VOL. 19, NO. 15, AUGUST 2024 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

EXPLAINABLE AI FRAMEWORK FOR SKIN CANCER CLASSIFICATION, AND MELANOMA SEGMENTATION

K. Srilatha and N. Satheesh Kumar
Computer Science and Engineering, Chaitanya (Deemed to be University), Warangal, Telangana, India
E-Mail: csephdscholar2021@gmail.com

ABSTRACT

Identification and classification of skin diseases are two critical challenges faced in diagnosing and treating patients suffering from them. Deep learning models have been created to best identify and classify skin problems to detect and identify them correctly and effectively. This paper proposes a comprehensive framework for accurate skin cancer prediction, classification, and melanoma surgical lesion extraction. Primarily, a comprehensive extraction method leveraging the unique approach of DenseNet201 and the Local Interpretable Model-Agnostic Explanation offers accurate insight into model decision making and prediction. Secondly, the model has a mid-extraction phase that utilizes advanced convolutional neural network levels to detect the boundaries of melanoma lesions correctly. The framework results in terms of IOU, accuracy, precision, recall, and other metrics compared to existing models like FPN, MAN, and U-Net. The framework presented in our model is smart, easy to use, and can provide functional and accurate information, which means it can be used in clinical practice.

Keywords: skin cancer detection, deep learning, melanoma segmentation, explainable AI, convolutional neural networks (CNNs).

Manuscript Received 21 June 2024; Revised 18 August 2024; Published 31 October 2024

1. INTRODUCTION

Millions of people across the globe suffer from skin diseases, making their diagnosis and treatment two most complicated issues. The rapid development of deep learning and CNNs has contributed to the emergence and further improvement of artificial systems capable of identifying and classifying various skin diseases [1]. Researchers and practices intend to change the established situation in dermatology and transform the approach to defining cases, making the most proactive assistance possible [2]. The development of deep learning models has revolutionized the diagnosis of skin disease. Previously, most diagnostics involved a certain degree of subjectivity and visual inspection by dermatologists using photographs. For comparison, the diagnostic performance of deep learning models is much more reliable due to their ability to analyze significantly more data than any human experts. Thousands of images and clinical notes enable the identification of the most subtle and hard-to-understand correlates via labeled data. Thus, they can identify various skin conditions in distinction to the accuracy levels of medical experts or even outperform them [3].

These deep learning technologies simplify the diagnostic process itself, which means that many different skin conditions can be detected faster and more accurately. Deep learning models analyze images of dermatological conditions and related clinical data, and the results may be limited or specific rights [4]. In areas where there are few specialists or dermatologists or their facilities, automatic systems can be used for the initial diagnosis and thus not to delay, but to prevent the disease in due time. Despite its promising use cases of deep learning-based models, various constraints make it difficult to implement and enhance efficiency in detection and classification. The limitations of such applications are primarily due to insufficient and unvarying datasets. Furthermore, even many training sets are limited, leading to inconsistent findings in various demographic populations. Despite advancements, the performance of algorithm development and evaluation is occasionally stymied by a shortage of annotated dermatologic imagery [5]. Furthermore, the imbalance and skewness of skin disease data are a major obstacle. Imbalance data with a skewness that is some disease classes represented far more than others could disrupt model generalization and result in poor classification of smaller classes. As a result, reducing the impact of skewed class distribution through meticulous data preprocessing and specialized loss functions is mandatory [6]. A serious issue with deep learning for skin conditions is the generalization of models to conditions not seen in the training set, which are often too rare for a model specific to a large subset of diseases to develop robust diagnostics. Most approaches alleviate this limitation with transfer learning, although it still is a topic of active research. Incorporating the use of explainable AI into the detection and classification of skin diseases in deep learning models provides a potential approach to mitigating numerous underlying long-term problems. The explainable AI approaches offer critical bases on which to understand how the models make their decisions and contribute to the fortification of the models' interpretability and trust from medical practitioners and patients. Since the stakeholder can track the essential occurrences of predictions made by the model, there exists an opportunity to focus on the classification results. Explainable AI techniques are critical for enhancing the interpretability of deep learning models. Saliency maps, gradient-based attribution methods, occlusion analysis, and other approaches facilitate the visualization of significant features in an image or signal that contribute to

the model's classification [7]. A further critical advantage for practitioners is that explainable AI helps constraints associated with model generalization and capacity that model focus cannot adequately address. Through a review of decision boundaries and feature representations, it is straightforward to determine where the model has difficulty generalizing between various skin conditions. On the other hand, accurate delineation of resection boundaries in melanoma cases has critical implications for clinical practice. Deep-learning-based segmentation offers exact lesion margins delineation allowing dermatologists and oncologists to perform accurate disease diagnosis and choose recommended treatment routines. It enables quantitative lesion size and growth rate measurement and monitoring of disease development and response to treatment over time. Furthermore, accurate identification of melanoma boundaries can help to design effective treatment plans such as surgical resection mark identification, especially on wide excisions, or radiation therapy target delineation ultimately improving therapeutic efficiency and cutting recurrence risk [8]. In this paper, we also discuss the use of deep learning-based image segmentation to achieve an accurate delineation of melanoma boundaries to improve diagnosis and treatment planning. The organization of the paper is as follows: The section-II describes the Literature survey, and the proposed methodology is explained in section-III. The simulation results are discussed in section-IV.

2. LITERATURE

To improve the classification accuracy of skin lesions, Natasha Nigar *et al* [9], presented a new classification system for skin lesions based on explainable Artificial Intelligence. It was introduced to help dermatologists make better and faster diagnoses, especially in the early stages of skin cancer. The proposed XAI model was validated using the International Skin Imaging Collaboration 2019 dataset. The results revealed that the developed model correctly classified eight types of skin lesions, such as dermatofibroma, squamous cell carcinoma, benign keratosis, melanocytic nevus, vascular lesion, actinic keratosis, basal cell carcinoma, and melanoma, with classification accuracy, precision, recall, and F1 score of 94.47%, 93.57%, 94.01%, and 94.45%, respectively. Additionally, the predictions were fed into the local interpretable model-agnostic explanations framework to elicit human-readable visual explanations that were aligned with prior expectations and adhered to the principles of general explanation best practices. The developed classification system was expected to be more useful in clinical settings.

Naveed Ahmad *et al* [10] proposed a new framework for skin lesion recognition using data augmentation, deep learning, and Explainable Artificial Intelligence. The dataset size was initially expanded using data augmentation in this work. The authors also utilized two pretrained deep learning models, Xception, and ShuffleNet. Finally, they fine-tuned and trained the models using deep transfer learning by using the global average pooling layer for deep feature extraction. They discovered that essential information was missing and decided to combine the two models. Although post-fusion computational time has increased, the enhanced Butterfly Optimization Algorithm selects only the best features to improve classification using machine learning classifiers. Additionally, the authors utilized GradCAM-based visualization to pinpoint essential regions in the images. This study validated the framework using ISIC2018 and HAM10000 datasets, thereby obtaining an increased accuracy of 99.35%. And 91.5% respectively. This work demonstrates that better accuracy can be achieved over the existing approach while taking less computational time.

The research work [11] trained and validated a ResNet18 model, the final result of the CVDL process. The Grad-CAM model explainability technique was applied to enable dermatologists to better understand the model's predictions. As a result, the classification model accurately showed 96%. The paper reviewed scientific research on CVDL and the explainability of deep learning with structural photos of skin disease to supplement the grounding evidence. Another aspect of the CVDL process, the discrepancy between the dermatologists' expected explainability and the available methods was provided. Further ways of solving this problem were suggested. In the research work [12], modifications have been proposed to the pre-trained MobileNetV2 and DenseNet201 deep learning models to help recognize skin cancer more efficiently. And the pre-trained MobileNetV2 and DenseNet201 Convolutional layers' models three more convolutional layers have been incorporated at the end of them. The comparative study determination foretold that the modified model overtook the existing pre-trained MobileNetV2 and DenseNet201 models. The proposed one seems to be more advantageous, as it recognizes both the benign and the malign class. According to the outcomes gathered, the Modified DenseNet201 model obtains an accuracy of 95.50% and is very close to being the best when equated to the antecedent works. Finally, the Modified DenseNet201 yields a sensitivity of 93.96% and a high specificity of 97.03%.

Khalid M. Hosny *et al* [13] developed a deep innate learning strategy, shown to be useful in detecting seven varieties of skin lesions. To verify the efficacy, several explanation strategies were employed. Explainable AI was used to clarify what contributes to the decision to create locally and globally, and visual aids were offered to boost physicians' trust. The proposed approach was subsequently assessed on the challenging HAM10000 dataset to assess its impact. With the help of the authors' easily realized stage-oriented X-AI architecture, physicians might better comprehend the internal operation of a black-box AI model. Alternatively, they may have confidence in it after it has utilized a reasonable answer.

Xinrong Lu *et al* [14], proposed a novel model which is built on top of the improved XceptionNet with the swish activation function and depthwise separable convolutions and demonstrated improved classification accuracy in comparison with the original Xception and other models. The simulated performance of the proposed method was compared to that of other state-of-the-art models for skin

cancer diagnosis. The results revealed its superior accuracy when compared to relative methods. Ghadah Alwakid *et al* [15], proposed DL as a tool to more accurately obtain the lesion zone. First, the image was enhanced, which raised its quality due to the Enhanced Super-Resolution Generative Adversarial Networks method. Then, ROI from the general area was isolated via segmentation. Further, the data augmenting solution was utilized due to the data's uneven representation Several CNNS and modified networks like ResNet-50 were applied to analyze the image and classify the skin lesion. The study used different-skinned people due to an unequal sample, which included seven types of skin cancer observed.

An automated Deep Learning framework was used in the research work [16], which relied on a CAD model called DLCAL-SLDC which is a computer-aided diagnosis with the class-attention layer. The main purpose of that model was to diagnose and categorize different types of skin cancers through dermoscopic images. Image preprocessing includes hair elimination using the dull razor method and noiselessness by the average median filter. A Tsallis entropy segmentation was used to isolate their lesion based on the dermoscopic image at this stage. Moreover, the DLCAL-based feature extractor was applied to the segmented lesions to extract features using a Capsule Network along with a Class Attention Layer and the Adagrad optimizer. The CAL layer in a CapsNet was created to learn discriminative class-specific features to capture class dependency and permit the CapsNet to be processed further. Eventually, the classification was done by the Swallow Swarm Optimization -Convolutional Sparse Autoencoder CSAE, better known as the SSO-CSAE model. An ISIC benchmark dataset validated the proposed DLCAL-SLDC approach.

Amina Bibi *et al* [17] presented a novel approach based on the integration of traditional and deep learning methods. The suggested framework is accompanied by two main tasks, specifically lesion segmentation and classification. In the case of the first task, the basis for contrast enhancement was the use of two filtering approaches, which after allowed applying a color transformation to differentiate changes in colors by pixels of the lesion area. Further, the optimal channel was chosen, and the lesion map was produced, which was, consequently, transformed into a binary map utilizing a threshold function. The process of lesion classification included the modification of two pre-trained CNN models and further training using the transfer learning approach. After the extraction of deep features using both networks, the merged features were derived through CCA. However, some redundant features were included in the fusion process, which, in turn, reduced the accuracy of classification. A novel approach of maximum entropy score-based selection was developed to eliminate unnecessary features. The extracted features were then fed into a cubic support vector machine.

Mohamed Yacin Sikkandar *et al* [18] proposed a new classification model for skin lesion diagnosis based on segmentation using the integrated GrabCut algorithm and ANFC. This model included four stages: preprocessing, segmentation, feature extraction, and classification. First, the preprocessing was carried out using the Top hat filter and inpainting method. After that, the integrated GrabCut-based algorithm was used to segment the preprocessing image. Then the deep network-based Inception model was implemented for extracting features. Finally, the ANFC system classified dermoscopic images into different classes. The proposed approach was tested on the benchmark International Skin Imaging Collaboration ISIC dataset, and the obtained results were measured in terms of accuracy, sensitivity, and specificity. The study showed that the proposed algorithm produced better results in the context of skin cancer identification and classification. To ensure the approach's efficiency, it was widely compared with other approaches.

3. PROPOSED MODEL

The proposed model consists of three separate phases to advance skin cancer prediction, improve classification outcomes, and segmented melanoma lesions. As for phase one, skin cancer prediction is pursued via a novel CNN model. More specifically, the Modified DenseNet201 architecture is used to classify the disease. Although the stated architecture has been adjusted and changed specifically for skin lesion classification, it stands to be noted that it is still enhancing the model needed. Secondly, the classification outcomes generated by the Modified DenseNet201 architecture are improved in the enhancement phase. An alternative or equivalent explainable AI model is utilized to add more insights into the determination basis of the classification model, which enhances transparency and understandability on the side of clinicians and stakeholders. In this manner, the model boosts the trust and transparency of the classification outcomes using explainable AI tactics, such as visual explanations, and feature attributions, which promote more informed decision-making in medical scenarios. The proposed framework is shown in Figure-1.

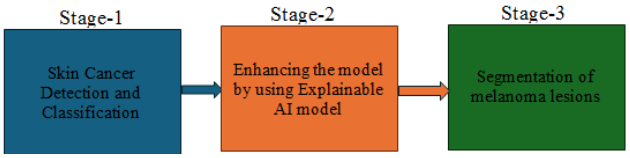


Figure-1. Proposed framework.

Finally, in the third phase, the proposed model seeks to segment the melanoma lesions existing within the dermoscopic images. Using sophisticated image processing capabilities and deep learning methodologies, the model will accurately identify the melanoma lesion boundaries. This is important as the confirmed segment can be analyzed to determine the extent of the melanoma lesion, a factor that is critical when considering different treatment approaches.

a. Skin cancer prediction and classification

In the proposed skin cancer prediction model, CNNs are used to extract hierarchical important features from the skin images. The proposed model consists of blocks, which are named "Block A", with three main layers which are depicted in Figure-2.

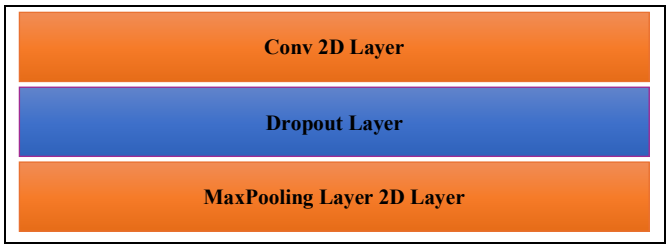


Figure-2. Block A.

Conv2D layer is the layer that makes use of the set of convolutions that are used for the extraction of features out of the input images. The Conv2D layers scan the input images to detect important visuospatial patterns important for the identification of all the types of skin datasets. Dropout was performed on the output layer of the CNN combined with the fully connected connection used to avoid overfitting the CNN model. MaxPooling2D is used in reducing the spatial dimensions to retain the most important features. The proposed CNN model for skin cancer prediction is shown in Figure-3.

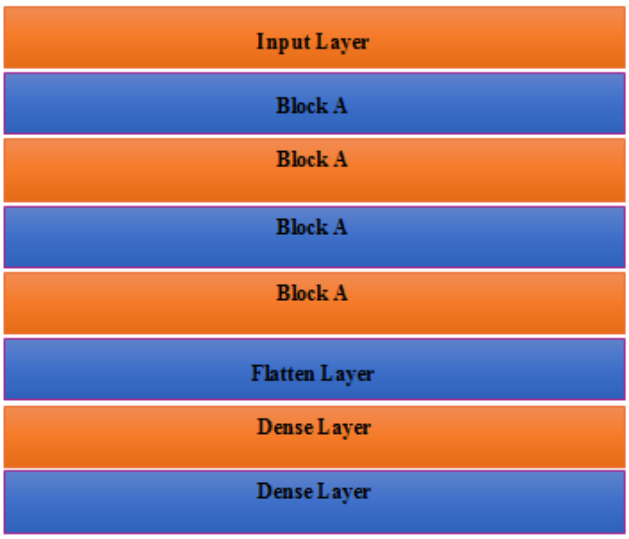


Figure-3. Proposed CNN Architecture for skin cancer prediction.

The architecture is made up of several Block A layers that enable the model to learn increasingly complex features from the data. Then, the data is flattened into a vector of dimension 1 which is passed through several fully connected dense layers. The dense layers are responsible for interpreting the most extracted features and making a conclusion regarding the possibility of skin cancer. The learning process of the model is facilitated by an Adam optimizer which aided in the updating of the model's weights by ensuring swift learning and convergence.

The model for the skin cancer classification task is based on a modified DenseNet201 architecture. DenseNet201 is a pretrained on a relatively large dataset with densely connected layers which ensures the effective reuse of features and the flow of gradients. This allows the model to successfully learn from limited amounts of data. The modified DensNet201 architecture is depicted in Figure-4.

Figure-4. Proposed model Architecture for skin cancer Classification.

Finally, after DenseNet201 learns multiple features, a Flatten Layer is utilized to convert the 2D feature maps into a 1D vector. After this, the vector is passed through a sequence of Dropout Layers and fully connected Dense Layers. Dropout Layers are used to avoid overfitting; this is done by setting a fraction of neuron outputs to zero during the training interval Dense Layers record complicated patterns and finally help to do classification. The SGD optimizer is used to modify the model's parameters, which helps CNN to learn a high amount of data efficiently.

b. Proposed explainable AI model

The second phase of the concept focuses on using explainable Artificial Intelligence techniques to improve the obtained classification results based on training the Modified DenseNet201 architecture. Explainable AI is fundamental in ensuring that the machine learning models' decisions are transparent and can be interpreted by clinicians. Explainable AI approaches attempt to elucidate the reasoning for the prediction provided by the classifier. Generating explanations intelligible to humans allows the clinician to understand why the classification decision was made and has to be made and to feel more confident and informed in the clinical practice. In this work LIME (Local Interpretable Model-Agnostic Explanations) technique is considered an Explainable AI. The steps involved in the LIME technique are shown in Figure-5.

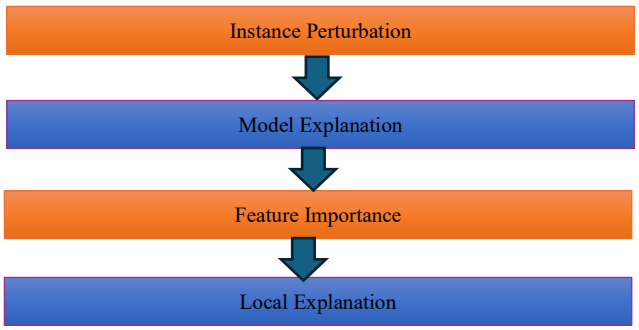


Figure-5. LIME Technique.

LIME is by far one of the most popular Explainable AI techniques, providing local interpretability for complex machine learning ML models, such as deep learning models. The main goal of LIME is to explain what predictions the black-box model makes in a way that is understandable to humans. Essentially, LIME works by creating explainable explanations of a specific instance or prediction by estimating the decision boundary of a black-box model in the area surrounding that instance.

- **Instance perturbation:** Instance perturbation refers to the generation of perturbed instances by modifying slightly selected instance's feature values, but the label remains constant. A model's decision is used to draw and sample the local neighborhood about which one wants the explanation. This involves simulating the local approximation of the black-box model near and about the selected instance. Generate variants of the input data with its ground truth labels intact. LIME aims to capture the decision boundary of the black-box model in the neighborhood of the instance being explained. This approach enables LIME to mimic how the black-box model would perform if provided with similar instances that vary slightly in their feature values. After the perturbed instances are obtained, and assuming a surrogate model, said instances are then materialized in the subsequent moves of the LIME. Instance perturbation allows LIME to achieve its aims of approximating the decision boundary of the black-box model by sampling the local neighborhood around the instance to be explained to generate human-intelligible rationales for individual predictions.

- **Model surrogate layer:** The perturbed instances are next passed to the surrogate model which is usually a simpler and more interpretable model than the original black-box model. Surrogate models can be constructed using linear regression algorithms, decision trees, or even fancy shallow neural networks. They are trained to learn about the functionality of the black-box model in some locality around the chosen instance. After training a surrogate model on the perturbed instances, it can be leveraged to model the black-box model's behavior around the point in question. Next, by inspecting the coefficients or feature importance of this surrogate model, one can also reveal the reasons behind their behavior on the black-box model's output values.

- **Feature importance layer:** The Feature Importance Layer of LIME is tasked with analyzing the coefficients or feature importance scores of the surrogate model developed in the Model Surrogate Layer. Therefore, the purpose of this layer is to ascertain the extent to which each feature contributes to the prediction done by the surrogate model regarding the instance in question. Obtained through the surrogate model, the feature importance scores can be utilized to determine the most important and least essential features towards the black-box model's prediction. The features with higher importance scores are thought to influence the prediction, while those with lower scores are assumed to impact it less. The Feature Importance Layer is an essential part of the LIME method as it draws attention to the reasons behind the black-box model's predictions. Understanding which features are key to predictions of an individual model can prove useful to analyze the decision-making procedure of complex machine learning models and build more trust in their predictions.

- **Local explanation layer:** Local Explanation Layer is the key component that generates local explanations for single predictions of B-level black-box machine-learning systems. Its purpose is to shed light on why a specific prediction was made, studying local behaviour around the instance of interest. Given that the surrogate model is trained with the perturbed instances and the feature importance scores have been obtained, the Local Explanation Layer then uses these to explain the prediction of the black-box model. The Local Explanation Layer is so critical in the LIME methodology since it helps explain the individual predictions of a black-box machine learning model. As it explains the active factors driving the machine learning model prediction, the user learns the base behind the prediction and builds more trust in the model.

The architecture of LIME is quite flexible. It can be used in conjunction with different types of machine learning models and data domains. Therefore, LIME can easily allow for the provision of LIEs which will help

users understand how complex models make decisions, increase their faith in the validity of the predictions, and detect and correct potential biases or mistakes.

c. Proposed segmentation model

Since melanoma, a type of skin cancer is characterized by abnormal borders and an uneven distribution of color, accurately segmenting it and determining its extent is crucial to diagnosing and planning treatment for the patient. In this Phase, the proposed CNN model concludes its representations to demarcate the edges of melanoma lesions identified in dermoscopic images. Consequentially, the model conducts feature extraction and spatial investigations to determine areas on the dermoscopic image within which melanoma may be growing. The proposed CNN model for melanoma segmentation is shown in Figure-6.

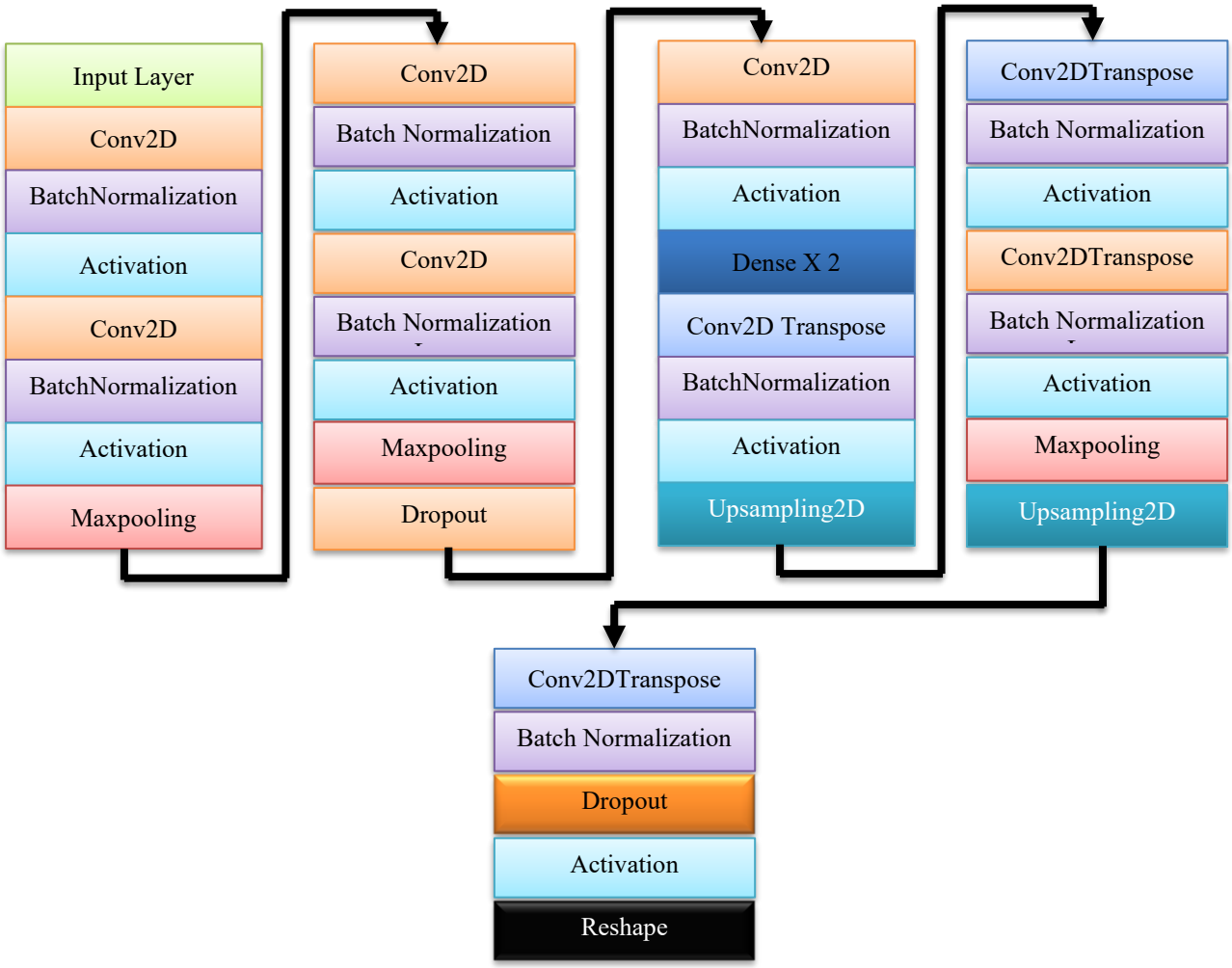


Figure-6. Proposed CNN Architecture.

Conv2D: The Conv2D layer is the basic building block in convolutional neural networks which are used to capture complex patterns and features from the input images. The Conv2D layer functions like a sliding window and slides a kernel or a filter over the input image domain with element-wise multiplication and sum. These operations help to retrieve different features from various spatial positions in the image, such as edges, textures, or patterns. By adjusting and learning the filter weights during the training phase, the network learns to identify more complicated representations of the input data. Many parameters of the Conv2D layer function as hyperparameters that define its behavior and the network architecture. The size of the filters defines the spatial extent of features blended. The quantity of filters defines the depth of the feature maps produced. Other parameters, including padding and strides, also define the output dimensions of the feature maps, ultimately determining the receptive field of the network and its computational efficiency. Artificial neural networks excel at learning representations and features in a hierarchical manner. Consequently, the Conv2D layer forms the basis of many CNN architectures because it allows them to achieve the top results on various computer vision applications, including image classification, object detection, and semantic segmentation.

Batch normalization layer: Batch Normalization layer is used to stabilize and speed up the

training. Specifically, BatchNormalization normalizes the activations of each layer across mini batches during training. This prevents the problem of internal covariate shift, where the distribution of activations within a layer changes during training significantly. If not prevented, internal covariate shift makes it difficult for the network to converge and causes the use of lower learning rates to prevent the model from diverging. Nonetheless, BatchNormalization addresses internal covariate shift by allowing the activations to be within a normalized range, hence a smooth and faster training. Furthermore, BatchNormalization serves as a regularizer which means it can eliminate the need for other techniques such as dropout or weight decay. Normalizing activations increases the stability of the training process by introducing a form of noise that helps to prevent overfitting and improves the generalization ability of models. Moreover, BatchNormalization makes it possible to use higher learning rates which further increases the convergence speed, reducing the training time required for optimal performance. Thus, the integration of BatchNormalization layers into neural network structures is now ubiquitous due to technological advancements and has provided a significant boost to the stability, efficiency, and performance of deep learning in almost any domain and task.

Activation layer: This layer is responsible for bringing non-linearity to the network's computations. The activation layer applies element-wise activation functions to the output of the previous layer, serving complex dependencies and assisting the network in learning and representing the complex patterns within the data. One of the most frequently used activation functions is the Rectified Linear Unit, or ReLU, which turns negative values to zero and positive values to themselves. ReLU is the simplest of activation functions, and its computation is easy, while during training, it solves the vanishing gradient problem; therefore, being broadly used.

MaxPooling2D layer: To effectively reduce the spatial dimensions of feature maps and computation, the MaxPooling2D layer down samples feature maps while ensuring that the most salient information is retained. The input is partitioned into non-overlapping regions, and the maximum value in all these regions is output. This enables to specially reduce the computation to avoid overfitting. However, it also has the effect of translation invariance that allows the model to extract the most discriminative features while ignoring replaying the irrelevant features. MaxPooling2D also enables increased computational efficiency by downsampling. This makes CNN architectures scalable to larger data sets and deeper architectures since the number of parameters to train is significantly reduced. The max pooling also helps the CNN to capture hierarchical statistics by aggregating higher level local patterns in the convoluted representation of the input data.

Dropout Layer: The main purpose of the Dropout layer is to reduce the propensity to overfit by dropping out or excluding a certain fraction of input units from the training phase. However, since for each new input the Dropout edge uses a different random mask of units that are removed, it acts as an almost ensemble averaging of different neural architectures. Thus, when some of the units are excluded, unity develops, and the whole is enhanced, preventing units from co-adapting to do this too great. It significantly diminishes harmful connections between neurons and especially complicates training, mostly reducing the capabilities of the network by forcing the training to converge more achieved, but up to the end of the training, more than simply throwing everything into the mixer and stopping. Also, Dropout deals with learning to generalize by covering the eternal dependence between one input and other features provided the proper feature set, but the network learns and hence utilizes the other learned features more easily, decreasing dependency.

Dense layer: The Dense layer permits high-level abstraction and complex decision-making. Every neuron within a dense layer is attached to each neuron in the preceding and succeeding layers, ensuring the flow of information across the entire network. By learning weights linked with each connection, the dense layer conducts a series of transformations that transform the input data into a higher-dimensional feature space, simplifying the extraction of complicated patterns and correlations. Dense layers are often used in the last set of a neural network to allow the network to predict or classify utilising the learned features drawn from the input data. Furthermore, as a result of achieving non-linear relationships within features, the Dense layer can be adapted to solve various problems across several domains, including image classification, natural speech language comprehension, and regression analysis.

Conv2DTranspose layer: The Conv2DTranspose layer is commonly known as a deconvolutional layer, and it is instrumental in several tasks such as image segmentation, generative modeling, and image super-resolution. The typical Conv2D layer downsamples the input through a series of convolutional operations, and Conv2DTranspose upsamples the input by increasing the spatial dimensions of the input feature maps. The expansion is obtained by using a layer of learnable parameters, expressed as filters, which convolved at the input to rebuild the original spatial resolution. Since such filters are learned during backpropagation, the Conv2DTranspose layer can recover fine-grained details lost during the downsampling phases above thus allowing the network to generate higher-resolution outputs. This same property of the Conv2DTranspose layer enables the creation of novel images with generative models through the transformation of low-dimensional feature vectors into high-dimensional ones. This mechanism gives tools for the generation of realistic and diverse samples in generative tasks like image generation or image-to-image translation.

UpSampling2D layer: UpSampling2D layer, which enables expanding the feature maps into higher resolutions. This is done by replicating the rows and columns of the input feature maps, hence increasing the size of the spatial dimension of the data, allowing more

fine-grained spatial patterns and details to be retained and passed through the network. Unlike MaxPooling or Conv2DTranspose layers that are trained to perform downsampling and upsampling, UpSampling2D simply replicates the input data and does not have any parameters; therefore it is computationally efficient and easy to use. UpSampling2D is often combined with the Conv2DTranspose layers to upsample the spatial dimension to be of the same size as the feature maps before applying CNNs, allowing the network to capture spatial patterns and relationships. Aside from that, UpSampling2D also helps to increase feature map resolution which enhances the network's performance in tasks such as image segmentation, image generation, and high-resolution pictures.

Reshape layer: The role of the Reshape layer is to change the shapes of the tensors to match a specific given formulation or dimension. Reshaping the input tensor into a desired form implies that it will be used to operate with another layer within the same network, which ensures the transmission of information and computation. Such a layer is suitable in cases where the current shape of an outlay must equal that of an input layer, especially during the last phase of the network when the output is being prepared for classification or regression. Moreover, the Reshape layer allows for changing the shape of the dimensions of a tensor and building decision trees based on the dimensions. This offers a flexible design of the model and eases the integration of different architectural forms. Instead, the shape of the data is determined and redefined in the convolution and pooling chapters, making for its efficient and effective processing in the workflow of specific tasks and applications.

4. SIMULATION RESULTS

This section describes the simulation results of the proposed model. The fundamental element of the proposed model is the explainable AI approaches using the LIME technique which is incorporated to clarify classification decisions. The local surrogate models identify the most significant features of each prediction as described above, which helps explain the model's predictive rationale. This level of detail allows clinicians to produce more educated judgments; there is a stronger alignment with the model's predictions and boosting confidence, supporting expert systems to apply within a clinical scenario. The classification results obtained from the DenseNet201 model are considered as input for the LIME technique. From the classification results melanoma images were taken and each image has been transformed into a grid of pixels. The CNN model is used to interpret these images. The sample images considered as input are depicted in Figure-7.

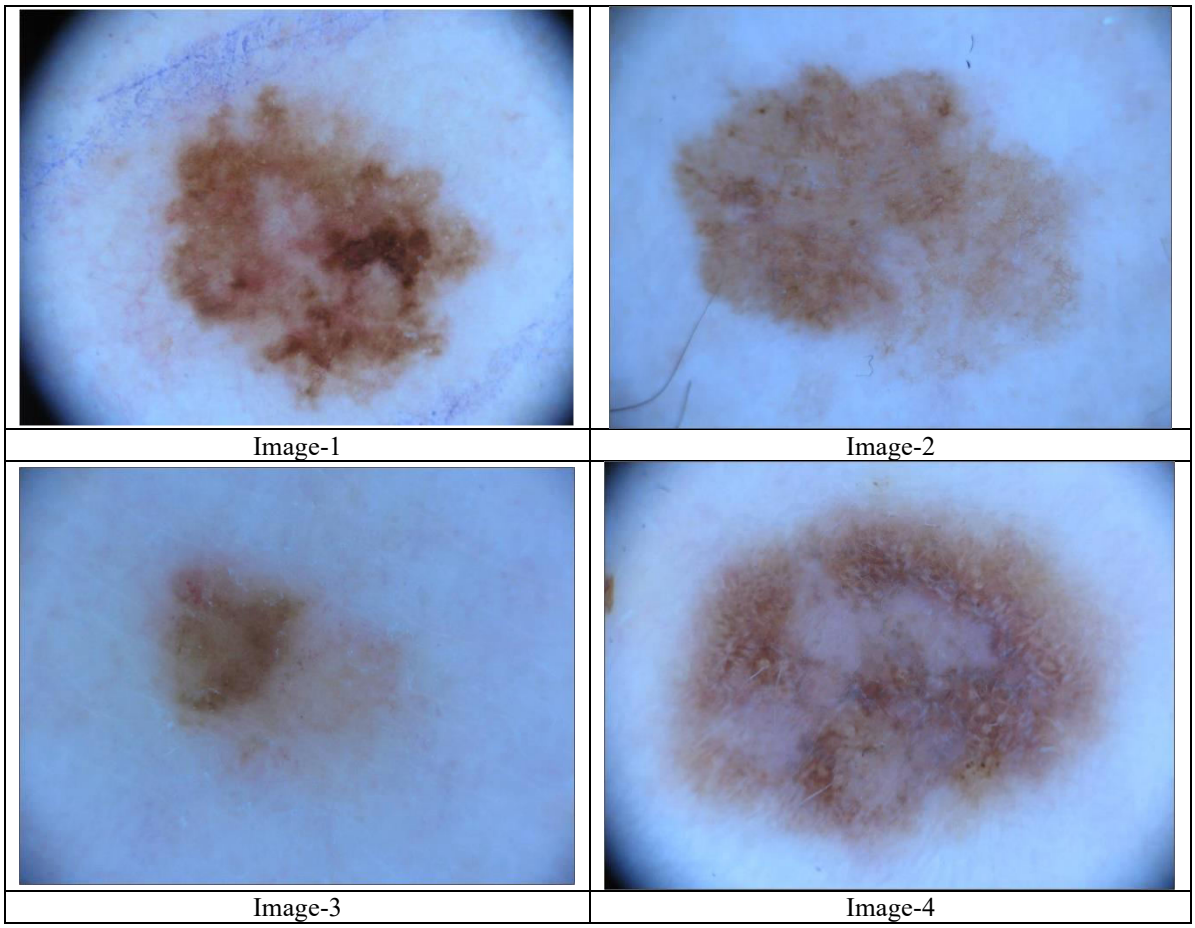
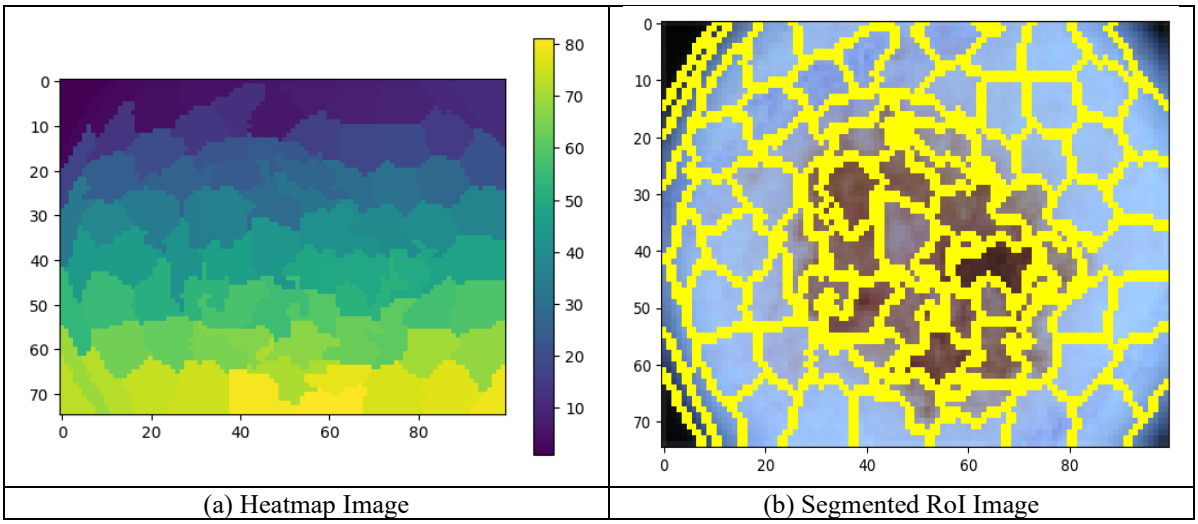


Figure-7. Sample input images.

The first step in the LIME technique consists of choosing an image to explain. These perturbed samples are intended to give an approximation of how the model will act in the vicinity of the given instance: slight variations are applied to its input features while preserving the original label. For an image, the input features may be pixel values perturbed by noise or color intensities.



(a) Heatmap Image (b) Segmented RoI Image

Figure-8. Heatmap and segmented RoI of input image.

Figure-8(a) represents the heatmap of the input image, which displays the data across different colors on the grid with the unique value of the data points. This plot easily identifies the patterns or gradients for the data points. This plot would be applicable for LIME visualization on the data points for the feature importance and localized explanation for the image-based data. LIME works by perturbing the input image to come up with an array of perturbations, which can be viewed as differently colored or gradient regions of the plot. In other words, the perturbations represent different areas of interest in the image. From the newly formed image, namely the perturbed instances, predictions are obtained through the model to indicate the areas where the predictions change drastically. Figure-8(b) is a visualization of a segmented region of interest on an input image. The yellow line in Figure-8(b) that denotes the segmented lines giving the boundaries between different regions. The dark field at the center may be the lesion or the area of interest, while the rest of the regions may imply portions of the analyzed object that are less of a factor for the model's decision.

Afterward, these perturbed samples are used as input for the original black-box model, resulting in a prediction for each of them. As each prediction was made for the perturbed sample, the difference in the input features can be mapped to see the difference in the model's prediction. Subsequently, the predictions can be collected, and the point of interest can be changed to focus on explaining the feature's importance. This explanation is done by finding a simpler model, like a linear model or decision tree. The selected model is called a surrogate model and is trained using the perturbed samples and resulting output to mimic how the black-box model behaves in the neighborhood of the picked instance. The prediction result of the LIME technique is shown in Figure-9.

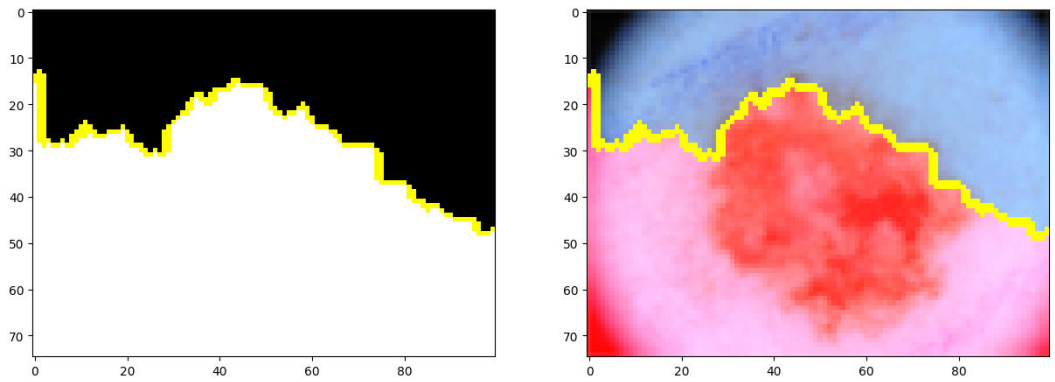


Figure-9. Prediction output of LIME technique.

Figure-9 shows a visual representation of image segmentation and the feature importance of detection. The leftmost is the binary mask that locates the region of interest, followed by the original image with that region highlighted for contextualization. The rightmost image is the heatmap, which depicts the importance scores/confidence against the entire image giving a deeper concept of which areas in the image are important for the model's decision. After completion of the Explainable AI LIME technique, the melanoma images are accurately segmented by using the proposed CNN model. By the proposed CNN-based model, segmentation capabilities, especially proper boundaries of melanoma lesions, were tested. With the help of Conv2DTranspose and UpSampling2D layers, the model detected abnormal borders and color distribution of melanoma. Consequently, segmented images yielded a more thorough and informative diagnosis and ensured accurate treatment options, and thus, its usability in clinical applications was revealed. The input images along with mask images are augmented by rotating and flipping. The corresponding augmented images are depicted in Figure-10.

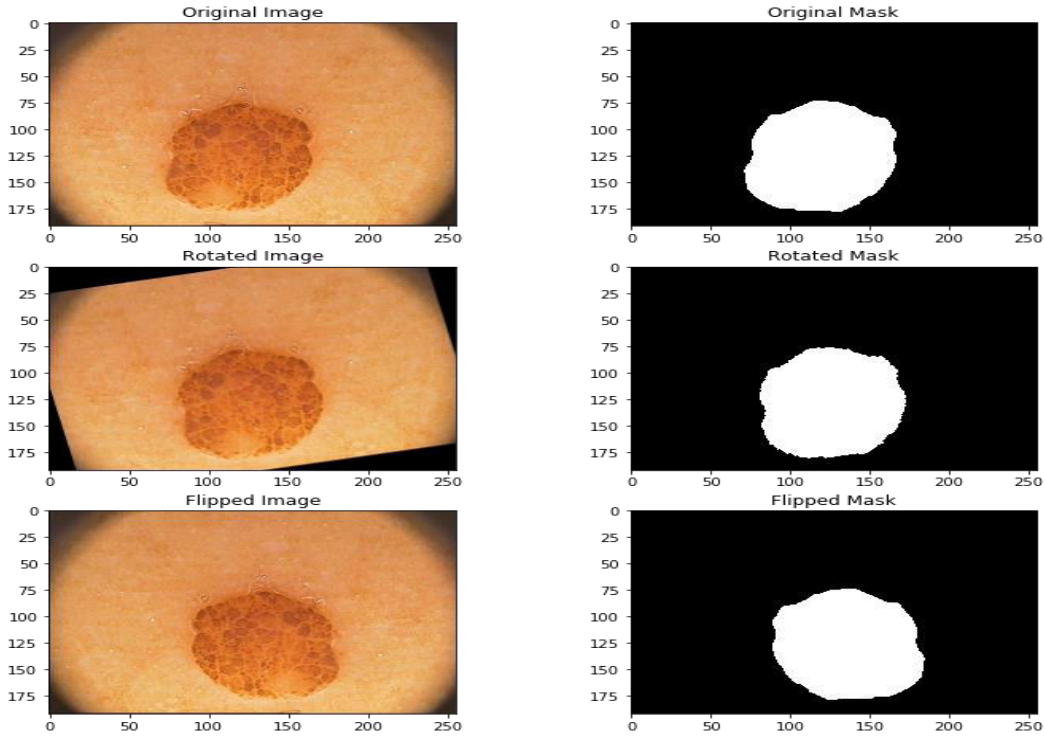
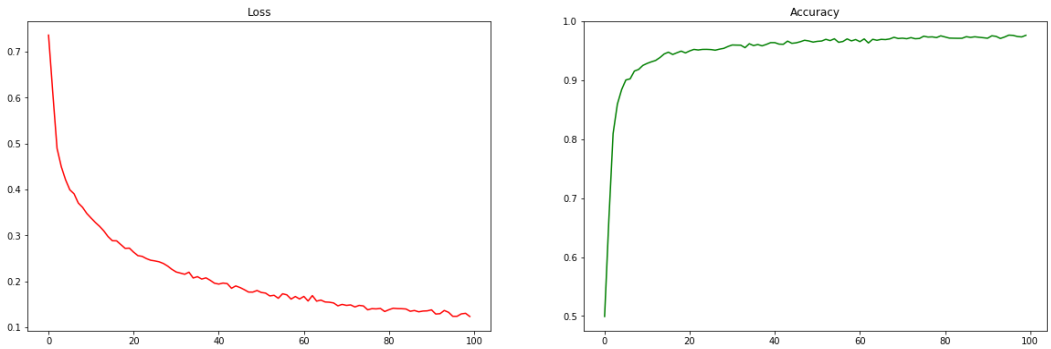


Figure-10. Augmented input and mask images.

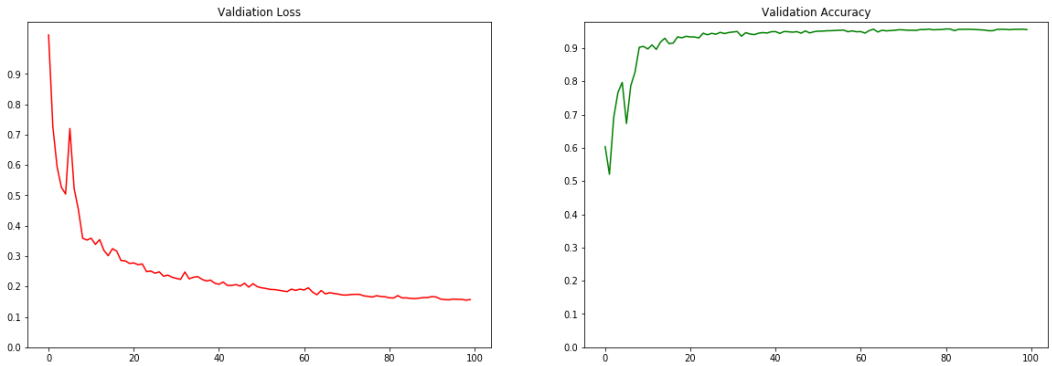
The training loss plot depicted in Figure-11(a), visualizes how the loss value of the model behaves on each epoch or training step. The commonly-used metric is a loss, which is a quantification of a model's prediction alignment with the correct labels. Thus, low loss is an indicator of a better model. The x-axis, in turn, captures the epochs or iterations while the y-axis visualizes the loss values.



(a) (b)

Figure-11. Training loss and accuracy plots.

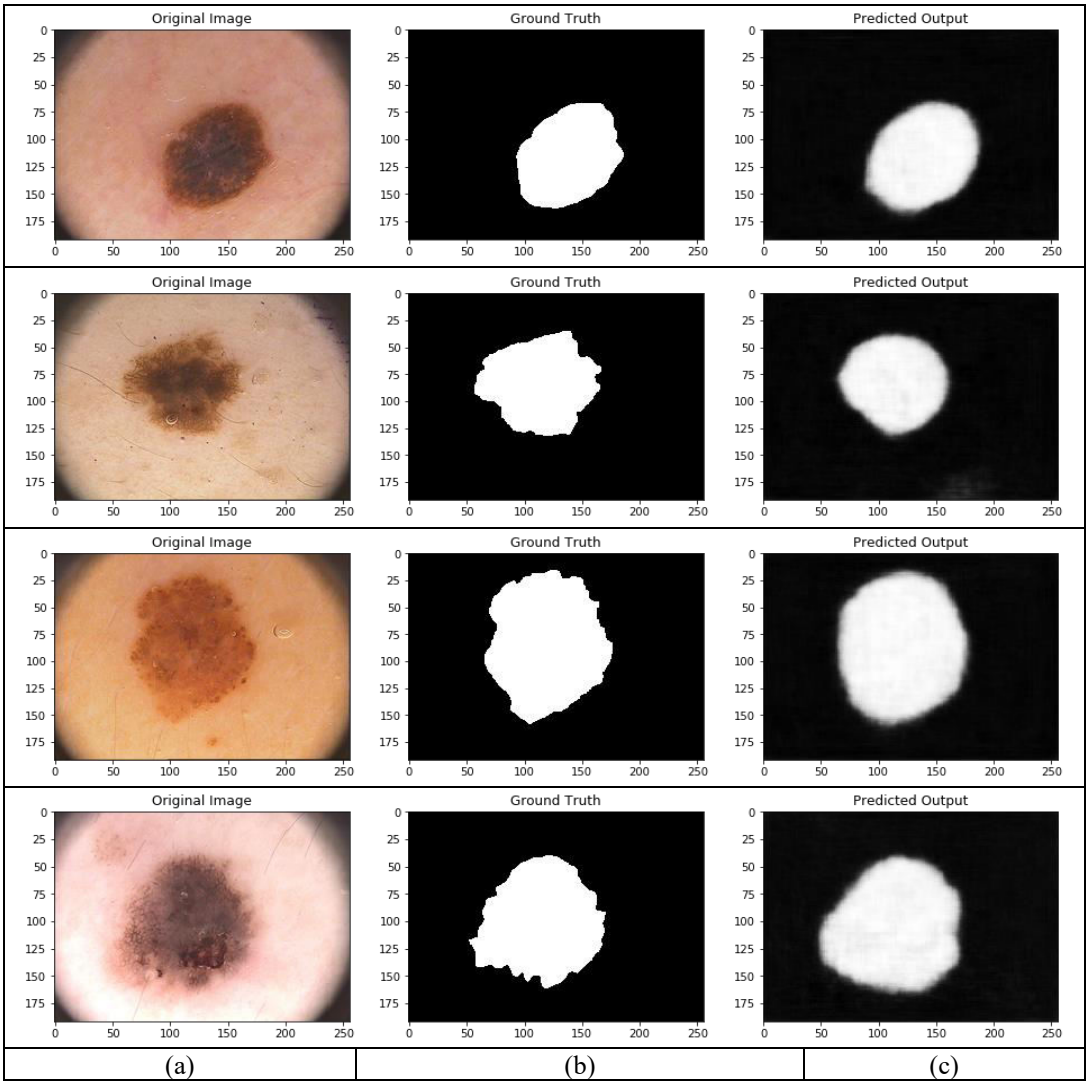
The Training accuracy plot shown in Figure-11(b) indicates how the model's accuracy develops through the training epochs. The training accuracy is compared to the validation accuracy to understand how the model performs on unseen data.



(a) (b)

Figure-12. Validation loss and accuracy plots.

The Validation accuracy plot shown in Figure-12(b) indicates how the model's accuracy after performing the validation. After training the segmented results are shown in Figure-12.



(a) (b) (c)

Figure-13. Segmented results of proposed CNN model.

Figure-13 represents the segmented results of the proposed CNN model. The Figure-13(a) is the original image of melanoma and Figure-13(b) is the binary mask/ ground truth image of the original image. The Figure-13(c) represents the segmented/ predicted output of the proposed CNN segmentation model. In the final stage, the segmented outputs are enhanced and correlated with the ground truth images. The corresponding enhancing images are shown in Figure-14.

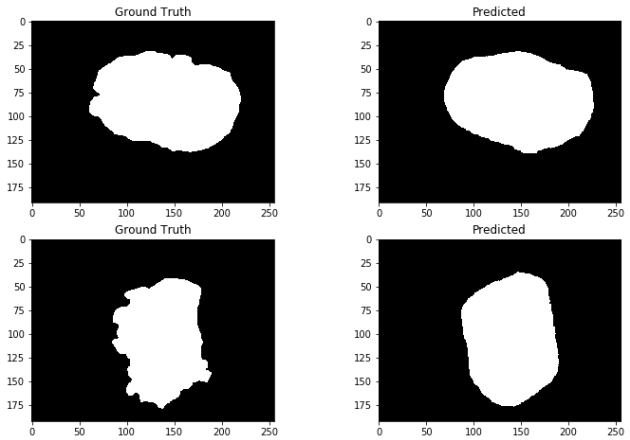


Figure-14. Enhanced segmented results.

Table-1. Performance metrics of the proposed model.

Parameter	Train Set	Test Set
IOU	97.20	94.32
Dice Coef	85.50	81.39
Precision	96.55	91.05
Recall	97.31	92.47
Loss	11.09	17.04
Accuracy	98.06	94.90

By applying enhancement techniques to the segmented regions, their visibility and clarity are improved. The techniques may involve sharpening the boundaries, improving contrast, or removing noise to make the features more distinct. These enhanced processes are undertaken to produce images that are clearer and interpretable and can be analyzed more easily and accurately. The enhanced images in visual form are presented in Figure 14, which makes readers and researchers judge the segmentation and enhancement quality. The enhanced image is compared with the ground truth, allowing a visual assessment of the efficiency and accuracy of the image analysis. The performance metrics of the proposed model are reported in Table-1.

Table-1 illustrates the performance of the proposed model by presenting several metrics for both the training set and the test set. The IOU values for the training set and test set are 97.20% and 94.32%, respectively. Based on this metric, the degree of overlap between the predicted and actual segments is indicated. The IOU recognition signifies that the model achieves higher accuracy in predicting the segment boundary for both datasets. The Dice Coefficient values with percentages of 85.50% and 81.39% for the training set and test set affirm the production of predictions that significantly match the actual segments.

The Precision metric, which depicts the percentage of properly predicted positive segments out of the overall predicted positives is 96.55% and 91.05% on the training and test set, respectively. Therefore, the higher rate indicates the model's efficiency in properly identifying positive segments. Similarly, the Recall metric, which shows the percentage of properly predicted positive segments out of the whole positive segment, is approximately 97.31% and 92.47% on the training and test set, respectively. Hence, this implies that the proper identification of true positives is greater in both datasets. Loss is the measure of error of the model's prediction with a loss value of 11.09 on the training set and 17.04 on the test set. Where lower values imply better performance, and the model's performance on the test set is poor, it is not necessary to worry about it as it is natural behavior. It often reveals that the model has learned some pattern in the training dataset. Accuracy is the ratio of correct predictions per number of all samples. Here, the accuracy is 98.06% and 94.90% for the training and the testing dataset, respectively. This performance level indicates that the model generalizes properly, reaching high correctness in its predictions. Although the correctness of the test data is lower than that of the training one, the model fulfills its purpose of task performance. The proposed model is compared with other models and the corresponding comparison table is reported in Table-2.

Table-2. Comparison of performance metrics.

Model Name	IoU	Accuracy
FPN [19]	89.78	93.7
MANet [20]	89.0	93.3
Unet [21]	83.31	91.9
Proposed CNN	94.32	94.90

Thus, in Table-2, a comparison is shown where two key performance metrics are taken and compared for different models. Each model has both its approach to the task and a certain level of "effectiveness". So, Feature Pyramid Network [19] has an IoU of 89.78% and an accuracy of 93.7%. The second model displayed in the table is MANet [20]. It has an IoU of 89.0% and an accuracy of 93.3%. The U-Net [21], which has IoU = 83.31% and accuracy = 91.9%. The results of the proposed CNN model, which has IoU 94.32 and an accuracy of 94.90%, are represented in the last lines of the table. It should be noted that in both cases, the proposed model has the highest values, i.e. it most accurately identifies the overlap and maximally true predictions.

5. CONCLUSIONS

To conclude, the proposed framework for improving the prediction, classification, and segmentation of skin lesions, specifically melanoma, has great potential. The proposed model using a modified DenseNet201 has been shown to predict and classify skin conditions accurately and the explainable AI method of LIME has been shown to enhance the transparency of the model for interpretability by the clinicians. Particularly, the segmentation module of the boundary of the melanoma lesion provides valuable information about lesion size and its extent for planning therapy. The performances of the proposed model with those of other models in prior literature and proved their superiority. Therefore, the proposed framework has a high potential for practical applications in clinical settings. By making highly accurate and explainable predictions and segmentation, the current framework can be highly conducive to skin disease diagnosis and treatment, which can lead to favorable outcomes for medical practitioners and patients alike.

REFERENCES

[1] Singh Satya P., Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan and Balázs Gulyás. 2020. 3D deep learning on medical images: a review. Sensors. 20(18): 5097.

[2] Heibel Haley D., Leah Hooey, and Clay J. Cockerell. 2020. A review of noninvasive techniques for skin cancer detection in dermatology. American journal of clinical dermatology. 21(4): 513-524.

[3] Dildar Mehwish, Shumaila Akram, Muhammad Irfan, Hikmat Ullah Khan, Muhammad Ramzan, Abdur Rehman Mahmood, Soliman Ayed Alsaiari, Abdul Hakeem M. Saeed, Mohammed Olaythah Alraddadi and Mater Hussen Mahnashi. 2021. Skin cancer detection: a review using deep learning techniques. International journal of environmental research and public health. 18(10): 5479.

[4] Goceri Evgin. 2021. Diagnosis of skin diseases in the era of deep learning and mobile technology. Computers in Biology and Medicine. 134: 104458.

[5] Adegun Adekanmi and Serestina Viriri. 2021. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of the state-of-the-art. Artificial Intelligence Review. 54(2): 811-841.

[6] Alam Talha Mahboob, Kamran Shaukat, Waseem Ahmad Khan, Ibrahim A. Hameed, Latifah Abd Almuqren, Muhammad Ahsan Raza, Memoona Aslam and Suhuai Luo. 2022. An efficient deep learning-based skin cancer classifier for an imbalanced dataset. Diagnostics. 12(9): 2115.

[7] Hauser Katja, Alexander Kurz, Sarah Haggenmüller, Roman C. Maron, Christof von Kalle, Jochen S. Utikal, Friedegund Meier et al. 2022. Explainable artificial intelligence in skin cancer recognition: A systematic review. European Journal of Cancer. 167: 54-69.

[8] Duggani Keerthana and Malaya Kumar Nath. 2021. A technical review report on deep learning approach for skin cancer detection and segmentation. Data Analytics and Management: Proceedings of ICDAM. 87-99.

[9] Nigar Natasha, Muhammad Umar, Muhammad Kashif Shahzad, Shahid Islam and Douhadji Abalo. 2022. A deep learning approach based on explainable artificial intelligence for skin lesion classification. IEEE Access. 10: 113715-113725.

[10] Ahmad Naveed, Jamal Hussain Shah, Muhammad Attique Khan, Jamel Baili, Ghulam Jillani Ansari, Usman Tariq, Ye Jin Kim and Jae-Hyuk Cha. 2023. A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. Frontiers in Oncology. 13: 1151257.

[11] Ballari Gayatri Shrinivas, Shantala Giraddi, Satyadhyan Chickerur and Suvarna Kanakareddi. 2022. An Explainable AI-Based Skin Disease Detection. In ICT Infrastructure and Computing: Proceedings of ICT4SD 2022, pp. 287-295. Singapore: Springer Nature Singapore.

[12] Zia Ur Rehman, Muhammad, Fawad Ahmed, Suliman A. Alsuhibany, Sajjad Shaukat Jamal, Muhammad Zulfiqar Ali and Jawad Ahmad. 2022. Classification of skin cancer lesions using explainable deep learning. Sensors. 22(18): 6915.

[13] Hosny Khalid M., Wael Said, Mahmoud Elmezain and Mohamed A. Kassem. 2024. Explainable Deep Inherent Learning for Multi-Classes Skin Lesion Classification. Applied Soft Computing. 111624.

[14] Lu Xinrong and Y. A. Firoozeh Abolhasani Zadeh. 2022. Deep learning-based classification for melanoma detection using XceptionNet. Journal of Healthcare Engineering 2022.

[15] Alwakid Ghadah, Walaa Gouda, Mamoona Humayun and Najm Us Sama. 2022. Melanoma detection using

deep learning-based classifications. In Healthcare. 10(12): 2481. MDPI.

[16] Adla Devakishan, G. Venkata Rami Reddy, Padmalaya Nayak and G. Karuna. 2022. Deep learning-based computer aided diagnosis model for skin cancer detection and classification. Distributed and Parallel Databases. 40(4): 717-736.

[17] Bibi Amina, Muhammad Attique Khan, Muhammad Younus Javed, Usman Tariq, Byeong-Gwon Kang, Yunyoung Nam, Reham R. Mostafa and Rasha H. Sakr. 2022. Skin lesion segmentation and classification using conventional and deep learning-based framework. Comput. Mater. Contin. 71(2): 2477-2495.

[18] Yacin Sikkandar, Mohamed, Bader Awadh Alrasheadi, N. B. Prakash, G. R. Hemalakshmi, A. Mohanarathinam and K. Shankar. 2021. Deep learning based on an automated skin lesion segmentation and intelligent classification model. Journal of ambient intelligence and humanized computing. 12: 3245-3255.

[19] Rehman Hafeez Ur, Nudrat Nida, Syed Adnan Shah, Wakeel Ahmad, Muhammad Imran Faizi, and Syed Muhammad Anwar. 2022. Automatic melanoma detection and segmentation in dermoscopy images using deep RetinaNet and conditional random fields. Multimedia Tools and Applications. 81(18): 25765-25785.

[20] Sun Yongheng, Duwei Dai, Qianni Zhang, Yaqi Wang, Songhua Xu and Chunfeng Lian. 2023. MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation. Pattern Recognition. 139: 109524.

[21] Alahmadi Mohammad D. 2022. Multiscale attention U-Net for skin lesion segmentation. IEEE Access. 10: 59145-59154.