DFU-VGG, a Novel and Improved VGG-19 Network for Diabetic Foot Ulcer Classification

Francisco Santos*, Elineide Santos*, Luís Henrique Vogado*, Márcia Ito[†], Andrea Bianchi[‡],
João Manuel Tavares[§] and Rodrigo Veras*
*Universidade Federal do Piauí, Teresina, Brasil
[†] Mestrado Profissional em Sistema Produtivo/CEETEPS, São Paulo, Brasil
[‡] Universidade Federal de Ouro Preto, Minas Gerais, Brasil
[§]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
Email: {fsantos, rveras}@ufpi.edu.br, {elineide.silva.inf,lhvogado,marciaito2000}@gmail.com,
andrea@ufop.edu.br, tavares@fe.up.pt

***Abstract*—A complication caused by diabetes mellitus is the appearance of lesions in the foot region called Diabetic Foot Ulcers (DFU). Delayed treatment can lead to infection or ulcer ischemia, leading to lower limb amputation in an advanced stage. This article proposes the DFU-VGG, a convolutional neural network (CNN) inspired by convolutional blocks of VGG-19 but with smaller dense layers and batch normalizations operations. To specify the DFU-VGG parameters, we fine-tuned seven different CNN architectures using two image datasets containing 8,250 images with different color, contrast, resolution, and texture features. The proposed evaluation identifies four classes: none, ischemia, infection, and both. Our approach achieved 93.45% of accuracy and an "excellent" Kappa index of 89.24%.**

I. Introduction

An estimated 536.6 million people will live with diabetes by 2021 [1]. This number is expected to increase by 46% in 2045. This disease can cause many blindness, cardiovascular disease, kidney failure, and diabetic foot ulcers.

Ulcers result in wounds in the foot region, usually caused by trauma, repetitive mechanical stress, or continuously applied mechanical stress. Diabetic foot ulcers (DFU) need proper treatment, as they can lead to the amputation of infected limbs in an advanced stage. Thus, an early diagnosis can delay the development of the disease and prevent adverse scenarios.

Severe injuries can be classified as infection or ischemia. Infection, Figure 1b, is recognized by the presence of inflammation or purulence, as well as increased redness around the ulcer. Ischemia, Figure 1c, is the inadequate circulation of blood through the lesion, being visually identified by the presence of poor reperfusion in the gangrened foot or toes. In some cases, as in Figure 1d, the ulcer has both ischemia and infection. However, after treatment, the ulcers reach a healing state and resemble healthy skin, as shown in Figure 1a.

Monitoring diabetic foot injuries is usually done by visually inspecting the injured areas and observing the signs and symptoms of diabetes. Thus, the assessment relies on the specialist's subjective criteria. In this context, using a diagnostic assistance system can support the specialist and enable automatic monitoring of injuries.

978-1-6654-9578-3/22/$31.00 ©2022 IEEE

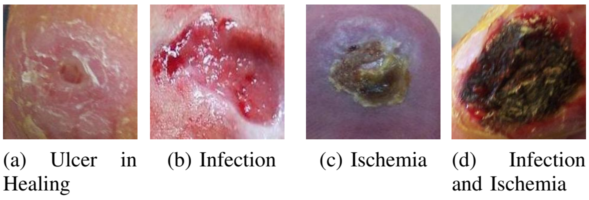
(a) Ulcer in Healing (b) Infection (c) Ischemia (d) Infection and Ischemia

Fig. 1. Examples of diabetic foot ulcer images.

Therefore, this work proposes a neural network for classifying diabetic foot injuries. To do this, we evaluated and refined the architecture of seven general-purpose CNNs. The seven CNNs were analyzed in five different scenarios: (1) their original architectures, (2) changing the dense layers, (3) addition of dropout layers (DP), (4) addition of batch normalization layers (BN), and (5) addition of dropout and batch normalization.

This article is organized as follows: in Section II recent works and methodologies on the problem under study are presented. Section III presents the proposed method, the image datasets, the applied techniques, and the evaluation metrics adopted in the development work. In Sections IV, the results and their discussion are presented. Finally, conclusions and future work are presented in Section V.

II. Related Works

The problem of classifying diabetic foot injuries in recent years has been addressed by some studies in the literature using automatic methodologies. Table I summarizes the found works in terms of year of publication, used classification technique(s), number of images, number of classes, and the performance achieved, which can be as to accuracy (A), sensitivity (S), specificity (E), precision (P) and area under the curve (AUC).

Analyzing the works indicated in Table I, one can realize that the approaches, in the majority, combine neural networks with data augmentation techniques. Furthermore, there is no standard regarding the number of images, the number of classes, or the choice of evaluation metrics.

TABLE I
SUMMARY OF THE IDENTIFIED STATE-OF-THE-ART WORKS.

Work	Classification technique(s)	Images	Classes	Performance(%)
[2]	Neural Networks, Bayesian Classifiers	113	5	A: 91.50
[3]	SVM	100	2	S: 73.30 S: 94.60
[4]	DFUNet	1,423	2	A: 92.50
[5]	DFU-QUTNet, SVM, KNN	754	2	P: 95.40
[6]	InceptionV3, ResNet50, InceptionResNetV2, SVM	1,459	2	A Isc: 90.00 A Inf: 73.00
[7]	Neural Networks, Naive bayes, Neural Networks, Decision tree	15,762	2	A Isc: 97.90 A Inf: 99.60
[8]	DFU_SPNet	1,679	2	A: 96.40
[9]	AlexNet, Sliding window, MLP	400	3	A max: 91.90 A average: 87.70
[10]	BiT-ResNeXt50	15,683	4	AUC: 88.49 P: 60.53

III. MATERIALS AND METHODS

This section presents the DFU-VGG, a CNN capable of classifying four patterns of diabetic foot ulcers. We refined seven CNN architectures, evaluated different combinations of fully connected layers, and the use of dropout and batch normalization operations.

A. Proposed Method

DFU-VGG is a convolutional neural network that uses VGG-19 as a backbone. Batch normalization operations were introduced after convolutional blocks. In addition, we opted for new dense layers with lower dimensionality. Figure 2 presents the changes made to the VGG-19 architecture that led to DFU-VGG.

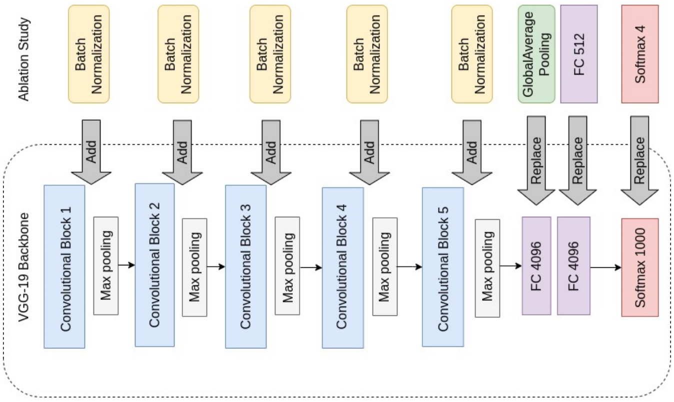

Fig. 2. Changes done to the original architecture of VGG-19 for the effective classification of diabetic foot ulcers.

B. Image dataset

In the experiments, we used the public image dataset of diabetic foot ulcers named Diabetic Foot Ulcer (DFU) in two versions: 2020 and 2021. The images of patients' feet with DFU at Lancashire Teaching Hospitals were captured for five years with three cameras after debridement (removing necrotic and devitalized tissue). The image diagnosis (ground truth) was developed with the help of two specialist physicians. The images were captured centered on the lesion area and had different dimensions; the smallest has 34×31 pixels, while the largest has 1103×1127 pixels.

The images were split into four classes: (1) none, containing images of healthy skin, ulcers in the process of recovery, and ulcers without infection or ischemia; (2) infection; (3) ischemia; and (4) both, with ulcers images containing infection and ischemia at the same time. The DFU 2020 and DFU 2021 datasets contain 3,833 images of the none class, 253 ischemia, 3,334 infections, and 830 both.

C. Data augmentation

It is common sense that CNNs need a large amount of data to be properly trained because they have millions of parameters. Even a tiny CNN requires thousands of images to be trained. One of the possibilities to reduce CNN overfitting and improve the generalization of the trained models is the application of data augmentation techniques.

In this way, we use the data augmentation operations provided by the Keras API. In this strategy, operations on images are applied randomly and at runtime. Our experiments used the following parameters: rotation interval was 40º; shear, zoom, vertical, and horizontal, translation interval was equal to 0.2; we applied vertical and horizontal flip. In addition, we apply the reflection fill technique to replace the black pixels resulting from rotation and translation operations. Finally, we normalized the image pixels to range from 0 to 1. Applying this set of operations generated an image dataset 20 times larger than the initial.

D. Evaluated Architectures

We evaluated CNNs models developed for the ImageNet Large Scale Visual Recognition Challenge. Kornblith et al. [11] concluded that the better the performance of a CNN in the ImageNet dataset, the better the transfer of learning to other datasets. The evaluated architectures are indicated in Table II, being referred in terms of topological depth of the network, number of parameters, and year of publication.

TABLE II
OVERVIEW OF THE EVALUATED DEEP LEARNING MODELS.

CNN	Depth	Parameters	Year	Reference
VGG-16	23	138,357,544	2014	[12]
VGG-19	26	143,667,240	2014	[12]
ResNet50	168	25,636,712	2015	[13]
InceptionV3	159	23,851,784	2016	[14]
DenseNet201	201	20,242,984	2017	[15]
MobileNetV2	88	3,538,984	2018	[16]
EfficientNetB0	240	5,330,571	2019	[17]

E. Transfer Learning

The transfer learning technique is widely applied when it is not desired (or is not feasible) to train all the parameters of a CNN from scratch.

After analyzing the shallow fine-tuning (SFT) and deep fine-tuning (DFT) approaches, we chose to use the DFT. The DFT approach refines all parameters of the CNN, from the shallowest to the deepest layers. Although it has a higher computational cost and requires a more significant amount of data than SFT, it can benefit applications where the problem domain is different from that used in network training. In the case of this work, in particular, the natural photographic images from the ImageNet dataset and the images of diabetic foot ulcers are from very different domains.

F. Dropout and batch normalization

Overfitting and extended training time are two critical challenges in CNNs. Dropout and batch normalization are two well-recognized strategies to tackle these challenges.

The dropout is a regularization technique. Its main feature is to disable, temporarily, some neurons. This effect provides the equivalent of different training architectures since different neurons will be disabled during the training (in other iterations). The use of dropout tends to reduce CNN complexity and overfitting.

The time for a network to converge depends on initializing the hyperparameters and using small learning rates. Also, a layer depends on previous layers, so minor modifications in one layer can be intensified as they flow to the subsequent layers. Batch normalization normalizes the input of each layer of the network. Thus, the training time can be reduced to allow higher learning rates.

G. Evaluation Metrics

To evaluate the performance of CNNs, we calculated the metrics Accuracy (A), Precision (P), Recall (R), F1-score (F) from the confusion matrix values.

In addition, we calculate the kappa index (K). This index considers all elements of the confusion matrix, not just those on the main diagonal, which occurs, for example, with the accuracy of the global classification. In this way, it adequately represents the confusion matrix and is commonly used as an appropriate evaluation measure.

The K values are between 0 and 1 and could qualified in five "status": $K \leq 0.2$: Bad; $0.2 < K \leq 0.4$: Fair; $0.4 < K \leq 0.6$: Good; $0.6 < K \leq 0.8$: Very Good and $K > 0.8$: Excellent [18].

IV. Results and Discussion

The CNNs in their original settings were fine-tuned with input images with 224×224 pixels. The training of networks in all configurations was performed with 500 epochs.

We applied the stratified *k*-fold cross-validation technique with k=5. A confusion matrix was computed for each fold, and the arithmetic average and standard deviation of the five folds achieved from each evaluated CNN was calculated. In addition, Kappa (K) was multiplied by 100 to facilitate understanding of the table. Table III presents the better results for the five evaluated scenarios (one per line).

Initially, we fine-tuned the VGG-16, VGG-19, InceptionV3, ResNet50, DenseNet201, MobileNetV2 and EfficientNetB0 networks in their original configurations. The first row of the table shows that DenseNet201 achieved the best results in all used metrics.

The second row presents the best result of the second experiment. In this experiment, we preserved the CNNs convolutional layers, inserted a Global Average Pooling layer, and then the fully connected layers in two scenarios: (1) a connected layer with the number of neurons assuming the following values: 256, 512, and 1024, and (2) two fully connected layers with configurations of 512-256, 1024-256 and 1024-512 neurons. These configurations led to fewer neurons than original CNNs and, consequently, to a smaller number of weights to be trained. VGG-19 with a dense layer of 512 neurons obtained the best results and a slight standard deviation, indicating CNNs stability in classifying diabetic foot ulcers.

The third row presents the best result of the third experiment. After each block of dense layers, we inserted a dropout layer. Although the Kappa values could be considered "excellent", and there has been a gain in training time, the results obtained were lower than those obtained without dropout.

The fourth row presents the best result of the fourth experiment. We added batch normalization layers in the VGG-16 and VGG-19 (since the other CNNs have normalization layers in their original architecture) in each convolutional block and before the MaxPooling. . VGG-19 networks, with the addition of the normalization layer, obtained accuracy, precision, recall, and F1-Score values above 93.45% and Kappa index 89.24%, which is considered excellent. The normalization layer in each block of VGG-19 models significantly increased the classification success rate.

The results of VGG-19 with a fully connected layer with 512 neurons and with the addition of batch normalization layers were the best ones found in this study. Therefore, this configuration is the proposed solution model, and we named it DFU-VGG.

The fifth row presents the best result of the fifth experiment. Batch normalization layers were added in the sequential networks VGG-16 and VGG-19, in each block of convolutional layers, and before the MaxPooling layers. Dropout layers were added after MaxPooling layers. We can realize that the metrics' values were lower than those of the DFU-VGG (line four).

Table IV presents the DFU-VGG confusion matrix with the average of the five folds. It can be observed that 93.60% of the ischemia images and 91.50% of the infection images were correctly classified. Among the results, the most worrying issue is that 7.65% of the images of ulcers with infection were classified as ulcers without infection and without ischemia, which would make this percentage of patients left without adequate treatment, believing that these ulcers were in the healing process.

Figure 3 shows the heat maps with the activation regions that DFU-VGG considered most important during feature extraction and, consequently, classification.

In the activation maps shown (Figure 3), the blue tones mean low activation and indicate that the corresponding regions are of minor relevance for the final classification; on the other hand, the red tones are associated with the regions whose characteristics most contributed to the classification. Interpreting results from CNNs is still a challenge for researchers, but the DFU-VGG activation maps indicate that it gives more importance to regions with ischemia and infection patterns.

TABLE III
Classification results obtained with the original CNNs (best values in bold).

Experiment	CNN	$A(\%)$	$P(\%)$	$R(\%)$	$F(\%)$	$K(\%)$
Original	DenseNet201	75.06±0.71	66.58±0.26	75.06±0.71	70.24±0.58	56.03±1.10
FC Layers	VGG-19 512	91.40±0.35	91.45±0.37	91.40 ±0.35	91.40±0.37	85.88±0.59
Dropout	VGG-19 + DP	88.86±0.40	89.00±0.39	88.86±0.40	88.87±0.41	81.70±0.65
Batch Normalization	DFU-VGG	**93.45**±0.34	**93.56**±0.30	**93.45**±0.34	**93.46**±0.34	**89.24**±0.58
Dropout + Batch Normalization.	VGG-19	92.36±0.56	92.46±0.54	92.36±0.56	92.36±0.56	87.43±0.93

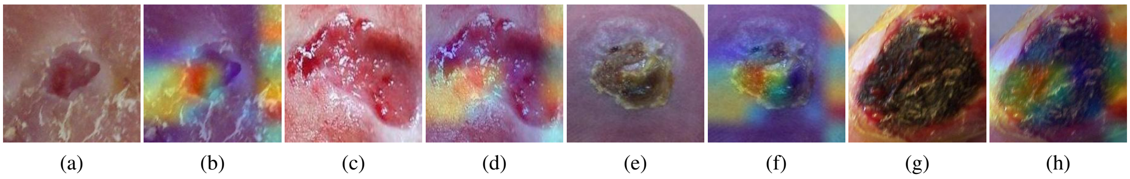

Fig. 3. Ulcer and heat map correspondent with activation regions for classes: none (a, b), infection (c, d), ischemia (e, f) and both (g, h).

TABLE IV
DFU-VGG confusion matrix.

	Predicted			
Actual	**None**	**Ischemia**	**Infection**	**Both**
None	95.17%	4.5%	0.20%	0.07%
Ischemia	7.65%	91.50%	0.42%	0.42%
Infection	1.20%	0.43%	93.60%	4.80%
Both	0.36%	2.40%	3.75%	93.45%

V. Conclusions and Future Work

This work presented a new CNN architecture, based on VGG-19, to classify diabetic foot ulcers, considering four classes. We evaluated several CNNs models, fine-tuning schemes, and other parameters to set up the proposed approach. We developed a more accurate and robust DFU classification method than the strategies presented in state-of-the-art works.

The results obtained were promising, but we believe they can be improved. In particular, we aim to increase classification accuracy and reduce the percentage of infection class images classified as none. We will carry out experiments with other CNNs and investigate the formation of multilevel CNNs. Finally, we intend that specialist physician analyze the computational results.

References

[1] K. Ogurtsova, L. Guariguata, N. C. Barengo, P. L.-D. Ruiz, J. W. Sacre, S. Karuranga, H. Sun, E. J. Boyko, and D. J. Magliano, "Idf diabetes atlas: Global estimates of undiagnosed diabetes in adults for 2021," *Diabetes Research and Clinical Practice*, vol. 183, p. 109118, 2022.

[2] F. Veredas, H. Mesa, and L. Morente, "Binary tissue classification on wound images with neural networks and bayesian classifiers," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 410–427, 2010.

[3] 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.

[4] 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.

[5] 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, no. 21, pp. 15 655–15 677, 2020.

[6] 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," *Computers in Biology and Medicine*, vol. 117, p. 103616, 2020.

[7] 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. 228 586–228 597, 2020.

[8] 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 Express*, 2021.

[9] B. Rostami, D. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, "Multiclass wound image classification using an ensemble deep cnn-based classifier," *Computers in Biology and Medicine*, vol. 134, p. 104536, 2021.

[10] A. Galdran, G. Carneiro, and M. Á. G. Ballester, "Convolutional nets versus vision transformers for diabetic foot ulcer classification," *Lecture Notes in Computer Science*, pp. 21–29, 2022.

[11] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 2661–2671.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2818–2826.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4700–4708.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[17] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

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