Biomedical Materials & Devices
https://doi.org/10.1007/s44174-024-00165-5

ORIGINAL ARTICLE

DFU_XAI: A Deep Learning-Based Approach to Diabetic Foot Ulcer Detection Using Feature Explainability

Shuvo Biswas[1] · Rafid Mostafiz[2] · Bikash Kumar Paul[1,3] · Khandaker Mohammad Mohi Uddin[4] · Md. Abdul Hadi[5] · Fahmida Khanom[6]

Received: 5 December 2023 / Accepted: 24 January 2024

Abstract
Diabetic foot ulcer (DFU) is a potentially fatal complication of diabetes. Traditional techniques of DFU analysis and therapy are more time-consuming and costly. Artificial intelligence (AI), particularly deep neural networks, has demonstrated remarkable effectiveness in medical applications. Despite this, the lack of explainability of deep learning models is currently viewed as a key hurdle to using these approaches in actual clinical settings. In this research, we present the DFU_XAI framework for assessing the interpretability of explainable-driven deep learning (DL) models. DFU_XAI evaluates five DL models (Xception, DenseNet121, ResNet50, InceptionV3, and MobileNetV2) to establish a transparent DL framework using three state-of-the-art explanation methods: Shapley additive explanation (SHAP), local interpretable model-agnostic explanations (LIME), and gradient-weighted class activation mapping (Grad-CAM). ResNet50 outperformed the other four models with remarkable results: 98.75% accuracy, 99.2% precision, 97.6% recall, 98.4% F1-score, and 98.5% AUC. For the most part, it can locate diabetic foot ulcers precisely on a diabetic foot and discriminate between diabetic foot ulcers and healthy feet in the DFU dataset. A heat map will indicate the precise location of the ulcer that needs care.

Keywords Diabetic foot ulcers · Deep learning · Explainable AI · Convolutional neural networks

Introduction

Diabetes is also known as DIABETES MELLITUS (DM). It is a chronic disorder brought on by hyperglycemia (high blood sugar levels), which can lead to serious, even deadly, consequences such as lower limb amputation, cardiovascular disease, blindness, and kidney failure [1]. DFUs, often referred to as sores, are open ulcers that can form on a diabetic person's foot. According to global data, only 108 million individuals had diabetes in 1980, but by 2014, there were more than 422 million cases worldwide. For instance, when looking at people over the age of 18, the frequency

✉ Rafid Mostafiz
rafid.iit@nstu.edu.bd

Shuvo Biswas
it21620@mbstu.ac.bd

Bikash Kumar Paul
bikash.k.paul@ieee.org

Khandaker Mohammad Mohi Uddin
jilanicsejnu@gmail.com

Md. Abdul Hadi
hmdabdul076@gmail.com

Fahmida Khanom
fahmida.khanom@aiub.edu

[1] Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh

[2] Institute of Information Technology, Noakhali Science and Technology University, Noakhali, Bangladesh

[3] Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

[4] Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh

[5] Department of Computer Science: Information Technology, University of Nebraska Omaha, Omaha, USA

[6] Department of Mathematics, American International University – Bangladesh, Dhaka, Bangladesh

Published online: 07 March 2024

rose overall from 4.7 to 8.5% between 1980 and 2014 [2]. Furthermore, it is anticipated that by the end of 2035, there will be 600 million people worldwide who are diabetic. Due to a lack of healthcare infrastructure and low levels of knowledge, it is also critical to note that around 80% of these individuals come from developing nations [3]. Furthermore, if proper care is not available, 15–25% of diabetic patients will experience DFU at an advanced stage of the disease, which could necessitate lower limb amputation [4].

Over a million diabetics with "high-risk feet" lose a limb [5] due to a lack of recognition and inadequate treatment each year. Individuals with diabetes need to take extra care of themselves by practicing good hygiene, taking their medications in time, and visiting the doctors regularly for checkups. As a result, people with diabetes and their families will face a significant financial burden, especially in developing countries where the cost of treatment is about 5.7 times the annual income [6]. It is anticipated that as diabetes becomes more common, the number of people at risk for diabetic foot ulcers will keep rising [7]. This will place an even greater strain on the limited resources and lack of specialized knowledge to treat the problem. Every year, more than a million diabetics lose a key body part—like their foot—due to inadequate treatment.

It takes a thorough examination of medical data for physicians to arrive at the right diagnosis. Computer-aided diagnosis techniques offer the potential to reduce costs and increase efficiency compared to manual diagnostic methods, which leave a greater chance for error. Modern wearable health and mobile technologies can regulate diabetes and related problems by monitoring and feeling harmful foot stress and inflammation, improving patients' quality of life, and prolonging remission [8]. An automated DFU diagnosis technique that is noninvasive, remotely deployable, extremely dependable, and reasonably priced is desperately needed. With the introduction of deep learning (DL), intelligent systems, and the expansion of computer vision (CV) applications, DFU pathology detection has emerged as a dynamic field of research with notable advancements. Current clinical practice has led to the inclusion of several critical mechanisms for early diagnosis in the DFU test. The patient's medical history is assessed; a wound or diabetic foot specialist thoroughly examines the DFU; other tests, such as an X-ray, CT scan, or MRI, may be carried out as necessary.

The last few years have seen a significant improvement in the biomedical sector, including protein structures explained, biomedical image classification, and genomic sequence alignment, owing to the progress made in extremely advanced technologies. The accumulation of biomedical information requires efficient and reliable computing systems to explain, analyze, and store such information. DL-based approaches can tackle these complex issues. DL strategies, specifically neural networks, have the capability to efficiently acquire hierarchical DL features from unprocessed data. This empowers such strategies to comprehend intricate representations and relationships within intricate biomedical data. Since more effort and emphasis have been placed on it in current times, DL has flourished and been widely implemented in industry. The utilization of DL methodologies has become increasingly important in the biomedical field owing to their capacity to evaluate large-scale datasets, retrieve task-relevant patterns, and make prognosis. Medical imaging plays an important role in the diagnosis of a variety of health problems in today's digital healthcare system [9, 10]. Traditional ML and DL methods achieve their goals in medical imaging by selecting and extracting features that are more sensitive to image forms, colors, and sizes. Previous research has shown that DFUs may be accurately identified and detected using ML and DL techniques. Despite the fact that many researchers have examined these difficult problems, no workable solutions have been found. However, it is challenging to understand how the model is making predictions or decisions and to identify potential errors or biases due to the complexity and black box nature of many ML and DL algorithms. Errors in models can have serious repercussions in areas including medicine, economics, and law enforcement. Another drawback of black box models is their sensitivity to hyperparameter choice, which makes it tough to tune and generalize to new data. Overfitting is another issue that can plague black box models, leading to subpar results on new data. While black box models have seen widespread success across many different use cases, there is cause for concern over their lack of interpretability and transparency. New techniques, like explainable machine learning (XML), can help solve these problems by making black box models more open and easier to understand.

While DL models have played a significant role in diabetic foot ulcer (DFU) analysis, they have yet to lack explainability in DFU analysis. Regulatory and ethical issues still remain with regard to the implementation of DL models in DFU analysis. A few of these problems are: (1) the potential for biases; (2) the lack of DL model transparency, explainability, and trustworthiness; (3) the privacy difficulties of the data utilized to evaluate DL models; and (4) potential liability and safety concerns when using DL models in DFU analysis. These problems can be solved with the use of an explainable AI (XAI) system. The XAI-based system is able to transform the black box DL models into white box DL models. The use of this white box DL model has increased due to its significant success in intelligence-based systems. It should be noted that the efficacy of the aforementioned methods is limited because machines are now unable to provide explanations for their outputs and communicate with medical experts. XAI-based DL models aid medical professionals and junior practitioners to appropriately trust, understand, and efficiently communicate with these approaches. So, our suggested XAI-based DFU_XAI framework solves the problem of DL models that are hard to understand, especially when it comes to DFU analysis.

In terms of advancements in DL and XAI techniques, the DFU_XAI framework has made many contributions to the field of medical research. In the DL technique, this framework deals with a very sensitive topic in the field of medical research. This framework extracts intricate features and patterns from complex medical data to create an accurate DFU diagnosis. On the other hand, in the XAI technique, the integration of the XAI algorithm with the AI models makes this framework more transparent and interpretable. This transparent framework helps healthcare professionals understand the model's decisions, the exact location of the DFU, and the importance of this location. Thus, by providing interpretable insights, this framework is exploited as a valuable tool for clinical decision support that assists clinicians in making decisions about DFU diagnosis and treatment procedures.

In this study, we have improved the transparency and interpretability of diabetic foot ulcer detection and classification by creating an explainable DL framework called DFU_XAI. InceptionV3, DenseNet121, Xception, MobileNetV2, and ResNet50—five CNN classifiers—are used alternatively as the backbone of DL models to select the best-fit one. When tested on the working dataset, the ResNet50 model produces the highest accuracy (98.75%) for separating ulcer from healthy-type data. Using backpropagating XAI approaches—LIME, SHAP, and Grad-CAM—this research focuses on the explainability of diabetic foot ulcer identification. These techniques are used in this scenario to visually grasp and explain the predictions of the proposed model. The backpropagation saliency map, Grad-CAM, and Grad-CAM++ find the exact locations for classification; LIME finds the segmented feature map; and SHAP assigns the score value for each pixel in the predicted images. In the root cause analysis, the DFU_XAI framework addresses the challenge of the model being a "black box" by applying XAI techniques. XAI converts these "black box" models to more transparent and explainable "white box" models. The DFU samples are then fed into the XAI-based visual explanation unit of the DFU_XAI framework to indicate the regions with ulcers and assist doctors during the diagnosis in an efficient way. The factors that contribute to the potential failures of this "black box" nature model are as follows: (1) more layers and more parameters; (2) high-dimensional input data; and (3) nonlinearity properties of the "black box" model.

The DFU_XAI framework aligns with current trends and advancements in the broader field of artificial intelligence, especially in healthcare applications. XAI-based ML and DL approaches aid medical professionals and junior practitioners to appropriately trust, understand, and efficiently communicate with these approaches. These approaches completely resolve the complexities associated with healthcare applications. Thus, our proposed XAI-based DFU_XAI framework aligns with current trends and advancements in the broader field of artificial intelligence. Considering the continuous evolution of AI technologies, our proposed XAI-based DFU_XAI framework entirely addresses the medical industry's challenges. This framework might adapt to the XAI-based ML and DL approaches in future developments that aid medical practitioners to appropriately trust, comprehend, and efficiently communicate with these approaches. However, the DFU_XAI framework has many contributions and implications for the field of medical image analysis. This framework contributes to pinpointing the precise DL features in medical images that influence the AI model's forecasts. This forecast aids in comprehending the qualities of an image that contribute to a medical diagnostic system. In future research, it might influence the improvement of the transparency, interpretability, usability, trustworthiness, and explainability of AI models. Using this framework to analyze medical images in future clinical practices may help clinicians make better decisions and improve patient outcomes.

The DFU_XAI framework contributes to advancing the field of diabetic foot ulcer analysis by increasing cooperation between endocrinologists and AI experts. The incorporation of AI models with explainability gives more reliable and accurate predictions. The highlight of explainability is that it fosters trust among medical practitioners, allowing them to interpret and understand models' decisions. However, in the broader medical imaging community, the DFU_XAI framework will collect more information than the limited medical dataset. The DFU_XAI framework trains this collected information and produces better predictions that will help to diagnose DFU at the initial stage. The following key contributions will help to understand how the model works and build trust in its predictions.

The DFU_XAI framework proposes a transparent and reliable model in diabetic foot ulcer detection.

The proposed model obtains a significantly improved outcome in classifying ulcer and healthy images with an acceptable false classification rate.

The implemented backpropagating XAI techniques detect the affected region with segmentation and pixel score which explain the model in both subjective and objective manner.

The paper is broken up into four parts, "Literature review" section, of which is an analysis of the literature on various writers' perspectives on DFU detection and categorization. Methods, including image labeling and augmentation, the

DFU image set description, CNN models, and the proposed DFU_XAI framework for enhancing DFU classification performance, are outlined in "Research Methodology" section. In addition, the publication describes three XAI techniques—LIME, SHAP, and Grad-CAM—that may be used to make sense of the predictions made by DL models. The methodology, analysis of results, and comparison to the state of the art are discussed in "Experimental Design and Result Computation" section. In "Discussion" and "Conclusion" section present the discussion and conclusion parts of this research.

Literature Review

Recent years have seen the development of a number of different classification systems for diabetic foot ulcers (DFUs). A literature review is a study of the existing literature. This section serves as a synopsis of the primary literature that supports the proposed action.

A detailed literature analysis on AI-assisted ulcer monitoring approaches was provided by Kaselimi et al., showing the benefits and difficulties of using AI in remote patient care. With the patient's physiology and the features of the sensors in mind, they zeroed in on the imaging methodologies and optical sensors [11] used to identify diabetic foot ulcers. The research also found that the data source determines the monitoring method, which in turn determines the artificial intelligence techniques that may be used [12].

Recent advancements were described by Das et al. [13]. Based on their research, they propose a new kind of CNN architecture they term DFU_SPNet, which significantly improves state-of-the-art performance metrics for DFU classification. The design of this network is made up of three stacked, parallel convolution layers with varying kernel sizes. A batch normalization layer and leakyRELU activation serve as a transition layer between each convolutional layer in this architecture. Although the DFU_SPNet model performed exceptionally well, this paper's use of the same filter (32 × 64 × 128) in each parallel convolution block is a significant drawback. Using a varied number of filters, as opposed to a constant number, might boost the accuracy score.

Alzubaidi et al. [14] created a hybrid CNN model specifically for DFU classification. Using their proposed framework, four deep networks were built with a total of six parallel convolutional modules by mixing multi-branched parallel layers with conventional convolutional layers. The networks were separated into groups according to the number of branches in the parallel convolutional layer, which ranged from two to five. The branches are joined with varying filter widths to improve the feature map extracted. Although the model presented in this paper achieves impressive performance, the large number of parallel convolution blocks required to do so significantly increases both the required computing time and the associated costs.

Alzubaidi et al. [15] present a novel deep convolutional neural network for DFU picture classification; they refer to it as DFU_QUTNet. The network was planned to expand in width while maintaining depth, as is typical of today's networks. Using the retrieved features from the proposed DFU_QUTNet, support vector machine (SVM) and *K*-nearest neighbors (KNN) classifiers are trained. The F1-score value, therefore, rose dramatically. Each parallel convolutional layer suffers from the same kernel size (3 × 3), which is a limitation of this approach. Using convolutional kernels of varying sizes can boost accuracy results compared to using kernels of a constant size.

Goyal et al. [16] introduced an innovative deep CNN approach called DFUNet. DFUNet incorporated parallel convolution blocks for the fusion of features obtained from conventional single convolutional stages and convolution operations. Additionally, to establish accurate results for DFU prediction, they employed Linde–Buzo–Gray (LBP) as a low-level extractor for extracting low-level features. Also, traditional LeNet, GoogleNet, and AlexNet have been employed as high-level extractors. Finally, these extracted features were subsequently utilized as inputs for ML models in the task of DFU identification. However, a potential pitfall of this architecture is the absence of sufficient transition layers between these parallel blocks. Properly designed transition layers can contribute to extracting more features with the help of a multilayer concatenated feature map.

The area of diabetic foot ulcers (DFUs) was calculated in two stages by Wang et al. [17], who used a capture box and support vector machine (SVM) classification. A two-stage classification was used to first segment the picture using superpixels and then extract features from the image.

To examine whole pictures of feet, Manu et al. proposed a two-scale transfer learning architecture based on FCNs (fully convolutional networks) to effectively segment the surrounding skin and DFU [18]. This novel architecture obtained a Dice similarity coefficient (DSC) of 0.851 (± 0.148) against surrounding skin, 0.794 (± 0.104) against DFU, and 0.899 (± 0.072) against both. Despite the impressive outcomes, the method is not without its drawbacks. It has problems processing massive datasets, which is one of its drawbacks. Another restriction is that it is not allowed in healthcare settings due to infection control concerns to have patients lay their feet on the perimeter of a box during data gathering.

Table 1 A summery table of all DFU detection approaches

Paper	Dataset	Outcomes/strengths	Weaknesses
Das et al. [13]	3827	Stacked parallel Conv. layers with suitable TLs	Use the same filter pattern (32, 64, and 128) for each stacked PCB
Alzubaidi et al. [14]	17,053	Combine traditional Conv. layers with multi-branch parallel Conv. layers	Computation time
Alzubaidi et al. [15]	17,053	Enhance the width of the model with lower computing cost	Utilize the same filter sizes (3×3) for the PCB
Goyal et al. [16]	22,777	Reduce the depth of the model, but enhance the size of filters in the PCB	The lack of sufficient TLs between PCB
Wang et al. [17]	100	Identify the wound area on an ulcer skin sample taken with a sample capture box	Slight running speed
Manu et al. [18]	705	Predict the pixel level for DFU sample segmentation	Small size of the dataset

PCB—refers to parallel convolutional block and Conv.—refers to convolution layer. TL stands for transition layer

The previously proposed investigations are more precise, but they are not as transparent, interpretable, or explainable. Our motivation for this study comes from the fact that current DL frameworks lack the reliability and explainability necessary to help doctors understand and trust the performance of DL algorithms during clinical operations. Table 1 describes an overview of the published paper.

Research Methodology

This paper proposes a novel deep learning method for recognizing healthy skin and ulcers in DFU images. Traditional pre-trained models are retrained to reduce the expense of training the model from scratch. These models were created with the explicit purpose of locating and extracting important information from images, such as forms and edges. A CNN model is used to input a single

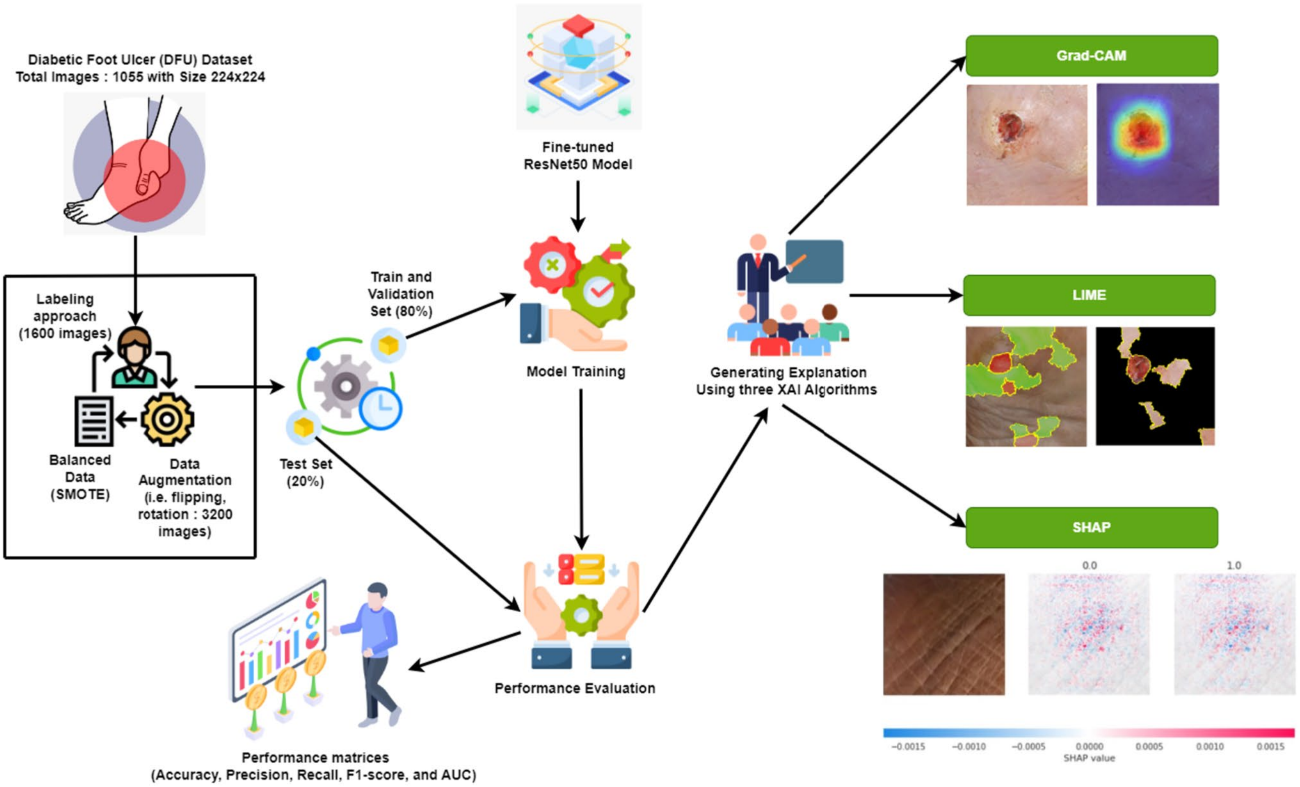

Diabetic Foot Ulcer (DFU) Dataset
Total Images : 1055 with Size 224x224
Labeling approach (1600 images)
Balanced Data (SMOTE)
Data Augmentation (i.e. flipping, rotation : 3200 images)
Train and Validation Set (80%)
Test Set (20%)
Fine-tuned ResNet50 Model
Model Training
Performance Evaluation
Performance matrices (Accuracy, Precision, Recall, F1-score, and AUC)
Generating Explanation Using three XAI Algorithms
Grad-CAM
LIME
SHAP
0.0
1.0
−0.0015
−0.0010
−0.0005
0.0000
0.0005
0.0010
0.0015
SHAP value

Fig. 1 The proposed DFU_XAI framework

diabetic foot picture and provide a prediction output that tells us whether or not the image has an ulcer. Then, using the explainable approaches, heat maps or region segmentation are created, emphasizing the diabetic foot image's superpixel areas that are most important to the model's prediction. Figure 1 illustrates the proposed DFU_XAI framework for DFU classification. The proposed framework looked into the patch labeling approach, data augmentation technique, which increases the volume of the dataset, model training process with five fine-tuned deep CNN models to select the best performing model, performance metrics, which are utilized to evaluate the model's performance, and finally three XAI techniques with LIME, SHAP, and Grad-CAM, which are used to explain the DL model and enhance the transparency and interpretability of the model's decision-making process. Using this XAI-based transparent framework, researchers can show how input samples are utilized and detect biases that need to be addressed. Thus, by enhancing the transparency of the black box model, the DFU_XAI framework contributes to addressing biases in AI models, especially in the context of medical image analysis. For SHAP explanations, a gradient explainer is employed, whereas for LIME, perturbation is computed. On the other hand, for Grad-CAM, the gradient score is calculated. The major challenges that were encountered during the implementation of the DFU_XAI framework are as follows: (1) limited availability of the DFU dataset; (2) imbalanced dataset; (3) lack of feature explainability. To address these issues, an augmentation approach is applied to enhance the volume of data, the SMOTE technique is used to balance the dataset, and finally, three XAI algorithms are integrated with the ResNet50 model to explain the features of the image. Algorithm 1 demonstrates the step-by-step classification task of the DFU_XAI framework, which is provided in "Experimental Design and Result Computation" section.

DFU Dataset

The dataset comprises four folders listed in Kaggle online data repositories [19], and the source of this dataset is the works [10, 20]. This dataset was taken from the Nasiriyah Hospital's diabetic department in southern Iraq [15]. As medical image data involves sensitive patient information, ethical approval and written consent were obtained from all pertinent patients and persons. These samples were captured by iPad and Samsung Galaxy devices at various angles and brightness's. To train and test our model, we selected the patch folder among these four folders, which included a total of 1055 skin patches. Out of these patches, 512 were identified as abnormal (ulcers), while the remaining 543 were classified as normal (healthy skin).

Image Labeling Approach

However, a deep learning model, especially a CNN-based architecture, requires a significant amount of labeled data to achieve optimal performance. At the same time, obtaining a large amount of medical data can be a costly and arduous process. Therefore, to enhance the efficiency of DL models and tackle overfitting, some approaches like image labeling, data augmentation, transfer learning, and regularization can be employed. To address this issue, a patch labeling approach is used to increase the size of the dataset. It is a sample labeling strategy in which significant areas are cropped across a large sample and the cropped samples are then labeled with the relevant class. In our study, a sliding window of 224 pixels in height and 224 pixels in width is moved from top left to bottom right of each sample, and we initially extracted the Region of Interest (ROI) from the samples. Based on the regions that include ulcer skin and normal skin, the patches are labeled in either ulcer or healthy classes. However, through labeling relevant areas (i.e., ulcer-related areas) within larger samples, the DFU_XAI framework focused on specific DL characteristics rather than the entire sample, which contributes to the efficiency of DL models. Using this strategy, we cropped pertinent DL information for categorization. Use of pertinent DL information reduces memory allocations and computational burden during DL model training, which contributes to dealing with overfitting. Finally, with the help of this strategy, we produced a total of 1600 image patches (740 healthy images and 860 ulcer images). Figure 2 shows a few examples of sample images that were cropped from the original images.

Data Augmentation

Some data augmentation techniques, like random scaling and flipping, rotation, and contrast enhancement using new color spaces, are applied to increase data variation and size. In our study, we performed the rotation technique, which rotates 1600 skin patches at two angles: 90° and 180°. After this technique, we created a new dataset, called the DFU dataset, which contains a total of 3200 skin patches containing 1720 ulcer (abnormal) images and 1480 healthy (normal) images. The DFU dataset is split into two sets, train (80%) and test (20%), with the help of the "train_test_split" function. The Python "sklearn.model_selection" package is used to import

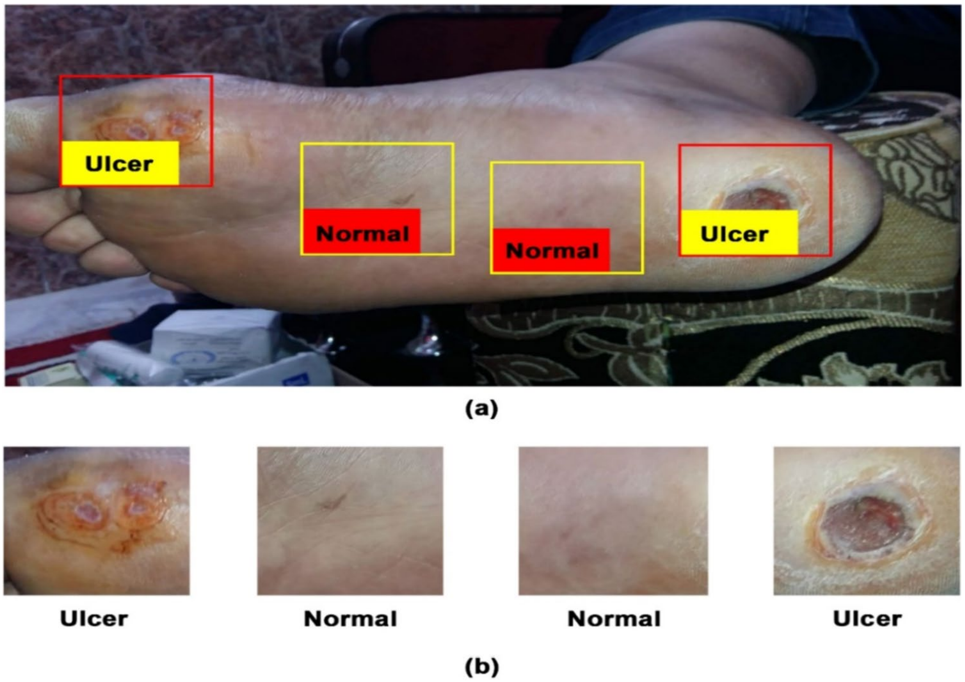


Fig. 2 Process of taking healthy (normal) and ulcer (abnormal) patches. **a** Sample images generated process, utilizing image cropping (sliding window) technique. The red squares represent the ulcer (abnormal) class, and the yellow squares represent the healthy (normal) class. **b** Samples of two types of patches that generated after applying the cropping process

Table 2 Working dataset

Dataset	Label	Original image	After cropping	After augmentation
DFU	Abnormal (Ulcer)	512	860	1720
	Normal (Healthy)	543	740	1480
	Total images	1055	1600	3200

this function. With the aid of the same splitting function, 10% of the data from the training set is divided again for validation. Table 2 summarizes the dataset used in this study. Synthetic Majority Oversampling Technique (SMOTE) is employed here to handle the problem of imbalanced DFU datasets. It creates synthetic instances within the minority class, and this is achieved by connecting randomly selected G minority class examples with their k-nearest neighbors (KNN) from the same class [21]. The parameters G and k denote the number of minority instances to generate and the user-specified value of k. However, this approach helps tackle the overfitting challenge caused by random oversampling. Imbalanced datasets can lead to biased DL models. SMOTE reduces this bias by balancing the data distribution. Thus, by minimizing bias and collecting significant characteristics of the minority label, it impacts the overall performance of the models. However, SMOTE may not always provide the improved results of the models. Sometimes, it can generate unrealistic synthetic samples. These unrealistic samples may not precisely display the actual distribution of the minority label. This might result in information leakage from the majority label, particularly where both minority and majority samples overlap.

Fine-Tuned ResNet50 Model

In addition to taking a long time, training a CNN model with a large number of training parameters requires high-performance hardware. Transferring pre-trained weights and parameters from models trained on diverse datasets to the newly constructed model is a frequent solution to these problems [22, 23]. When a new layer is added, the transfer learning method is used in addition to the transferred parts to move from one layer to another. This approach provides a quicker and less computationally intensive technique to get results and has been shown to be successful even with limited datasets [24]. The traditional AI approach requires a large amount of data to train its model, but collecting a large dataset is time-consuming, costly, and challenging. To address this challenge, the DFU_XAI framework used pre-trained weights from the ImageNet database based on the transfer learning concept. This approach offers some benefits, as follows: (1) makes the training time of the DFU_XAI framework faster; (2) leads the DFU_XAI framework to a higher possibility of success; (3) allows adding more data and modifying the weights during model training to further refine the framework.

This experiment used five cutting-edge CNN-based models: DenseNet121 [25], ResNet50 [26], InceptionV3 [27], Xception [28], and MobileNetV2 [29] for classification purposes. These cutting-edge models were selected based on their architectural structures, performance on classification tasks, and the explainable ability of the model's prediction. Each model has a different set of architectures with unique modules. For instance, Xception uses depthwise separable convolution blocks, ResNet has residual modules, DenseNet has dense modules, InceptionV3 allows inception modules, and MobileNetV2 uses inverted residual modules. These networks have exhibited high performance on the ImageNet database. Their pre-trained weights retrieved DL information from the input image to contribute to the overall performance of the model. These models have previously demonstrated high performance when trained on large datasets. These CNN models used pre-trained transfer learning, whereas the weights had already been tuned through prior training. This strategy allows the transfer of knowledge and weights learned from one model to another, resulting in faster and more efficient training of new models. Performance is improved when various fine-tuned layers are concatenated with the pre-trained model. These five pre-trained models are trained and tested using the same data and split ratio. In addition, six fine-tuning layers are added to these models to improve their performance. These additional layers adjusted the pre-trained weights of the model to refine the entire DFU_XAI framework. Finally, these fine-tuning layers process the extracted DL features for the classification task. Thus, with the help of these additional fine-tuned layers, the model's performance improved. Among all these, ResNet50 exhibits better performance. The overall architecture of this ResNet50 model is illustrated in Fig. 3. All layers of the pre-trained ResNet50 model and fine-tuning layers are described in the following subsections.

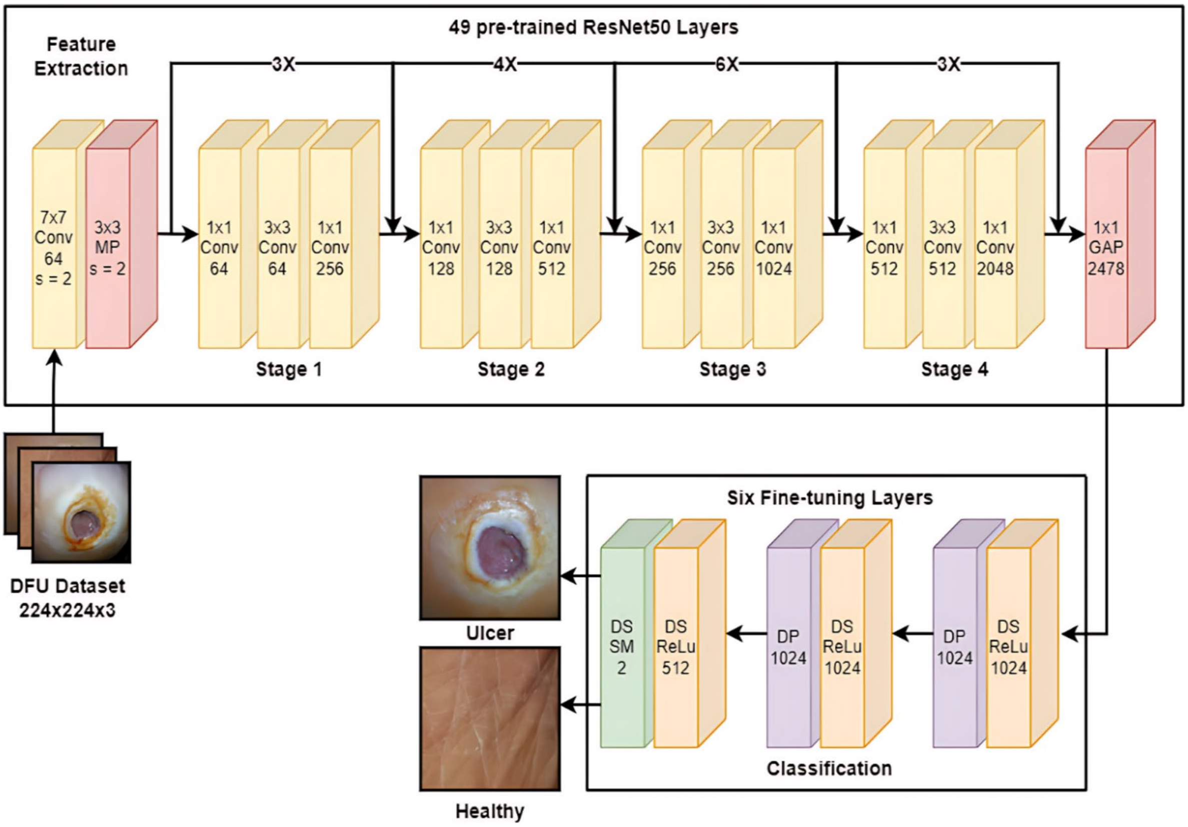


Fig. 3 Fine-tuned ResNet50 for DFU detection. Here, Conv indicates convolution layer, MP indicates max pooling layer, s indicates stride, GAP indicates global average pooling layer, DS indicates dense layer, DP indicates dropout layer, and SM indicates softmax activation function

Residual Network (ResNet)

He et al. [26] proposed ResNet, a DL model that was established by winning the popular ImageNet competition named ILSVRC. ResNet contains 26 million parameters, which makes it a more complex network compared to the others. Different versions, such as ResNet18, ResNet50, and ResNet101, were developed with varying numbers of layers and stages for the purpose of classification tasks. In this work, we utilize the ResNet50 pre-trained model as our DL model explainer. This model consists of a total of 50 deep CNN layers, comprising 49 convolution layers, followed by a max pooling layer for feature extraction and one fully connected (FC) layer for classification. The model's accuracy can be improved by increasing the number of layers during evaluation and training. For this reason, we concatenate some extra layers at the top of the ResNet50 model. This method is called "fine-tuning," which can handle a small dataset and improve overfitting issues.

Since InceptionV3, the DL-based model has been deepening, and MobileNetV2, Xception, and DenseNet have 53, 71, and 121 deep neural layers, respectively. With the enhancement of model depth, these models encounter various challenges like "vanishing gradient" and model complexity. The ResNet50 network, developed based on VGG-Net, has less complexity and fewer filters. This network has more than 23 million training parameters. The gradient issue was tackled by the skip connection blocks, which enable gradients to pass through other channels. This approach is the root idea in the residual modules of ResNet50 to handle the vanishing gradient issue. The characteristics that make it more effective compared to other pre-trained models in this context are as follows: (1) reduces the vanishing gradient issue using residual blocks; (2) contains a smaller number of layers; and (3) reduces the model size using the global average pooling layer.

Convolutional layer The convolutional layer has a number of learnable filters that can recognize features depending on the weights assigned to each filter. Each filter in this layer is known as a kernel in the CNN design. The ResNet50 network used in our experiment contains 49 convolution layers, followed by various kernel sizes that quicken the feature extraction process. Only three filter sizes—1 × 1, 3 × 3, and 7 × 7—are employed to construct the refined ResNet50 network, and these layers are made up of 64, 128, 256, 1024, and 2048 kernels, respectively.

Batch Normalization (BN) Layer The BN layer is responsible for applying a mini batch to each input channel, which helps to decrease the sensitivity and enhance the speed of the DL network. In our model, the BN layer is placed between the convolution and ReLU layers.

Rectified Linear Unit (ReLU) Each convolution layer uses ReLU as an activation function to produce nonlinearity in the output layer. It may be described mathematically as $f(x) = \max(0, x)$, where x is the input to the neuron, and $f(x) = \max(0, x)$ means that it sets any negative input to zero and allows any nonnegative input to pass through unaffected. Except for the final FC layer, all convolution layers and FC layers in this experiment incorporate ReLU.

Pooling Layer The pooling layers of CNN models are configured with a 2D filter that slides the feature map over each channel and then compiles the results into a feature summary. Multiple pooling layers are used by the CNN network, including global pooling, L2-norm pooling, average pooling, and max pooling (MP). The maximum value from the input region is chosen by the max pooling at the final pooling layer, and it is then applied to the layer below. This test employs a single MP layer with a 3 × 3 kernel size and a stride of 2.

Global Average Pooling (GAP) Layer GAP is a pooling method technique employed in the CNN network in place of a fully connected (FC) layer. It combines the sample average values to create the final image and sends it straight to the softmax (SM) layer. One GAP layer is used in this investigation, and its output is 2478.

Fully Connected (FC) Layer The FC layer in CNN is a crucial part that connects the flow of each neuron coming from the preceding and following layers. It helps to anticipate how well each value fits in a certain class. The FC layer does this task by using a variety of activation functions. The primary purpose of these FC layers in CNNs is to categorize DFU samples into two classes—ulcer and healthy—by integrating the image features extracted by the preceding layers.

The following key features of the DFU_XAI framework provide insights into the decision-making process of the ResNet50 model: (1) development of the DFU_XAI framework using the fine-tuned ResNet50 model rather than from scratch to mitigate the "vanishing gradient" challenge; (2) integration of XAI algorithms with the pre-trained CNN model to eliminate the lack of explainability of existing models; and (3) automatic explanation of ulcer cases using the XAI-based heatmap implementation system from the DFU dataset to assist endocrinologists and medical practitioners. The above information enhances the transparency of the model in clinical practice and permits medical professionals to trace and validate the logic behind the framework's prognosis. This transparency fosters trust in the

Table 3 Proposed fine-tuned ResNet50 model layers in detailed

Name of layer	Output shape	Param #
Input image	224×224×3	0
Conv1	112×112×64	9472
Max pool	56×56×64	0
Conv2	56×56×256	16,640
Conv3	28×28×512	66,048
Conv4	14×14×1024	263,168
Conv5	7×7×2048	1,050,624
Global average pool	2048	0
"Fine-tuned layers"	–	–
FC1 (ReLu)	1024	2,098,176
FC2 (Dropout)	1024	0
FC3 (ReLu)	1024	1,049,600
FC4 (Dropout)	1024	0
FC5 (ReLu)	512	524,800
FC6 (Softmax)	2	1026
Total params: 27,261,314		
Trainable params: 3,673,602		
Non-trainable params: 23,587,712		

framework's predictions, enhancing confidence in its clinical decisions.

Fine-Tuning Process

Four FC (fully connected) layers and two dropout layers are coupled at the base of the ResNet50 during the fine-tuning phase (see Fig. 3). First, new features have been extracted from the input data by updating the top layer of the original ResNet50. The pre-trained ResNet50 is then concatenated with the fine-tuned layers for DFU classification. Finally, the suggested ResNet50 model is created using these refined layers. We use two dropout layers that reject 50% and 30% of the data during training to address the overfitting issue. Additionally, this method shortens the training period. In this experiment, we use four FC layers, with the final FC layer's DFU classification using two neurons. The predicted label is generated in this layer, followed by the softmax activation function. Using the characteristics of the anticipated class, this function produces a predicted value between 0 and 1. Table 3 lists the parameters for the fine-tuned layers and the pre-trained ResNet50 model. The following key features of the DFU_XAI architecture contribute to its respectable receptive field: (i) Development of the DFU_XAI framework using the fine-tuned ResNet50 model rather than from scratch to mitigate the "vanishing gradient" challenge; (ii) Integration of XAI algorithms with the pre-trained CNN model to eliminate the lack of explainability of existing models; and (iii) Automatic explanation of ulcer cases using the XAI-based heatmap implementation system from the DFU dataset to assist endocrinologists and medical practitioners. The DFU_XAI framework enhances the detection of ulcers at various scales in diabetic foot ulcer images by providing transparency and feature explainability in the decision-making process of AI models.

Explainability of Deep Learning Model

Artificial intelligence (AI) advancements have created new opportunities for human existence in a number of sectors, including business, health care, and education [30]. AI-derived deep learning algorithms aid in the classification of medical images in the field of medical research [31]. By inferring attributes at each level, explainable AI algorithms offer tools that may be used to comprehend the outcomes of deep learning [32]. Biases can emerge in the models far before the training and testing periods. The model's training data may have its own biases. As a result, every AI technique must recognize and control any bias in the training dataset. Explainable AI should aim to provide models that are trustworthy, transparent, and devoid of prejudice. It has been feasible to precisely evaluate different medical picture collections using AI algorithms, such as distinguishing healthy skin from skin afflicted by ulcers. Black box models, on the other hand, are another name for AI models since they take the essential steps to incorporate AI imaging methods into routine clinical practice while concealing logical justifications. A technique called explainable AI aggregates the top-level attributes of a trained model. The three widely used XAI techniques for visual analysis—LIME, Grad-CAM, and SHAP—are described in the subsections that follow.

Local Interpretable Model-Agnostic Explanations (LIME)

The LIME method, proposed by Ribeiro et al. [33], tries to give detailed justifications for each prediction made by a black box model. LIME's core idea is to provide a local approximation of the behavior of the black box model using a simpler, more transparent glass box model, making interpretation easier. The LIME approach generates perturbations by selectively activating and deactivating certain superpixels within an image. This method aims to determine the significance of continuous superpixels in the original picture of the output class and to convey the results in a way that is understandable to humans. LIME contributes to improving the interpretability of the model, increasing transparency, and fostering confidence in machine learning systems by showing how the input properties of a CNN model affect its predictions. The first step in applying LIME to an input image is to split it up into superpixels. The number of superpixels determines how finely the region is segmented. A superpixel is a linked collection of pixels with the same color and position. This procedure produces a finer and more comprehensive segmentation, allowing for better identification of the areas essential for predicting the output class.

Gradient-Weighted Class Activation Mapping (Grad-CAM)

A classification block, in addition to a feature extraction block, combines to form a conventional CNN model. A fully connected layer is presented in the classification block, and it uses the collected features to calculate a probability score from the softmax layer. The final classification output of the model is then calculated using the probability score with the greatest value. The accuracy and performance of the model are increased by ensuring that it chooses the category most likely to match the input image. Grad-CAM [34] is a localization method that is class-specific and produces visual explanations without altering the network architecture or training procedure. It locates the pertinent area of an image and uses the gradient of the feature map in the final convolutional layer of the network to emphasize the regions that have the biggest influence on the outcome prediction. Grad-CAM enhances the interpretability and transparency of the model by localizing the crucial areas of an image, allowing users to better comprehend the justification for the model's prediction. Large gradients have a significant impact on image prediction outcomes in Grad-CAM. By detecting the unique parts of an image and acquiring qualitative and quantitative insight into its inner workings, Grad-CAM and Grad-CAM++ are visualization techniques used to examine the convolutional layer of a CNN. To solve the low-resolution heatmap problem of Grad-CAM, Grad-CAM++ was created. Our tests have shown that by precisely localizing the presence of ulcers in pictures, Grad-CAM and Grad-CAM++ can improve the network's performance in ulcer analysis tasks.

Shapley Additive explanations (SHAP)

The SHAP technique, created by Lundberg et al. [35], can offer local explainability for a variety of dataset formats, including tabular, picture, and text datasets, to solve this. By computing the average marginal contribution from the feature space, it determines the additive feature relevance. The contribution of each input feature to the model's output for a certain prediction may be explained using SHAP values. The DFU_XAI framework leverages explainable AI (XAI) techniques such as LIME, SHAP, and Grad-CAM to enhance transparency and interpretability in the decision-making process of DL models. In this framework, LIME extracts the superpixel from the predicted samples. These extracted pixels make the DL model more transparent. SHAP creates negative and positive values to explain the decision of the DL model in diverse areas of the given samples. These values enhance transparency and interpretability in the decision-making process of the DL model. Finally, Grad-CAM creates the heatmap, leveraging the information from the last convolution layer of the DL model. After that, it highlights these heatmaps to make the model's decision more clear.

Experimental Design and Result Computation

The working environment setup, hyperparameters utilized, and subjective and objective evaluation results obtained using the DFU_XAI framework are presented in this section. A comparative analysis of the proposed framework with the existing methods is also presented. In an experimental environment, the following considerations were taken into account when selecting the learning parameters and sample split for the training, validation, and testing phases of the DFU_XAI framework: (1) an efficient learning rate; (2) an optimum number of epochs; (3) an effective optimizer; (4) a moderate batch size; and (5) an ideal train/test split ratio (80:20).

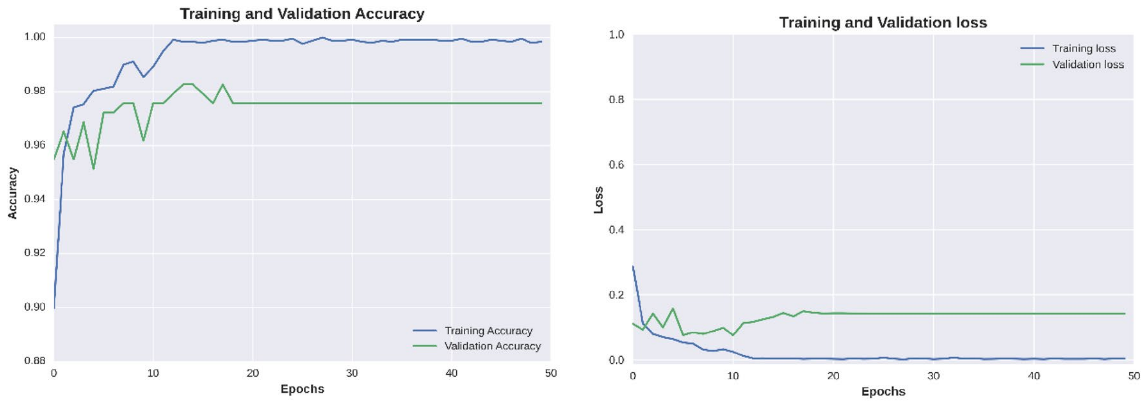
```
Algorithm 1: Proposed DFU_XAI framework for DFU detection.
Input: DFU dataset (Break down in ratio 7:1:2) with size 224x224;
  Ƞ = Learning rate;
  δ = Epochs;
  λ = Batch size;
  σ = the number of images covered in one batch size;
Output: ω = pre-trained CNN model weight;
Start
  1. Enhance the image data using image labeling approach and create DFU dataset;
  2. Implement data augmentation strategy for increasing the DFU dataset size;
  3. Extract the features from the DFU dataset using ResNet50 pre-trained CNN model;
  4. Set the fine-tuned layers: $CNN^{dense}$, $CNN^{dropout}$, $CNN^{dense}$, $CNN^{dropout}$, $CNN^{dense}$, $CNN^{softmax}$;
  5. Initialize the ResNet50 model parameters: Ƞ, δ, λ and σ;
  6. Train the ResNet50 model for evaluating the initial weights;
  7. for δ = 1 to δ do
  8.    Select a mini batch size σ for training data;
  9.    Forward propagation and determine the loss function;
  10.   Backpropagation and update ω;
  11.   Calculate loss, accuracy, validation loss, validation accuracy;
  12. end for
  13. Perform LIME, SHAP and Grad-CAM three XAI algorithms for the model;
End
```

Experiment Setup

Several existing models, including Xception, InceptionV3, DenseNet121, MobileNetV2, and Resnet50, were used in the training process of the DL model with various modified layers, utilizing the Keras [36] library. This simulation involves different resources. Table 4 describes the materials utilized for developing this DFU_XAI framework. Instead of creating random weights, the models were trained using pre-trained weights from the ImageNet [37] database. The last layers of the relevant models were modified to enable the differentiation between samples showing ulcers and those not showing them. The softmax activation function was employed in the new layers to enhance the model's performance. Initial considerations for parameter tuning include batch size, number of epochs, and learning rate [38]. The settings of 50 epochs, binary cross-entropy loss, 32 batch sizes, and an Adam [39] optimizer with a learning rate of 0.0001 were used to train the suggested model. Training (70%), validation (10%), and testing (20%) were done on the dataset. All fine-tuned models were trained using the same training settings to find the optimal model. The criteria that were considered during the selection of ResNet50 are: (1) an optimal learning rate (LR); (2) a moderate number of epochs; (3) an effective optimizer; and (4) an ideal batch size (BS). Higher LR may oscillate the model, and lower LR may slow down the model. Hence, we select an optimal

Table 4 Environment setup of the DFU_XAI framework

Resource	Details
CPU	Intel Core i5-12600 K @ 3700 MHz
RAM	64 GB
GPU	Tesla K80
Platform	Google Colab
Language	Python

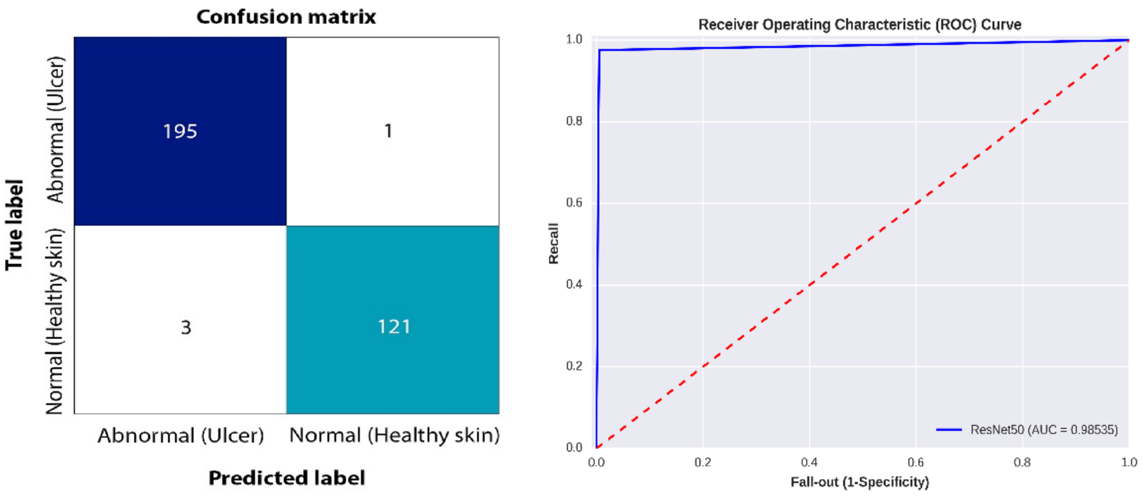


(a) **(b)**

Fig. 4 Training and validation phases of the final ResNet50 model **a** Shows the training and validation accuracy achieved by the model was 89.97% and 95.49% after the first epoch, with the *X*-axis indicating the number of epochs and the *Y*-axis indicating the accuracy from the final ResNet50 model. **b** Shows the training and validation loss achieved by the model was 0.2865 and 0.1107 after the first epoch, with the *X*-axis indicating the number of epochs and the *Y*-axis indicating the loss from the final ResNet50 model

LR that trains the model perfectly. Adam is an ideal choice as an optimizer due to its outstanding adaptive LR abilities compared to the other optimizers. We select the ideal BS due to faster training time and less memory. A larger BS requires more memory, and a smaller BS reduces the training time. A lower number of epochs may create an underfitting issue, and conversely, a higher number of epochs may create an overfitting issue. Hence, we chose a moderate number of epochs to tackle both issues. In the decision-making process, the above considerations are involved in selecting ResNet50 as the optimal model based on performance metrics. Finally, at the end of the training process, the optimal model was selected based on the performance metrics.

The DFU_XAI framework was developed using pre-trained CNN models and then fine-tuned. These fine-tuned models can use pre-trained weights and knowledge from the ImageNet dataset during the training time of the data. Thus, the DFU_XAI framework contributes to reducing the training time of the model. However, the DFU_XAI framework has several implications for its practical deployment and scalability, especially in healthcare domains where it ensures comprehensibility, transparency,



(a) **(b)**

Fig. 5 Receiver operating characteristic (ROC) curve and confusion matrices (CM) for the final pre-trained ResNet50 model on the classification tasks. **a** ROC curve obtained on recall/specificity, with *X*-axis indicating the fall-out (1-specificity) and *Y*-axis indicating the recall score from the final ResNet50 model. The blue color indicating the AUC curve obtained by the model was 0.98535, best AUC score from all other models. **b** The final CM was achieved on the DFU test set, with the *X*-axis indicating the predicted class labels and the *Y*-axis indicating the true class labels from the final ResNet50 model

and trustworthiness for clinical trials of medical professionals. Apart from that, this framework can be used to identify and address biases in AI models during training. The simulation outcome for the proposed DFU_XAI framework during training on the DFU dataset is shown in Fig. 4. Each backbone CNN's hyperparameter settings used to train the framework are fine-tuned. The optimizer function and the gradient descent loss function are two important hyperparameters for training a model. Since Adam combines the essential characteristics of the AdaGrad and RMSProp optimizers and can handle sparse gradients on a large dataset, we selected Adam as an optimizer function. Since our work is based on binary classification in the DFU dataset, we selected binary cross-entropy as a loss function. A reasonable generalization of the model may be shown by the use of a modest batch size of 32. The suggested framework received 50th-epoch training. However, the model was able to achieve more than 97 percent training accuracy and 98 percent validation accuracy after only the 29th epoch of training. To overcome the over-optimistic issue, we used a cross-validation approach during oversampling. In this approach, the model never saw or oversampled the test samples during the training phase, thus measuring a correct estimation of the model's efficiency based on the training samples [40]. We can see from Fig. 4a that the overfitting problem is not present when the model is being trained. It is evident from the loss function curve in Fig. 4b that the curve begins to decline the loss value dramatically.

Evaluation Metrics

Confusion matrices are frequently employed in the context of classification tasks to assess the efficacy of DL models. It shows the relationship between the true label and the expected label. The model may forecast two labels (yes or no) for binary classification, leading to up to four possible results. Here are some of them:

True positive (TP) is the term used when a model properly predicts that a picture with an ulcer label actually contains an ulcer. False positive (FP) describes the situation in which a picture without an ulcer label is mistakenly identified as having an ulcer. False negative (FN) is the term used to describe when a picture with an ulcer label is mistakenly anticipated to be ulcer-free. When a picture with no ulcer label is accurately identified as not having an ulcer, it is referred to as a true negative (TN). The following definitions apply to the various evaluation metrics:

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

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

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

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

In DFU classification, the evaluation metrics play a vital role in the DFU_XAI framework's effectiveness. High accuracy reflects the overall correctness of the DFU_XAI framework, which aims to make strong predictions. By minimizing FP, precision helps to predict the presence of a foot ulcer. In the same way, by minimizing FN, recall reflects the majority of actual foot ulcers, preventing instances where a potentially serious condition is overlooked. F1-score ensures that the DFU_XAI framework accurately identifies ulcers while keeping FP and FN at the same level. AUC suggests that the framework is efficient at assigning higher likelihoods to ulcer cases compared to healthy cases.

Result Analysis

The ROC curve and confusion matrix for the DFU dataset in the DFU_XAI framework are shown in Fig. 5. ResNet50, a transfer learning model, serves as the foundation of the suggested framework. Explainable methodologies are then used to determine if the foot tissue is normal or an ulcer by using CNN. 195 and 221 DFU samples in total are categorized for ulcer and normal images, respectively, as shown in Fig. 5a. The suggested approach simultaneously misclassified just one sample of an ulcer. The DFU_XAI framework performs explainability, which makes the model dependable in some way, which is the most significant benefit. Additionally, this framework obtains an area value of 0.98535 (see Fig. 5b), demonstrating the model's consistency and stability. Additionally, each transfer learning model is run separately on the DFU dataset in order to better understand the effectiveness of the DFU_XAI framework. All of the pre-trained CNN models utilized in this experiment are compared in Table 5. The greatest accuracy, precision, recall, F1-score, and AUC score of 98.8%, 99.2%, 97.6%, 98.4%,

Table 5 Comparison of the DFU_XAI framework with ultra-modern networks

Model	Accuracy	Precision	Recall	F1-score	AUC
InceptionV3 [26]	0.863	0.812	0.839	0.825	0.858
DenseNet121[24]	0.928	0.917	0.895	0.906	0.922
Xception [27]	0.844	0.856	0.718	0.781	0.821
MobileNetV2 [28]	0.903	0.897	0.847	0.871	0.893
ResNet50 [25]	0.988	0.992	0.976	0.984	0.985

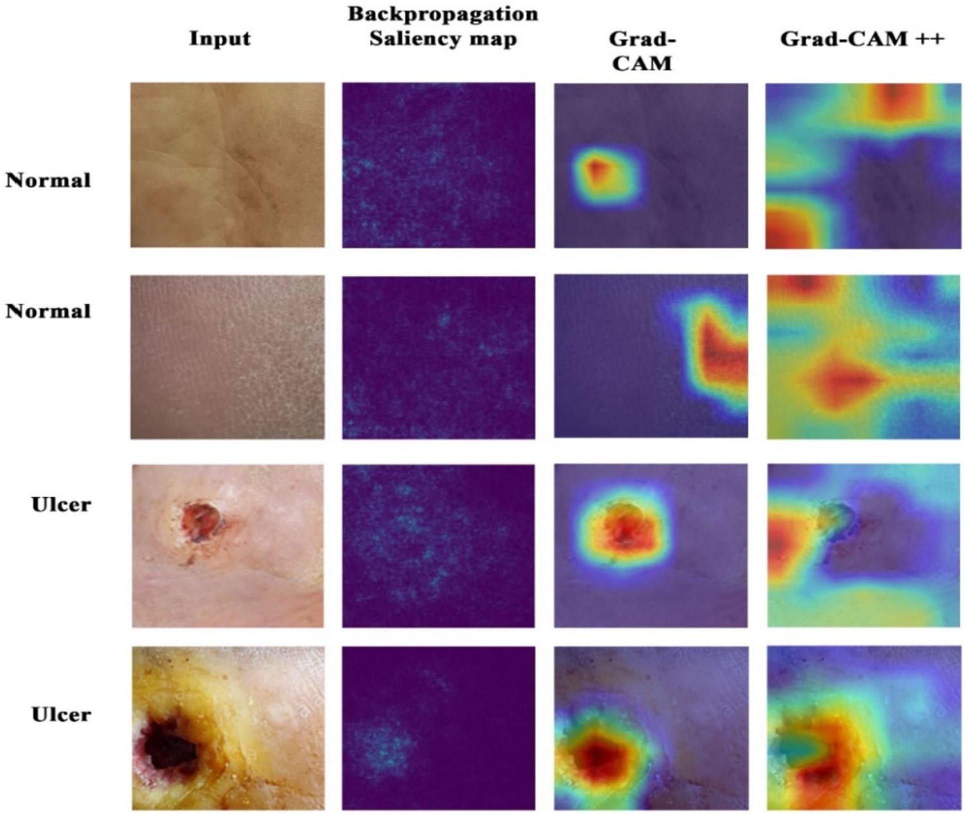


Fig. 6 Activation maps are generated by the three gradient-based algorithms (backpropagation saliency map, Grad-CAM, and Grad-CAM++) from ulcer, and normal DFU images. These algorithms used the last convolutional layer of the ResNet50 model to generate the heat maps. However, red and yellow areas indicate the large pixel areas of the DFU image that focuses on while a CNN model makes predictions

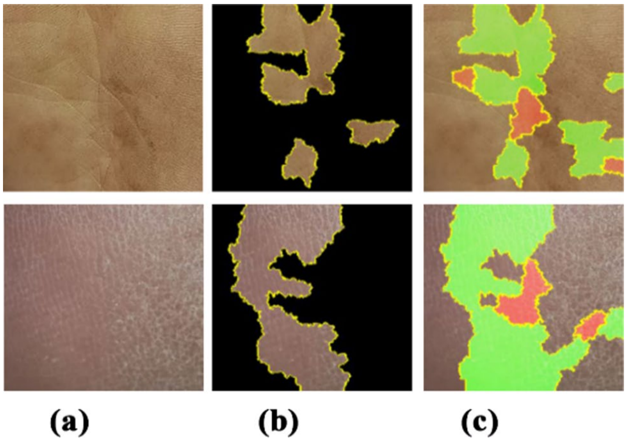


Fig. 7 Interpretations generated by LIME for normal DFU image. **a** Sample of the normal image **b** superpixels generated from the sample for segmentation **c** final perturbed image for the sample

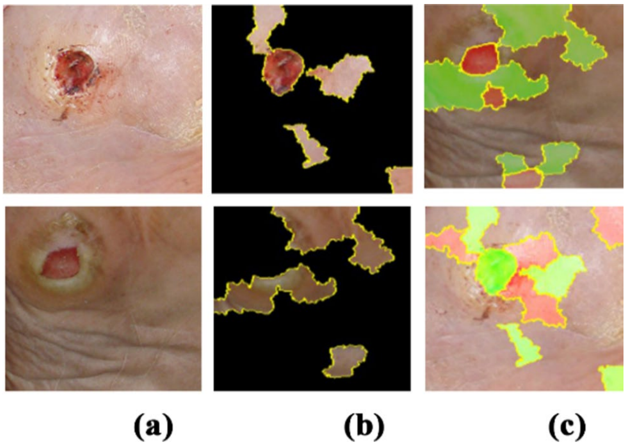


Fig. 8 Interpretations generated by LIME for ulcer DFU image. **a** Sample of the ulcer image **b** Superpixels generated from the sample for segmentation **c** final perturbed image for the sample

and 98.5% were attained by the ResNet50. The ResNet50 model was used as the foundation for the development of our suggested DFU_XAI framework since it produced the best results. Indicating the model's efficacy, the FNR is close to 0 and the TNR is nearly 1.

Our formulated DFU_XAI framework based on the transfer learning (TL) concept leverages pre-trained features on the ImageNet database and tunes these features by fine-tuning layers to improve the performance of the DFU_XAI framework compared to training traditional methods from scratch. TL is useful for a limited amount of labeled data. Our ResNet50 network pre-trained on ImageNet uses previous information that helps the network provide optimal performance metrics (i.e., accuracy, precision, recall, F1-score, and AUC) on a small volume of samples. On the other hand, a traditional model (TM) trained from

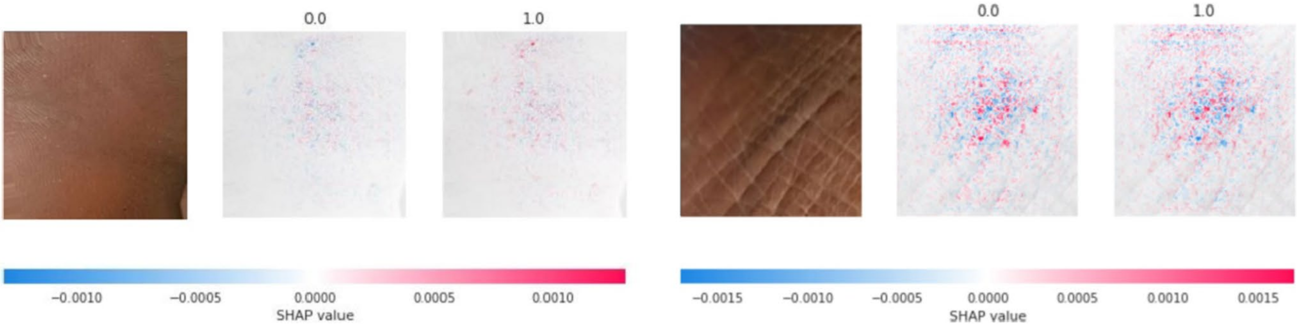


Fig. 9 Based on the Shapley values, we can conclude that the DFU image appears to be normal

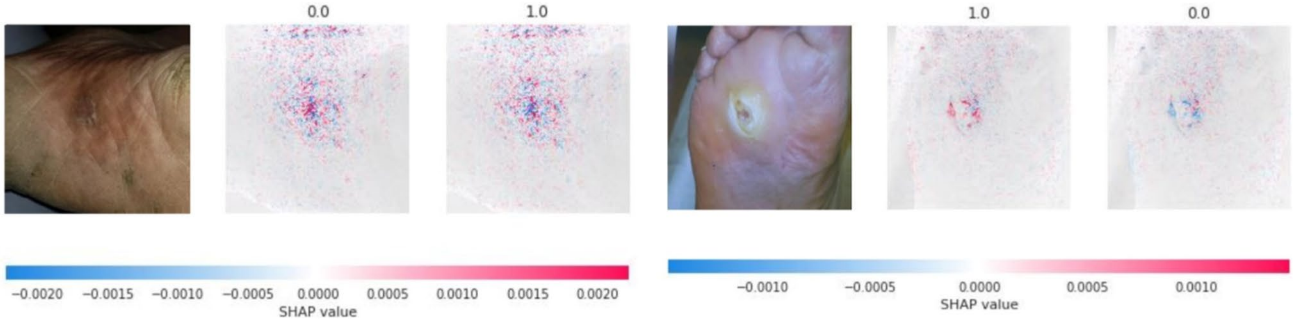


Fig. 10 Based on the Shapley values, we can conclude that the DFU image holds ulcer. The red points indicate positive SHAP values that increase the likelihood of the predicted label, on the other hand, blue points indicate negative SHAP values that decrease the likelihood of the predicted label

scratch on a small dataset may provide suboptimal results due to the limited amount of DFU data. This framework uses pre-trained weights that make it faster during training. This framework already has low-level features and can easily adjust its parameters for DFU classification. Training a TM from scratch might require more iterations that slow down the training time of the model. Thus, the DFU_XAI framework performs optimally in comparison to traditional methods for DFU analysis in terms of accuracy, precision, recall, F1-score, and AUC.

The numerical performance measures presented in Table 3 are certainly helpful in evaluating the ability of the models to differentiate between ulcer and healthy images. However, these measures only provide limited insight into how the results were obtained and can create challenges when conducting root cause analysis. Since the CNN model results are generated by a complex network, it can be difficult to trace how they were derived, turning the results into a "black box" for analysis. While the performance parameters can provide some understanding of the model's performance, they are not sufficient for root cause analysis as they do not provide a complete picture of the underlying processes. Therefore, when conducting a root cause analysis, it is essential to consider other factors beyond just the numerical performance measures to fully understand the factors that may have contributed to any potential failures. Despite its high accuracy, the ResNet50 model still exhibits a level of inaccuracy in 1.25% of cases. To address this issue, the study employed Grad-CAM, a technique that generates heat maps to highlight the image regions where the model focuses its efforts. These heat maps provide human experts with a visual representation of the areas used by the model to determine the predicted class, enabling them to evaluate its accuracy.

The accuracy of the localized regions indicated by these XAI techniques can be evaluated by skilled medical professionals. These experts can evaluate the indicated areas for accuracy and clinical relevance, ensuring that the model's predictions align with the actual clinical interpretations. To further test the model's performance in real-world clinical circumstances and confirm the efficacy of the XAI techniques, case studies including actual patient data can be examined. There are many practical implications of using Grad-CAM, LIME, and SHAP to visualize and interpret the decision-making process of the ResNet50 model, especially in a clinical setting. Healthcare professionals can visually show the ulcer-related areas in the image that are predicted by the ResNet50 model. This helps in understanding which

Fig. 11 ROC curves for proposed ResNet50 model and others pre-trained model, with the *X*-axis indicating the false positive rate (FPR) and the *Y*-axis indicating the true positive rate (TPR). The orange, green, LightSeaGreen, lime, and red colors show the MobileNetV2, Xception, ResNet50, InceptionV3, and DenseNet121, respectively. The highest AUC score achieved by the ResNet50 model was 0.985 and the lowest AUC score achieved by the Xception model was 0.821

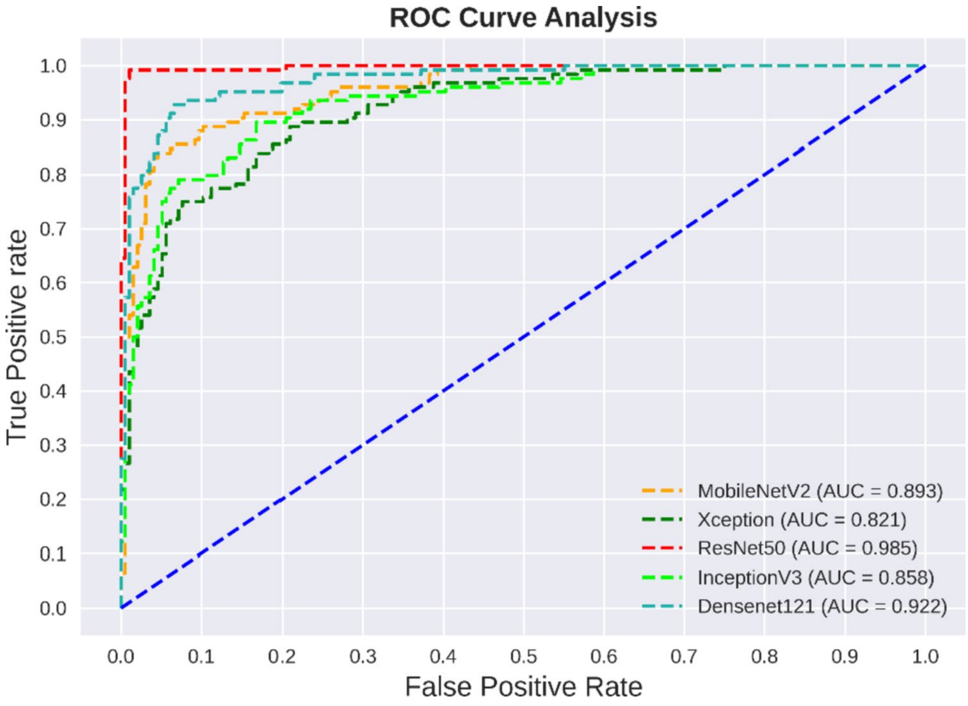


Table 6 Comparative analysis of the proposed DFU_XAI framework with other existing studies, where, except for the recall score, DFU_XAI exhibited outstanding performance compared with others

Papers	Accuracy	Precision	Recall	F1-score	AUC	Method
Das et al. [13]	0.964	0.926	**0.984**	0.954	0.974	DFU_SPNet
Alzubaidi et al. [14]	–	0.973	0.945	0.958	–	Hybrid CNN
Alzubaidi et al. [15]	–	0.954	0.936	0.945	–	DFU_QUTNet + SVM
Goyal et al. [16]	0.925	0.945	–	0.939	0.961	DFUNet
Proposed Framework	**0.988**	**0.992**	0.976	**0.984**	**0.985**	**DFU_XAI**

Obtaining high accuracy, the DFU_XAI framework can detect subtle changes in DFU skin and simultaneously classify both normal and ulcer skin

Bold values indicate the highest values of all performance metrics obtained from the proposed framework and all existing papers

regions are deemed significant for the model's decision, aiding in clinical validation. Healthcare professionals can gain insights into why the model made a specific prediction for a particular case. These algorithms can visualize and interpret the irregular structures and uncertain outer boundaries of the DFU.

The Grad-CAM algorithm utilizes activation maps from the last convolution layer of a model to create heat maps that highlight critical locations for class labels on an image. The study used the last convolution layer of the ResNet50 model as a reference to create heat maps. The resulting heat maps were compared to areas marked by DFU specialists for ulcers on multiple test images, and Fig. 6 displays the three gradient-based algorithm outputs. This procedure helps the specialist predict the image, which is the exact location of an ulcer.

Another XAI technique, called LIME, is employed to interpret the ResNet50 model and extract the important features that greatly affect the prediction. An illustration of region segmentation through superpixels is presented in Figs. 7 and 8, providing an example of how the input image is divided into smaller, interconnected regions with similar color and location.

On the other hand, the SHAP XAI algorithm is particularly useful for DFU classification tasks, where scores can be assigned to each pixel in the predicted image. In order to achieve local accuracy and consistency for model interpretation, we employed SHAP which analyzes all possible input combinations to determine Shapley values. To illustrate this, Figs. 9 and 10 show test images with their corresponding SHAP values.

In a clinical setting, there are many practical implications of using Grad-CAM, LIME, and SHAP as XAI algorithms to visualize and interpret the decision-making process of the ResNet50 model. These XAI algorithms play a role in gaining medical professionals' trust and confidence in their predictions. They are able to generate heatmaps that provide transparent and understandable insights for medical professionals' clinical decisions, such as prognosis, treatment, and diagnosis. They enhance the transparency of the ResNet50 model, permitting medical professionals to trace and validate the logic behind the model's prognosis. This transparency fosters trust in the model's predictions, enhancing confidence in its clinical decisions.

However, the DFU_XAI framework was evaluated using the receiver operating characteristic (ROC) curve, with the area under the curve (AUC) used as a measure of performance. The ROC curve depicted in Fig. 11 illustrates a comparative analysis among five commonly utilized CNN models like DenseNet121, ResNet50, InceptionV3, Xception, and MobileNetV2, all evaluated on the same learning parameter and sample split. The results indicate that the proposed ResNet50 outperformed the conventional CNN designs for the classification challenge of distinguishing between healthy DFU and ulcer cases.

Table 6 demonstrates a comparison between the proposed DFU_XAI framework and the most recent model, with DFU_XAI demonstrating outstanding results. The results in Table 4 showed that the DFU_XAI framework outperformed all other models such as DFUNet [19], hybrid CNN [17], DFU_QUTNet [18], and DFU_SPNet [16], in respect to the performance metrics. Due to a slightly larger value of false negatives (FN), the DFU_XAI framework has a lower recall score. When Table 4 is examined in more detail, it becomes clear that the DFU_XAI framework has good recall, precision, and accuracy values. F1-score and AUC prove the effectiveness of DFU_XAI in DFU ulcer classification. The excellent F1-score (98.4%) and accuracy (98.75%) ratings of DFU_XAI have led to its designation as an efficient prediction system.

The DFU_XAI framework works by combining XAI algorithms with the traditional AI model for diabetic foot ulcer analysis, while other existing models use only traditional AI models or trained data from scratch. The unique features that contribute to its outstanding performance are:

(i) Development of the DFU_XAI framework using a fine-tuned ResNet50 model rather than from scratch to mitigate the "vanishing gradient" challenge,
(ii) Integration of XAI algorithms with the pre-trained CNN model to eliminate the lack of explainability of existing models,
(iii) Automatic explanation of ulcer cases using an XAI-based heatmap implementation system from the DFU dataset to assist endocrinologists and medical practitioners.

As we use a more generalizable Adam optimizer and fewer epochs for the network to converge, that ultimately reduces the duration of the image training process. Our network has the potential to gain knowledge from a significantly larger dataset, resulting in the model being used for practical deployment and better generalization. By using parallel convolutional modules with the right amount of depth and breadth, the DFU_XAI architecture makes a good receptive field. The detection of ulcers from the DFU images at various scales is further improved by the use of many distinct kernel numbers and sizes. At the same time, our explainable model provides clear insights into the decision-making process, making it easier for endocrinologists to determine their areas of focus. This system can also serve as a useful adjunct tool for trainee or junior endocrinologists, enabling them to make accurate diagnoses without the need for segmentation techniques. Thus, the proposed methodology has the potential for clinical application in detecting ulcer-related areas on DFU images.

The DFU_XAI framework addresses the interpretability and trustworthiness of AI models (black box models) in a clinical context by converting them into more explainable and transparent "white box" models. Significant drawbacks of these "black box" models include the inability to (1) assess the prevalence of ulcer cases and (2) offer adequate insights into model intricacies. The most widely used XAI tools to clearly understand these "black box" models are LIME, Grad-CAM, and SHAP. The DFU_XAI framework ensures the credibility and interpretability of AI models in a clinical setting, particularly when compared to black box models, through the utilization of these XAI tools. This framework presents an XAI-based heatmap interpretation system to find and detect ulcers in the DFU dataset.

Discussion

The DFU_XAI framework has been proposed only for diabetic foot ulcer (DFU) analysis. This framework is not generalizable to different datasets and medical scenarios beyond DFU analysis. In the future, the DFU_XAI framework can be expanded by combining multiple models instead of the ResNet50 model. We believe that this customized DFU_XAI framework will be more generalizable to different datasets and medical scenarios than the proposed DFU_XAI framework.

The DFU_XAI framework encountered many challenges when deployed in real-world clinical settings. This framework may fail to explain the rapid spread of ulcers, the overall health condition of the patients, and their diabetic status. It must be considered that the explanations generated by the XAI-based model are clinically acceptable and pertinent to the skills of clinicians, especially in the context of real-world clinical settings. It must be considered that the security and privacy of patient details are important when providing explanations of the prognosis to ensure framework reliability and safety.

The DFU_XAI framework offers various benefits to endocrinologists and medical practitioners in terms of making accurate diagnoses and treatment decisions. This XAI-based framework aids endocrinologists in comprehending the reasons behind the conventional AI model's decisions. By emphasizing risk factors, this framework helps medical practitioners know how likely they are to develop an ulcer. As well as transparency, this framework could increase endocrinologists' confidence by providing a higher degree of accuracy (98.75%).

While the AI models significantly contribute to clinical application, they have yet to provide a significant outcome in diabetic foot ulcer detection and treatment. Regulatory and ethical problems still remain with regard to the implementation of existing medical practices in clinical applications. These problems might be addressed by integrating the DFU_XAI framework into existing medical practices. The use of this AI-based framework has increased due to the significant success of intelligent decision-based systems. The XAI-based framework aids medical professionals and junior practitioners to appropriately trust, understand, and efficiently communicate with these approaches.

The explainability of the DFU_XAI framework plays a role in gaining medical professionals' trust and confidence in its predictions. The XAI-based heatmap explanation provides transparent and understandable insights for medical professionals' clinical decisions such as prognosis, treatment, and diagnosis. This XAI-based framework enhances the transparency of the model and permits medical professionals to trace and validate the logic behind the framework's prognosis. This transparency fosters trust in the framework's predictions, enhancing confidence in its clinical decisions. However, this framework permits medical professionals to examine ulcer samples where the framework may have given an incorrect prognosis. Thus, the DFU_XAI framework might impact its adoption in clinical practice.

Limitations of the Study

Considering the success of the DFU_XAI framework in diabetic foot ulcer prediction, there are some potential challenges or considerations that might arise when applying it to other skin lesions or medical conditions. During the training period of the DFU_XAI framework, full data is transferred or accessed, but without permission, accessing private information breaches patient confidentiality. So information privacy is the potential limitation in this framework that might arise when applying it to other skin lesions or medical conditions. During the evaluation of the DFU_XAI framework, this framework transfers or accesses full data for training the model, but accessing a patient's private information breaches patient confidentiality. So the privacy of data is a potential limitation for improvement identified during the evaluation of the DFU_XAI framework. Recently, the federated learning (FL) approach has gained popularity as it offers an efficient solution for centralized computation, high computation capacity, and data privacy. As a result, the FL approach might be able to address this limitation in subsequent iterations or research.

Conclusion

This study suggested a successful deep learning framework integrating the explainable method to automatically identify foot ulcers. The DFU image patches of 1655 are utilized obtaining the best prediction accuracy for ResNet50 as a backbone CNN. Among the examined pre-trained models, the ResNet50 obtained an accuracy of more than 98%, along with a satisfied F1-score and AUC. The minimal number of skin patches was utilized, which was the primary drawback of this study and was successfully overcome. The Grad-CAM, LIME, and SHAP approaches, on the other hand, were also used to visualize the model's decision-making process and to understand the salient characteristics that distinguish between normal and pathological skin. The outcome indicates that the suggested model is more reliable and can quickly pinpoint the precise position of the ulcer, which enhances medical faith. Additionally, the proposed model may help medical providers with the correct diagnosis of foot ulcers brought on by illnesses other than diabetes, improving patient treatment and results. Due to the success of DFU_XAINet in predicting DFU, this new method will be used to find different skin lesions, such as infections like chickenpox or shingles, wound classification, skin lesions like moles and freckles, pimples, and spotting marks when compared to normal skin. In the future, there is a plan to integrate multiple pre-trained CNN models to create a customized multiscale transfer learning architecture. This approach aims to not only reduce training time but also enhance specificity and sensitivity, paving the way for more accurate and efficient skin lesion classification. In future research directions, the DFU_XAI framework may be expanded to detect different skin lesions, infections, and other dermatological conditions. This expansion can be done by combining multiple pre-trained CNN networks instead of a single network. However, when the extended version of the DFU_XAI framework is trained on different datasets, it will transfer or access full data, but without permission, accessing private information breaches patient confidentiality.

So data privacy challenges might be associated with this extension.

The DFU_XAI framework could help improve the accuracy and sensitivity of classifying skin lesions, especially when more than one pre-trained CNN models are used in future studies. This customized architecture allows for the unique lesion feature extraction of each pre-trained CNN model independently. So, this customized architecture might help improve the specificity and sensitivity of skin lesion classification by pulling out more lesion features and combining the best parts of different networks. In the field of medical image analysis, the potential future applications of this framework are to predict and analyze complications associated with diabetes problems other than foot ulcers, like retinopathy or neuropathy. The framework can be extended by training the ResNet50 model on different datasets comprising other skin lesions like eczema, melanoma, or psoriasis.

Future Work

In terms of future research directions, we have a plan to build up a customized multiscale transfer learning architecture by combining multiple pre-trained CNN networks instead of a single network. To build up this customized architecture, a global average pooling layer must be added to each individual network to compute the average of all features. Then, these average features will be combined using a concatenation layer. Finally, a fully connected layer must be used for classification tasks. We have explained the concept of multiscale transfer learning in another article [41] named DFU_MultiNet in more detail. However, this customized architecture allows for the unique feature extraction of each pre-trained CNN model independently. Thus, by extracting more features and combining the strengths of individual networks, this customized architecture might enhance the overall framework's performance.

The DFU_XAI framework plays a crucial role in shaping the landscape of AI advancements in health care, particularly in the domain of medical image analysis. This XAI-based framework allows healthcare practitioners to analyze and explain AI-based insights. This framework also contributes to progress in AI in health care by identifying biases in AI models and giving clear explanations of model predictions to patients. It might play an effective role in shaping future developments in medical image analysis, such as improved diagnostic decisions, patients' health conditions, diabetic levels, and treatment plans.

In the future, the DFU_XAI framework might help with patient care and outcomes, particularly in cases where diseases other than diabetes might be to blame for foot ulcers. Various types of infections, such as bacterial, viral, or fungal, affect the tissue of the foot and may contribute to the creation of foot ulcers. In future research, the DFU_XAI framework should be extended to classify and predict such infections that are responsible for foot ulcers, improving patient treatment and outcomes.

Funding There is no funding for this research work.

Data Availability The works used publicly available datasets mentioned in reference [19].

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. R. Mostafiz, Diagnosis of diabetes: a machine learning paradigm using optimized features. Netw. Biol. **11**(3), 222 (2021)
2. Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 12 May 2023
3. A.J. Boulton, L. Vileikyte, G. Ragnarson-Tennvall, J. Apelqvist, The global burden of diabetic foot disease. Lancet **366**(9498), 1719–1724 (2005)
4. C. Liu, J.J. van Netten, J.G. Van Baal, S.A. Bus, F. van Der Heijden, Automatic detection of diabetic foot complications with infrared thermography by asymmetric analysis. J. Biomed. Opt. **20**(2), 026003 (2015)
5. D.G. Armstrong, L.A. Lavery, L.B. Harkless, Validation of a diabetic wound classification system: the contribution of depth, infection, and ischemia to risk of amputation. Diabetes Care **21**(5), 855–859 (1998)
6. P. Cavanagh, C. Attinger, Z. Abbas, A. Bal, N. Rojas, Z.-R. Xu, Cost of treating diabetic foot ulcers in five different countries. Diabetes Metab. Res. Rev. **28**, 107–111 (2012)
7. F. Aguiree et al., IDF diabetes atlas, 6th edn., ed. L. Guariguata, T. Nolan, J. Beagley, U. Linnenkamp, O. Jacqmain (International Diabetes Federation, Brussels, 2013)
8. G. Litjens et al., A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
9. F.J. Veredas, R.M. Luque-Baena, F.J. Martín-Santos, J.C. Morilla-Herrera, L. Morente, Wound image evaluation with machine learning. Neurocomputing **164**, 112–122 (2015)
10. L. Alzubaidi, M.A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, Robust application of new deep learning tools: an experimental study in medical imaging. Multimed. Tools Appl. **81**(10), 13289–13317 (2022)
11. B. Najafi, H. Mohseni, G.S. Grewal, T.K. Talal, R.A. Menzies, D.G. Armstrong, An optical-fiber-based smart textile (smart socks) to manage biomechanical risk factors associated with diabetic foot amputation. J. Diabetes Sci. Technol. **11**(4), 668–677 (2017)
12. M. Kaselimi, E. Protopapadakis, A. Doulamis, N. Doulamis, A review of non-invasive sensors and artificial intelligence models for diabetic foot monitoring. Front. Physiol. **13**, 924546 (2022)
13. S.K. Das, P. Roy, A.K. Mishra, DFU_SPNet: a stacked parallel convolution layers based CNN to improve diabetic foot ulcer classification. ICT Express **8**(2), 271–275 (2022)
14. L. Alzubaidi, A.A. Abbood, M.A. Fadhel, O. Al-Shamma, J. Zhang, Comparison of hybrid convolutional neural networks models for diabetic foot ulcer classification. **16** (2021)

15. L. Alzubaidi, M.A. Fadhel, S.R. Oleiwi, O. Al-Shamma, J. Zhang, DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network. Multimed. Tools Appl. **79**(21–22), 15655–15677 (2020)
16. M. Goyal, N.D. Reeves, A.K. Davison, S. Rajbhandari, J. Spragg, M.H. Yap, Dfunet: convolutional neural networks for diabetic foot ulcer classification. IEEE Trans. Emerg. Top. Comput. Intell. **4**(5), 728–739 (2018)
17. L. Wang, P.C. Pedersen, E. Agu, D.M. Strong, B. Tulu, Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification. IEEE Trans. Biomed. Eng. **64**(9), 2098–2109 (2017). https://doi.org/10.1109/TBME.2016.2632522
18. M. Goyal, M.H. Yap, N.D. Reeves, S. Rajbhandari, J. Spragg, Fully convolutional networks for diabetic foot ulcer segmentation, in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE, 2017), pp. 618–623
19. Dataset: diabetic foot ulcer (DFU), Available link: https://www.kaggle.com/laithjj/diabetic-foot-ulcer-dfu
20. L. Alzubaidi et al., Towards a better understanding of transfer learning for medical imaging: a case study. Appl. Sci. **10**(13), 4523 (2020)
21. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. Artif. Intell. Res. **16**, 321–357 (2002)
22. S.J. Pan, Q. Yang, A survey on transfer learning IEEE transactions on knowledge and data engineering. **22**(10), 1345 (2009)
23. J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: a survey. Knowl.-Based Syst..-Based Syst. **80**, 14–23 (2015)
24. Z. Huang, Z. Pan, B. Lei, Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. Remote Sens. **9**(9), 907 (2017)
25. G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Honolulu, 2017), pp. 2261–2269
26. Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn.Recogn. **90**, 119–133 (2019)
27. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Las Vegas, NV, 2016) pp. 2818–2826
28. F. Chollet, Xception: deep learning with depthwise separable convolutions, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Honolulu, HI, 2017) pp. 1800–1807
29. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Salt Lake City, UT, 2018) pp. 4510–4520
30. G. Battineni, G.G. Sagaro, N. Chinatalapudi, F. Amenta, Applications of machine learning predictive models in the chronic disease diagnosis. JPM **10**(2), 21 (2020)
31. E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019)
32. A.M. Antoniadi et al., Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl. Sci. **11**(11), 5088 (2021)
33. M.T. Ribeiro, S. Singh, C. Guestrin, 'Why should I trust you?' Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 1135–1144
34. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626
35. S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30** (2017)
36. N. Ketkar, *Deep Learning with Python* (Apress, Berkeley, 2017). https://doi.org/10.1007/978-1-4842-2766-4
37. O. Russakovsky et al., ImageNet large scale visual recognition challenge (2015). Accessed 23 Sep 2023. Preprint at http://arxiv.org/abs/1409.0575
38. A. Alqahtani, X. Xie, M.W. Jones, Literature review of deep network compression. Informatics **8**(4), 77 (2021)
39. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014). Preprint at http://arxiv.org/abs/1412.6980
40. M.S. Santos, J.P. Soares, P.H. Abreu, H. Araujo, J. Santos, Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [Research Frontier]. IEEE Comput. Intell. Mag.Comput. Intell. Mag. **13**(4), 59–76 (2018). https://doi.org/10.1109/MCI.2018.2866730
41. S. Biswas, R. Mostafiz, B.K. Paul, K.M. Mohi Uddin, M.M. Rahman, F.N.U. Shariful, DFU_MultiNet: a deep neural network approach for detecting diabetic foot ulcers through multi-scale feature fusion using the DFU dataset. Intell.-Based Med. **8**, 100128 (2023)