Biomedical Materials & Devices
https://doi.org/10.1007/s44174-025-00434-x

ORIGINAL ARTICLE

SkinFLNet: A Federated Learning Approach for Skin Cancer Detection Utilizing Skin Dermoscopy Images

Sajeeb Saha[1] · Shuvo Biswas[1,2] · Sneha Sarkar[3] · Md. Abdul Hadi[4] · Nilanjana Basak[1] · Mst. Zakia Sultana[1,2]

Received: 1 March 2025 / Accepted: 10 June 2025

Abstract

Skin cancer is among the most widespread types of cancer globally, and detecting it can be difficult, even for experienced dermatologists. Early detection is key to successful treatment, and deep learning methods, especially deep convolutional neural networks (DCNNs), have demonstrated significant potential in this area. However, achieving high accuracy with these models requires large datasets, which individual medical institutions often lack. Sharing medical data directly are also challenging due to privacy and legal concerns. To address this problem, we propose a federated learning approach to develop a privacy-preserving and accurate system for classifying skin cancer, helping dermatologists make better decisions. In this manuscript, we used five Deep CNN models (i.e., Densenet169, VGG16, InceptionV3, Xception, and InceptionResNetV2) to identify key characteristics from skin lesion images. Then a fine-tuning layer is used to refine the entire model and reduce complexity. A fully connected layer combined with a softmax activation function is applied to perform the classification task. Finally, we apply the federated learning approach with the Deep CNN model for the privacy-preserving of the patient information. We evaluate our method using the ISBI2016 dataset containing 1279 skin lesion images. The FL-VGG16 model performed the best among all models, achieving 92.08% accuracy, 76.92% F1-measure, 90.91% precision, and 98.36% specificity. Additionally, we also used the Local Interpretable Model-Agnostic Explanations (LIME) technique to make the predictions easier to understand. We believe our model can help healthcare professionals make accurate predictions about skin cancer and support better treatment decisions.

Keywords Skin cancer · Deep learning · Federated learning · Explainable deep learning

Introduction

Skin cancer (SC) is one of the most common cancers worldwide, and its cases have been increasing steadily over the years [1]. The World Health Organization (WHO) reports that about 3 million cases of non-melanoma SC occur each year and over 132 cases of melanoma are diagnosed worldwide each year [2]. In addition, SC leads to a considerable number of deaths annually, with melanoma accounting for around 66,000 fatalities in 2020 alone [3].

SC occurrence differs significantly between regions and populations, with higher rates found in countries that have

✉ Shuvo Biswas
shuvo.ict13@gmail.com

Sajeeb Saha
sajeebsaha998@gmail.com

Sneha Sarkar
snehasarkar19112004@gmail.com

Md. Abdul Hadi
hadim3@mail.montclair.edu

Nilanjana Basak
nilanjanabasak96@gmail.com

Mst. Zakia Sultana
sultanajakia628@gmail.com

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

[2] Department of Computer Science and Engineering, The People's University of Bangladesh, Dhaka, Bangladesh

[3] Department of Accounting and Information Systems, Bangladesh University of Professionals, Dhaka, Bangladesh

[4] Department of Information Technology, Montclair State University, New York, USA

Published online: 10 July 2025

fair-skinned populations and elevated levels of ultraviolet radiation. [4]. Australia has the world's highest rate of SC, with melanoma cases occurring at two to three times the rates seen in the United States, Europe, and Canada. Likewise, New Zealand has the highest global melanoma incidence [5]. Therefore, early detection and precise SC diagnosis are essential for enhancing patient outcomes and lowering mortality rates. Despite this, diagnosing SC can be difficult [6], and visual skin examinations by dermatologists can take a lot of time and may vary depending on individual judgment [7].

Machine learning (ML), mainly deep convolutional neural networks (DCNNs), has shown promise in correctly detecting and categorizing skin lesions in medical data [8]. Many studies highlight the continuous research efforts in SC classification using ML techniques, underscoring their potential to advance early detection and diagnosis [9]. However, these models need extensive data for training, which is often not available in typical medical settings, as individual institutions usually lack sufficient information. Additionally, more studies are needed to improve the accuracy and dependability of SC classification using ML. A major challenge in this area is designing a powerful and reliable DCNN that can correctly classify SC from medical images [10]. Exploring various datasets for training and testing DCNNs is crucial to ensure that these models are robust and generalizable across diverse populations and regions.

Federated Learning (FL) is a method in ML that enables models to learn from data distributed across various locations without the need to share the actual training data [11]. FL enables the training of ML algorithms on multiple local datasets without the need to share data, allowing healthcare organizations to create a single global model while ensuring that the data remains decentralized. This method is particularly valuable in scenarios such as SC detection, where hospitals are restricted from sharing patient data due to privacy regulations. Nonetheless, deploying FL in real-world medical environments presents several challenges, including managing data variability, addressing communication overhead, and balancing model complexity with computational efficiency. Furthermore, ensuring data privacy and security, complying with regulatory requirements, and integrating the model into clinical workflows require careful attention. Implementing FL in real-world hospital settings is essential but poses various practical challenges that must be addressed to ensure feasibility. One major concern is network latency, as FL requires frequent communication between client servers and a central server for model updates. High latency can impact the training time of the proposed model and make the system less responsive. Again, hardware constraints in some hospitals may impact the installation of complicated DL models due to low processing capacity or obsolete infrastructure. To address these pitfalls, lightweight model frameworks, efficient smart methodologies, and edge computing systems can be implemented. Additionally, a secure and stable network infrastructure is needed to ensure a real-time synchronization mechanism while maintaining data privacy. Addressing these problems is vital for making FL an effective and practical solution in several clinical settings. Apart from these obstacles, the potential advantages of a privacy-preserving AI-based diagnostic system justify the investment in strong FL infrastructure, automatically detecting SC in the early stage while maintaining privacy-preserving.

The novelty of this manuscript lies in the combination of FL with multiple DCNN algorithms to present a privacy-preserving and accurate SC identification framework. Unlike conventional methods that need centralized data management, our strategy permits collaborative training across diverse platforms without exposing medical information, solving information privacy with healthcare concerns. Additionally, the implement of diverse DCNN architectures—DenseNet169, VGG16, InceptionV3, Xception, and InceptionResNetV2—boosts feature extraction and framework strength. The fusion of fine-tuning layers minimizes complexity and computational cost of the proposed architecture. However, most of the papers fail to show the transparency and explainability of their proposed architecture. In this paper, we applied XAI-based LIME strategy permits transparency in final decisions, helping trust and approval among medical experts. This novel amalgamation of FL, DL, and interpretability indicates a noteworthy improvement toward reliable, secure, and clinically applicable AI tools in dermatological diagnostics. The following are the manuscript contributions:

1. We employ five deep CNN models (i.e., Densenet169, VGG16, InceptionV3, Xception, and InceptionResNetV2) to identify rich characteristics of skin lesions.
2. Used a fine-tuning layer to refine the entire architecture and reduce the computational complexity.
3. Utilization of the federated learning approach with deep CNN model to create a secure environment for classifying skin cancer.
4. We also applied the LIME XAI method to interpret the predicted results of the suggested system.

This manuscript is structured as follows: In "Literature Review" Section presents the literature review, In "Methods and Methodology" Section outlines the proposed methods, In "Experiments and Results Analysis" Section details the experimental setup and results, In "Discussion" Section provides a comprehensive discussion, and In "Conclusion" Section concludes the study with future directions.

Literature Review

Several approaches have been offered in the literature for the identifying of skin lesions.

Yu. et al. [12] offered a system for predicting the ISBI2016 skin dataset, which can work with or without a segmentation module. For segmentation, a fully convolutional residual network (FCRN) was used to accurately detect lesions. This FCRN consists of 16 residual blocks, with each block containing one 3×3 convolutional layers and two 1×1 convolutional layer. The classification process combines two methods: a SoftMax function and a support vector machine (SVM) ML classifier. When segmentation was applied, the classification accuracy reached 85.5%, and without segmentation, it was 82.8%. Ali et al. [13] discovered a novel model named LightNet to classify melanoma images as either benign or malignant. The model used a CNN that was adjusted to work with the ISBI2016 skin challenge dataset. LightNet achieved an accuracy of 81.6%, a sensitivity of 14.9%, and specificity of 98%. Its architecture includes 5 convolutional layers, 4 ReLU activation layers, 3 max-pooling layers, and 4 dropout layers.

In [14], the authors proposed a two-class classifier that takes skin lesion images labeled as benign or malignant as input and builds a model using a CNN. This method showed strong results, achieving an accuracy of 81.33%, a sensitivity of 78.66%, and a precision of 79.74% when tested on the ISBI2016 challenge dataset. Demir et al. [15] presented that skin lesion images were divided into two types: melanoma and benign, using the ResNet-101 and Inception-v3 models trained on the ISIC archive. The classification resulted in accuracy rates of 87.42% with Inception-v3 and 84.09% with ResNet-101.

Khan et al. [16] described a new approach for SC segmentation and classification was developed, using a hybrid framework for the segmentation task. This framework combined a 20-layer and a 17-layer CNN to segment skin lesion images. For classification, a 30-layer CNN was used to retrieve features from the lesion images. A feature fusion technique, based on Summation Discriminant Correlation Analysis (SDCA), and a feature selection method using Regular Falsi (RF) were also applied. The proposed methods were tested on several datasets, including ISBI2019, ISIC2018, ISBI2017, ISIC2016, and HAM10000. The segmentation model achieved an accuracy of 92.70% on the ISIC2018 dataset, and the classification model achieved 87.02% accuracy on the HAM10000 dataset. In [17], the authors provided a transfer learning approach that was used to classify skin cancer, where six different models— MobileNet, InceptionV3, Xception, ResNet50, and InceptionResNetV2, VGG19—were tested. Among these models, Xception showed the best performance, achieving an accuracy of 90.48%. Fraiwan and Faouri et al. [18] reviewed a system using AI that was created to classify skin cancer, testing thirteen pre-trained deep CNN models with skin lesion images from the HAM10000 dataset. Of all the models, DenseNet-201 achieved the best performance, with an F1-score of 74.4% and an accuracy of 0.829. Since the HAM10000 dataset is imbalanced, the F1-score is a better measure of performance. However, the F1-score of 0.744 in this study was still relatively low.

Aljohani and Turki et al. [19] proposed 8 CNN models—DenseNet201, GoogleNet, VGG19, Xception, ResNet152V2, ResNet50V2, VGG16, and MobileNetV2—were tested for SC classification. These models were trained on 7,164 images from the ISIC2019 dataset. Of all the models, GoogleNet achieved the best test accuracy of 76.09%. Gouda et al. [20] used ESRGAN [21] to increase the dataset for training the CNN model for SC classification, and synthetic samples were created. The CNN, which was trained on the ISIC2018 dataset, attained an accuracy of 0.832. This is similar to the performance of more advanced models like ResNet50, Inception-ResNet and InceptionV3. The offered method was tested on the ISIC2018 dataset, with the best classification accuracy of 0.8576 achieved using InceptionV3. Keerthana et al. [22] provided two new hybrid CNN classifiers with a SVM classifier at the output layer were proposed to classify skin images as either melanoma or benign lesions. The features retrieved from these two models were combined and then fed into the SVM ML classifier for the final result. These models were tested on the ISBI2016 dataset. The first hybrid model, which combined MobileNet and DenseNet201 with the SVM ML classifier, achieved the best accuracy of 88.02%. The second hybrid model, which used DenseNet201 and ResNet50 with the SVM ML classifier, reached an accuracy of 87.43%.

Bassel et al. [23] combined CNN approach using the Stacked Cross-Validation (CV) method was suggested for classifying SC, with models trained on the ISIC2019 dataset. This Stacked CV method worked at three levels, combining DL techniques with traditional ML methods like SVM, logistic regression (LR), K-Nearest Neighbors (KNN), Neural Networks (NN), and Random Forest (RF). The approach used three different models for feature extraction: ResNet50, EfficientNet, and VGG16. Among these, EfficientNet gave the best results, with an accuracy of 90.9% and an F1-measure of 89%. Gajera et al. [24] focused on using a pre-trained CNN-based system to automatically classify melanoma, combining different classifiers and eight popular CNN models across four skin cancer datasets: ISIC2016, ISIC2017, PH2, and HAM10000. The results showed that using DenseNet121 with a multi-layer perceptron (MLP) gave the best results. The model achieved accuracies of 98.33% on

PH2, 80.47% on ISIC2016, 81.16% on ISIC2017, and 81% on HAM10000, outperforming other CNN classifiers and the latest methods available. Yu and Wang et al. [25] offered a new approach was introduced that combines DCNN, Fisher Vector (FV), and feature encoding techniques to improve melanoma detection. The model, trained using the ISBI2016 dataset, reached an accuracy of 86.54%. Several researchers have proposed alternative frameworks for cancer therapy. For instance, the study in [26] introduced a metal–organic method using lanthanum metal, while the authors in [27] designed a drug delivery framework utilizing silymarin-loaded nanovesicles.

While these methods are designed for automatic SC prediction, gathering medical images from multiple medical centers or hospitals remains challenging due to the sensitivity of patient information. Security concerns often lead patients to be reluctant to share their data with third parties.

In the existing works, most research on SC detection primarily relies on traditional DL approaches. Few studies have explored ensemble learning [14, 15] to enhance model accuracy, while specific methods have been tailored for particular tasks [5, 16]. Due to limited training data, many papers have utilized conventional data augmentation techniques [13, 24]. Additionally, some studies have looked into privacy issues related to data [25]. However, to the best of our knowledge, no studies have yet proposed methods for identifying SC using FL.

In conclusion, this paper introduces a secure system aimed at solving the challenges mentioned earlier. By combining FL with DL methods, the system removes the need for sharing data between different parties. To augment the limited dataset of SC lesions, we employ data augmentation techniques to expand the dataset size. However, the overview of the existing papers with strengths and weaknesses are provided in Table 1.

Methods and Methodology

This part presents our recommended method and the techniques applied to identify SC. The overall architecture of the offered system is illustrated in Fig. 1. The offered method

Table 1 Overview of strengths and weaknesses of key DL techniques for skin lesion classification

References	Strengths	Weaknesses
Yu et al. [12]	The model offers flexibility by functioning with or without image segmentation, enhancing adaptability	Classification performance drops slightly when segmentation isn't applied, showing few dependency on segmentation
Ali et al. [13]	LightNet achieves high specificity, effectively identifying benign cases with minimal false positives	The model has low recall and is struggling to correctly identify many malignant melanoma cases
Lopez et al. [14]	The classifier demonstrates balanced performance with good sensitivity and precision for detecting malignant cases	Overall accuracy could be improved, as 81.33% leaves room for better classification performance
Demir et al. [15]	Inception-v3 model outperforms ResNet-101, providing higher accuracy for classifying skin lesion images	ResNet-101 shows slightly lower accuracy, indicating it may not be as effective for this task
Khan et al. [16]	The hybrid framework achieved high segmentation accuracy, with 92.70% on the ISIC 2018 dataset	Classification accuracy, though decent at 87.02%, could be further improved on the HAM10000 dataset
Jain et al. [17]	The Xception model achieved the highest classification accuracy of 90.48%, demonstrating excellent performance in skin cancer classification	Other models did not perform as well as Xception, indicating potential limitations in their effectiveness
Fraiwan and Faouri [18]	DenseNet201 performed well, achieving the highest accuracy of 82.9% among evaluated models	The F1-score of 0.744 is low, indicating potential issues with model robustness and balance
Aljohani and Turki [19]	GoogleNet achieved the highest test accuracy of 76.09%, demonstrating effective skin cancer classification performance	Overall accuracy of 76.09% is relatively low, suggesting room for improvement in classification methods
Gouda et al. [20]	The use of ESRGAN effectively expanded the dataset, improving CNN training for skin lesion classification	The overall accuracy of 83.2% is still lower than the highest performance of InceptionV3
Keerthana et al. [22]	The hybrid CNN models effectively combined feature extraction and SVM classification, achieving high accuracy	The second hybrid model's accuracy of 87.43% is lower than the first model's performance
Bassel et al. [23]	The exception model demonstrated excellent performance with 0.909 accuracy and an F1-measure of 89%	The complex Stacked CV method may increase computational requirements and training time significantly
Gajera et al. [24]	DenseNet121 with MLP achieved impressive accuracy, particularly 98.33% on the PH2 dataset	The performance on ISIC datasets is lower, indicating variability in effectiveness across different datasets
Yu and Wang et al. [25]	The novel method successfully combines deep CNN and FV techniques, achieving an accuracy of 86.54%	An accuracy of 86.54% suggests there is still room for improvement in melanoma recognition

for predicting SC involves several key steps. First, it balances the data using the Synthetic Minority Over-sampling Technique (SMOTE) and increases the dataset with data augmentation. Then, DL models are applied to classify SC. To ensure privacy and evaluate performance, we also tested these DL models in a FL setup. These strategies are designed to improve the accuracy of the classification model. The results show that using DL models for SC detection is a promising approach. The overall predicting approach of the proposed system is presented in Algorithm 1.

Algorithm 1 Automated skin cancer classification and detection.

Input:

D_{Tr} : Training data (70%)

D_{Vl} : Validation data (10%)

D_{Ts} : Testing data (20%)

b : Batch size

l : Learning rate

o : Optimizer

e : No of epochs

m : Mini-batch size

Output:

w : Optimized weights of pre-trained DCNN algorithms

Steps:

1. ***Preprocessing****: Resize all images into 224×224 pixels.*
2. ***Data Augmentation****: Utilize data augmentation method to enhance the training samples.*
3. ***Feature Extraction****: Extract feature maps utilizing pre-trained DCNN algorithms: Densenet169, VGG19, InceptiobV3, Xception, and InceptionresnetV2*
4. ***Feature Pooling****: Flatten the extracted maps using a Global Average Pooling layer.*
5. ***Fine-Tuning****: Add and configures tuning layers: dense, dropout, and softmax.*
6. ***Parameter Initialization****: Initialize training parameters: b, l, e, m, and o.*
7. ***Model Initialization****: Initialize DCNN models and find the primary weights.*
8. ***Model Training****:*

for *i = 1 to e* ***do***

a. Divide D_{Tr} into mini-batch size, m.

b. Perform a forward pass to calculate the binary loss function.

c. Perform a backpropagation to update the weights w.

end for

End

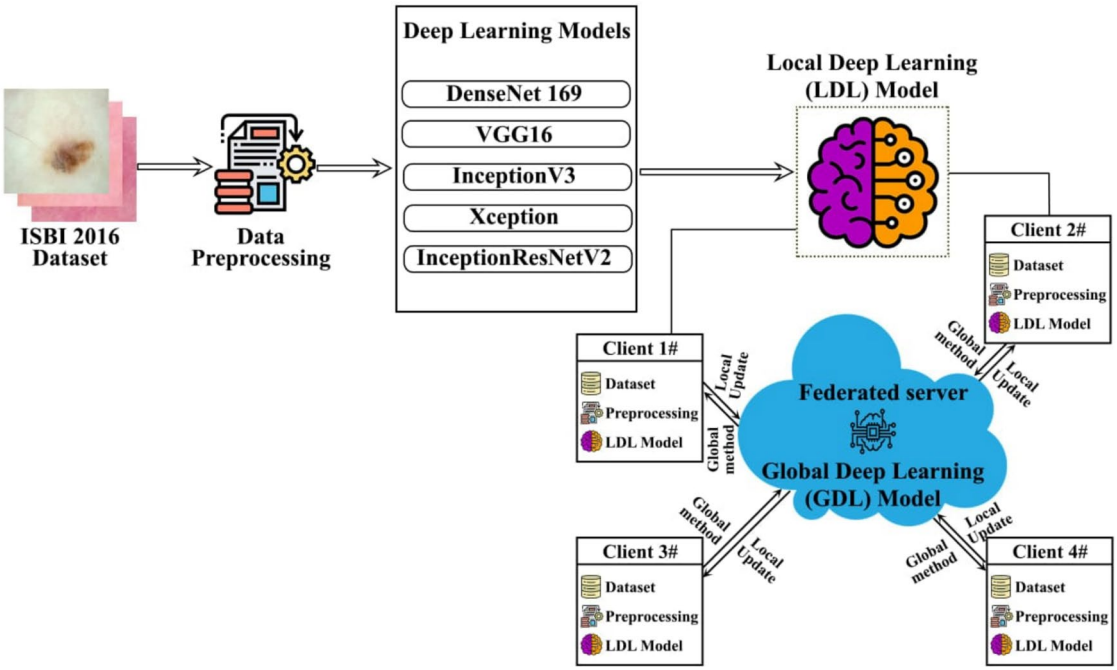

Fig. 1 Structure of the proposed system

Dataset Description and Data Pre-processing

In early 2016, the International Symposium on Biomedical Imaging (ISBI) introduced a challenging dataset for skin lesion analysis aimed at detecting melanoma. This dataset was created using images from the International Skin Imaging Collaboration (ISIC), which has one of the largest collections of skin images taken with dermoscopy. For this study, five DL classifiers were trained and evaluated using the ISBI2016 dataset [28]. This dataset is a part of the larger ISIC collection and contains 1279 skin lesion images. Out of these, 900 images are used for training (173 melanoma and 727 benign), and 379 images are used for testing (75 melanoma and 304 benign). The images have varying resolutions, ranging from 1022×767 to 4288×2848 pixels. Each image is labeled as either melanoma or benign, with 1 representing melanoma and 0 representing benign. Figure 2 provides the sample skin images of the final dataset.

Before inputting the ISBI2016 images into the DL models, we performed several pre-processing steps. As per transfer learning guidelines, we resized all images to a consistent size of 224×224 pixels in RGB format and normalized the pixel values to a range of 0 to 1. Challenges like a small dataset and uneven class distribution can affect the accuracy of the model. Effective data pre-processing is essential for optimal results. Although the dataset didn't have missing values, the training data was imbalanced. To overcome this, we applied the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for the minority class. SMOTE creates new examples by combining randomly chosen samples from the minority class with their k-nearest neighbors [29]. Here, G refers to the number of synthetic examples to create, and k is a user-defined value. After balancing the data, we increased the training set from 900 to 1454 images and distributed these images into validation and training sets (see Table 2). The training set included 1308 images (about 90% of the 1454 total), consisting of 654 melanoma images and 654 benign images. The remaining 146 images (about 10% of the total) formed the validation set, with 73 melanoma images and 73 benign images.

DL algorithms need a large number of training samples since these algorithms have many trainable parameters. To increase the amount of data, we use a technique called data augmentation. This method involves making small changes to the images using various parameters. For parameters like rotation angle, shifts, zoom, or shearing, we typically choose values randomly from specific ranges or distributions. For instance, when rotating an image, we can randomly pick an angle anywhere between 0 and 360 degrees. When zooming, the image is usually resized within a range of [1 − zoom range] to [1 + zoom range]. For shearing, we randomly select a floating-point value from a uniform range between 0 and 1. On the other hand, flipping a sample is a yes/no decision, indicated by either False or True.

After applying data augmentation, we increased the training set from 1308 to 10,464 images, with 5232 being melanoma images and 5232 being benign images.

Similarly, the validation set grew from 146 to 1168 images, consisting of 584 melanoma images and 584 benign images. The augmented parameters with values applied in this paper are listed in Table 3. Table 4 provides the total number of data after applying data augmentation technique. We converted the images into Numpy arrays to speed up training.

The Fine-Tuning Mechanism of Skin Cancer Identification

In this paper, we used a fine-tuned DL network for SC detection. This network comprises a pre-trained CNN classifier, a global average pooling (GAP) layer, and a series of fine-tuning layers with Softmax. These classifier are Densenet169, VGG16, InceptionResNetV2, Xception, and InceptionV3 used for extracting DL features from skin dermoscopy images. These classifiers overcome the challenge of training the fine-tuned model on a limited dataset based on the transfer learning (TL) concept. By applying TL, we can use the knowledge gained from the large ImageNet dataset to enhance our smaller, specialized dataset [30]. We evaluated both unique and lightweight CNN models to see which model works best. The overall architecture of a fine-tuned DL network for SC prediction is illustrated in Fig. 3. The following section explains each pre-trained CNN model, along with the GAP layer and the fine-tuning layers.

DenseNet169

DenseNet169 is a CNN classifier with 169 layers, respectively. This classifier comes in pre-trained versions, which have been evaluated on over 1 M (million) images from the ImageNet database. The pre-trained model can identify 1,000 different categories of objects in images, which means it has learned rich features from a variety of images. The network can process images with a resolution of up to 224 by 224 pixels [31].

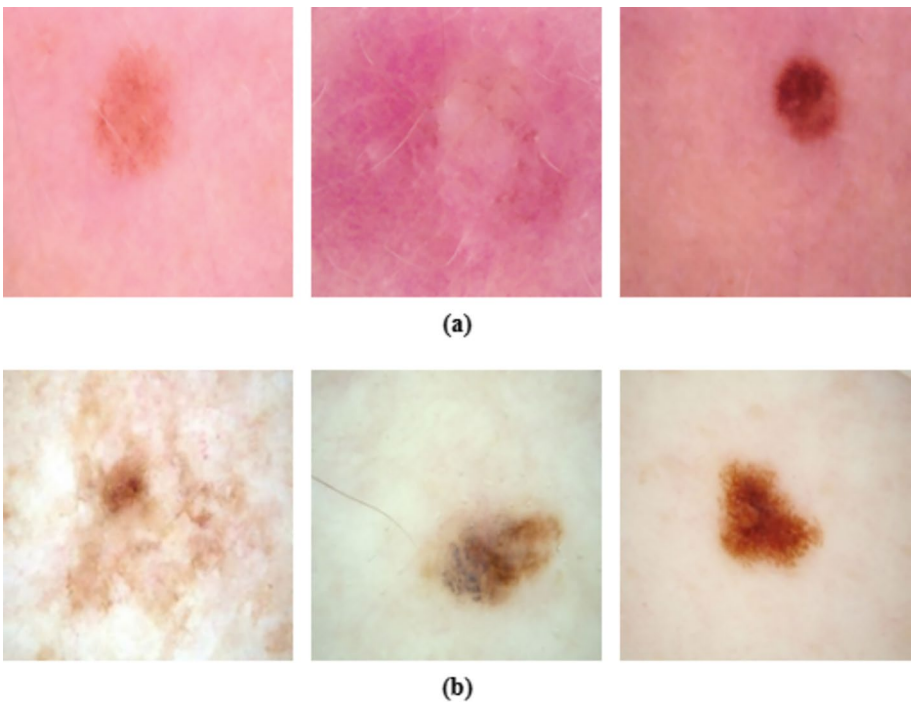

Fig. 2 Some examples of skin lesions images: **a** melanoma and **b** benign

Table 2 Data distribution before data augmentation

Dataset	Class	Train set	Train set after SMOTE	Train set	Validation set	Test set
ISBI2016	Benign	727	727	654	73	304
	Melanoma	173	727	654	73	75
	Final	900	1454	1308	146	379

Table 3 Several data augmentation methods with values

SI No	Methods	Values
1	Vertical flip	True
2	Shearing	0.4
3	Horizontal flip	True
4	Rotation	90
5	Height shift	0.2
6	Zooming	2
7	Width shift	0.2

Table 4 Data distribution after data augmentation

Dataset	Class	Train set	Validation set
ISBI2016	Benign	5232	584
	Melanoma	5232	584
	Final	10,464	1168

Xception

The Xception network was designed to improve upon the tasks previously handled by the Inception network. It represents an extreme version of the Inception model, known as XceptionNet. In XceptionNet, traditional convolutional layers are substituted with depth-wise separable convolution layers. These layers allow the model to separately handle spatial and cross-channel information, which is a key feature of its design. XceptionNet eventually replaced Inception's core architecture. This classifier has 36 convolutional layers, organized into 14 blocks. Even after removing the first and last layers, the remaining layers are still linked together. To classify an image, XceptionNet first processes the image to calculate probabilities across different input channels. It then uses 11 depth-wise convolutions to retrieve characteristics from the image. Finally, this method helps user to replace the three-dimensional maps for showing the relationships between features [32].

InceptionV3

In 2015, researchers at Google introduced InceptionV3. This model is designed to classify and recognize images. When it was released, InceptionV3 was considered the most advanced model based on the outcomes of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). InceptionV3 uses an Inception module, which combines convolution filters of different sizes. This approach enables the model to collect data at various levels, improving its executing capability. As the model becomes deeper, the stacked Inception blocks allow it to learn more intricate patterns. During training, InceptionV3 enhances the flow of gradients by combining convolutional layers, fully connected layers, pooling layers, and additional layers. This model can classify, identify, and segment images, making it highly effective for deep learning tasks. Its balance of performance and efficiency makes it a popular choice in deep learning research and applications [33].

InceptionResNetV2

The InceptionResNetV2 model was evaluated on more than 1 M (million) images from the ImageNet dataset. This deep architecture, which has 164 layers, is capable of classifying images into 1000 different categories. As a result, it has learned to recognize and capture features from a wide range of images. It works with images that are 299 by 299 pixels in size [34].

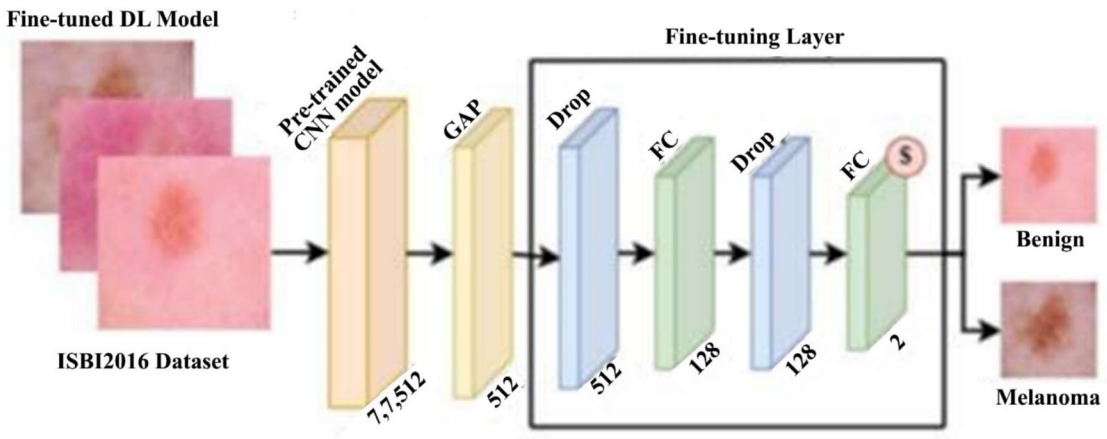

Fig. 3 The architecture of the fine-tuned DL mechanism. GAP means Global Average Pooling and Drop stands for Dropout

VGG16 VGG16 is a DL model developed by the Visual Geometry Group at Oxford University. It consists of 16 deep layers, with thirteen convolutional layers and three fully connected layers. VGG16 is commonly utilized for tasks such as classifying images, detecting objects, and segmenting images. When applying TL with VGG16, the model can use its previously learned knowledge from other tasks to help create new models for different applications [35].

Fine-Tuning Mechanism

In this mechanism, a Global Average Pooling2D layer is added after each pre-trained CNN classifier. This layer averages the features from the input images to create a vector. The vector then passes through several other layers, such as dense layers, batch-normalization layers, and dropout layers, to enhance the efficiency of the proposed approach. Finally, a SoftMax activation function is used to predict skin cancer. A more detailed explanation of each layer is provided below.

In DL models, overfitting happens when the model becomes too focused on the training data, which makes it struggle to perform well on new and unseen data. To address overfitting, we use two dropout [36] regularization layers in our model. These layers discard 20 percent and 40 percent of the neurons during training, helping to enhance the classifier's performance. This method also reduces the training time significantly.

However, two batch-normalization (BN) [37] layers are used in this paper to normalize the extracted features. These layers help adjust and standardize the skin images, improving the model's stability and reliability.

The dense layer called the fully connected (FC) layer, connects all the neurons between the previous and current layers. Its main role is to execute the input data and generate the final output. In this approach, two FC layers are used to complete this execution. The first layer uses the ReLU activation function [38], while the second layer utilizes SoftMax. The last layer makes the prediction and gives the skin cancer diagnosis. SoftMax helps identify the key features to determine whether the image is benign or melanoma, producing an output between 0 and 1, and activating the neurons based on this. This can be given by the mathematical formula 1.

$$\mathrm{SoftMax(x)_q} = \frac{\exp\left(x_q\right)}{\sum_{v=1}^{u} \exp\left(x_q\right)} \tag{1}$$

Table 5 presents the outcomes of merging several CNN layers with fully connected layers. It was generated while developing the proposed model for binary classification. Therefore, the final fully connected layer contains two neurons.

Skin Cancer Identification with Federated Learning Mechanism

In this section, we set up FL environment using the Flower FL framework. We chose five DL models for the image classification task: Densenet169, VGG16, InceptionV3, Xception, and InceptionResNetV2. These models are used to classify the images. Table 6 shows the summary of hyperparameters that are used to train each DL algorithms in FL environment. The central server first creates the global model and waits for input from clients. Clients connect to the server, download the global model, and then train it on their experimental data. Rather than sharing experimental data, the clients send updates to the global model back to the server. Once all updates are received, the server combines them using the FedAvg method, as shown in Eq. 2.

$$\varphi_{R+1}^{v} = \frac{1}{e_l} \sum_{l=1}^{e_l} x_l * \varphi_R^l \tag{2}$$

In this equation, φ_{R+1}^{v} shows the update of the global model at that moment (R + 1), c_k is the number of participants involved in the averaging process., x_l is the importance is given to each client during the averaging process, and φ_R^l refers to the model parameters that are stored and updated on the local device l at time R.

Model Explainability with Explainable AI Mechanism

XAI mechanism are improving quickly, especially in fields like medical image analysis, where making quick and accurate decisions is very important [39–41]. We explain the prediction results from the DCNN algorithm to medical experts to improve their understanding and interpretation. This helps them make accurate and timely diagnoses of skin cancer and other conditions [42, 43]. This experiment uses LIME, one of the most widely known XAI algorithms.

Local Interpretable Model-Agnostic Explanations (LIME)

To clearly interpret the root visualization of a sample $y \in \mathbb{R}^l$ using LIME, a binary feature map $y \in \{1, 0\}^l$ was applied to indicate whether a constant region of super-pixels was active (1) or inactive (0). In the proposed system, $n \in N$, taking values in $\{1, 0\}^l$, was used to represent the DL features. This n indicates the active or inactive of these explainable features. However, it was found that each feature in n provided limited explanation. Therefore, a complexity measure

Table 5 Layer-wise summary of the proposed architecture

Layer name	Output	Parameters
Input	224×224×3	0
Vgg16 (base network)	7×7×512	14,714,688
GlobalAveragePooling	512	0
Dropout	512	0
Dense	128	65,664
Dropout	128	0
Dense (softmax)	2	258

Final parameters: 14,780,610

Non-trainable parameters: 7,635,264

Trainable parameters: 7,145,346

Table 6 Configured hyperparameters for training DL models within a FL setup

Hyperparameter name	Selected value
Loss	Sparse Categorical Crossentropy
Training Epochs	10
Learning rate	0.0001
Optimization algorithm	Adam
Batch Size	32

(n) was introduced to assess the overall interpretability of the explanation. The LIME explanation is formalized in Eq. 3.

$$\delta(y) = \arg^{\max}_{n \in N} \delta(f, g, \varphi_y) + \Omega(n) \tag{3}$$

In this formulation, $f(y)$: $\mathbb{R}^l \rightarrow \mathbb{R}$ represents the probability for a class, and $\varphi_y(x)$ indicates the proximity between the instance x and the instance y. The fidelity function, $\delta(f, g, \varphi_y)$, measures how much n differs from f in the area defined by φ_y. To improve the feature explanation, it should be minimized the fidelity error.

Experiments and Results Analysis

Experiment Environment

This part explains the hardware and software setup for the DL network used to predict pneumonia. In this experiment, Keras python library was used to connect the DCNN with the Python programming language. The experimental resources are tabulated in Table 7.

Assessment measure

The performance evaluation metrics predicting SC from images include accuracy, ROC curve, precision, F1-measure and specificity. These evaluation metrics are measured using the values of false negatives (FN), true positives (TP), true negatives (TN), and false positives (FP). The mathematical formula for the above evaluation metrics are given in Eqs. 4–7.

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision(Pre) = \frac{TP}{TP + FP} \tag{5}$$

$$Specificity(Spe) = \frac{TN}{TN + FP} \tag{6}$$

$$F1 - \text{measure } (F1) = 2 * \frac{Pre * Rec}{Pre + Rec} \tag{7}$$

Results Analysis with DL Along with FL

In this approach, we applied five DL models: DenseNet169, InceptionV3, VGG16, Xception, and InceptionResNetV2 to perform the classification task. Table 8 displays the classification outcomes of these models on the final dataset without a FL environment. The table shows that DenseNet169 achieved the maximum accuracy at 83.3%, while VGG16 achieved the minimum accuracy at 78%. However, VGG16 also achieved the best precision at 97.7% and specificity at 98.7% among all the models. Despite these strengths, VGG16 recorded the lowest F1-measure at 72.3%.

To compare the results, we also tested these models in a FL environment. We used four clients, each with the same dataset, over three rounds. Table 9 shows the classification results of these DL models on the final dataset in the FL setting. The results indicate that the global model attained the maximum accuracy of 92.08% and the best F1-measure of 76.92% using VGG16 within the secure FL framework. Besides, the minimum accuracy was 88.13% with InceptionV3, while Xception obtained the lowest F1-measure at 61.26%. Nonetheless, Xception also recorded the highest precision at 94.44% and the best specificity at 99.34% among all the models.

Figure 4 demonstrates the performance indicators of the FL-VGG16 model. Figure 4a demonstrates the obtained F1-measure, specificity, and precision values for each class. The chart compares the performance of F1-measure, precision, and specificity for detecting melanoma and benign skin lesions. For melanoma, precision is high, close to 90%,

Table 7 Environmental parameters used in the experiment

Resource name	Specification
GPU	Tesla M100
CPU	Intel Core i5-7700 M @ 8100 MHz
Platform	Google Colab
RAM	64 GB

but the F1-measure and specificity are lower, around 65% and 80%, respectively. This suggests that the model is good at identifying true positive melanoma cases but less effective at minimizing false positives and achieving balanced performance. For benign lesions, all three metrics—precision, specificity, and F1-measure—are much higher, approaching or exceeding 90%, indicating the model's stronger and more consistent accuracy in detecting benign cases. Overall, the model performs better for benign lesions than for melanoma. Figure 4b demonstrates the confusion matrix for the FL-VGG16 model. In Fig. 4b, there are 75 melanoma images and 304 benign images. Out of these, 299 images were predicted correctly, while 5 were incorrectly identified as melanoma. Among the melanoma images, 25 were mistakenly labeled as benign, and 50 were correctly recognized as melanoma. Figure 4c shows the accuracy vs loss curve. Figure 4c describes the training performance of a DL model used for detecting SC through benign and melanoma images. The accuracy improves over time, reaching nearly 98% by the final epoch, indicating that the model is increasingly successful at distinguishing between benign and melanoma cases. Meanwhile, the loss, which measures prediction errors, decreases significantly, showing that the model makes fewer mistakes as training progresses. This pattern suggests that the model is learning effectively and could be well-suited for SC prediction utilizing FL.

Experimental Result Analysis with XAI

The XAI algorithm assists an expert in making decisions that are described in Fig. 5. The LIME algorithm creates altered versions of the skin dermoscopy images by masking specific areas. Each altered image is then input into the CNN to get prediction probabilities. LIME constructs a simple model based on the predictions of these altered samples, highlighting the importance of pixels near the original image. The coefficients of this simple model show how diverse parts of the sample contribute to the VGG16 model's prediction. Therefore, the LIME algorithm offers clear insights into which areas of the image were most important in influencing the VGG16 model's decision.

Figure 6 shows the results generated by the LIME XAI algorithm to explain the predicted outcomes for skin images from the proposed model. In Fig. 6, the LIME algorithm divides the predicted benign image into small regions for easier understanding and interpretation, with each region connected to others of the same color. This XAI method highlights the most important areas that are included in the proposed model during the evaluation process. In Fig. 6a, the red area represents the chances of a benign result, and the green area in Fig. 6b shows the chances of melanoma. By highlighting the key regions in the SC detection system, the LIME algorithm helps doctors understand how the model arrives at its decisions.

Table 8 Classification results for all DL models without FL environment

DL model	Accuracy (%)	Precision (%)	Specificity (%)	F1-measure (%)
DenseNet169	83.3	96.3	97.3	80.6
VGG16	78.0	97.7	98.7	72.3
InceptionV3	80.0	95.9	97.3	75.8
Xception	80.7	92.6	94.7	77.5
InceptionResNetV2	81.3	89.8	92.0	79.1

Table 9 Classification results for all DL models under the FL environment

DL-FL model	Accuracy (%)	Precision (%)	Specificity (%)	F1-measure (%)
DenseNet169	89.71	84.62	97.37	69.29
VGG16	92.08	90.91	98.36	76.92
InceptionV3	88.13	78.85	96.38	64.57
Xception	88.65	94.44	99.34	61.26
InceptionResNetV2	89.18	88.64	98.35	65.55

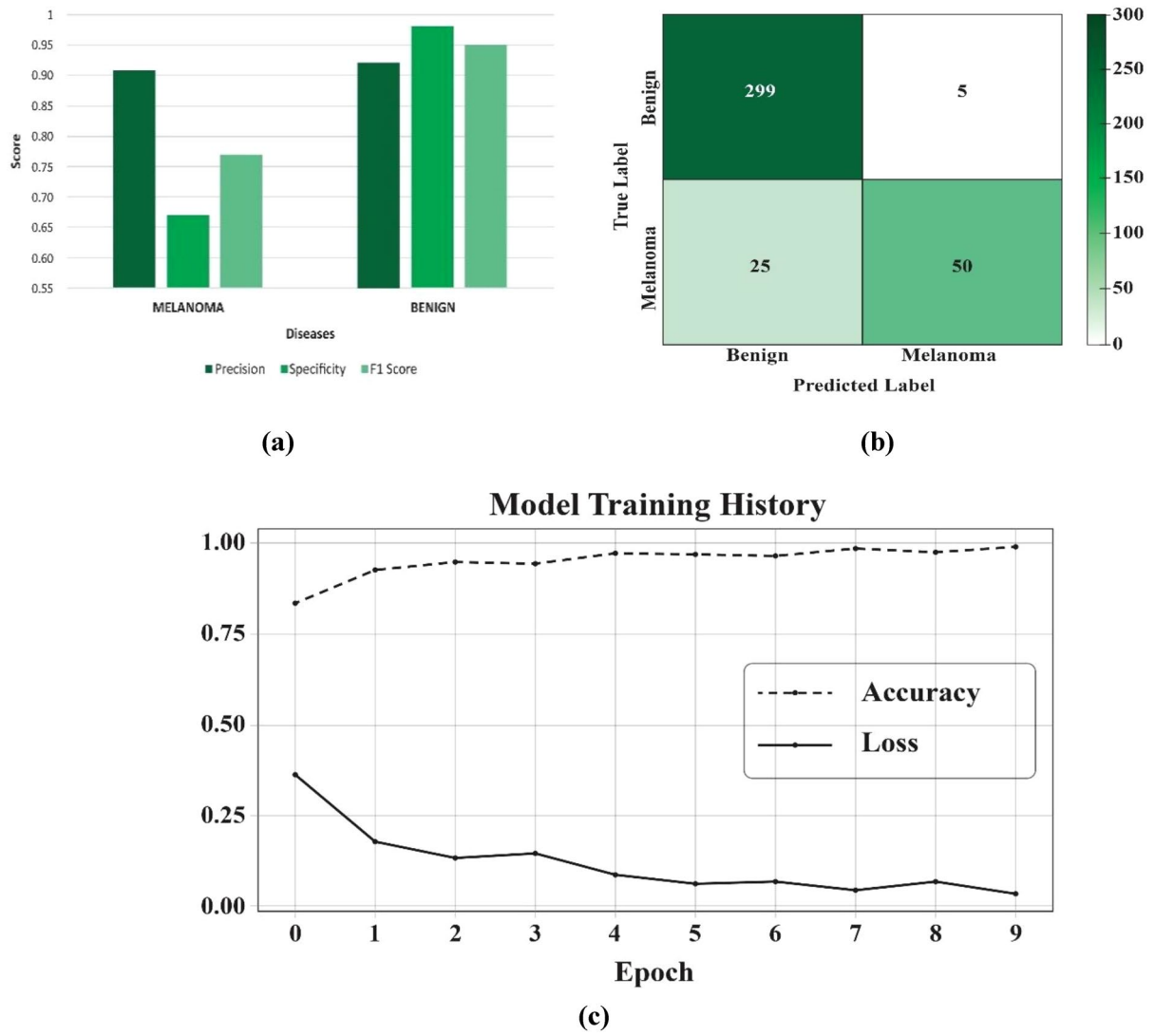

Fig. 4 Performance indicators of the FL-VGG16 model: **a** obtained performance for each category, **b** confusion matrix, and **c** accuracy and loss graph

Discussion

In this research, we present a novel approach for detecting SC by utilizing DL models within a FL framework. By testing our approach on the ISBI2016 dataset, we achieved impressive classification accuracy, showing that FL can improve privacy and accuracy in medical diagnosis systems. The use of FL tackles important issues related to data privacy and access, demonstrating a scalable model for collaborative, decentralized ML without risking patient data.

Our results show a promising path forward, with accuracy rates that are competitive with leading methods in SC classification, as seen in Table 10. For example, Keerthana et al. [22] attained an accuracy of 88.02% on the ISBI2016 dataset using a Hybrid CNN method with an SVM predictor. Although their method had a higher accuracy, our FL approach adds an important layer of data privacy and security, addressing key concerns in handling medical data. In another case, Gajera et al. [24] reached a lower accuracy of 80.47% using DenseNet121 and MLP models on the same dataset. In contrast, our FL-VGG16 model with four clients achieved nearly 92.08% accuracy, demonstrating the effectiveness of FL in utilizing data from different clients to enhance model performance.

The uniqueness of our approach is that it maintains high accuracy while protecting data privacy, which is an important factor often missed in centralized ML systems. Furthermore, FL has proven to be adaptable and scalable through multiple experiments, showing that it can be used effectively across diverse institutions and with various data sizes. This is a major advantage over traditional DL models, which need to gather data in one central location.

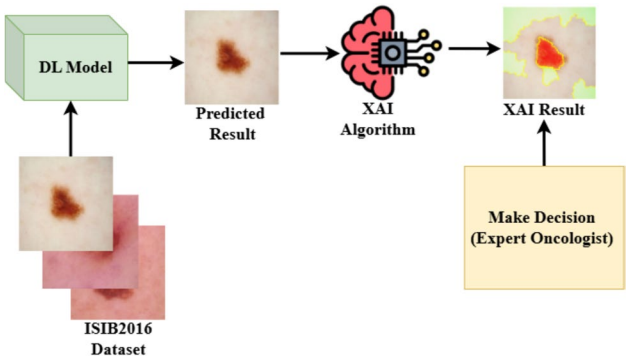

Fig. 5 An example shows how the XAI algorithm helps an expert to make a decision

Despite these encouraging results, our study faces limitations related to FL and the specific architecture used. First, the differences in data from various institutions can be a challenge, possibly causing the model to be biased or perform poorly on certain types of skin lesions that are not well represented in the training datasets. This is a common problem for ML models that depend on diverse data sources, highlighting the need for better data collection strategies that are more inclusive and comprehensive.

Second, the computational and communication demands of FL, particularly when many clients with different data sizes are involved, can impact efficiency. It is important to optimize the processes for training the model and combining updates to address these challenges, ensuring that the model remains scalable and practical for real-world use.

Third, although FL improves data privacy, it doesn't completely remove all privacy and security issues. There are still potential weaknesses that could be targeted through model inversion attacks or by inferring information from combined updates. This highlights the need for continued research into stronger privacy-protecting methods.

Lastly, incorporating our FL model into clinical workflows presents a challenge. For it to be effectively used, it needs to be compatible with existing technology and also gain the trust and acceptance of medical professionals. Additional research on user-friendly interfaces and testing in clinical environments is important to connect technical advancements with real-world use.

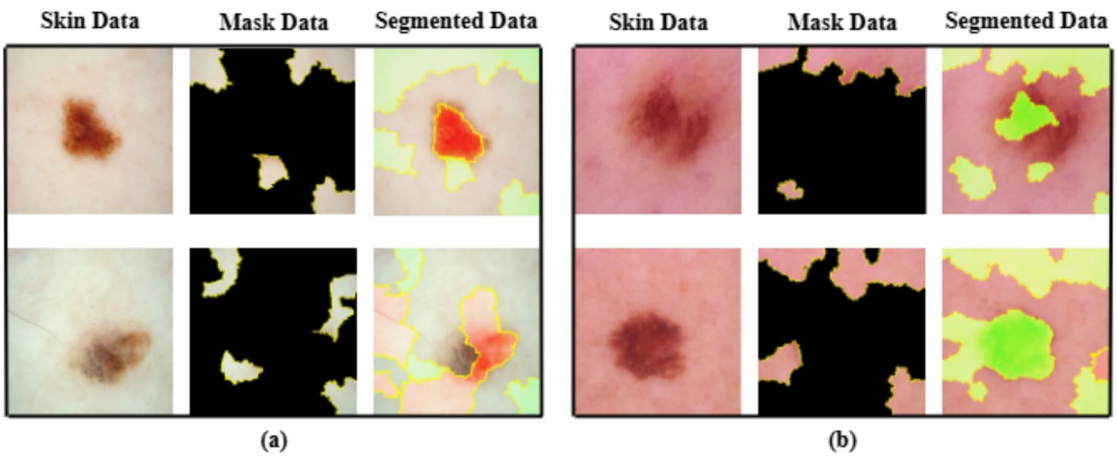

Fig. 6 XAI result analysis by LIME: **a** benign skin image and **b** melanoma skin image

Table 10 Results analysis of our work with prior works using ISBI2016 dataset

Study	Dataset	Model	Performance
Yu et al. [12]	ISBI2016	Deep residual model (DRM)	Accuracy = 85.50%, Precision = 62.40%, Recall = 54.70%
Ali et al. [13]	ISBI2016	DenseNet121 + MLP	Accuracy = 80.47%
Lopez et al. [14]	ISBI2016	VGG16	Accuracy = 81.33%, Recall = 78.66%, Precision = 79.74%
Keerthana et al. [22]	ISBI2016	Hybrid model + SVM	Accuracy = 88.02%
Gajera et al. [24]	ISBI2016	DenseNet121 + MLP	Accuracy = 80.47%
Wang et al. [25]	ISBI2016	DL model + FV	Accuracy = 86.54%
Proposed	ISBI2016	VGG16-FL	Accuracy = 92.08%, Precision = 90.91%, Specificity = 98.36%, F1-measure = 76.92%

Conclusion

The rising number of SC cases emphasizes the need for better early identification methods to improve treatment results. This study uses AI, especially DL models, to tackle the challenge of identifying SC. Our research reflects how DL models can improve diagnostic accuracy and introduce FL as an important tool to address concerns about data privacy and access to medical research data. We assess five DL models Densenet169, VGG16, InceptionV3, Xception, and InceptionResNetV2 on the ISBI2016 dataset to detect SC. Among these models, Densenet169 achieved the highest accuracy of 83.3%, while VGG16 obtained the lowest accuracy of 78%. However, FL experiments showed promising results, with the VGG16 model achieving a notable accuracy of 92.08% when applied across multiple clients. This result highlights the potential of advanced AI techniques in privacy-preserving and distributed data environments.

In future direction, we plan to enhance our framework's accuracy by testing it on larger and more varied datasets. Additionally, integrating these technologies into clinical practice and enhancing the transparency of AI-generated insights will be crucial. Such efforts will not only improve diagnostic tools but also support the development of AI solutions that are innovative and ethically responsible, ultimately leading to better patient outcomes. This study underscores the importance of AI in medical diagnostics and lays the groundwork for further research that could significantly influence the field of medical imaging and AI applications.

Funding There is no funding for this research work.

Data Availability This work used publicly available dataset named ISIC. Dataset link: https://challenge.isic-archive.com/data/

Declarations

Conflict of Interest The authors declare that there are no conflicts of interest.

Human and Animal Rights Not applicable.

Informed Consent I would like to extend my heartfelt gratitude to my co-authors for their insightful suggestions and meaningful contributions during the formulation and execution of this study.

References

1. J.M.H. Abarca, A.J.P. Chávez, Malignant Nail Melanoma in a case report. J. Pharm. Negative Results **14**(2), 67–72 (2023)
2. Z.W. Yu, M. Zheng, H.Y. Fan, X.H. Liang, Y.L. Tang, Ultraviolet (UV) radiation: a double-edged sword in cancer development and therapy. Mol. Biomed. **5**(1), 1–24 (2024)
3. E.R. Parker, The influence of climate change on skin cancer incidence—A review of the evidence. Int. J. Women's Dermatol. **7**(1), 17–27 (2021)
4. S. Vaccarella, D. Georges, F. Bray, O. Ginsburg, H. Charvat, P. Martikainen, H. Brønnum-Hansen, P. Deboosere, M. Bopp, M. Leinsalu, B. Artnik, Socioeconomic inequalities in cancer mortality between and within countries in Europe: a population-based study. Lancet Regional Health–Europe (2023)
5. K. Venugopal, D. Youlden, L.T. Marvelde, R. Meng, J. Aitken, S. Evans, I. Kostadinov, R. Nolan, H. Thomas, K. D'Onise, Twenty years of melanoma in Victoria, Queensland, and South Australia (1997–2016). Cancer Epidemiol. **83**, 102321 (2023)
6. S. Bibi, M.A. Khan, J.H. Shah, R. Damaševičius, A. Alasiry, M. Marzougui, M. Alhaisoni, A. Masood, MSRNet: Multiclass skin lesion recognition using additional residual block based fine-tuned deep models information fusion and best feature selection. Diagnostics **13**(19), 3063 (2023)
7. E.-G. Dobre, M. Surcel, C. Constantin, M.A. Ilie, A. Caruntu, C. Caruntu, M. Neagu, Skin cancer pathobiology at a glance: a focus on imaging techniques and their potential for improved diagnosis and surveillance in clinical cohorts. Int. J. Mol. Sci. **24**(2), 1079 (2023)
8. R.A. Mehr, A. Ameri, Skin cancer detection based on deep learning. J. Biomed. Phys. Eng. **12**(6), 559 (2022)
9. E.H. Houssein, M.M. Emam, A.A. Ali, An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm. Neural Comput. Appl. **34**(20), 18015–18033 (2022)
10. S.M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, M.K. Khan, Medical image analysis using convolutional neural networks: a review. J. Med. Syst. **42**, 1–13 (2018)
11. A. Moglia, K. Georgiou, B. Marinov, E. Georgiou, R.N. Berchiolli, R.M. Satava, A. Cuschieri, 5G in healthcare: from COVID-19 to future challenges. IEEE J. Biomed. Health Inform. **26**(8), 4187–4196 (2022)
12. L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks in *IEEE Transactions on Medical Imaging* (vol. PP, ed, 2016) pp. 1–1.
13. A. A. Ali and H. Al-Marzouqi, "Melanoma detection using regular convolutional neural networks," in *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, (Ras Al Khaimah: IEEE, Nov. 2017), pp. 1–5
14. A.R., Lopez, X. Giro-i-Nieto, J. Burdick, O. Marques, Skin lesion classification from dermoscopic images using deep learning techniques. In *2017 13th IASTED international conference on biomedical engineering (BioMed)*. (IEEE, 2017). pp. 49–54
15. A. Demir, F. Yilmaz, O. Kose, Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3. in *Proceedings of the 2019 Medical Technologies Congress (TIPTEKNO)*, (Izmir, Turkey, 3–5 October 2019). pp. 1–4
16. M.A. Khan, K. Muhammad, M. Sharif, T. Akram, S. Kadry, Intelligent fusion-assisted skin lesion localization and classification for smart healthcare. Neural Comput. Appl. **36**(1), 37–52 (2024)
17. S. Jain, U. Singhania, B. Tripathy, E.A. Nasr, M.K. Aboudaif, A.K. Kamrani, Deep learning-based transfer learning for classification of skin cancer. Sensors **21**(23), 8142 (2021)
18. M. Fraiwan, E. Faouri, On the automatic detection and classification of skin cancer using deep transfer learning. Sensors **22**(13), 4963 (2022)
19. K. Aljohani, T. Turki, Automatic classification of melanoma skin cancer with deep convolutional neural networks. Ai **3**(2), 512–525 (2022)
20. W. Gouda, N.U. Sama, G. Al-Waakid, M. Humayun, N.Z. Jhanjhi, Detection of skin cancer based on skin lesion images using deep learning. In *Healthcare* (Vol. 10, No. 7, p. 1183, 2022). MDPI.

21. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, (Munich, Germany, 8–14 September 2018).
22. D. Keerthana, V. Venugopal, M.K. Nath, M. Mishra, Hybrid convolutional neural networks with SVM classifier for classification of skin cancer. Biomed. Eng. Adv. **5**, 100069 (2023)
23. A. Bassel, A.B. Abdulkareem, Z.A.A. Alyasseri, N.S. Sani, H.J. Mohammed, Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. Diagnostics **12**(10), 2472 (2022)
24. H.K. Gajera, D.R. Nayak, M.A. Zaveri, A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features. Biomed. Signal Process. Control **79**, 104186 (2023)
25. A. Adegun, S. Viriri, Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif. Intell. Rev. **54**(2), 811–841 (2021)
26. M. Safinejad, A. Rigi, M. Zeraati et al., Lanthanum-based metal organic framework (La-MOF) use of 3,4-dihydroxycinnamic acid as drug delivery system linkers in human breast cancer therapy. BMC Chem. **16**, 93 (2022)
27. M.R. Hajinezhad, M. Roostaee, Z. Nikfarjam et al., Exploring the potential of silymarin-loaded nanovesicles as an effective drug delivery system for cancer therapy: in vivo, in vitro, and in silico experiments. Naunyn-Schmiedeberg's Arch. Pharmacol. **397**, 7017–7036 (2024)
28. Data available link: https://challenge.isic-archive.com/data/
29. S.A. Alex, J.J.V. Nayahi, S. Kaddoura, Deep convolutional neural networks with genetic algorithm-based synthetic minority oversampling technique for improved imbalanced data classification. Appl. Soft Comput. **156**, 111491 (2024)
30. 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)
31. M.K. Awang, J. Rashid, G. Ali, M. Hamid, S.F. Mahmoud, D.I. Saleh, H.I. Ahmad, Classification of Alzheimer disease using DenseNet-201 based on deep transfer learning technique. PLoS ONE **19**(9), e0304995 (2024)
32. X. Lu, Y.A. Firoozeh Abolhasani Zadeh, Deep learning-based classification for melanoma detection using XceptionNet. J. Healthc. Eng.202(1), 2196096 (2022)
33. K. Dwivedi, A. Gupta, A. Rajpal, N. Kumar, Deep learning-based NSCLC classification from whole-slide images: leveraging expectation-maximization and inceptionv3. Procedia Comput. Sci. **235**, 2422–2433 (2024)
34. M. Humayun, M.I. Khalil, S.N. Almuayqil, N.Z. Jhanjhi, Framework for detecting breast cancer risk presence using deep learning. Electronics **12**(2), 403 (2023)
35. T. Fatima, H. Soliman, Application of VGG16 transfer learning for breast cancer detection. Information **16**(3), 227 (2025)
36. H.I. Lim, A study on dropout techniques to reduce overfitting in deep neural networks. in *Advanced Multimedia and Ubiquitous Engineering: MUE-FutureTech 2020* (Springer Singapore, 2021). pp. 133–139
37. M. Awais, M.T.B. Iqbal, S.H. Bae, Revisiting internal covariate shift for batch normalization. IEEE Trans. Neural Netw. Learn. Syst. **32**(11), 5082–5092 (2020)
38. G.E. Dahl, T.N. Sainath, G.E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing* (IEEE, 2013). pp. 8609–8613
39. B.H.M. Van Der Velden, H.J. Kuijf, K.G.A. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. **79**, 102470 (2022)
40. S. Biswas, R. Mostafiz, M.S. Uddin, B.K. Paul, XAI-FusionNet: Diabetic foot ulcer detection based on multi-scale feature fusion with explainable artificial intelligence. Heliyon **10**(10), e31228 (2024)
41. F. Khanom, S. Biswas, M.S. Uddin, R. Mostafiz, XEMLPD: an explainable ensemble machine learning approach for Parkinson disease diagnosis with optimized features. Int. J. Speech Technol. **27**(4), 1055–1083 (2024)
42. S. Biswas, R. Mostafiz, B.K. Paul, K.M.M. Uddin, M.A. Hadi, F. Khanom, DFU_XAI: a deep learning-based approach to diabetic foot ulcer detection using feature explainability. Biomed. Mater. Devices **2**(2), 1225–1245 (2024)
43. 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)