IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 5, May, 2025, pp. 1349-1362
Journal homepage: http://iieta.org/journals/isi

Activation Heatmap-Guided FT-MultiCNN: Advancing Skin Cancer Classification Through Transfer Learning

Nivedita Shimbre*, Ram Kumar Solanki

School of Computer Science & Engineering, Sandip University, Nashik 422213, India

Corresponding Author Email: nivedita.shimbre@gmail.com

https://doi.org/10.18280/isi.300520

Received: 2 March 2025
Revised: 12 May 2025
Accepted: 22 May 2025
Available online: 31 May 2025

Keywords:
skin cancer detection, multi-modal, deep learning, transfer learning, class activation heat-map, medical imaging

ABSTRACT

Cancer of the skin is one of the most prevalent and malignant diseases known to mankind. The most effective treatment for skin cancer is early and accurate detection. Skin lesions which are often complex in structure and diverse in appearance make it not easy for traditional machine learning methods to precisely obtain features and recognize them. We propose a Multi-modal CNN for Skin Cancer detection and classification to address these difficulties. Our approach used a Fine-tune Custom CNN multi-model (FT-MultiCNN) to handle and fuse image data and metadata. The processed images were then employed for input to the FT-MultiCNN model beside patient's metadata as well as the extracted Class Activation Heat-map (CAM) features. To identify skin lesions, a CNN was used using parallel processing architecture and multimodal fusion. This model was trained and tested on a public ISIC dataset, and its performance was assessed using cross-validation and compared to other leading approaches. Our model beat existing machine learning and transfer learning models in accuracy and recall, with 0.96 ± 0.25 accuracy and 0.94 ±.34 recall, indicating robust performance. The experiments show exceptional classification capabilities, producing cutting-edge outcomes in identifying diverse skin cancer forms. This work provides a promising line for automated non-invasive screening of skin cancer, and paves the way for the promise of multimodal deep learning in dermatology.

1. INTRODUCTION

Skin cancer is the most frequent and lethal dermatological disease and results in millions of new cases being diagnosed worldwide annually [1]. Treatment is successful only if the disease is detected early and accurately, because a late diagnosis could lead to serious effects and higher death ratios. Dermatologists' skills are the basis of traditional diagnostic methods but hand examination may be arduous, subjective, and subjective to human error. The deep learning approach, which integrates large scale medical imaging data into training to enhance the classification performance, has been shown to be powerful in protocol automation for skin cancer detection in recent years, since they have the capacity to extract spatial features, as a kind of artificial intelligence technique, the CNNs have been widely applied in the field of medical imaging analysis [2]. However, they often struggle with understanding the bigger picture globally.

A general deep learning model, known as a Multimodal CNN, is created to process multiple types of input data, which concurrently considers various input data and obtaining increased model performance in difficult categorization tasks. A dual input CNN uses patient metadata and dermoscopic images for skin cancer recognition, where the model is allowed to take advantage of additional inputs to have a more accurate classification. In particular, convolutional neural networks (CNNs) can distinguish between benign and malignant lesions due to their learn local patterns such as texture, edges, and color changes through the use of convolutional layers to learn spatial characteristics from images [3].

Transfer learning is well suited to medical image analysis, as it permits models pre-trained on a large annotated image dataset (e.g., ImageNet) to be fine-tuned to a specific medical task, even if only few annotated data are available. The local spatial information of the ground glass opacity was successfully learned using CNNs, which is crucial for the diagnosis of the disease. In this study we propose a multi-modal DL model that leverages transfer learning on CNN in order to work on dermoscopic images and to work with other neural network on patients' meta-data including, age, gender and location of the lesion. To achieve high diagnostic accuracy, fewer false positives and enhanced clinical decision support, the model integrates both image and metadata modalities in this study. Some example images from the ISIC dataset are shown in Figure 1.

1.1 Motivation

Skin cancer is among the most frequent malignancies in all populations, and therefore represents a major public health concern. Globally there are approximately 2-3 million nonmelanoma skin cancers and 132,000 melanoma skin cancers diagnosed annually [4] according to the World Health

Organization (WHO). In India SC comprises 1-2% of all CAs diagnosed, and is increasing as a result of continuous exposure to ultra violet radiation, changing lifestyle, and low knowledge about sun protection. The literature describes a constant increase in the number of reported skin carcinomas in India [5], thousands of new instances are diagnosed annually. Although skin cancer is not as common as in the West, it is daunting in India because of its delayed presentation and restricted access to specialized care especially in rural India. It reinforces the necessary of more attention, earlier diagnosis, and effective interventions for this emerging health concern.

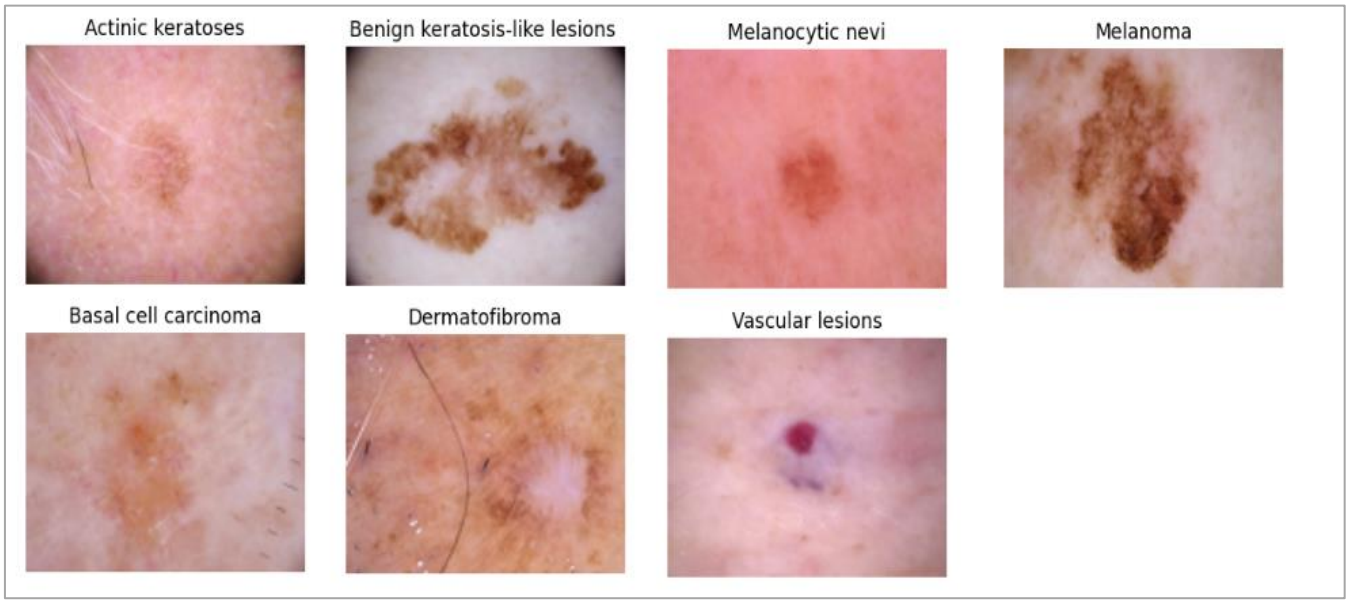


Figure 1. ISIC dataset sample

1.2 Role of deep learning and transfer learning in skin cancer diagnosis

CNN are capable to capture spatial hierarchies in images for patterns; and have revolutionized image processing. In a related context of skin cancer diagnostic the CNNs can be trained to recognize and classify skin lesions based on attributes like texture, color, shape, etc [6]. This allows to make the diagnosis of dermoscopic lesions with high reliability, including identification of benign and malignant lesions. CNNs are the ideal deep learning architecture for end-to-end learning since they can be trained end-to-end and don't need manually created features to produce a diagnosis from unprocessed visual data. It is a more efficient and effective process. Higher Performance CNNs have been proven superior to conventional machine learning methods in the detection and categorization of skin cancer from images in a number of studies. Nevertheless, they demand a huge labeled dataset for training, which is the primary challenge in adoption.

Transfer learning is a process to reuse a pre-trained model, typically trained on a large dataset for a different, but similar, task to refine it for a new, often smaller task [7]. This permits the model to make use of learned features and transfer them to new tasks with minimum retraining requirements. Advantages for Diagnosis of Skin Cancer With scarce labelled datasets in skin cancer, pre-training can be useful to obviate the need for collecting huge amounts of data. By transferring learned models of large-scale image datasets (such as ImageNet) to a small skin cancer dataset, the model can be fine-tuned and delivers high performance without the need for large quantities of data. Obtained Superior Accuracy with Less Data Transfer learning mitigates the overfitting issue for deep learning networks on small or imbalanced datasets, which is among the difficulties in analyzing medical images.

1.3 Multimodal approaches: Combining diverse data for more accurate diagnosis

The use of multiple types of information (like images, medical records, and histopathology information) to provide a holistic understanding of the problem is referred to as multimodal learning [8]. When matching the dermoscopic image with clinical information, including the patient's age, sex, medical history, and location of the lesion, it could increase the accuracy of diagnosing skin cancer. Multi-modal allows the model to process more than one input at a time and making it more robust. For instance, the clinical context can aide in interpreting the images, and lesion attributes not directly perceived from the images can be incorporated into the diagnostic process.

The combination of global and spatial features to enhance effect of skin cancer categorization is achieved by the proposed multi-scale CNN and transfer learning based model; Fine-tune Custom CNN Multi-model (FT-MultiCNN). This model has potential to address the shortcomings of conventional approaches by using the generated activation heat-maps from a pre-trained CNN instead of the original dermoscopic images. The use of heat maps allows for the activation of significant portions within an image that are pertinent to the classification decision. This can be accomplished by improved visualization of significant patterns, as opposed to concentrating on the entire image, which may contain background data that is not necessary.

The FT-MultiCNN model combines a multimodal approach where patients' information (age, gender, lesion location) was added to the features (CNN, multi-scale images, and bone subtraction images). Moreover, the inclusion of visual (dermoscopic) and nonvisual (patient information) parameters enables a broader overview of the patient condition to be obtained, leading to greater diagnostic accuracy.

1.4 Advantages of activation heat-maps over regular images

Feature Focus: Activation heat-maps serve to visibly emphasize the most relevant portions of an image, so that the model can concentrate on region that convey the most useful information for classification, rather than being distracted by less relevant features.

Reduced Noise: By focusing on relevant attributes and reducing background noise or irrelevant image components, the heat-maps turn out to be capable of assisting the model in making more efficient and accurate predictions.

Improved Interpretability: Heat-maps provide better interpretability of the model decision by making clear which areas of the image contribute to the model's prediction. This can support the development of trust in the model output and makes it easier to diagnose the model, based on the areas we know what we should be outputting.

Enhanced Transfer Learning: The use of pre-trained network's activation heatmaps as an input feature however enables the model to take advantage of knowledge learned from large datasets, even if the new dataset is smaller or has reduced variability.

By leveraging these heatmaps alongside patient-specific data, the FT-MultiCNN model offers a more accurate, interpretable, and holistic approach to skin cancer classification. The proposed contributions are mention below;

1.5 Key objective

- Development of a Multimodal Deep Learning Model: This work aims to create a deep learning model that integrates clinical data (e.g., patient demographics, medical history, lesion attributes) with skin cancer images.
- Activation Heatmap Integration for Enhanced Model Interpretability: The use of activation heatmaps as input features, rather than the original images, allows the model to focus on critical areas of the skin lesions, thereby increasing classification accuracy.
- Comparison with Deep Learning and Transfer Learning Models: To validate the effectiveness of the proposed FT-MultiCNN model, it is compared with both traditional deep learning models and transfer learning models.

The structure of the paper is as follows: The literature evaluation of related work is presented in Section 2, with a focus on the methods and advantages and disadvantages of each. Section 3 Related Work provides a models used for comparative analysis in current work. Section 4 Methodology provides detailed explanation of the proposed framework developed in this study named as FT-MultiCNN, including the mathematical modeling and important steps to classify skin cancer images. Results and Parameter Analysis of the Model are described in addition to detailed model parameters in section 5. Last, Section 6 conclude the study and suggests future research directions.

2. LITERATURE REVIEW

Recent advances in deep learning and medical image processing have made skin cancer classification using deep learning in medical imaging increasingly popular and promising. This paper reviews deep learning and machine learning algorithms for skin lesion classification. The studies covered here include techniques across different paradigms such as conventional CNNs, transfer learning-based strategies using pre-trained models, multimodal models that embed clinical metadata for joint learning of dermoscopic images, and most recently transformer-based networks designed to attend over global contextual information. This review will compare the strengths and limitations and outcomes of these approaches to facilitate a comparative understanding of current trends and to direct attention towards successful strategies for enhancing diagnostic performance in the detection of skin cancer.

2.1 Deep learning and transfer learning techniques

Recent advances in artificial intelligence have revolutionized dermatology, notably in skin cancer detection and categorization. Several deep learning methods have been studied to increase diagnostic efficiency and accuracy. One such approach combines discrete wavelet transform (DWT) with convolutional neural networks (CNNs) for enhanced feature extraction [7]. By leveraging both spatial and frequency domain representations, this method enables more precise recognition of subtle lesion characteristics that are crucial for early diagnosis. Given its robustness and performance gains, this fusion technique shows strong potential for clinical adoption.

The recent development of transfer learning has significantly enhanced the performance of skin cancer classification models, particularly when dealing with small or imbalanced datasets. A multimodal framework leveraging both clinical characteristics and dermoscopic images through transfer learning has demonstrated that models using multi-input or multi-source data can more effectively distinguish common skin lesions compared to those relying on a single input or dataset [8]. This approach not only improves performance in data-scarce settings but also enhances generalization across diverse patient phenotypes. The framework has proven effective in managing data imbalance and shows promise for real-world diagnostic applications.

Building on this foundation, the integration of transfer learning with federated learning has been explored to improve melanoma classification accuracy while preserving patient data privacy [9]. By training models across decentralized datasets without centralizing sensitive information, this method addresses the twin challenges of data scarcity and privacy—critical issues in medical AI. Federated learning, when combined with transfer learning, enables the deployment of robust diagnostic models across varied populations and locations, supporting the broader adoption of AI-driven tools in dermatological care.

Extensive exploration of deep learning, particularly CNNs, has been conducted in the context of medical image-based cancer diagnostics [10, 11]. These investigations confirm the strong performance of CNNs in identifying and categorizing skin lesions from dermoscopic images. They also emphasize key challenges such as overfitting, limited data availability, and the critical need for large, well-annotated datasets. Additionally, integrating supplementary information from medical sources, such as clinical notes, has been suggested to improve model generalization and support clinical decision-making.

Emerging research further highlights the promise of CNN-

based solutions for smartphone-acquired skin lesion images, underscoring the potential of mobile platforms to enhance accessibility and promote early detection, especially in underserved or resource-constrained settings [12]. Broader analyses of AI in dermatology have also underscored challenges including model interpretability, data diversity, and regulatory concerns [13]. While the benefits—such as faster and more accurate diagnosis—are evident, these findings stress the necessity of developing robust, transparent models and ensuring responsible clinical deployment.

These studies reflect the growing maturity of deep learning and transfer learning approaches in skin cancer classification.

2.2 Transformer models

Following the advancements brought by deep learning and transfer learning, Transformer-based architectures have recently emerged as a powerful evolution in medical image analysis. These models address key limitations of convolutional neural networks (CNNs), particularly their inability to capture long-range spatial dependencies. Recent studies have demonstrated that Transformers, through their self-attention mechanisms, can introduce spatial inductive biases that limit viable spatial configurations without the need for additional refinement, achieving high accuracy on benchmark skin lesion datasets. This positions Transformers as highly effective for complex dermatological image classification tasks.

In addition, enhanced Transformer architectures have been proposed to overcome CNNs' restricted contextual understanding. By leveraging global attention, these models effectively encode complex lesion patterns and have achieved state-of-the-art results on large-scale skin cancer datasets [14]. Together, these developments highlight the growing potential of Transformer-based models in dermatology, offering high-precision, scalable, and fully automated solutions for skin cancer detection and diagnostic support.

2.3 Hybrid models

While Transformer-based models excel at capturing long-range dependencies and enhancing classification accuracy, earlier methods like CNNs and transfer learning also have their strengths and limitations. CNNs, particularly when combined with wavelet transforms and multimodal fusion, effectively detect local lesion patterns but struggle with global context. Transfer learning helps address limited annotated data, yet faces challenges in generalizing across diverse lesion types and varying image scales.

To overcome these limitations, recent research has shifted toward more holistic and hybrid solutions that integrate deep learning with scale-adaptive and generative strategies. One such approach involves a multi-scale deep learning framework that leverages transfer learning to improve classification performance across varied skin lesion datasets [15]. Its multi-resolution design enhances the model's ability to handle lesions of different sizes and types while requiring fewer annotated samples due to the use of pre-trained models.

In addition, ensemble strategies combining transfer learning with conditional generative adversarial networks (CGANs) have been proposed to address class imbalance and data scarcity [16]. By generating synthetic training samples, these models effectively expand dataset diversity, reduce overfitting, and improve generalization. Group-wise methods further enhance prediction robustness across lesion categories. Similarly, mixed strategies that fine-tune pre-trained CNNs have proven effective in capturing subtle skin lesion features, particularly in resource-constrained settings, achieving high accuracy even with limited annotated data [17].

2.4 Multimodal fusion techniques

Deep learning and transfer learning methods have shown notable improvements in skin lesion classification, especially when leveraging efficient architectures like EfficientNet and pre-trained CNNs [18]. These techniques enable high classification accuracy even with limited labeled data. Transfer learning enhances generalization across lesion types, while lightweight architectures reduce computational demands without compromising performance. However, such methods typically rely solely on visual features from dermoscopic images, which may not fully capture the clinical context—particularly when visual cues are subtle or ambiguous.

To overcome these limitations, recent efforts have focused on multimodal fusion techniques that integrate dermoscopic images with clinical metadata such as age and medical history [19]. This approach enhances decision-making by linking visual patterns with relevant patient information. Further studies have demonstrated that multimodal data fusion not only boosts diagnostic accuracy but also improves robustness, especially in the presence of class imbalance [20, 21]. By incorporating data augmentation, customized loss functions, and enriched input representations, these models achieve more reliable performance in distinguishing benign from malignant lesions. Collectively, these findings highlight the growing significance of multimodal strategies in building accurate, robust, and clinically applicable skin cancer diagnostic systems.

Table 1 displays a thorough comparison of the literature in which the reviewed studies are compared according to different methodologies, dataset, pros and cons and results.

Table 1. Comparative analysis of literature reviews

References	Methodology	Dataset(s) Used	Advantages	Results
[7]	CNN with Discrete Wavelet Transformation (DWT)	HAM10000	Improved precision and effectiveness in skin cancer identification	Sensitivity of 94% and specificity of 91%
[8]	Framework for transfer learning for multimodal analysis of skin lesions	ISIC 2018	Effective integration of multimodal data for enhanced analysis	Demonstrated improved classification performance
[9]	Federated and transfer learning methods	HAM10000 and BCN20000	Preserved data privacy and leveraged knowledge transfer	Achieved high classification accuracy
[10]	Deep learning for medical image-based cancer diagnosis	Various medical imaging	Comprehensive analysis of deep learning applications	Provided insights into model performance
[11]	Deep learning for skin cancer classification	Collected skin image	Systematic review of deep learning methods	Highlighted challenges and opportunities

[12]	CNNs for smartphone image-based diagnosis	Smartphone-acquired skin images	Comparative study on CNN performance	Provided insights into model effectiveness
[13]	Review of AI-based image classification methods for skin cancer, covering CNNs, transfer learning, and hybrid models	ISIC datasets	Comprehensive overview of state-of-the-art techniques	Highlighted opportunities and challenges in applying AI for skin cancer diagnosis
[14]	Improved transformer network for classification	Collected skin lesion dataset	Captured long-range dependencies effectively	Demonstrated superior performance in classification
[15]	Multi-scale deep learning and transfer learning	ISIC 2018	Enhanced detection through multi-scale analysis	Achieved an accuracy of 94.42%
[16]	Ensemble of transfer learning models with GANs	ISIC 2019	Improved prediction through ensemble learning	Demonstrated enhanced performance
[17]	Enhanced transfer learning-based classification	Dermoscopic images from ISIC archive	Improved diagnosis through transfer learning	Achieved high classification accuracy
[18]	Deep neural network using modified EfficientNet	Dermoscopic images from ISIC archive	Improved detection in dermoscopic images	Achieved high performance in skin cancer detection
[19]	Combining clinical data and skin pictures using a multimodal fusion technique	Collected clinical data	Enhanced diagnostic accuracy through data fusion	Achieved an accuracy of 80.42%
[20]	Multimodal evaluation of dermatological data that is not balanced	HAM10000	Addressed data imbalance and improved recognition	Demonstrated effective skin cancer recognition
[21]	Multimodal deep learning approach combining clinical metadata with dermoscopic images	ISIC 2017 Challenge dataset	Improved diagnostic accuracy and robustness through integration of multimodal data	outperforming unimodal counterparts in lesion classification accuracy

3. METHODOLOGY

ISIC, which stands for the International Skin Imaging Collaboration dataset, is a dataset that is frequently utilized and has a substantial amount of documentation. It is typically utilized for the identification and categorization of skin cancer. Dermatologists are equipped with the knowledge and skills necessary to read high-resolution dermoscopic pictures of skin lesions, which can range from benign to specific types of melanomas.

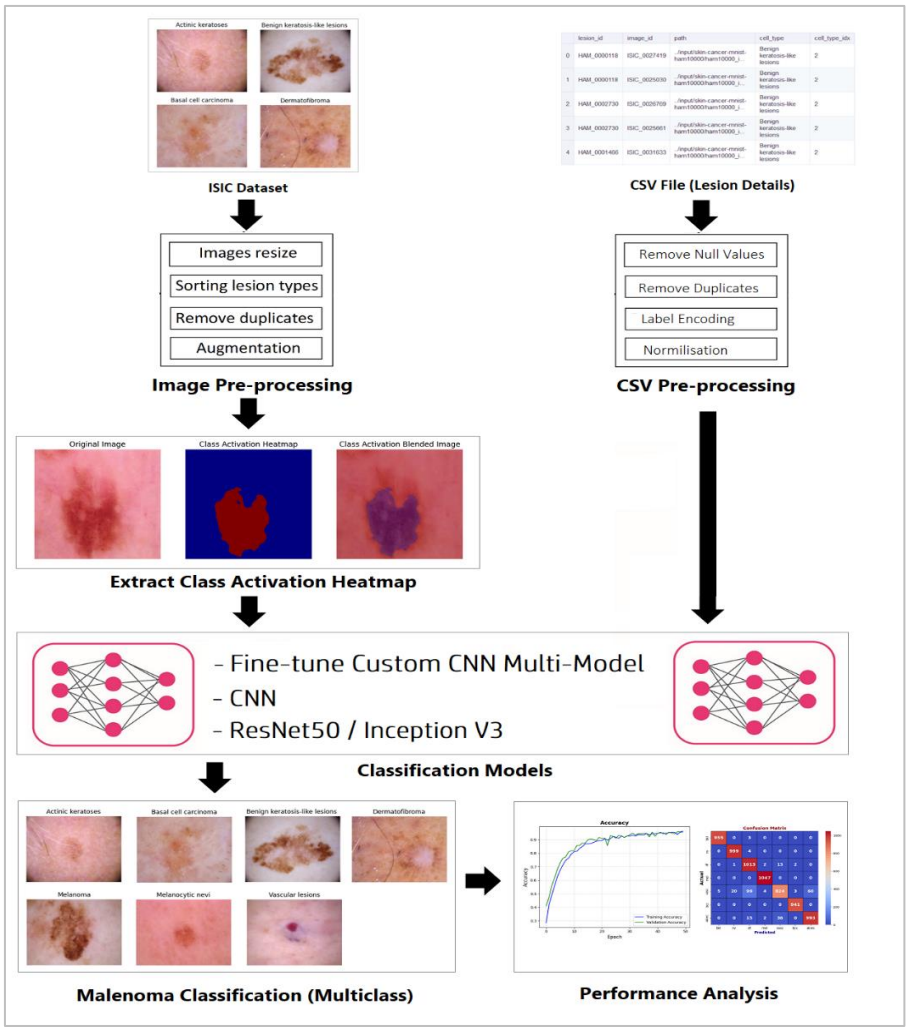


Figure 2. Proposed system architecture

	lesion_id	image_id	dx	dx_type	age	sex	localization	path	cell_type	cell_type_idx
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2

Figure 3. ISIC dataset CSV file sample

Nevus, benign keratosis, actinic keratosis, dermatofibroma, squamous cell carcinoma, melanoma, and base-cell carcinoma are the seven patterns that correspond to the aforementioned conditions. An important advantage of the dataset is that it often comes with clinical metadata, besides photos, which enables development of multimodal learning. Skin cancer diagnosis machine learning models have been established with the help of the ISIC dataset that has enabled countless numbers of deep learning research in dermatological stream. It is frequently seen in computer vision and AI research, particularly when training Transformer-based models and CNNs. The dataset is part of global competitions and is published in the ISIC Archive to support research and development of medical image analysis. Figure 2 shows the architecture diagram of the suggested system.

3.1 Load dataset

The process of loading the ISIC dataset involves obtaining dermoscopic images and corresponding labels, which categorizes skin lesions as benign or malignant. First, the dataset is downloaded from the ISIC Archive or Kaggle, followed by loading image files and clinical metadata. Images are resized (e.g., 224×224 pixels) and normalized to improve model efficiency. After that, the dataset is partitioned into training and validation sets by employing stratified sampling in order to achieve a balanced distribution of classes.

In the final step, the images and labels are transformed into NumPy arrays or TensorFlow datasets, which prepares them for the training of DL models.

Figure 3 illustrated the ISIC dataset, which is commonly employed for skin cancer classification, features a sample metadata table displayed. The lesion ID, image ID, diagnosis (dx), diagnostic type, patient age, sex, localization (body part), image path, cell type, and cell type index are among the important details regarding each skin lesion that are included in the table. Benign keratosis-like lesions, or "bkl", are indicated in the diagnosis (dx) column of this sample. Additionally, the dataset offers histopathological diagnosis ("histo"), which is necessary for AI model training.

By enabling models to include both image and clinical features for increased skin cancer diagnosis accuracy, such metadata aids in multimodal learning.

The suggested multimodal model is trained and assessed using data from the ISIC. Dermoscopic images of various skin lesions, both benign and malignant, are added in the dataset.

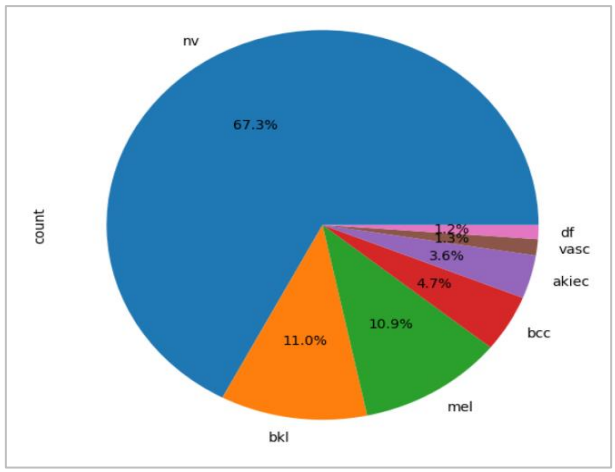


Figure 4. ISIC dataset samples per label

Figure 4 shows the dataset class distribution across classes mention above. It covers a variety of classes, including:

- Melanoma (MEL)
- Squamous Cell Carcinoma (SCC)
- Actinic Keratosis (AKIEC)
- Benign Keratosis (BKL)
- Basal Cell Carcinoma (BCC)
- Dermatofibroma (DF)
- Nevus (NV)

3.2 Data pre-processing

The preprocessing steps include image resizing, normalization, augmentation, and handling class imbalance using oversampling, under sampling, or weighted loss functions. The dataset is preprocessed to ensure high-quality input for the model. In order to get the ISIC dataset suitable for being used for deep learning model training, the pre-processing of the data is a crucial step. To guarantee high-quality inputs, it entails cleaning the image data and CSV metadata files. A thorough explanation of the data pre-processing workflow is explained below:

3.2.1 CSV file processing

Prior to combining with image data, the patient and lesion-related information in the CSV metadata file needs to be cleaned.

Remove Null Values Data from CSV files. Missing values in patient age, sex, or lesion localization can affect training, so we drop or impute them.

Remove Duplicates. Duplicate entries can cause data leakage and affect model performance, so we drop them.

3.2.2 Image file processing

Remove Corrupted Images from Image Data. All photographs are properly formatted, labeled, and scaled before being fed into the model thanks to pre-processing image data.

Label Encoding. Some image files may be corrupted or unreadable, which can cause issues during training. We check for such images and remove them.

Image Resizing. Deep learning models require numeric labels instead of text. We convert diagnosis labels into integer class indices.

3.2.3 Apply data augmentation (Image data)

A key method in deep learning for increasing dataset variety, decreasing overfitting, and boosting model generalization is data augmentation. It entails altering images in different ways while keeping their class designations intact. The enhancement methods used on images of skin lesions are listed below:

Transpose(p=0.5). The transpose operation swaps the x and y axes of the image, effectively rotating it by 90 degrees or flipping it diagonally.

Mathematically, given an image matrix I (x, y), the transposed image I'(y, x) is obtained as:

$$I'(y,x) = I(x,y)$$

Probability p=0.5 means this transformation is applied to 50% of the images.

VerticalFlip(p=0.5). A vertical flip inverts the image along the horizontal axis.

Given a pixel coordinate (x, y), the transformation results in:

$$I'(x,y) = I(x,H-y)$$

where, H is the image height.

HorizontalFlip (p=0.5). A horizontal flip inverts the image along the vertical axis.

Mathematically, it is defined as:

$$I'(x,y) = I(W-x,y)$$

where, W is the image width.

Rotate (p=0.5). Random rotation applies clockwise or counterclockwise rotation to the image. The transformation follows a rotation matrix:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

where, θ is the random angle (in degrees), and p=0.5 ensures a 50% probability of applying rotation.

RandomBrightness (limit=0.2, p=0.5). Adjusts the brightness by scaling pixel values within a given limit. If I (x, y) represents the original pixel intensity, brightness adjustment is:

$$I'(x,y) = I(x,y) + \beta$$

where, β∈[-0.2, 0.2] is a random factor.

RandomContrast (limit=0.2, p=0.5). Modifies the contrast of the image by adjusting pixel intensities. The transformation is:

$$I'(x,y) = \alpha \cdot (I(x,y) - \mu) + \mu$$

where, α is a random contrast factor within [1-limit, 1+limit], and μ is the mean pixel intensity.

GaussianBlur (blur_limit=5, p=0.25). Applies Gaussian blurring to smooth the image and reduce noise. The Gaussian kernel function is:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where, σ (standard deviation) determines the blurring intensity.

Normalize (max_pixel_value=255.0, p=0.5). Normalization scales pixel values between 0 and 1 to stabilize training. Given an original pixel intensity I (x, y):

$$I'(x,y) = \frac{I(x.y)}{255.0}$$

ShiftScaleRotate(shift_limit=0.1, scale_limit=0.1, rotate_limit=15, border_mode=0, p=0.85). This applies a random shift, scaling, and rotation transformation simultaneously. Shifting moves the image in x and y directions:

$$I'(x,y) = I(x+\Delta x, y+\Delta y)$$

where, Δx, Δy are random shifts within [-0.1, 0.1] of the image size.

Scaling resizes the image by a factor s:

$$I'(x,y) = I(sx,sy)$$

where, s∈ [0.9, 1.1]. Rotation follows the same rotation matrix used earlier with a limit of ±15 degrees.

By adding artificial variations to the dataset, these augmentation strategies strengthen DL models for the classification of skin cancer. The model learns discriminative features and improves generalization by incorporating transformations like flipping, rotation, brightness shifts, and blurring, which lowers the possibility of overfitting.

3.3 Extract class activation heatmap from images

Class Activation Maps (CAMs) visually identify the regions of an image that are most influential in the model's prediction [22]. CAM features—such as the feature maps from the gradient-based weights, and their weighted combinations—are critical as they highlight the spatial areas that have the biggest impact on the categorization result. By confirming that the model focuses on meaningful features, CAMs enhance diagnostic reliability and build trust in AI-assisted decision-making. Figure 5 illustrates the workflow for extracting Class Activation Heatmaps from images.

3.3.1 Get image data

- To generate a heatmap, we first retrieve the image

and preprocess it to match the model's expected input format.

- The image is loaded, resized, and normalized as required by the trained deep learning model.
- The image is then converted into a tensor so it can be fed into a neural network for inference.
- If the model requires batch inputs, the image is expanded to include a batch dimension (e.g., shape (1, 224, 224, 3)).

3.3.2 Compute heatmap

- To compute the Class Activation Map (CAM), we extract the feature maps from the last convolutional layer of a CNN-based model (e.g., ResNet, EfficientNet).
- The model's prediction is obtained, and the gradient of the output class is computed with respect to the final convolutional layer.
- The gradients are weighted and integrated with the feature maps to create a heatmap that highlights significant areas in the image.
- These gradients demonstrate how essential each feature map is for the projected class.
- The heatmap matrix is normalized to values between 0 and 1 to enhance visualization.

Mathematically, the Class Activation Map CAM(x,y) is computed as:

$$CAM(x,y) = ReLU\left(\sum_k w_k f_k(x,y)\right)$$

where,

$f_k(x,y)$ are feature maps from the final convolutional layer.

w_k is weights obtained from the class-specific gradients.

$ReLU$ ensures only positive activations contribute to the heatmap.

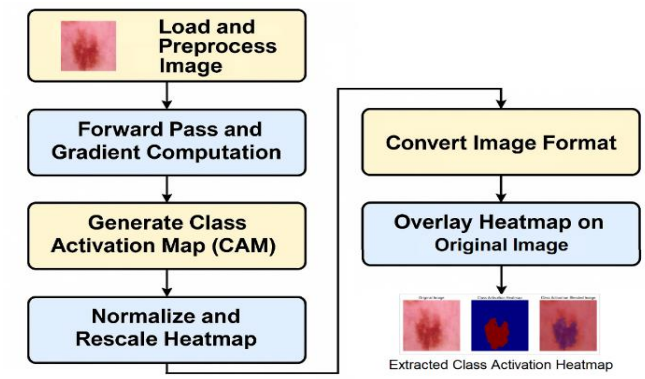


Figure 5. Workflow of extraction of class activation heatmap from images

3.3.3 Rescale heatmap to a range 0-255.

Since the heatmap is originally scaled between 0 and 1, it must be rescaled to 0-255 (pixel intensity range) for proper visualization.

- This is done by multiplying the normalized heatmap by 255.
- This transformation enhances contrast, making the high-activation areas (important regions) more visible in the final overlay.
- A colormap (e.g., Jet, Plasma) can be applied to enhance interpretability by coloring high-importance areas in red and low-importance areas in blue.

The rescaling formula is:

$$Heatmap_{scaled}(x,y) = \begin{pmatrix} Heatmap(x,y) - min \\ max - min \end{pmatrix} \times 255$$

where, min and max are the minimum and maximum values of the heatmap.

3.3.4 Convert the image from BGR format to the RGB format

By default, the majority of image processing libraries, including OpenCV, load images in BGR format. Nevertheless, visualization tools like Matplotlib and deep learning models like TensorFlow and PyTorch require RGB format.

- To ensure correct visualization, we convert BGR images to RGB before overlaying the heatmap.
- This step ensures that colors are displayed correctly when blending the original image with the heatmap overlay.

This process is critical in medical imaging, where accurate visualization of lesion areas can aid in clinical interpretation and model explain ability.

Class Activation Maps (CAMs) offer visual information about CNN-based models and aid in the explanation of how they classify skin cancer. Dermatologists and researchers can confirm that the model is focused on pertinent lesion locations by extracting and superimposing CAMs on photos. This enhances the interpretability and reliability of AI-assisted medical diagnosis.

Figure 6 shows The Class Activation Map (CAM) visualization for a skin lesion classification model. The original dermoscopic image is depicted on the left, a heatmap emphasizing the key areas utilized for classification is displayed in the middle, and the heatmap is combined with the original image on the right to improve interpretability. This method aids in comprehending model choices and guarantees that the AI concentrates on clinically significant regions for the diagnosis of skin cancer.

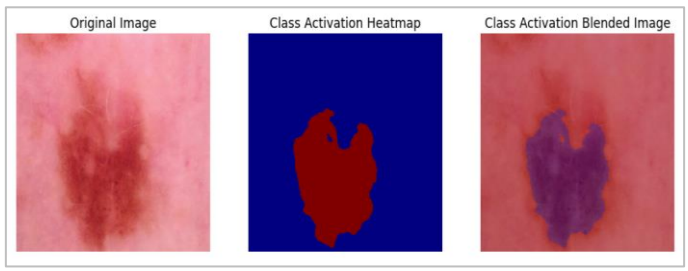


Figure 6. Class activation heatmap of images

3.4 Deep learning and transfer learning model on image data

3.4.1 CNN

CNNs are essential for diagnosing skin cancer and are commonly employed in image classification tasks in deep learning applications [23]. CNN architectures typically have convolutional layers that extract spatial data, pooling layers that reduce dimensionality, and fully connected classification layers. Convolutional filters help identify skin cancer by catching edges, textures, and abnormalities. CNNs can identify complicated and delicate medical picture patterns

because they learn hierarchical data representations. CNNs are essential to AI-powered dermatological diagnostic systems because they can distinguish benign from malignant skin lesions.

3.4.2 ResNet50

To address the issue of disappearing gradients in deep designs, a deep convolutional neural network known as ResNet50 (Residual Network with 50 layers) [24] use residual learning. Skip connections are used to improve the stability of the training since they allow the gradients to flow along the network uninterrupted. ResNet50 can extract complex features from dermoscopic images efficiently, so it is widely adopted in medical imaging. ResNet50 can be trained for skin cancer classification with transfer learning from pre-trained weights of ImageNet, and still achieve high accuracy with fewer labeled data. It performs medical AI very effectively due to its deep hierarchical feature extraction.

3.4.3 InceptionV3

A deep learning model called InceptionV3 is tuned for multi-scale feature extraction and computational efficiency. It presents inception modules, which use several convolutional filters (1×1, 3×3, 5×5) in tandem to capture both large-scale structures and fine-grained details at the same time. By improving model efficiency, this architecture lowers computing expenses without sacrificing accuracy. Pre-trained on ImageNet, InceptionV3 can be improved for the identification of skin cancer, hence increasing generalization. It is especially helpful for seeing subtle yet important patterns in medical imaging because of its adaptable receptive fields. InceptionV3 is a popular AI tool in dermatology for early melanoma diagnosis and lesion classification.

3.5 Fine-tune custom CNN multimodal (FT-MultiCNN)

Multimodal learning enhances model performance by integrating many data sources, such as images and information. A multimodal DL model for skin cancer identification integrates clinical metadata (e.g., patient age, sex, and lesion location) with CNN-based image features to provide a more thorough diagnosis.

The proposed model, Fine-tune custom CNN multimodal (FT-MultiCNN), integrates image and clinical metadata through a dual-branch architecture that combines CNN-based visual feature extraction with metadata processing, followed by feature fusion via concatenation for robust multiclass skin cancer classification

After independently processing the image data through a CNN branch and the clinical metadata through a dense neural network branch, the extracted feature representations from both modalities are merged using concatenation. This fusion step is critical for enabling the model to jointly reason about both visual patterns in the skin lesion and patient-specific contextual information. From the CNN branch, the output from the final convolutional block is flattened into a one-dimensional feature vector that captures high-level spatial and textural patterns from the dermoscopic image. Simultaneously, the clinical metadata—such as patient age, sex, lesion location, and other relevant variables—is passed through a fully connected layer to produce a separate, dense feature embedding. This embedding captures relationships between different metadata attributes and learns a compact representation that is informative for classification.

In the concatenation layer, these two feature vectors—one from the image and one from the metadata—are combined into a single unified vector. This operation does not perform any mathematical transformation but simply aligns the two vectors side-by-side, allowing the subsequent layers to access and learn from both sources of information simultaneously. The fused representation is then passed through one or more fully connected (dense) layers, which are responsible for learning the joint feature space and making the final multiclass classification decision.

The given architecture (Figure 7) is an example of a custom multimodal CNN model that concurrently processes clinical metadata and dermoscopic images. An analysis of the functional model summary can be found below:

3.5.1 CNN-based image feature extraction

- The input layer accepts 128×128×3 images, representing RGB dermoscopic images.
- Three convolutional blocks extract spatial features with increasing depth:
- Conv2D layers (32, 64, 128 filters) learn low-to-high-level patterns.
- MaxPooling layers reduce spatial dimensions, retaining important features.
- The Flatten layer converts extracted features into a dense representation.

3.5.2 Clinical metadata processing

- A separate InputLayer (7 features) processes clinical metadata.
- A Dense layer (64 neurons) learns feature embeddings from metadata.

3.5.3 Feature fusion via concatenation

- Features from the CNN image branch and the clinical metadata branch are concatenated.
- A final Dense layer (7 neurons) performs multiclass classification.

3.5.4 Fine-ttuning for improved performance

- Pre-trained CNN backbones (e.g., ResNet50, EfficientNet) can replace the current CNN layers for better feature extraction.
- Dropout and Batch Normalization layers are added to prevent overfitting.
- Hyper-parameter tuning (learning rate, batch size) further optimizes model performance.

In comparison to image-only models, this customized multimodal CNN model effectively combines image-based and metadata-based learning, increasing the accuracy of skin cancer detection. Further improving diagnostic performance can be achieved through fine-tuning using data augmentation, feature engineering, and transfer learning. Figure 7 presents the architecture summary of a multimodal deep learning model.

By integrating features from both modalities, this fusion strategy improves the model's ability to capture complex interactions between visual and clinical cues—something that would not be possible if the modalities were processed in isolation. As a result, the model can make more accurate and context-aware predictions, ultimately improving diagnostic performance.

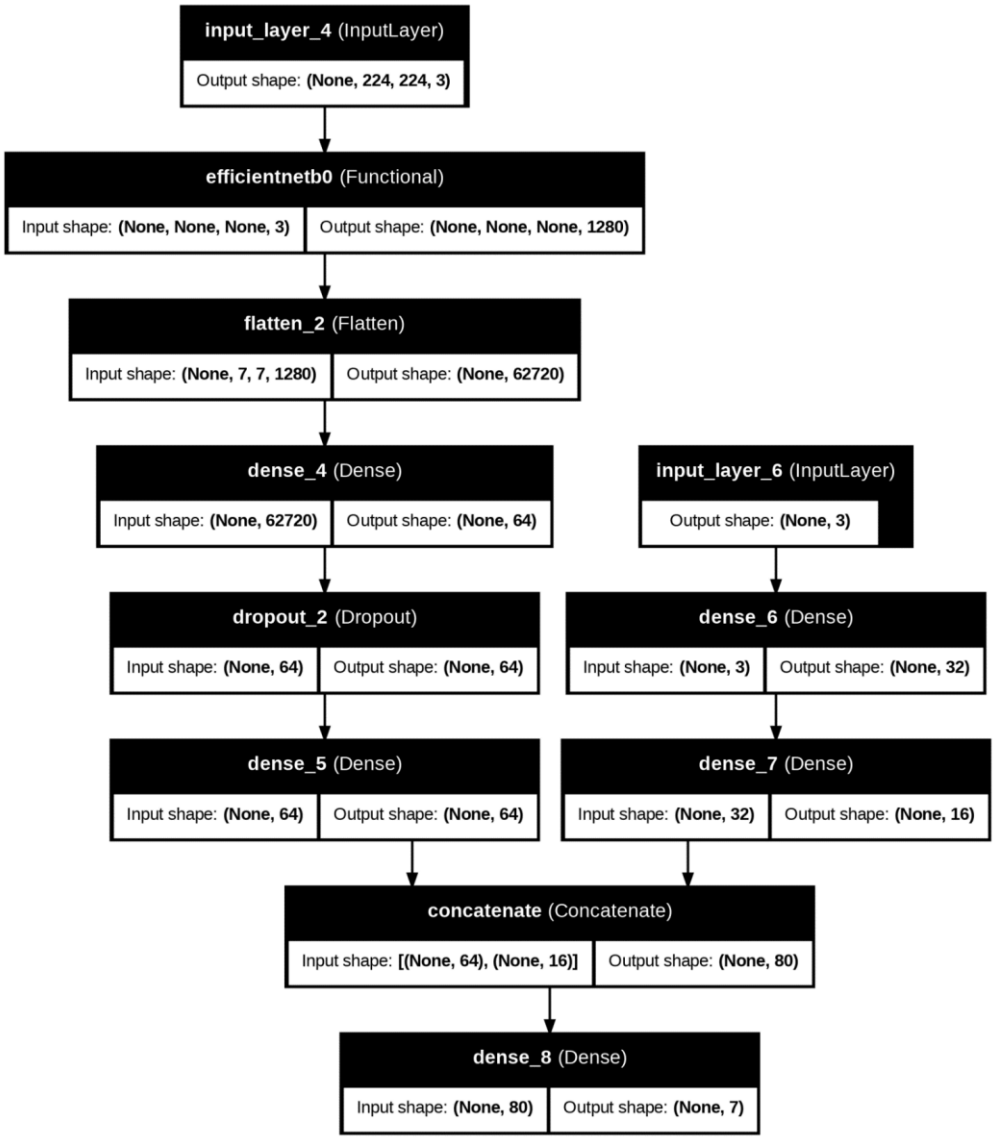


Figure 7. Fine-tune custom CNN multimodal (FT-MultiCNN)

4. RESULT ANALYSIS

4.1 Experimental setup

The experiments were conducted using PYTHON software on a Google Colab having ~13GB of RAM and ~15GB of GPU. The network's hyper-parameters included a batch size of 32, 50 epochs, the Adam optimizer, cross-entropy as the loss function, and ReLU and Softmax as the activation functions. This setup ensured a robust framework for evaluating the model's performance.

4.2 Performance parameters

In order to prove the overall success of categorization models, a number of metrics are utilized in the process of analyzing their performance. In this study, we have used accuracy, precision, recall, F1-score, loss, and the confusion matrix to assess the models. Below are the performance parameters formulas:

$$Accuracy = \frac{Number\ of\ Correctly\ Classified\ Sample}{Total\ Number\ of\ Sample}$$

$$Precision = \frac{Number\ of\ True\ Positive\ Sample}{Number\ of\ True\ Positive\ Sample + Number\ of\ False\ Positive\ Sample}$$

$$Recall = \frac{Number\ of\ True\ Positive\ Sample}{Number\ of\ True\ Positive\ Sample + Number\ of\ False\ Negitive\ Sample}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.3 Results

This analysis demonstrates the superiority of our proposed Fine-tune Custom CNN Multi-Model (FT-MultiCNN) over existing methods. We have compared our model with CNN, ResNet50, and InceptionV3, evaluating its performance against these cutting edge DL architectures. Figure 8 presents the comparative analysis of all models, demonstrating that the suggested Fine-tune Custom CNN Multi-Model (FT-MultiCNN) outperforms CNN, ResNet50, and InceptionV3 in terms of accuracy and F1-score.

In Figure 9(a), the graphs depict the loss (right) and training and validation accuracy (left) curves for a deep learning model

used to classify skin cancer. Both training and validation accuracy improve steadily, as seen by the accuracy curve, which stabilizes at 80.0%, suggesting successful learning. The loss curve indicates that the model is optimizing well because it shows a sharp decline in both training and validation loss. Minor variations in validation loss and accuracy, however, might point to a small amount of overfitting. Dropout, data augmentation, and early halting are some methods that can be used to enhance generalization.

Figure 9(b), the graphs displayed the Training and validation accuracy (left) and training and validation loss (right) curves spanning several epochs. With a consistent rise to over 92%, the accuracy curves demonstrate successful learning.

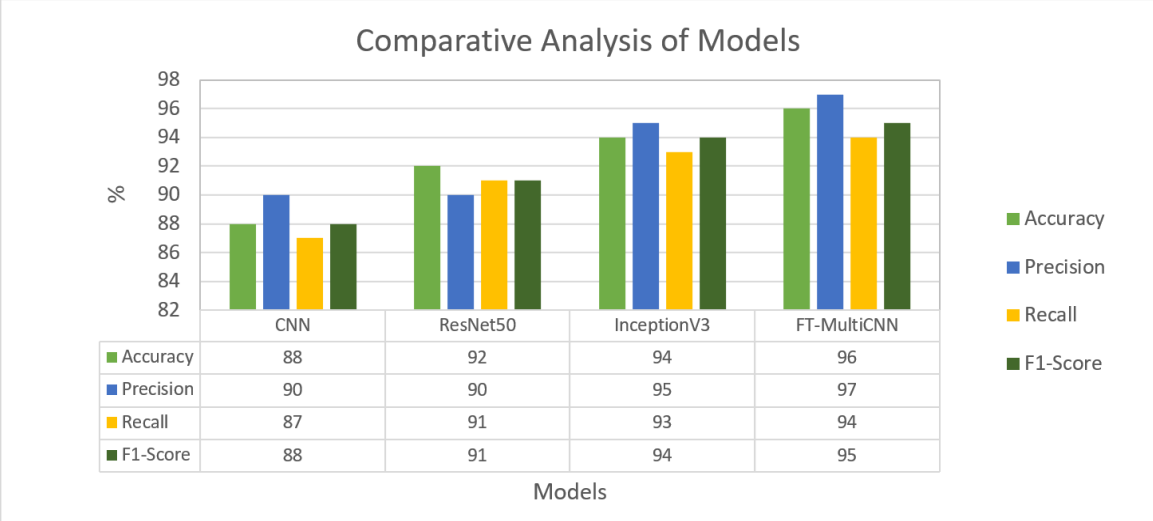
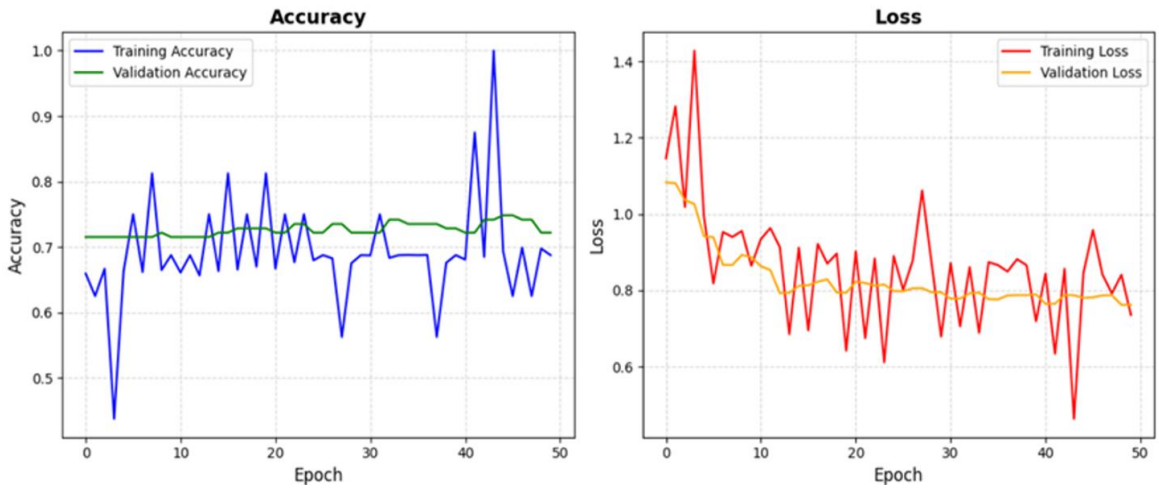
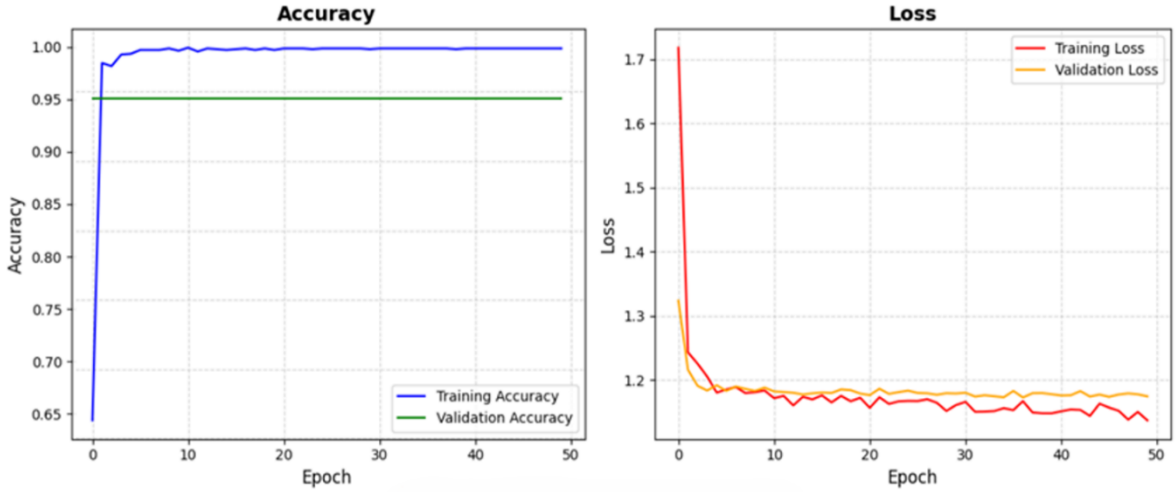


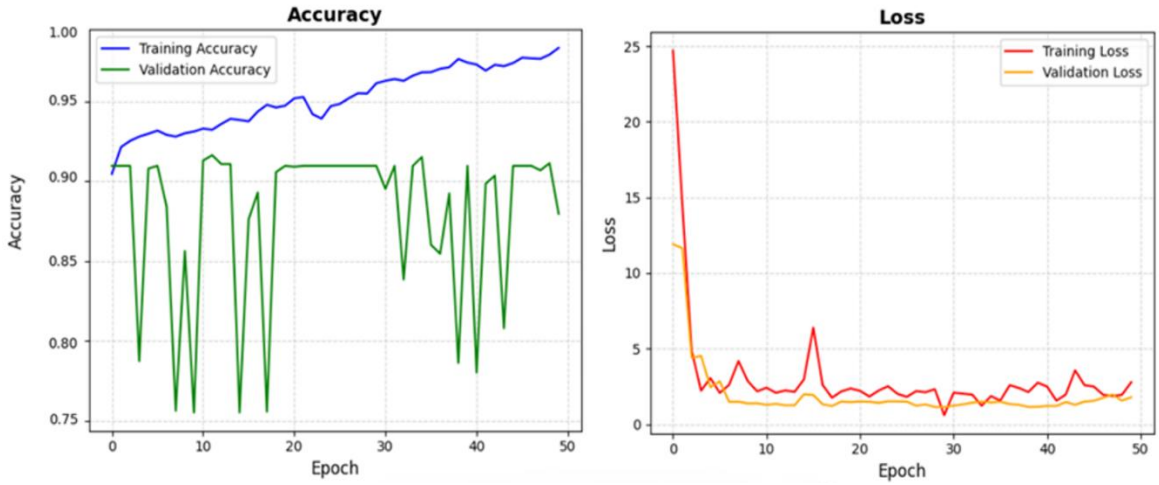
Figure 8. Comparative analysis of models



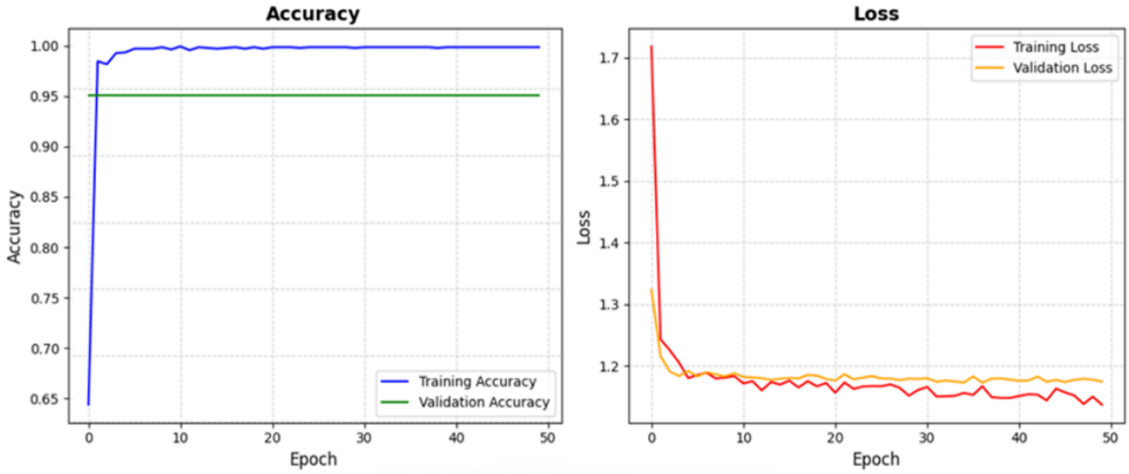
(a) CNN model accuracy and loss curve



(b) ResNet50 accuracy and loss curve



(c) InceptionV3 accuracy and loss curve



(d) Fine-tune custom CNN muti-model accuracy and loss curve

Figure 9. Accuracy and loss comparison graph

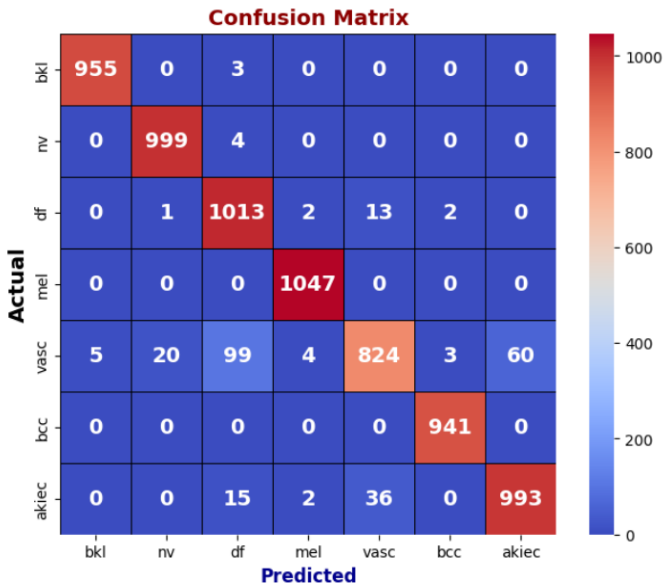


Figure 10. Confusion matrix of FT-MultiCNN

Figure 9(c), the accuracy and loss curves for the InceptionV3 model are displayed on the graph. Strong learning ability is demonstrated by the training and validation accuracy (left), which increases progressively to over 91%. Both curves closely follow one another, showing minimal overfitting and good generalization when the training and validation loss (right) steadily decline. This shows that InceptionV3 performs effectively for classifying skin cancer.

Figure 9(d) displays the accuracy and loss curves for the customized CNN multi-model that has been optimized. Training accuracy (left) steadily increases to around 96%. The test loss decreases as the number of epochs increases, suggesting that the model can effectively generalize to unidentified data.

The results highlight the effectiveness of our model in achieving superior classification performance for skin cancer detection. Figure 10 shows the confusion matrix of FT-MultiCNN.

5. CONCLUSIONS

Multimodal CNN for Skin Cancer Detection and Classification, integrating both image data and patient metadata to enhance diagnostic accuracy is proposed. By employing the Fine-tune Custom CNN Multi-Model (FT-MultiCNN) and leveraging Class Activation Heatmap features, our model effectively captured complex patterns in skin lesions. The multimodal fusion approach, coupled with parallel processing architecture, enabled superior classification performance. Experimental results demonstrated that our model outperformed traditional deep learning, transfer learning, and metadata fusion techniques, achieving 96%

accuracy and 94% recall. These findings demonstrate the potential of multimodal deep learning in dermatology, offering a highly accurate, automated, and non-invasive method for early skin cancer detection. This research underscores the promise of AI-driven solutions in improving skin cancer diagnostics and patient outcomes.

The proposed multimodal architecture based on CNN and Transformer–autoregressive models may be one of the steps in improving accuracy of automatic diagnosis. However, there are several aspects that are still in need of further study and enhancement. Firstly, increasing the variety of medical imaging modalities such as 3D imaging and histology might increase the accuracy of the classification by giving insight into the lesion architecture. Second, clinical adoption will depend on improving model interpretability using explainable AI (XAI) techniques, which will provide dermatologists confidence in AI-generated judgments. To ensure fairness and reduce prediction bias, future work should also consider raising the model's performance in different demographic subgroups. Furthermore, by enabling the training of models over the decentralized datasets and maintaining patient confidentially, federated learning can enhance data privacy. Development of lightweight AI models for mobile and edge devices to facilitate remote and real time skin cancer screening is also a major area to explore. Taken together, the continued advances in AI-based dermatology solutions holds great promise for revolutionizing early detection of skin cancer and enhancing patient outcomes everywhere.

REFERENCES

[1] Hasan, N., Nadaf, A., Imran, M., Jiba, U., Sheikh, A., Almalki, W.H., Almujri, S.S., Mohammed, Y.H., Kesharwani, P., Ahmad, F.J. (2023). Skin cancer: Understanding the journey of transformation from conventional to advanced treatment approaches. Molecular Cancer, 22(1): 168. https://doi.org/10.1186/s12943-023-01854-3

[2] Gouda, W., Sama, N.U., Al-Waakid, G., Humayun, M., Jhanjhi, N.Z. (2022). Detection of skin cancer based on skin lesion images using deep learning. In Healthcare. MDPI, 10(7): 1183. https://doi.org/10.3390/healthcare10071183

[3] Chakkarapani, V., Poornapushpakala, S., Suresh, S. (2025). Enhancing skin cancer detection with multimodal data integration: A combined approach using images and clinical notes. SN Computer Science, 6(1): 1-14. https://doi.org/10.1007/s42979-024-03601-x

[4] World Health Organization (WHO)-Skin Cancer Statistics.

[5] Indian Journal of Dermatology-Skin Cancer Trends in India.

[6] Gallazzi, M., Biavaschi, S., Bulgheroni, A., Gatti, T., Corchs, S., Gallo, I. (2024). A large dataset to enhance skin cancer classification with transformer-based deep neural networks. IEEE Access, 12: 109544-109559. https://doi.org/10.1109/ACCESS.2024.3439365

[7] Claret, S.A., Dharmian, J.P., Manokar, A.M. (2024). Artificial intelligence-driven enhanced skin cancer diagnosis: Leveraging convolutional neural networks with discrete wavelet transformation. Egyptian Journal of Medical Human Genetics, 25(1): 50. https://doi.org/10.1186/s43042-024-00522-5

[8] Remya, S., Anjali, T., Sugumaran, V. (2024). A novel transfer learning framework for multimodal skin lesion analysis. IEEE Access, 12: 50738-50754. https://doi.org/10.1109/ACCESS.2024.3385340

[9] Riaz, S., Naeem, A., Malik, H., Naqvi, R.A., Loh, W.K. (2023). Federated and transfer learning methods for the classification of Melanoma and Nonmelanoma skin cancers: A prospective study. Sensors, 23(20): 8457. https://doi.org/10.3390/s23208457

[10] Jiang, X., Hu, Z., Wang, S., Zhang, Y. (2023). Deep learning for medical image-based cancer diagnosis. Cancers, 15(14): 3608. https://doi.org/10.3390/cancers15143608

[11] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. Frontiers in Oncology, 12: 893972. https://doi.org/10.3389/fonc.2022.893972

[12] Medhat, S., Abdel-Galil, H., Aboutabl, A.E., Saleh, H. (2022). Skin cancer diagnosis using convolutional neural networks for smartphone images: A comparative study. Journal of Radiation Research and Applied Sciences, 15(1): 262-267. https://doi.org/10.1016/j.jrras.2022.03.008

[13] Goyal, M., Knackstedt, T., Yan, S., Hassanpour, S. (2020). Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. Computers in Biology and Medicine, 127: 104065. https://doi.org/10.1016/j.compbiomed.2020.104065

[14] Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., Zhou, Q., Wang, S., Li, L., Yang, F., Xu, S., Chen, H. (2022). An improved transformer network for skin cancer classification. Computers in Biology and Medicine, 149: 105939. https://doi.org/10.1016/j.compbiomed.2022.105939

[15] Hajiarbabi, M. (2023). Skin cancer detection using multi-scale deep learning and transfer learning. Journal of Medical Artificial Intelligence, 6. https://doi.org/10.21037/jmai-23-67

[16] Al-Rasheed, A., Ksibi, A., Ayadi, M., Alzahrani, A.I., Zakariah, M., Ali Hakami, N. (2022). An ensemble of transfer learning models for the prediction of skin cancers with conditional generative adversarial networks. Diagnostics, 12(12): 3145. https://doi.org/10.3390/diagnostics12123145

[17] Anand, V., Gupta, S., Altameem, A., Nayak, S.R., Poonia, R.C., Saudagar, A.K.J. (2022). An enhanced transfer learning based classification for diagnosis of skin cancer. Diagnostics, 12(7): 1628. https://doi.org/10.3390/diagnostics12071628

[18] Venugopal, V., Raj, N.I., Nath, M.K., Stephen, N. (2023). A deep neural network using modified EfficientNet for skin cancer detection in dermoscopic images. Decision Analytics Journal, 8: 100278. https://doi.org/10.1016/j.dajour.2023.100278

[19] Chen, Q., Li, M., Chen, C., Zhou, P., Lv, X., Chen, C. (2023). MDFNet: Application of multimodal fusion method based on skin image and clinical data to skin cancer classification. Journal of Cancer Research and Clinical Oncology, 149(7): 3287-3299. https://doi.org/10.1007/s00432-022-04180-1

[20] Lyakhov, P.A., Lyakhova, U.A., Kalita, D.I. (2023). Multimodal analysis of unbalanced dermatological data for skin cancer recognition. IEEE Access, 11: 131487-

131507. https://doi.org/10.1109/ACCESS.2023.3336289
[21] Yap, J., Yolland, W., Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. Experimental Dermatology, 27(11): 1261-1267. https://doi.org/10.1111/exd.13777
[22] Si, N., Zhang, W., Qu, D., Luo, X., Chang, H., Niu, T. (2021). Spatial-Channel attention-based class activation mapping for interpreting CNN-based image classification models. Security and Communication Networks, 2021(1): 6682293. https://doi.org/10.1155/2021/6682293
[23] Mane, D., Ashtagi, R., Suryawanshi, R., Kaulage, A.N., Hedaoo, A.N., Kulkarni, P.V., Gandhi, Y. (2024). Diabetic retinopathy recognition and classification using transfer learning deep neural networks. Traitement du Signal, 41(5): 2683-2691. https://doi.org/10.18280/ts.410541
[24] Patil, R., Bellary, S. (2021). Transfer learning based system for melanoma type detection. Revue d'Intelligence Artificielle, 35(2): 123-130. https://doi.org/10.18280/ria.350203