia require greater attention because of the increased risk of amputation. These cases may require vascular control and, in some instances, surgical intervention. In addition, identifying healthy feet is important for preventive monitoring and implementation of strategies to avoid future complications. Thus, each class has distinct therapeutic and prognostic implications, making these systems relevant for clinical decision support.

The dataset presents a class imbalance beacause some classes contain more samples, which can compromise the learning process. For example, the "none" class includes 3,828 images, whereas the "infection" class contains only 253. Hence, handling this issue when developing models is essential to ensure more reliable results. Furthermore, due to privacy constraints, patient-level information such as age, gender, and number of images per patient is not publicly available. Despite this, the DFUC 2020 challenge documentation [55] indicates that most participants are white and that multiple images may be derived from the same patient ulcer at different stages of healing; these characteristics can cause data limitations and impact the model's training and evaluation. Therefore, cross-dataset validation is necessary, using external sources with diverse populations, such as the Kaggle DFU dataset, to validate the proposed approach.

The Kaggle DFU dataset comprises foot images acquired at the Diabetes Center of Nasiriyah Hospital, located in southern Iraq, and was captured using Samsung Galaxy Note 8 and iPad under diverse illumination conditions and camera angles [13], [56]. The dataset includes ulcerated and non-ulcerated foot regions and is organized into four folds: original images, patches, testSet, and transfer-learning images.

This study employed patch folds, which consist of cropped regions extracted from the "Original Images". Each patch has a resolution of 224 × 224 pixels, which facilitate compatibility with common deep-learning architectures. The patches set contained a total of 1,055 image samples, divided into two classes: normal (healthy skin), comprising 543 patches, and abnormal (ulcerated regions), comprising 512 patches. Fig. 6 shows the examples of each class. All images in the dataset were anonymized according to ethical research standards and data protection regulations, ensuring that no personally identifiable information is present [13]. Consequently, patient-level identification is not possible.

TABLE 3. Meaning of the kappa coefficient.

Kappa value	Degree of agreement
Kappa ≤ 0.2	Bad
$0.2 <$ Kappa ≤ 0.4	Moderate
$0.4 <$ Kappa ≤ 0.6	Good
$0.6 <$ Kappa ≤ 0.8	Very good
Kappa > 0.8	Excellent

B. EVALUATION METRICS

The test dataset was input into the trained convolutional neural networks (CNNs), and the results were assessed based on the confusion matrix values. Four key values were derived from this matrix: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). We obtained inference times in milliseconds (ms) per image and calculated the classification metrics using these values, Equations 1 to 5 including accuracy, precision, recall, F1-score, and the Kappa coefficient.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

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

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

$$\text{Kappa} = \frac{observed - expected}{1 - expected}. \tag{5}$$

The Kappa coefficient, as described by Landis and Koch [57], assesses the degree of agreement in classification, considering all elements of the confusion matrix. It takes values between 0 (zero) and 1 (one), as in Table 3. The term "observed" refers to the proportion of agreement observed, defined as the number of times the evaluators agree, divided by the total number of observations. On the other hand, "expected" represents the proportion of agreement expected by chance.

C. RESULTS WITH ORIGINAL ARCHITECTURES

Initially, the results of the VGG-16, VGG-19, InceptionV3, ResNet-50, DenseNet-201, MobileNetV2, and EfficientNetB0 architectures without modification in their layers are analyzed. During the execution of this experiment, it was necessary to resize the input images to 224 × 224 pixels, since attempting to reduce the network input to a one-dimensional matrix with inputs of 112 × 112 pixels resulted in errors. Table 4 presents the obtained results.

Upon analyzing the results in Table 4, it can be realized that the DenseNet-201, ResNet-50, VGG-19, and VGG-16

TABLE 4. Classification results obtained with the studied CNNs in their original configuration and the respective average processing times (best values in bold).

CNN	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa (%)	Time (ms)
DenseNet-201	**75.11**±0.66	**66.62**±0.74	**75.11**±0.66	**70.28**±0.64	**56.12**±1.19	79.07±8.27
Resnet-50	74.21±0.62	65.35±0.59	74.21±0.62	69.33±0.57	54.44±1.09	72.43±14.59
VGG-19	73.11±0.94	64.97±0.65	73.11±0.94	68.46±0.84	52.63±1.63	831.83±4.61
VGG-16	71.90±2.50	63.37±2.67	71.90±2.50	67.17±2.50	50.33±4.58	689.12±16.11
MobileNetV2	61.51±7.37	53.78±6.42	61.51±7.37	55.15±11.22	30.84±14.95	**18.91**±5.67
InceptionV3	58.19±7.46	54.49±7.79	58.19±7.46	51.60±10.45	24.96±15.42	34.18±8.91
EfficientNetB0	46.55±0.10	42.76±16.93	46.55±0.10	29.91±0.59	0.12±0.24	199.96±26.11

TABLE 5. Best results found for each studied CNN and the respective average processing times (BN means with batch normalization; best values in bold).

CNN	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa (%)	Time (ms)
VGG-16 512 BN	**93.43**±0.33	93.46±0.29	**93.43**±0.33	93.42±0.33	**89.21**±0.56	87.73±2.51
VGG-19 512 BN	93.42±1.17	**93.53**±1.14	93.42±1.17	**93.43**±1.21	89.20±1.72	109.47±7.27
Resnet-50 512	90.20±0.67	90.34±0.63	90.20±0.67	90.21±0.67	83.93±1.08	35.62±0.70
DenseNet-201 1.204-512	90.01±0.69	90.21±0.75	90.01±0.69	90.00±0.70	83.58±1.12	43.24±9.21
InceptionV3 512	81.63±1.02	81.99±0.88	81.63±1.02	81.57±1.03	69.63±1.69	23.37±3.61
MobileNetV2 512−256	81.37±0.76	81.47±0.76	81.37±0.76	81.27±0.75	69.26±1.26	**12.47**±1.22
EfficientNetB0 512	76.56±1.02	76.79±0.95	76.56±1.02	76.47±1.08	61.25±1.77	24.18±4.88

architectures achieved accuracy and recall values above 70%. However, none of these architectures reached a precision above this threshold, with the highest precision recorded being 66.62%. Regarding the F1-score, only the DenseNet-201 architecture demonstrated satisfactory performance, achieving a value of 70.28%.

It is important to highlight that although CNNs in their original configurations achieved Kappa scores ranging from 40% to 60%, which are considered good according to Landis and Koch [57], the performance of EfficientNetB0 was substantially lower, registering a Kappa score of only 0.12%.

D. RESULTS WITH MODIFIED CNN ARCHITECTURES

We examined the results of modifying the architectures of the networks' layers in four distinct configurations, according to the following order:

1) The convolutional layers of the CNNs were kept, and then a Global Average Pooling layer was inserted. Subsequently, fully connected layers were added in two scenarios: (1) a connected layer with different numbers of neurons: 256, 512, and 1,024; and (2) two fully connected layers with configurations of 512−256, 1,024−256, and 1,024−512 neurons.
2) A dropout layer was added for the VGG-16 and VGG-19 CNNs. However, this procedure was not performed for the other networks because they already have many blocks, which requires a more in-depth study to determine where dropout layers would be included.
3) Batch normalization layers were included only in the VGG-16 and VGG-19 models, integrated into each convolutional layer block and before the MaxPooling layer. It is worth noting that the other CNNs already have normalization layers in their original architectures.
4) In the VGG-16 and VGG-19 models, dropout and batch normalization layers were simultaneously added, based on a previous study by Garbin et al. [58], which suggests that this combination can be effective under certain circumstances.

The results obtained with the best configuration for each CNN under study are given in Table 5. When comparing these results with those in Table 4, a slight improvement in the performance metrics is observed across all CNNs. Thus, the adjustments performed made to each architecture proved to be effective. After modifying their architectures, the VGG-16, VGG-19, ResNet-50, and DenseNet-201 networks achieved accuracy, precision, recall, and F1-score values above 90%, a significant increase compared to their original versions. Furthermore, these networks achieved a Kappa index greater than 83%, indicating an increase in result accuracy. Another important detail was the performance improvement of the MobileNetV2, InceptionV3, and EfficientNetB0 models, especially regarding Kappa, which was considered "very good".

E. RESULTS WITH MULTILEVEL

Each multilevel configuration follows previously established criteria in Table 2. Individual CNNs with a Kappa value greater than 80% were selected to identify the best configuration. The modified architectures of VGG-16 with 512 batch normalization (BN), VGG-19 with 512 BN, ResNet-50 with 512, and DenseNet-201 with 1,024−512 models were chosen to perform combinations using the multilevel approach. These architectures were combined into sets of two, three, and four members. After extracting the flattened layers, all multilevel models were trained for 50 epochs using a batch size of 32.

We analyzed the 88 multilevel architectures developed to address the classification problem. We evaluated each architecture based on performance metrics, allowing for a thorough comparison of their effectiveness. Table 6 highlights the seven best-performing architectures, showcasing

TABLE 6. Better results found with a multilevel approach and the respective average processing times (⇓ indicates a reduction of Kappa; best values in bold).

Multilevel	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa (%)	Time (ms)
M2_V16V19_S_2048	**95.91**±0.36	**95.92**±0.36	**95.91**±0.36	**95.93**±0.36	**93.28**±0.59	30.61±0.64
M3_V16V19Res50_S_256	95.81±0.24	95.84±0.25	95.81±0.24	95.81±0.24	93.12±0.40	80.75±0.63
M3_V19V16Res50Den201_S_256	95.26±0.41	95.33±0.41	95.26±0.41	95.27±0.41	92.23±0.68	88.30±0.36
M3_V16Res50_S_256	94.89±0.38	94.92±0.39	94.89±0.38	94.89±0.38	91.61±0.62	41.26±1.09
M3_V19Res50_N_256	94.85±0.74	94.90±0.76	94.85±0.74	94.86±0.75	91.56±1.23	50.64±0.63
M3_V16Den201_S_1024	94.38±0.37	94.45±0.38	94.38±0.37	94.37±0.38	90.77±0.62	42.03±1.09
⇓ M3_Res50Den201_N_512	92.23±0.67	92.38±0.60	92.23±0.67	92.24±0.65	87.25±1.08	**21.58**±0.03

TABLE 7. Better results found from cross-dataset evaluation using CNNs trained on the DFUC dataset, including respective averages (best values in bold).

Architecture	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	Kappa(%)	Time(ms)
M2_V16V19_S_2048	**79.79**	**85.42**	**79.78**	**79.02**	**59.80**	**10.09**
VGG-19 512 BN	79.14	85.15	79.14	78.05	57.73	157.68
M3_V19Res50_N_256	78.48	84.69	78.48	77.30	56.38	51.07
M3_V16V19Res50_S_256	78.10	84.64	78.10	76.84	55.60	78.84
DenseNet-201	77.63	84.11	77.63	76.32	54.64	12.01
M3_V16Res50_S_256	77.16	81.48	77.16	76.17	53.76	42.99
M3_Res50Den201_N_512	76.68	83.20	76.68	75.27	52.70	22.85

superior accuracy, robustness, generalization capability, and shorter training time.

Analyzing Table 6, one can observe a reduction of 1.96% in the Kappa achieved by the multilevel M3_Res50Den201_N_512 compared to the best value obtained by the individual CNN. Additionally, the best result was obtained by the multilevel M2_V16V19_S_2048 model, composed of two CNNs, VGG-19 and VGG-16 architectures, with the insertion of dropout and a fully connected layer with 2,048 neurons.

The M2_V16V19_S_2048 model achieved an accuracy of 95.91% and a Kappa of 93.28%, indicating a high agreement with the specialist. It is worth mentioning that combining CNNs with the multilevel approach improved Kappa compared to individual CNNs, with results above 83% for the four best performances. The M2_V16V19_S_2048 model presented an average runtimes of approximately 30.61 and 57.70 milliseconds in training and inference times, respectively, with a minimal standard deviation. This stability suggests that the execution duration remained consistent regardless of changes in the architecture.

Although other configurations achieved faster inference times, such as M3_Res50Den201_N_512, which recorded an inference time of 21.58 milliseconds, training time of 57.70 milliseconds, and a Kappa of 87.25%, this reduction in processing time did not correspond to superior predictive performance. Models with lower computational costs may not provide maximum agreement. Despite requiring slightly more time, the M2_V16V19_S_2048 model demonstrated the highest performance, with greater confidence in class prediction and reliability for clinical diagnosis.

To better understand the model performance, Fig. 7 presents a graphical representation of the best results obtained in experiments with multilevel CNNs. This figures highlights the performance improvements achieved through the multilevel approach compared to the best results obtained by individual networks. By integrating information more effectively than conventional models, this approach captures the complexities and nuances of the data with greater accuracy.

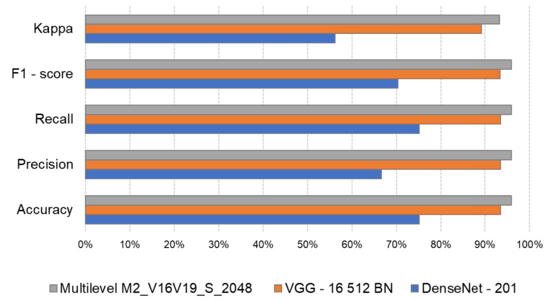


FIGURE 7. Comparison of the best results obtained by individual CNN configurations, modified CNN architecture, and multilevel.

Additional experiments were conducted using the Kaggle DFU dataset through a cross-dataset method to assess the generalization ability of the proposed multilevel model beyond the training conditions. This experiment enabled a more comprehensive evaluation, simulating a real-world scenario in which the model encounters images captured under different lighting conditions, camera angles, acquisition devices, and patient populations. As shown in Table 7, the M2_V16V19_S_2048 model maintained its performance in a scenario with architecture pre-trained using the DFUC dataset, achieving an accuracy and recall of approximately 79.7%, a precision of 85.42%, an F1-score of 79.02%, and a Kappa coefficient of 59.80%, which is considered "good" according to the literature. The model also achieved an average inference time of approximately 10 milliseconds per image. These findings highlight the potential of the multilevel approach to support diabetic foot ulcer diagnosis in other clinical and demographic settings.

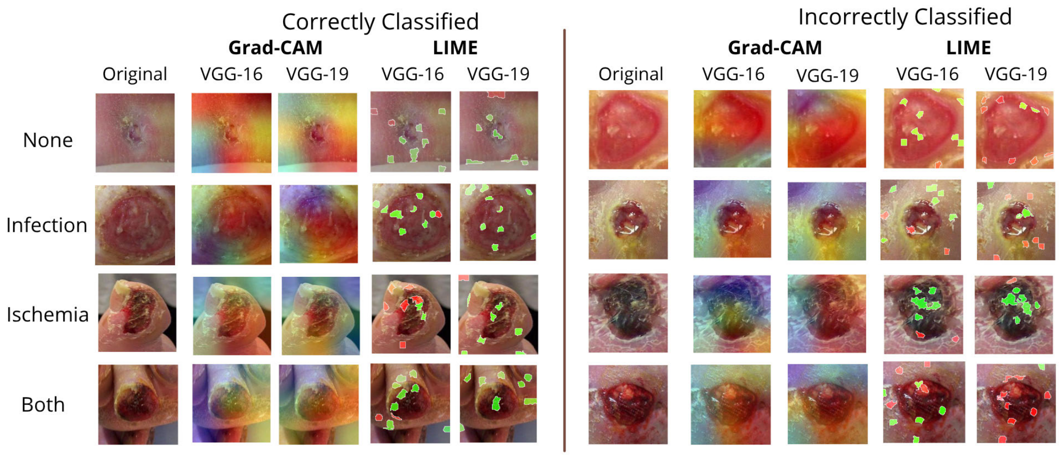


FIGURE 8. Samples with visual interpretations of classified images by CNNs for each class under study in DFUC dataset.

FIGURE 9. Samples with visual interpretations of classified images by CNNs for each class under study in the Kaggle DFU dataset.

The GRAD-CAM explainable artificial intelligence (XAI) method was applied to the test images to provide a clearer interpretation of the proposed model, highlighting the regions that most significantly influence classification. The heatmaps generated by GRAD-CAM illustrate the activation regions of the CNNs; the red tones in the maps are associated with regions that contributed significantly to the final classification, whereas the others contributed less. LIME was used to complement visual interpretability analysis. Unlike GRAD-CAM, which shows heatmaps, LIME perturbs the input image to evaluate its local impact on model predictions. The ten principal superpixels show regions that contribute positively (green superpixels) and negatively (red superpixels).

This combination of techniques provides a clearer insight into the decision-making process of the model by integrating both interpretability methods. Figures 8 and 9 illustrate the areas of interest highlighted by GRAD-CAM and LIME in the M2_V16V19_S_2048 model. GRAD-CAM generated different heatmaps across all examples, showcasing the multilevel model's ability to utilize diverse image features to make its final predictions. This variation further supports the robustness of the proposed method and demonstrates its effectiveness in harnessing multiple levels of information for DFU classification.

In Fig. 8 and Fig. 9, the images were randomly selected for analysis, showing correctly and incorrectly classified examples. In the correctly classified samples in CNNs, it can be observed that both Grad-CAM and LIME highlighteds the lesion regions as critical areas, even when these lesions are not visually evident, as in the cases of the second and four examples, in which VGG-19 focused on some skin areas around the lesions. The integration of these features is highly relevant to DFU classification.

Figures present examples of false positives and false negatives, illustrating samples incorrectly classified as DFU, None and Normal. False negatives are serious, as they can result in missed diagnoses or delayed treatment, posing a risk to the patient's health.

The example in Fig. 8, "none" illustrates a false positive case. In this instance, the image displayed characteristics similar to those highlighted in the correctly classified "infection" image. Grad-CAM and LIME identified the area of a healed ulcer as the region of interest. In the "both" example, a false negative case is highlighted. In this case, interpretations show that the networks considered skin regions the most relevant areas in the image, resulting in an incorrect classification. Factors contributing to this error may include low image quality or the presence of artifacts. Also, we observed that these errors predominantly occur in the infection image class due to the more significant presence of skin regions textcolorblack, as emphasized by the LIME analysis.

In ‘‘ischemia’’ and ‘‘both’’ classified images were incorrectly classified as both or infection. In these cases, the networks identified highly similar characteristics between the classes, making distinction more challenging. The interpretability analysis based on LIME supported this observation by highlighting lesion regions as the main contributors to the classification in the ‘‘ischemia’’ case. In the ‘‘both’’ example, skin and lesion regions influenced the decision-making process. The sample is misclassified since the ‘‘infection’’ class contains more images with similar characteristics.

The selected samples used to validate the model interpretability in the experiments conducted on the Kaggle DFU dataset are presented in Fig.9. Heatmaps highlighted the ulcer regions in the ‘‘abnormal’’ cases, whereas, in the ‘‘normal’’ cases, they focused on healthy skin areas. LIME explanations confirmed this information by focusing on the same region. In contrast, in the analysis of incorrect classification, in the ‘‘abnormal’’ case, Grad-CAM and LIME explanations identified healthy skin regions instead of the ulcer area, resulting in false negatives. In the ‘‘normal’’ case, Grad-CAM highlighted artifacts such as lighting variations as significant features, leading to false positives. These instances emphasize the need to incorporate an artifact removal stage. However, the confusion matrix for the Kaggle DFU dataset demonstrates the effectiveness of the multilevel approach: of the 512 ‘‘abnormal’’ cases, 339 were correctly classified, while among the 543 ‘‘normal’’ cases, only two images were classified incorrectly.

F. COMPARISON WITH STATE-OF-THE-ART MODELS

Based on Table 8 and Table 9, the M2_V16V19_S_2048 multilevel model was compared with state-of-the-art models regarding the architecture, dataset size, and performance metrics.

In Table 8, the analyzed studies address the challenge of multiclass classification of foot ulcers, considering the classes none, infection, ischemia, and both. The works of Ai et al. [23], [24] highlighted the scarcity of data and clinical subjectivity as the main obstacles. A common point between the two approaches is the adoption of asymmetric convolution modules (AC/ACM), which aim to improve the extraction of local features, emphasizing central regions of the image and reduce the redundancy and computational cost.

However, the basic architectures adopted diverge in studies: ACTNet combines the AC module with a structure based on Vision Transformer (ViT), enhanced by a sequence pooling layer, while the other work inserts ACM modules in residual blocks of ResNet34, using the LeakyReLU activation function to preserve gradients. In the tests performed with the DFUC 2021 dataset, ACTNet obtained an F1-score of 59.30%, whereas the ResNet-34 based network reached an F1-score of 58.50%.

In comparison, the proposed multilevel model demonstrated superior performance, achieving an F1-score of 95.93%, as indicated in Table 8. This model uses a larger dataset (DFUC 2020 and 2021) and employs pre-trained architectures VGG-16 and VGG-19 to extract image features. The feature vector was used to predict the test images. The combination of these architectures contributed to a greater generalization capacity in the classification task, highlighting the potential of the multilevel approach to methods based on transformers or residual convolutional networks mentioned.

The studies by Santos et al. [25] and Santos et al. [30] employed data augmentation on the training dataset, utilized convolutional neural networks (CNNs), and applied transfer learning techniques with models pre-trained on the ImageNet dataset. Regarding the adopted architectures, Santos et al. [25] proposed DFU-VGG, a modified version of VGG-19, whereas Santos et al. [30] implemented an ensemble-based approach, integrating the predictions of multiple networks to produce the final classifier output.

The DFU-VGG architecture incorporates batch normalization layers after the convolutional blocks to reduce the dimensionality of the dense layers. In contrast, the ensemble model consisted of adapted versions of VGG-16, VGG-19, ResNet-50, InceptionV3, and DenseNet-201. The final prediction was obtained through weighted majority voting among the individual classifiers. Both studies employed five-fold cross-validation (k=5), applying the models to DFUC 2020 and 2021 datasets. The results demonstrated a performance of above 89% in accuracy, precision, recall, F1-score, and Kappa.

Under comparable experimental conditions, the proposed multilevel model was evaluated using five-fold cross-validation ($k = 5$) and data augmentation of the training dataset. However, after removing duplicate images, the dataset was reduced to 8,242. Unlike the ensemble approach, which combines the predictions from different networks, the multilevel model performs feature concatenation from two distinct networks (VGG-16 and VGG-19), using concatenate representation as input to the classification stage. This strategy resulted in superior performance across all evaluated metrics, surpassing the results of the models proposed by Santos et al. [25] and Santos et al. [30] in the DFU classification tasks.

Table 9 presents additional studies that used the same Kaggle DFU dataset. The works of Biswas et al. [31] and Thotad et al. [32] highlight the use of data augmentation techniques during the training phase, which enhanced model performance. Both studies employed the same dataset for training and testing, which favors the adjustment of models to the specific characteristics of the dataset.

Biswas et al. [31] combined features from DenseNet201, VGG19, and NASNetMobile using Global Average Pooling (GAP) for binary classification, whereas Thotad et al. employed EfficientNet individually. In contrast, the proposed multilevel model combines features extracted from VGG-16 and VGG-19, trained on the DFUC dataset, to evaluate

TABLE 8. Comparison among the results achieved with the proposed multilevel model and those from the state-of-the-art models (best values in bold).

Work	Architecture	Images	Performance(%)
Ai, Yang and Xie [23]	ACTNet	5,734	F1-score: 59.30
Ai, Yang and Xie [24]	Asymmetric convolution, LeakyRelu, ResNet-34	5,734	F1-score: 58.50
Proposed Multilevel	Multilevel [VGG16, VGG19]	8,242	**F1-score: 95.93**
Santos F. *et al.* [25]	DFU-VGG	8,250	Accuracy: 93.45, Precision: 93.56, Recall: 93.45, F1-score: 93.46, Kappa: 89.24
Santos E. *et al.* [30]	CNN Ensemble [VGG-16, VGG-19, ResNet-50, InceptionV3, DenseNet-201]	8,250	Accuracy: 95.04, Precision: 95.06, Recall: 95.04, F1-score: 95.38, Kappa: 91.85
Proposed Multilevel	Multilevel [VGG16, VGG19]	8,242	**Accuracy: 95.91, Precision: 95.92, Recall: 95.91, F1-score: 95.93, Kappa: 93.28**

TABLE 9. Comparison among the results achieved with the proposed multilevel model and those from the state-of-the-art models in the Kaggle DFU dataset (best values in bold).

Work	Architecture	Images	Performance(%)
Biswas *et al.* [31]	DenseNet201, VGG19, NASNetMobile	1,055	**Accuracy: 99.05, Precision: 100, Recall: 98.18, F1-score: 99.08**
Thotad *et al.* [32]	EfficientNet	1,055	Accuracy: 98.97, Precision: 99, Recall: 98, F1-score: 98
Proposed Multilevel	Multilevel [VGG16, VGG19]	1,055	Accuracy: 79.79, Precision: 85.42, Recall: 79.78, F1-score: 79.02

the Kaggle DFU dataset without prior exposure to these images during training and without applying data augmentation techniques to this dataset. This evaluation followed a cross-dataset validation protocol adopted to assess the model's generalization capability when faced with data acquired under different conditions. Although the results were lower than those of models trained directly on the Kaggle DFU dataset, the accuracy, precision, recall, and F1-score remained above 79%, demonstrating the potential for application in scenarios involving different domains. Furthermore, in scenarios where the networks were trained using the DFUC dataset, the proposed multilevel model outperformed all the other models.

V. CONCLUSION AND FUTURE WORK

The DFUC and Kaggle DFU datasets comprised foot images collected from the two hospitals. They present various ulcer healing stages, image resolutions, capture angles, lighting conditions, and acquisition devices. Data augmentation techniques such as rotations and flips were applied exclusively to DFUC images to expand the dataset and reduce the developed model's overfitting. These methods, used in DFU classification, provide valuable insights to benefit researchers exploring similar topics. The Kaggle DFU dataset, with no data augmentation, was used to evaluate the proposed generalization capability of the proposed model.

Our experiments selected seven convolutional neural networks (CNN) architectures from existing literature: ResNet-50, InceptionV3, DenseNet-201, MobileNetV2, EfficientNetB0, VGG-16, and VGG-19. These networks performed well on the ImageNet dataset, achieving Top-1 accuracy exceeding 70%. Each network was individually analyzed in two scenarios: without altering the architecture and with architecture modification. Based on the individual performances, the networks with the best results were selected to compose the multilevel model.

Combining multiple CNNs enhances generalization and reduces variance in the proposed method. This combination produced excellent results, with the multilevel model achieving the highest classification performance for four and two image classes. The model was evaluated using accuracy, precision, recall, F1-score, and Kappa metrics. It achieved an accuracy of 95.91%, comparable to previously reported values, along with a Kappa value of 93.28% with four classes, indicating a high level of agreement with the ground-truth. The binary classification achieved an accuracy of 79.79% and a Kappa value of 59.80%. The proposed multilevel model also reduced standard deviation and improved other evaluation metrics by integrating more refined CNNs. These findings may contribute to the development of a clinically effective solution for classifying DFU images through the strategic combination of CNNs. In addition, we intend to integrate the autoencoder method after the feature concatenation stage to reduce dimensionality and emphasize the most informative features for classifying the image. This strategy will enhance the efficiency, interpretability, and scalability of the model in future developments.

ACKNOWLEDGMENT

The authors express gratitude for their financial support throughout this study. Funding was essential for the completion of this study.

REFERENCES

[1] I. D. Atlas. (Jun. 2024). *IDF Diabetes Atlas Reports*. [Online]. Available: https://diabetesatlas.org/atlas-reports/

[2] M. Gandhi, E. Fargo, L. Prasad-Reddy, K. M. Mahoney, and D. Isaacs, "Diabetes: How to manage diabetic peripheral neuropathy," *Drugs Context*, vol. 11, pp. 1–13, Jun. 2022.

[3] J. Golledge, "Update on the pathophysiology and medical treatment of peripheral artery disease," *Nature Rev. Cardiology*, vol. 19, no. 7, pp. 456–474, Jul. 2022.

[4] K. Alexiadou and J. Doupis, "Management of diabetic foot ulcers," *Diabetes Therapy*, vol. 3, no. 1, p. 4, Apr. 2012.

[5] J. C. Rodríguez and S. Guerrero, "Psychosocial factors of patients with venous leg ulcers and their association with healing," *ESTIMA, Brazilian J. Enterostomal Therapy*, vol. 18, p. e0720, Jun. 2020.

[6] E. Santos, R. Veras, H. Miguel, K. Aires, M. L. Claro, and G. B. Junior, "A skin lesion semi-supervised segmentation method," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2020, pp. 33–38.

[7] L. Shamir, S. Rahimi, N. Orlov, L. Ferrucci, and I. G. Goldberg, "Progression analysis and stage discovery in continuous physiological processes using image computing," *EURASIP J. Bioinf. Syst. Biol.*, vol. 2010, pp. 1–7, May 2010.

[8] B. Rostami, D. M. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, "Multiclass wound image classification using an ensemble deep CNN-based classifier," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104536.

[9] M. Kaselimi, E. Protopapadakis, A. Doulamis, and N. Doulamis, "A review of non-invasive sensors and artificial intelligence models for diabetic foot monitoring," *Frontiers Physiol.*, vol. 13, Oct. 2022, Art. no. 924546.
[10] 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 Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 5, pp. 728–739, Oct. 2020.
[11] S. K. Das, P. Roy, and A. K. Mishra, "DFU_SPNet: A stacked parallel convolution layers based CNN to improve diabetic foot ulcer classification," *ICT Exp.*, vol. 8, no. 2, pp. 271–275, Jun. 2022.
[12] C. Venkatesan, M. G. Sumithra, and M. Murugappan, "NFU-Net: An automated framework for the detection of neurotrophic foot ulcer using deep convolutional neural network," *Neural Process. Lett.*, vol. 54, no. 5, pp. 3705–3726, Oct. 2022.
[13] 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 Appl.*, vol. 79, nos. 21–22, pp. 15655–15677, Jun. 2020.
[14] 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. Healthcare Eng.*, vol. 2022, pp. 1–12, Apr. 2022.
[15] N. Al-Garaawi, R. Ebsim, A. F. H. Alharan, and M. H. Yap, "Diabetic foot ulcer classification using mapped binary patterns and convolutional neural networks," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105055.
[16] T. A. M. Devi and R. Hepzibai, "Clinical assessment of diabetic foot ulcers using GWO-CNN based hyperspectral image processing approach," *IETE J. Res.*, vol. 69, no. 12, pp. 8705–8716, Dec. 2023.
[17] T. Lan, Z. Li, and J. Chen, "FusionSegNet: Fusing global foot features and local wound features to diagnose diabetic foot," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106456.
[18] Z. Liu, J. John, and E. Agu, "Diabetic foot ulcer ischemia and infection classification using EfficientNet deep learning models," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 189–201, 2022.
[19] C. Y. Rubavathi and J. Diofrin, "Diabetes foot ulcer diagnosis using fast convolution neural network," in *Proc. Int. Conf. Netw. Commun. (ICNWC)*, Apr. 2023, pp. 1–5.
[20] V. S. Preiya and V. D. A. Kumar, "Deep learning-based classification and feature extraction for predicting pathogenesis of foot ulcers in patients with diabetes," *Diagnostics*, vol. 13, no. 12, p. 1983, Jun. 2023.
[21] Y. Xu, K. Han, Y. Zhou, J. Wu, X. Xie, and W. Xiang, "Classification of diabetic foot ulcers using class knowledge banks," *Frontiers Bioeng. Biotechnol.*, vol. 9, p. 1531, Feb. 2022.
[22] 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 Syst.*, vol. 30, no. 1, p. 36, Jan. 2024.
[23] L. Ai, M. Yang, and Z. Xie, "ACTNet: Asymmetric convolutional transformer network for diabetic foot ulcers classification," *Phys. Eng. Sci. Med.*, vol. 45, no. 4, pp. 1175–1181, Dec. 2022.
[24] L. Ai, M. Yang, and Z. Xie, "Improved residual connection network for diabetic foot ulcers classification," in *Proc. 6th Int. Conf. Comput. Sci. Appl. Eng.*, New York, NY, USA, Oct. 2022, pp. 1–5.
[25] F. Santos, E. Santos, L. H. Vogado, M. Ito, A. Bianchi, J. M. Tavares, and R. Veras, "DFU-VGG, a novel and improved VGG-19 network for diabetic foot ulcer classification," in *Proc. 29th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jun. 2022, pp. 1–4.
[26] N. Bansal and A. Vidyarthi, "DFootNet: A domain adaptive classification framework for diabetic foot ulcers using dense neural network architecture," *Cognit. Comput.*, vol. 16, no. 5, pp. 2511–2527, Sep. 2024.
[27] C. Gamage, I. Wijesinghe, and I. Perera, "Automatic scoring of diabetic foot ulcers through deep CNN based feature extraction with low rank matrix factorization," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioengineering (BIBE)*, Oct. 2019, pp. 352–356.
[28] C. Cao, Y. Qiu, Z. Wang, J. Ou, J. Wang, A. H. Hounye, M. Hou, Q. Zhou, and J. Zhang, "Nested segmentation and multi-level classification of diabetic foot ulcer based on mask R-CNN," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18887–18906, May 2023.
[29] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Comput. Biol. Med.*, vol. 117, Feb. 2020, Art. no. 103616.
[30] E. Santos, F. Santos, J. Dallyson, K. Aires, J. M. R. S. Tavares, and R. Veras, "Diabetic foot ulcers classification using a fine-tuned CNNs ensemble," in *Proc. IEEE 35th Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jul. 2022, pp. 282–287.
[31] S. Biswas, R. Mostafiz, M. S. Uddin, and B. K. Paul, "XAI-FusionNet: Diabetic foot ulcer detection based on multi-scale feature fusion with explainable artificial intelligence," *Heliyon*, vol. 10, no. 10, May 2024, Art. no. e31228.
[32] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Diabetic foot ulcer detection using deep learning approaches," *Sensors Int.*, vol. 4, Jan. 2023, Art. no. 100210.
[33] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019.
[34] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–25, Sep. 2022.
[35] L. Vogado, R. Veras, and K. Aires, "'LeukNet'—A model of convolutional neural network for the diagnosis of leukemia," in *Proc. SIBGRAPI*, Jun. 2020, pp. 119–125.
[36] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, May 2016.
[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.
[38] M. Izadyyazdanabadi, E. Belykh, M. Mooney, N. Martirosyan, J. Eschbacher, P. Nakaji, M. C. Preul, and Y. Yang, "Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization for neurosurgical CLE images," *J. Vis. Commun. Image Represent.*, vol. 54, pp. 10–20, Jul. 2018.
[39] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[40] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, and L. Shi, "$L1$-norm batch normalization for efficient training of deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2043–2051, Jul. 2019.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[43] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[45] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 6105–6114.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Jan. 2015, pp. 1–14.
[47] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2656–2666.
[48] J. Lyu and S. H. Ling, "Using multi-level convolutional neural network for classification of lung nodules on CT images," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 686–689.
[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[50] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 1135–1144.
[51] M. H. Yap, N. Reeves, A. Boulton, S. Rajbhandari, D. Armstrong, A. G. Maiya, B. Najafi, E. Frank, and J. Wu, Mar. 2020, "Diabetic foot ulcers grand challenge 2020," *Zenodo*.

[52] M. H. Yap, N. Reeves, A. Boulton, S. Rajbhandari, D. Armstrong, A. G. Maiya, B. Najafi, E. Frank, and J. Wu, Mar. 2021, "Diabetic foot ulcers grand challenge 2021," *Zenodo*.

[53] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[54] S. O. Oyibo, E. B. Jude, I. Tarawneh, H. C. Nguyen, L. B. Harkless, and A. J. M. Boulton, "A comparison of two diabetic foot ulcer classification systems: The Wagner and the University of Texas wound classification systems," *Diabetes Care*, vol. 24, no. 1, pp. 84–88, Jan. 2001.

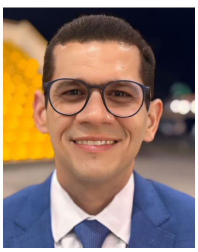
[55] B. Cassidy, N. D. Reeves, J. M. Pappachan, D. Gillespie, C. O'Shea, S. Rajbhandari, A. G. Maiya, E. Frank, A. J. M. Boulton, D. G. Armstrong, B. Najafi, J. Wu, R. S. Kochhar, and M. H. Yap, "The DFUC 2020 dataset: Analysis towards diabetic foot ulcer detection," *touchREV. Endocrinol.*, vol. 17, no. 1, p. 5, Jan. 2021.

[56] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, and Y. Duan, "Robust application of new deep learning tools: An experimental study in medical imaging," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 13289–13317, Apr. 2022.

[57] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[58] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. Batch normalization: An empirical study of their impact to deep learning," *Multimedia Tools Appl.*, vol. 79, nos. 19–20, pp. 12777–12815, May 2020.

ELINEIDE SILVA DOS SANTOS received the master's degree in computer science from the Universidade Federal do Piauí (UFPI). She has experience with digital image processing, computer-based diagnostics, medical informatics, deep learning, and artificial intelligence. She is currently involved in research projects focused on diagnostic assistance for diabetic foot lesions. In addition, she performed image segmentation for the skin lesions and glaucoma. She has published research papers in international journals and conferences.



RODRIGO DE MELO SOUZA VERAS received the bachelor's degree in computer science from the Universidade Federal do Piauí (UFPI), Brazil, in 2005, and the master's degree in computer science and the Ph.D. degree in teleinformatics engineering from the Universidade Federal do Ceará (UFC), Brazil, in 2007 and 2014, respectively.

He is currently an Associate Professor with the Department of Computing, UFPI, and a Permanent Member with the Graduate Program in Computer Science, UFPI, and the Doctoral Program in Computer Science (UFMA/UFPI). He has extensive experience in the fields of computer vision and medical image processing. Since 2007, he has published numerous research articles focusing on diseases, such as skin cancer, leukemia, cervical cancer, macular edema, diabetic retinopathy, glaucoma, corneal ulcers, diabetic foot ulcers, and kidney pathologies. His current research interests include deep learning applications in medical imaging, automated disease diagnosis, and the development of computational tools to assist clinical decision-making.

FRANCISCO DAS CHAGAS TORRES DOS SANTOS received the bachelor's degree in computer science from the Universidade Federal do Piauí—UFPI, in 2012, with a specialization in Databases from the Faculty of Technology of Piaui, in 2017, and the master's degree in computer science from the Universidade Federal do Piauí, in 2022. He has worked as an Information Technology (Development) Analyst at the Superintendence of Information Technology, since 2014.

MÁRCIA ITO is currently pursuing the Ph.D. degree in electrical engineering with EPUSP. She is a Medical Doctor with EPM/UNIFESP and a Data Processing Technologist with FATEC-SP. She is an Advisor with the Master's Program in Productive Systems, CPS, and a Professor with FATEC-SP. She is a fellow of the International Academy of Health Science Informatics and a member of the Telehealth Working Group of the International Medical Informatics Association. She was the General Coordinator of the Strategic Information Management at the Ministry of Health (DEMAS/SE/MS), from 2019 to 2020, a Researcher with the IBM Research Brazil, from 2012 to 2018, a member of the Management Committee of the Brazilian Computer Society's Special Commission on Computing Applied to Health, from 2008 to 2025, and the Editor-in-Chief of *Journal of Health Informatics*, Brazilian Society of Health Informatics, from 2020 to 2024. Also, she is the Editor-in-Chief of *Journal of Digital Health and Biomedical Informatics*, Brazilian Computer Society's Special Commission on Computing Applied to Health.

ANDREA GOMES CAMPOS BIANCHI received the B.Sc. degree in physics with an emphasis in computational physics, and the M.Sc. and Ph.D. degrees in applied computational physics from the Universidade de São Paulo, Brazil, in 1996, 1998, and 2003, respectively. She completed postdoctoral studies at the Universidade Federal de Ouro Preto (UFOP) and the Lawrence Berkeley National Laboratory (LBNL), USA. She is a Full Professor with the Department of Computing at UFOP, Brazil, and a Permanent Member of the Graduate Programs in Computer Science and Instrumentation and Automation in Mining Processes. Her research interests include computer vision, image processing and analysis, pattern recognition, and machine learning, focusing on developing intelligent systems for decision support in healthcare and industrial inspection. She has led interdisciplinary projects in Artificial Intelligence applied to Health, published and received recognition for her contributions to biomedical image analysis.

KELSON RÔMULO TEIXEIRA AIRES received the bachelor's degree in electrical engineering from the Universidade Federal da Paraíba (UFPB), Brazil, in 1999, the master's degree in electrical engineering and the Ph.D. degree in computer engineering from the Universidade Federal do Rio Grande do Norte (UFRN), Brazil, in 2001 and 2009, respectively.

His Ph.D. research included a sandwich period at the Universidade de Coimbra, Portugal. Since 2012, he has been a Permanent Member with the Graduate Program in Computer Science (PPGCC-UFPI), and since 2018, he has been affiliated with the Doctoral Program in Computer Science–UFMA/UFPI Association. He is currently an Associate Professor with the Universidade Federal do Piauí (UFPI), Brazil. He has extensive teaching experience in computer science, focusing on topics, such as graphics processing and computer vision. He has contributed significantly to the academic community through his involvement in research projects and publications in these fields. His research interests include image processing, computer vision, and intelligent systems.

JOÃO MANUEL R. S. TAVARES received the degree in mechanical engineering, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Universidade do Porto, Portugal, in 1992, 1995, and 2001, respectively, and the Habilitation degree in mechanical engineering, in 2015.

He is currently a Senior Researcher with Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial (INEGI) and a Full Professor with Departamento de Engenharia Mecânica (DEMec), Faculdade de Engenharia da Universidade do Porto (FEUP). Since June 2023, he has been the Head of DEMec. He has made significant contributions to his field, serving as the co-editor of more than 90 books and co-authoring more than 50 book chapters and 650 articles in international and national journals and conferences. He also holds three international and three national patents. He has also (co-)supervised numerous M.Sc. and Ph.D. theses and supervised several postdoctoral projects, participating in many scientific projects as both a researcher and scientific coordinator. His research interests include computational vision, medical imaging, biomechanics, biomedical engineering, and new product development.

Dr. Tavares is the Co-Founder and the Co-Chair of several international conference series, including the International Symposium on Computational Modeling of Objects Presented in Images, the ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, the International Conference on Computational and Experimental Biomedical Sciences, and the International Conference on Biodental Engineering. His editorial roles are extensive; he is a committee member for numerous international and national journals and conferences, the Co-Founder and the Co-Editor of *Lecture Notes in Computational Vision and Biomechanics* series published by Springer, the Founder and the Editor-in-Chief of *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* published by Taylor and Francis, and the Editor-in-Chief of *Computer Methods in Biomechanics and Biomedical Engineering*, also published by Taylor and Francis. More details are available at www.fe.up.pt/tavares.