IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.DOI

A Deep Learning approach based on Explainable Artificial Intelligence for Skin Lesion Classification

NATASHA NIGAR[1], MUHAMMAD UMAR [2], MUHAMMAD KASHIF SHAHZAD [3], SHAHID ISLAM [1], DOUHADJI ABALO [4]

[1]Department of Computer Science (RCET), University of Engineering and Technology, Lahore 39161, Pakistan
[2]TEKHQS INC, Irvine, California, USA
[3]Power Information Technology Company (PITC), Ministry of Energy, Power Division, Govt. of Pakistan, Lahore, 39161, Pakistan
[4]University of Lomé, P.O.Box 1515, Lomé, Togo

Corresponding author: Douhadji Abalo (e-mail: douhadjiabalo@gmail.com)

ABSTRACT The skin lesion types result in delayed diagnosis due to high similarity in early stages of the skin cancer. In this regard, deep learning algorithms are well-recognized solutions; however, these black box approaches result in lack of trust as dermatologists are unable to interpret and validate the decisions made by the models. In this paper, an explainable artificial intelligence (XAI) based skin lesion classification system is proposed to improve the skin lesion classification accuracy. This will help the dermatologists to make rational diagnosis in the early stages of skin cancer. The proposed XAI model is validated using International Skin Imaging Collaboration (ISIC) 2019 dataset. The developed model correctly identifies the eight types of skin lesions (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 as 94.47%, 93.57%, 94.01%, and 94.45% respectively. These predictions are further analyzed using the local interpretable model-agnostic explanations (LIME) framework to generate visual explanations that match a prior belief and general explanation best practices. The explainability integrated within our model will enhance its applicability in real clinical practice.

INDEX TERMS explainable artificial intelligence, skin lesion classification, deep learning

I. INTRODUCTION

The skin cancer is a type of cancer that affects the surface of the skin. More than 5 million people in the United States have been diagnosed with skin cancer [1]. Thus, the improvement in the diagnostic accuracy and the rate of early diagnosis is a crucial task. In this regard, both medical experts and researchers are putting their great efforts in advancing medical diagnosis, treatments, and examinations [2].

Skin lesion is the abnormal appearance or growth of skin compared to the skin area around it. Lesions can differ in type, texture, color, shape, affected location and distribution. They are classified into 2,032 categories that is organized into a hierarchy [3] as shown in Fig. 1. The first two main categories of this hierarchy are: melanocytic and non-melanocytic. Melanocytic (i.e., pigmented) or non-melanocytic (i.e., non-pigmented) is based on the presence or the lack of melanocytes and melanin pigment in the lesion, respectively. Melanocytic lesions have 8 global features which aid in the detailed classification of pigmented skin lesions, and 14 local features that give more accurate information about a given lesion [4]. Non-melanocytic lesions can appear yellow or orange due to keratin; or red, purple, blue and black due to hemoglobin [5]. Lesions could be cancerous (i.e., malignant) or non-cancerous (i.e., benign).

Dermoscopy is one of the most widely used skin imaging techniques to improve the diagnostic performance and reduce skin cancer deaths [6]. It is a non-invasive method in which a magnified and well illuminated picture of skin is taken to clearly see and understand the lesion area [7]. This technique is usually used to diagnose the skin cancer in early stages and enhances the diagnostic ability of the doctors. Usually, dermatologists analyze the dermoscopic images (aka biomedical images) through visual inspection, which requires a high degree of skill and concentration, and is time-consuming and prone to operator bias [8]. The reason is that the skin infected parts and normal moles are so similar that sometimes it is

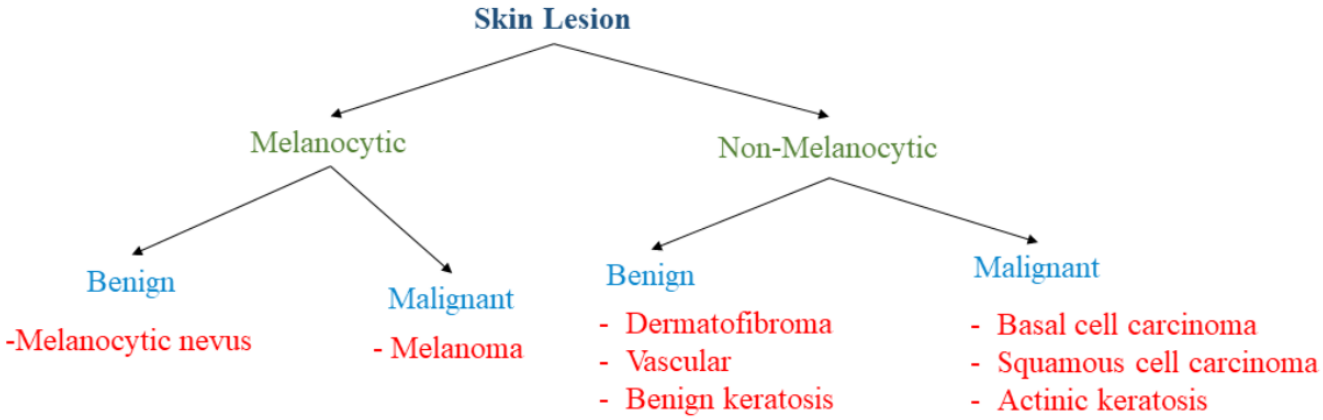


FIGURE 1: Skin Lesions Hierarchy

hard to make an accurate diagnosis.

In order to assist the dermatologists to diagnose the skin cancer, many computer aided diagnosis (CAD) systems [9]–[13] have been developed, not only bypassing aforementioned issues but also improving the accuracy, efficiency and objectivity of the diagnosis system. In this regard, deep learning (DL) algorithms have shown promising results and large potential for image processing and data analysis. DL has been widely used due to its popularity and unique features in many complex domains e.g., detection, identification, classification, and recognition of objects [14]. It is a machine learning (ML) technique that adds more 'depth' (complexity) into the model and transforms the data using various functions that allow data representation in a hierarchical way, through several levels of abstraction [15]. DL can solve more complex problems in a fast and efficient manner due to more complex models employed [16]. DL algorithm such as convolutional neural networks (CNNs) and image processing techniques are the most important part of common CAD systems [17].

However, the use of such CAD systems by dermatologists and patients remains doubtful because the processing cycle behind model learning and features encoding is not well understood. The DL model without a rational explanation is a barrier for dermatologists in accurate decision making. Occasionally, the experts find it difficult to understand the predictions made by the model. For example, a DL model with 87% accuracy result for the diagnosis of skin cancer, is frequently difficult to understand that why the DL model produces inaccurate results in the remaining 13% of cases, and how to improve these decisions. The DL models are not always similar or representational of dermatologists' decision-making processes. Hence, these models are often deemed as a 'black box' nature of ML algorithms, which does not give clear explanation for its conclusions. The lack of model transparency associated with DL algorithms in the complete cycle of decision-making cannot be neglected in skin cancer diagnosis. It is, therefore, needed to develop such robust approaches to better understand the black box decisions. Such approaches are commonly referred to as interpretable deep learning or XAI [18].

In this study, the state-of-the-art pre-trained deep learning algorithm ResNet-18 [19] is applied on on ISIC 2019 dataset classifying 8 skin lesions, using LIME as an explanation method with enhanced explanation and accuracy. We train the model deeply to resolve the problem of imbalance dataset and showed their effect on the accuracy of the model. In summary, we present a robust model with enhanced accuracy with the involvement of XAI techniques in the skin cancer diagnosis and makes the following main contributions:

- Model Transparency: An XAI model is developed employing LIME framework and ResNET-18 i) to explain that why a deep learning model is predicting particular skin lesion, and ii) to increase the model accuracy which can lead to increase the level of trust, thus, increasing the safety of the diagnostic system.
- Data Set: The developed approach is tested with 25,331 dermoscopic images using ISIC 2019 dataset.

This paper is organized as follows: Section II overviews the background and related work highlighting the merits and limitations of existing methods. The developed model is explained in section III followed by experimental analysis in section IV. The threats to validity of this study are in section V. Finally, section VI concludes this study with future directions.

II. LITERATURE REVIEW

A. BACKGROUND

1) Lime Framework

In this paper, LIME (local interpretable model-agnostic explanations) is used as an XAI (eXplainable AI) method. It is a post hoc method which is applied after the model is trained [31]. Moreover, model-agnostic refers to the group of explainers that are not specifically designed for a certain ML algorithm and has wide scope [31].

The LIME [32] is a popular technique for interpreting and explaining the black box decisions made by the ML algo-

TABLE 1: Research Matrix of Related Work

Study	Dataset(s)	Methodology	Results	Skin Classes	XAI method
Chowdhury et al. [20]	HAM10000	custom CNN	accuracy 82.7%	7 classes	CAM
Esteva et al. [21]	ISIC 2018	CNN	AUC 94%	7 classes	Backpropagation
Li et al. [22]	ISIC 2017	CNN	Wilcoxon's Sign Rank Test	7 classes	CAM
Li et al. [6]	ISIC 2018	VGG16+ResNet-50	accuracy 85%	7 classes	Occlusion
Nunnari et al. [23]	ISIC 2019	VGG16, ResNet-50	accuracy 72.2%, 76.7%	8 classes	GradCam
Sadeghi et al. [24]	1021 images	ResNet-50	accuracy 60.94%	4 classes	CBIR
Xie et al. [25]	ISIC 2017, PH2	Modified version of deep CNN	accuracy 90.4%	3 classes	CAM
Yang et al. [26]	ISIC 2017	ResNet-50	accuracy 83%	2 classes	CAM
Young et al. [27]	HAM10000	Inception	accuracy 85%	2 classes	GradCam, Kernal SHAP
Zunair et al. [28]	ISIC 2016	VGG16	sensitivity 91.76%, AUC 81.18%.	2 classes	CAM
Brinker et al. [17]	ISIC 2018	CNNs	specificity 86.5%	1 class	No
Kassem et al. [9]	ISIC 2019	Deep CNN	accuracy 94.92%	8 classes	No
Kasani et al. [29]	ISIC 2018	ResNet 50	accuracy 92%	7 classes	No
Salido et al. [12]	PH2	CNN	accuracy 93%	3 classes	No
Shahin et al. [10]	ISIC 2018	Inception V3 + ResNet 50	validation accuracy 89.9%	7 classes	No
Sherif et al. [13]	ISIC 2018	Deep CNN	accuracy 96.67%	2 classes	No
Ünver et al. [30]	PH2, ISBI 2017	YOLO, Grab Cut	accuracy 93.39%	3 classes	No

rithms. The objective of LIME is to train surrogate models locally and explain an individual prediction [32]. The high-level structure of LIME is presented in Fig. 2. At the first step, a synthetic data set is generated by permuting the samples around an instance from a normal distribution in a random manner. This perturbed dataset is used by LIME to train an interpretable model (e.g., linear regression) followed by corresponding predictions are gathered using the black box model.

Linear regression is used to estimate relationships amongst dependent variables and multiple independent variables by utilizing a regression line as shown in Eq. (1).

$$y = a + bx_i \tag{1}$$

where y is dependent variable and x is independent variable, a is intercept, b is slope of the line and $i = 1, 2, ..., n$. The main purpose of this equation is to predict the value of target variable from given predictor variables. Further, the number of important features is given as input (K) to LIME to generate the explanation. The model is easier to understand with lower value of K. There are many techniques to select the K important features e.g., backward or forward selection of features and highest weights of linear regression coefficients. The forward feature selection method is used by LIME for small datasets having less than 6 attributes. For higher dimensional datasets, it uses highest weights approach [33]. The mathematical formulation of LIME is stated in Eq. (2).

$$explanation(x) = argmin\ g \in GL(f, g, \pi x) + \Omega(g) \tag{2}$$

where x represents the instance to be explained and g represents the interpretable model, the loss function L, also known as the fidelity function (e.g., mean squared error), calculates the explanation's closeness to the original model's prediction. In addition, G refers to a group of potentially interpretable models, such as decision trees. The neighbourhood size around the initial instance x is defined by the proximity measure x.

B. RELATED WORK

According to World Health Organization (WHO) [34], cancer is expected to be the leading cause of death (13.1 million) by 2030. The skin cancer is common in human beings which arises from the skin due to the abnormal growth of the cells that can easily invade and spread to the other parts of the human body [35].

Different methods have been presented and implemented in healthcare domain with focus on skin lesion classification over recent years. In this regard, Chowdhury et al. [20] used a custom CNN identifying 7 classes of skin diseases using HAM10000 dataset [36]. They used CAM [37] as an XAI method and maximum achieved accuracy is 82.7% and 78% of precision. Esteva et al. [21] used CNN to identify 7 classes while using ISIC 2018 dataset and Backpropagation [38] as explainable method. They achieved 94% Area Under Curve (AUC). Li et al. [22] used CAM [37] as an explainable method using ISIC 2017 dataset to detect 7 classes of skin diseases. However, they used Wilcoxon's sign rank test [39] to differentiate their results. Li et al. [6] incorporated Occlusion [40] as explainable method using ISIC 2018 dataset to diagnose 7 classes of skin diseases with accuracy rate of 85%, while using an ensembled VGG16 [41] and ResNet-50 [19].

Nunnari et al. [23] utilized GradCAM [42] as an explainable method with ISIC 2019 dataset and classifying 8 skin classes. They also used VGG16 [41] and ResNet-50 [19] as explanation models with 72.2% and 76.7% accuracy, respectively. Sadeghi et al. [24] used ResNet-50 [19] to identify 4 skin classes with 1021 dermoscopic images. They incorporated Content-Based Image Retrieval (CBIR) [43] as explanation method, with accuracy rate of 60.94%. Xie et al. [25] used CAM [37] as an explanation method to classify 3 skin diseases with a modified version of deep CNN and achieved average accuracy rate of 90.4%. They used ISIC

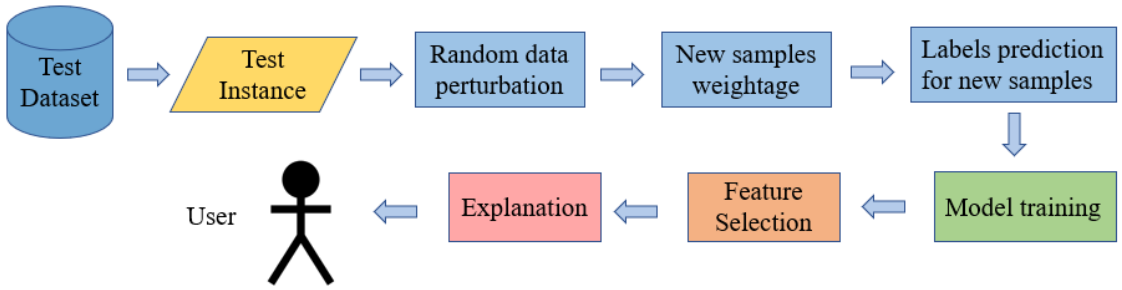


FIGURE 2: The workflow of LIME method

2017 and PH2 [44] datasets. Yang et al. [26] used ResNet-50 [19] along with CAM [37] as explanation method to classify 2 skin diseases using ISIC 2017 dataset with accuracy rate of 83%. Young et al. [27] used both GradCAM [42] and Kernel SHAP [45] as explanation methods using HAM10000 dataset [36], to identify 2 skin diseases with accuracy rate of 85%. Zunair et al. [28] used VGG16 [41] to classify 2 skin diseases using ISIC 2016 dataset and CAM [37] as explanation method with sensitivity 91.76% and AUC 81.18%.

In our work, we also compare our model accuracy with the studies who have not applied XAI method. In this context, Brinker et al. [17] deep learning system outperformed 136 out of 157 experienced dermatologists of the hospitals in a German university. When the system's results were compared to those of board-certified dermatologists, the system outperformed 136 of 157 in the melanoma detection challenge. They used 12,378 images from the ISIC dataset for training the network. A total of 100 images were utilised to compare the system's performance against that of human specialists. They used the Local Outlier Factor (LOF) approach to find outliers. The specificity of the network was 86.5% as compared with the human experts who got only 60%. The sensitivity was also 74.1% for both doctors and network system.

Kassem et al. [9] explained skin lesion classification into eight classes. In their research, they employed the ISIC 2019 dataset for testing and training. They demonstrated that image augmentation and transfer learning can improve classification rates. Their results show 94.2% accuracy, 74.5% sensitivity, 96.5% specificity, 73.62% precision and 74.04% F1 score using image augmentation techniques. When they applied additional image augmentation steps and modified GoogleNet architecture, the results obtained were 94.92% accuracy, 79.8% sensitivity, 97% specificity, 80.36% precision and 80.07% F1 score.

Kasani et al. [29] compared various deep learning architectures for melanoma diagnosis. They tested the most recent deep learning architectures for melanoma detection in dermoscopic images. They used image pre-processing to improve image quality and remove noise. Overfitting was reduced using the data augmentation approach. The data augmentation and picture preparation techniques considerably improve the classification rates, according to their research. They were able to reach 93% precision, 92% accuracy and 92% recall.

Salido et al. [12] proposed technique automatically segmented the skin lesion after pre-processing the photos by removing undesirable elements such as hair. They constructed a deep CNN after eliminating artifacts and noise from the images. Their tests revealed that the processed photos had a high level of categorization accuracy. They were able to reach 93% accuracy and sensitivity in the 84-94% range.

Shahin et al. [10] proposed a framework based on deep neural network that follows an ensemble method to skin lesion classification by integrating Inception V3 and ResNet-50 architectures. To train the algorithm, they used the ISIC 2018 dataset. On the same dataset of dermoscopic images, the system was tested and validated. The validation experimental results achieved an accurate classification rate with a validation accuracy of up to 89.9%. Sherif et al. [13] also employed deep CNN for melanoma classification and detection. To train the system, they used the ISIC 2018 dataset. The system was tested and validated on the same dermoscopic images dataset. They were able to reach 96.67% accuracy.

Ünver et al. [30] used latest deep learning algorithm for melanoma detection. You Only Look Once (YOLO) [46] and GrabCut algorithm [47] was used to detect and segment the melanoma affected body parts. The YOLO is used for detection purposes which has great detection results. It's very fast and computationally inexpensive [46]. After this GrabCut algorithm was applied to segment the detected area on image. They used PH2 and ISBI 2017 datasets and got an accuracy of 93.39%.

Table 1 presents the summary of these works. It can be observed that most of the researchers have used CAM [37] as model explainability method with not so high accuracy. Only 1 study has considered ISIC 2019 dataset (with large number of images). This motivates our research to develop a robust XAI based model with the goal to achieve AI model transparency, traceability, and improvement in skin lesion classification.

III. PROPOSED METHODOLOGY

In this section, we explain the proposed methodology. The flow is shown in Fig. 5 and steps are explained below.

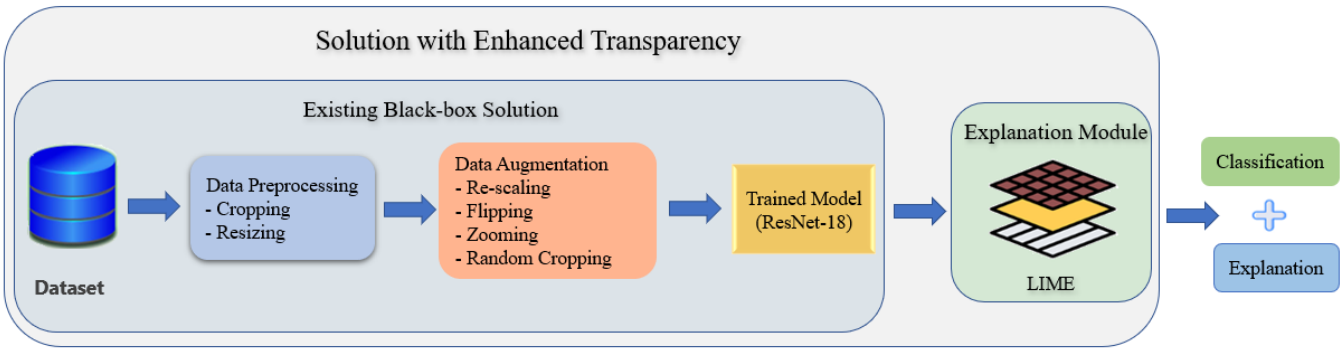


FIGURE 3: The workflow of Proposed Methodology

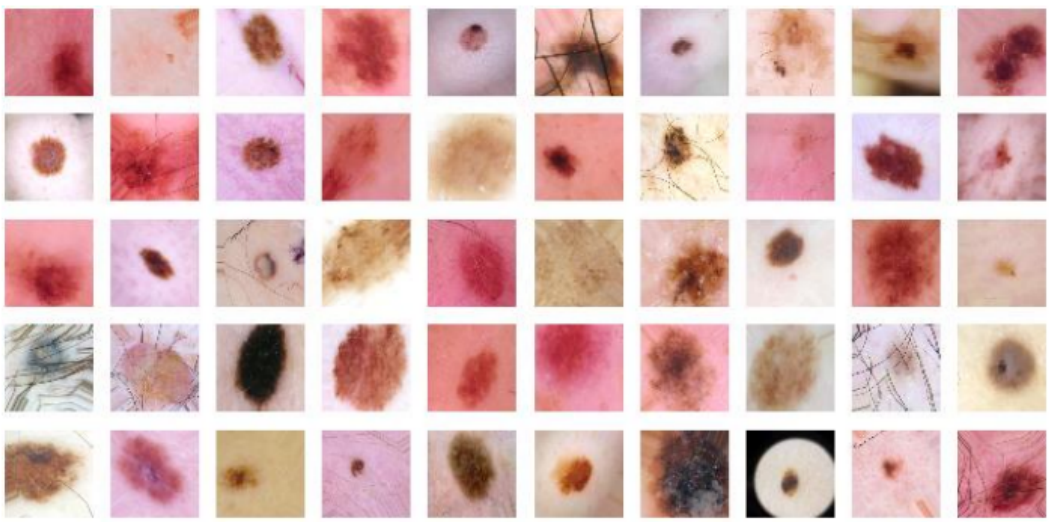


FIGURE 4: Data Augmented Preview

A. DERMOSCOPY IMAGE PRE-PROCESSING

Due to the intricacy of digital pictures, the detection of malignancy by visual evaluation becomes complicated. As a result, effective image processing techniques are required to assist clinicians in properly diagnosing skin lesions. In this study, the training set contained more than 25,000 skin lesion images of different resolutions [9]. As the resolution of all lesion images is greater than 299 $\times$ 299, it was necessary to extract the region of interest (ROI) and get rid of unnecessary/redundant regions from each image. Therefore, these images are cropped automatically and processed before using the images in classification algorithm. This preprocessing step is necessary to reduce the computation time and increasing the effective performance and reliability of the classifier.

1) Image Resampling and Cropping

This step applies image resampling and cropping to the images. Image resampling is a technique used to manipulate the size of an image. Increasing the size of the image is called upsampling while decreasing the size is called downsampling. These two techniques are essential for applications like image display, compression, and progressive transmission. During upsampling or downsampling processes, a two-dimensional (2D) representation is kept the same while the spatial resolution is reduced or increased, respectively. On the other hand, cropping is a technique used to find the ROI in an image by framing around and clipping the area.

2) Image Resizing with Adding Zero-Padding

The data obtained from the ISIC archive [9] is not always ready to directly feed into the algorithm which requires structured, clean, and meaningful data. To overcome this problem, all images are resized from the archive to 224 $\times$ 224 without losing any feature. The pseudo-code for this process is as follows:

1) Identify which side of the image is short.
2) Find the difference between two sides.
3) Take half of the difference.
4) Do padding by putting number of zeros to short sides by adding half of the difference.
5) Resize the image to 224 $\times$ 224.

B. DATA AUGMENTATION

In this step, the data augmentation technique is employed. In image classification, this equates to rotating, flipping, and cropping the picture. The ISIC dataset was supplemented with several random modifications to make the most of our limited training samples and improve the model's accuracy. Furthermore, data augmentation is intended to aid in preventing overfitting (a typical problem in ML with limited datasets in which the model learn patterns that do not apply to new data) and, as a result, improve the model's capacity to generalise. Model overfitting can also be avoided by using an early stopping criterion [48]. The Fig. 4 shows the data augmentation of few dataset instances.

C. FEATURES EXTRACTION USING RESNET-18

Existing algorithms require manual feature extraction, preprocessing and calculate only numeric values. To pass these cumbersome steps and make the algorithm to do the feature extraction itself, we use transfer learning algorithm ResNet-18 [19] which is a specialized version of CNN. The general architecture of the algorithm is shown in Fig. 5.

D. PREDICTION EXPLAINABILITY

In this step, the LIME framework is applied which is an approach for explaining individual predictions that uses a local, interpretable model to approximate any black box ML model. We perturbed the original data points, fed them into a black box model, and then observed the outcomes. The technique then weights the additional data points based on their distance from the original location. Finally, it uses those sample weights to train a surrogate model on the dataset, such as linear regression. The newly trained explanation model may then be used to explain each of the original data points.

IV. EXPERIMENTAL ANALYSIS

A. EXPERIMENTAL SETUP

1) Dataset

The developed model is evaluated on the skin lesion classification using ISIC 2019 dataset. This dataset is publicly available and comprises of 25,331 RGB images. It is divided into 8 classes namely: melanocytic nevus (NV), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), vascular lesion (VASC), dermatofibroma (DF), and actinic keratosis (AKIEC). The images are distributed as NV : 12,875, MEL : 4,522, BKL : 2,624, BCC : 3,323, SCC : 628, VASC : 253, DF : 239 and AKIEC : 867. All dataset images are labelled with one type of skin lesion (Table 3). In Fig. 6, we depict several forms of skin cancer. This dataset is one of the most difficult to categorise into eight classes with an uneven number of images in each class.

2) Parameters

Table 3 shows the hyperparameters of the Resnet-18 classifier used in the experiments.

TABLE 2: ISIC Dataset 2019 [9] Distribution

Type	Subset	Type	Subset
NV	12,857	SCC	628
MEL	4,522	VASC	253
BKL	2,624	DF	239
BCC	3,323	AKIEC	867
Total		25,331	

TABLE 3: Hyperparameters of the Resnet-18 Classifier

Parameters	**Values**
Input Size	224*224*3
Batch Size	32
Loss Function	Categorical Cross Entropy
Activation	ReLU
Optimizer	SGDM
Learning Rate	0.001
Momentum	0.9
Dropout	0.5
Train/Validation/Test Split	70% / 20% / 10%

3) Performance measures

To evaluate the performance of classifiers, common quantitative metrics are presented in this section. For classification problems, results are categorized as either normal case or abnormal, named as positive class or negative class, respectively. The prediction results can also be either true or false, implying correct prediction or incorrect prediction, respectively. Thus, we can categorize classification into below four possible states which is commonly known as confusion matrix [49].

i) True positive (TP) : Correct prediction of positive class
ii) True negative (TN) : Correct prediction of negative class
iii) False positive (FP) : Incorrect prediction of positive class
iv) False negative (FN) : Incorrect prediction of negative class

Based on the confusion matrix, the Accuracy, Precision, Recall and F1 score are calculated as below:

$$Accuracy = \frac{TP + TN}{FP + TN + TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The F1 score is the harmonic mean of precision and recall:

$$F1 = (\frac{recall^{-1} + precision^{-1}}{2})^{-1} = 2.\frac{precision.recall}{precision + recall} \tag{6}$$

4) Experimental environment

In our experiments, it took about 24 hours to train the ResNet-18 model with NVIDIA GeForce GTX 1650 GPUs. All the experiments are implemented in Python, running on a

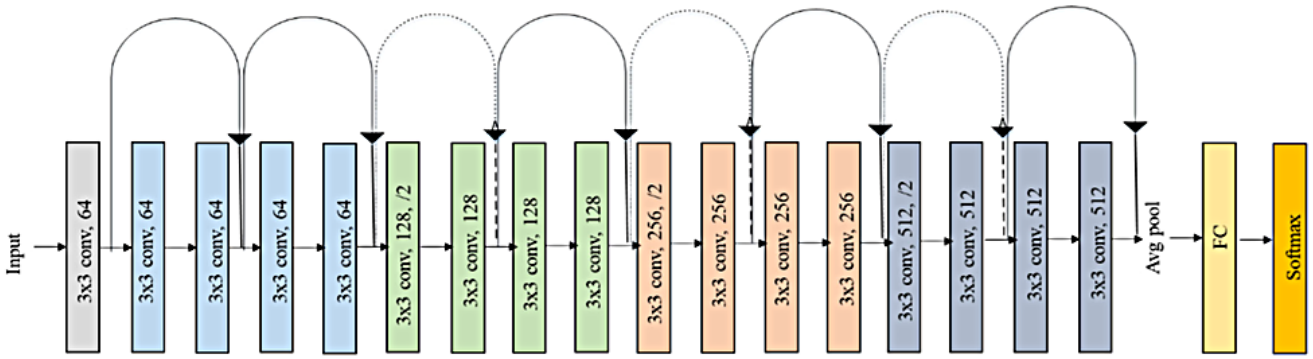


FIGURE 5: Resnet-18 Transfer Learning Algorithm Layers

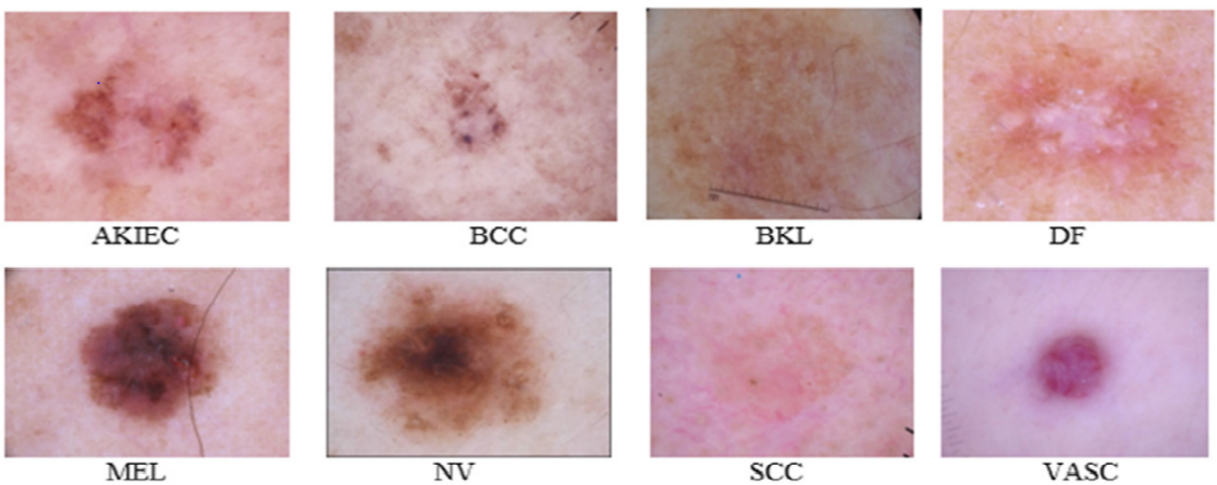


FIGURE 6: An illustration of ISIC 2019 Skin Lesions Instances

personal computer with Intel core i5, 3.2 GHz CPU and 16 GB RAM.

B. RESULTS AND DISCUSSION

The skin cancer detection is complicated by irregular forms of skin lesions, various types of colours on each skin, and defining the ROI on each dermoscopic picture. The detection of minute changes on the skin requires expertise in this field. However, the human eye may not always catch these tiny changes. Many lives can be saved by assisting doctors with computer vision and deep learning techniques. With this motivation, we studied skin cancer malignancy detection to classify skin lesions and identify malignant cases. The pre-training settings and post-training measurements of all experiments showed that the skin cancer malignancy detection is a difficult task and generalizing a model for all cases requires some image pre-processing techniques to apply before feeding into any deep learning algorithm. We did many experiments and tried various techniques to solve the complexity of skin lesions classes.

Regarding model selection, we compare ResNet-18 with Inception v3 using ISIC 2019 dataset. ResNet is one of the most powerful deep neural networks which has achieved fantabulous performance for classifications problems [19]. The Inception v3 [50] is a pre-trained model on the ImageNet datasets. It has also shown better performance for images classification tasks as compared to other deep learning algorithms [50]. The results indicate that ResNet-18 outperforms Inception v3 in terms of accuracy, precision, recall and F1 score as shown in Fig. 7. Therefore, we select ResNet-18 as our final model for training purposes.

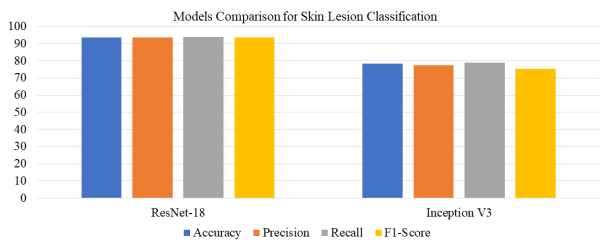


FIGURE 7: Models Comparison using ISIC 2019 Dataset

In first part of the experiments, 8000 images are used that were not pre-processed before feeding the algorithm.

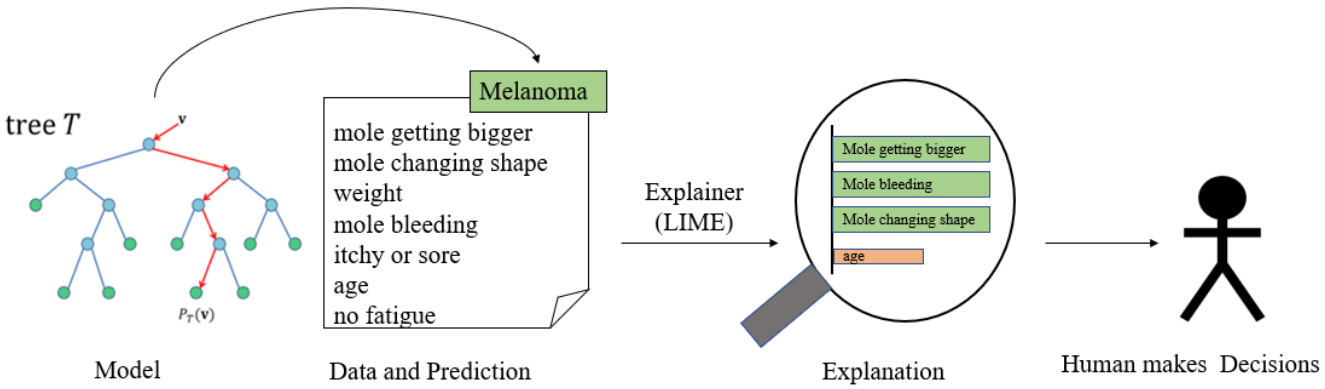


FIGURE 8: Developed Model Working Example

The purpose is to examine the performance of Resnet-18 algorithm based on the existence of the noise and other artifacts to see how much it tolerates the noise. Images were randomly split into training and testing subsets. We obtained 0.75 F1 score (75%) for classification accuracy.

In the second part of experiments, 1600 pre-processed and augmented images are used for training, and 240 pre-processed and augmented images for testing. After this, the classification algorithm is trained. The performance measures were accuracy, precision, recall and F1 score, while the values of these measures were 94.47%, 93.57%, 94.01%, and 94.45% respectively. As compared to the first part of experiment, it shows higher recall and F1 score average values. This indicates that the image pre-processing has a profound impact on the classification algorithm by making the ROI more clean, distinguishable, obvious, and easy to capture so that the algorithm could extract better features about the image and learn better.

The developed model working example is shown in Fig. 8. The visual representation of results (Fig. 11) show that our developed model detected each infected image correctly with 100% confidence. This result is a good indicator for the potential of such a technology to classify predictions

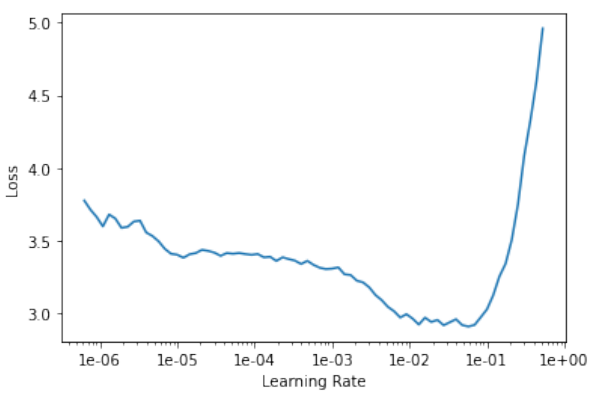


FIGURE 9: Model Learing rate vs Loss

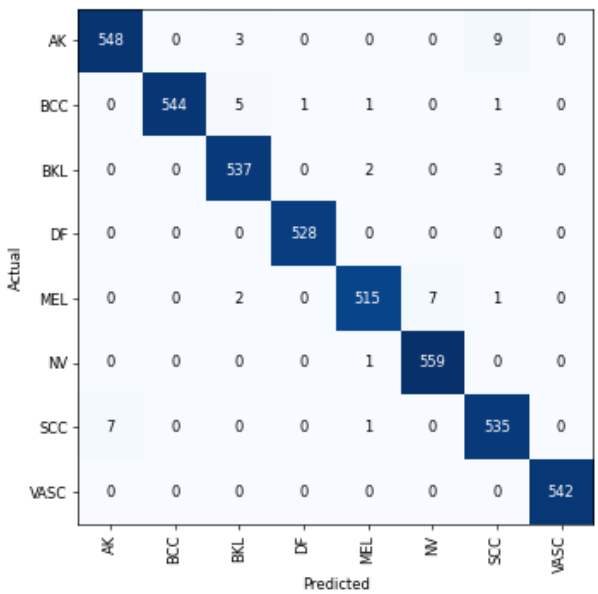
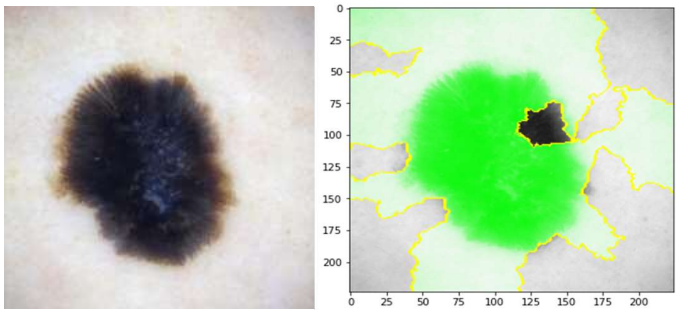


FIGURE 10: Confusion Matrix of our Developed Model

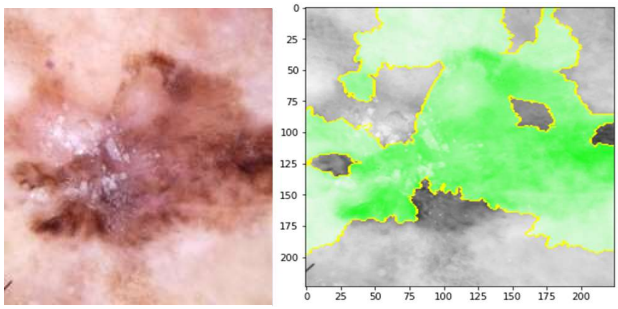
accurately and eventually help physicians increase their diagnostic prediction power. We also present the learning rate (log) against loss (Fig. 9); it can be seen that as the learning rate increase, there is a point where the loss stops decreasing and learning rate starts to increase. The confusion matrix is shown in Fig. 10. It can be observed that true positives for 8 classes are, AKIEC : 548, BCC : 544, BKL : 537, DF : 528, MEL : 515, NV : 559, SCC: 535, and VASC : 542. It means that our developed model has correctly identified the respective disease with good number of percentages.

V. THREATS TO VALIDITY

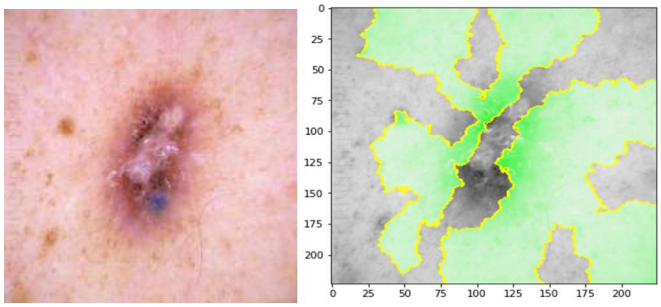
This study is to help the dermatologists in the early assessment of skin cancer. However, there are some limitations to this work. First, we have not considered the large or different datasets. Second, we have used only one pre-trained network in our work. The model extension to incorporate



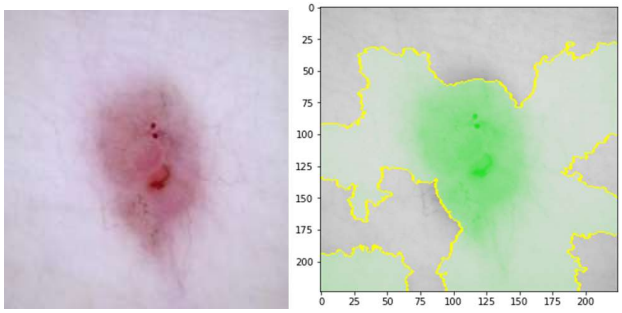
(a) Predicted Melanoma



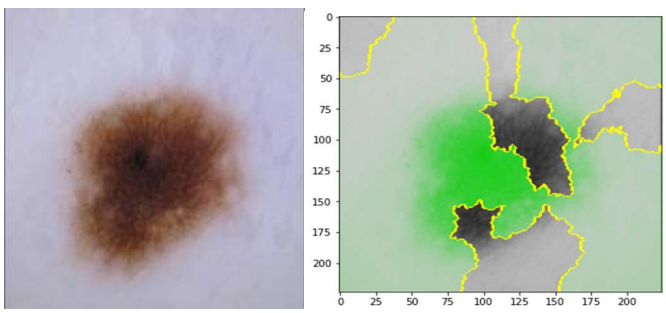
(b) Predicted Squamous Cell Carcinoma



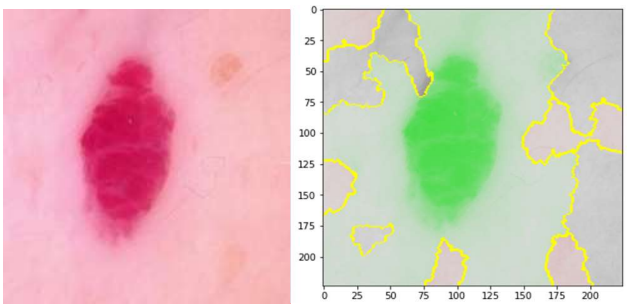
(c) Predicted Dermatofibroma



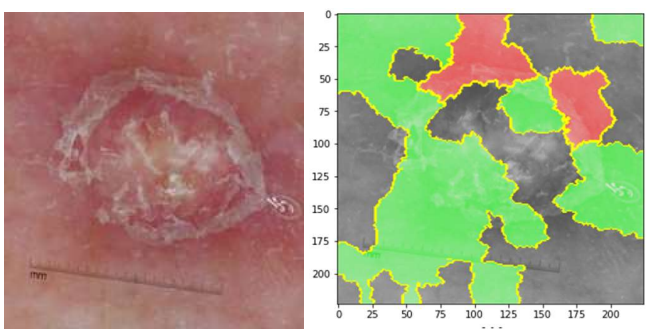
(d) Predicted Basal Cell Carcinoma



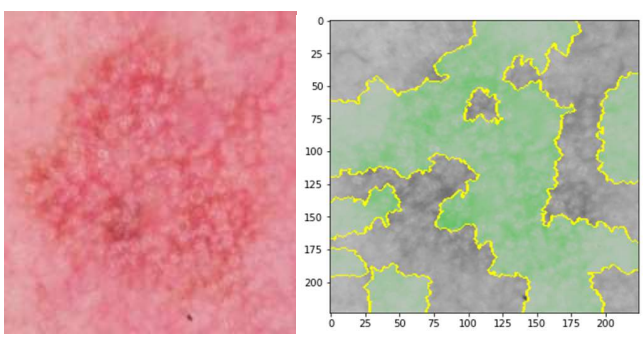
(e) Predicted Melanocytic Nevus



(f) Predicted Vascular Lesion



(g) Predicted Benign Keratosis



(h) Predicted Actinic Keratosis

FIGURE 11: Infected Images Detection by our Developed Model with 100% Confidence.

more advanced pre-trained models could result in improved classification performance. Third, the more training data could lead to better results. The resizing of the images to very small patches could affect the classifier's performance. It may deteriorate some useful information from the lesions when images are downsized. To balance the dataset, the classifier's performance could also be affected by decreasing the total number of samples available for training and validation.

VI. CONCLUSION AND FUTURE WORK

Skin cancer is the most common type of cancer and a major health and economic concern. The dermatologists examines patients individually with the naked eye or a magnifying glass for the skin cancer diagnosis. However, with the advancements in the field of ML, early skin cancer detection can be made more accurate through skin lesion classifiers. Consequently, the ML driven solution has the potential to save many lives by assisting in the early detection of malignant lesions, assisting in decision-making, reducing diagnostic costs, and reducing money spent on treatment. This offers great help for the doctors and patients and can be offered through smartphone apps, websites, or hospital stations.

In this paper, transfer learning and the pre-trained deep neural network Resnet-18 is used to develop the ML model using ISIC 2019 dataset. This model is capable of accurately classifying eight different types of lesions with accuracy, precision, recall, and F1 score as 94.47%, 93.57%, 94.01%, and 94.45%, respectively. Moreover, LIME framework is used to present the useful explanations to support rational decisions. The visual explanations are capable of demonstrating model's good generalisation as well as biases learned from the outlier images. Moreover, these insights enable researchers and field experts to better understand the rational associated with skin lesion classification resulting from the black-box model's inner working.

It's worth mentioning that the availability and quality of dataset is critical for training more accurate ML models. The ISIC 2019 dataset, used in this paper, comprise of 25,331 images with 8 skins lesions classes. Due to privacy, these datasets require continuous enrichment with patients consent which is not obvious. The proposed approach where ML model is complemented with XAI helps the dermatologist with a visual rational to identify new classes and enrich the existing datasets with good examples for improved performance in earliest skin lesions detection. This is a significant contribution in not only improving skin cancer detection accuracy but also in identifying the new classes.

As a future work, a more robust model can be developed that considers other diseases, as well as opposing examples such as healthy skin, fingers, hair, nose, eyes, and background objects. This addition will help the model in better generalising features association with a given lesion while ignoring adjacent features. Moreover, gathering written reports of lesion observations, both in technical and non-technical languages, is another task that would lead to the adoption of this model. This may also help to create a model to generate image captions to serve as an image explanation which is important for the decision being made.

REFERENCES

[1] R. L. Siegel, K. D. Miller, A. Goding Sauer, S. A. Fedewa, L. F. Butterly, J. C. Anderson, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2020," *CA: a cancer journal for clinicians*, vol. 70, no. 3, pp. 145–164, 2020.

[2] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, F. A. Oluwaseyi, and K. I. Okesola, "Multi-classification of alzheimer disease on magnetic resonance images (mri) using deep convolutional neural network (dcnn) approaches," *International Journal of Advanced Computer Research*, vol. 11, no. 53, p. 51, 2021.

[3] C. Barata, J. S. Marques, and M. Emre Celebi, "Deep attention model for the hierarchical diagnosis of skin lesions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[4] H. P. Soyer, G. ARGENZIANO, V. RUOCCO, and S. CHIMENTI, "Dermoscopy of pigmented skin lesions*(part ii)," *European Journal of Dermatology*, vol. 11, no. 5, pp. 483–98, 2001.

[5] B. S. Ankad, P. S. Sakhare, M. H. Prabhu *et al.*, "Dermoscopy of non-melanocytic and pink tumors in brown skin: A descriptive study," *Indian Journal of Dermatopathology and Diagnostic Dermatology*, vol. 4, no. 2, p. 41, 2017.

[6] X. Li, J. Wu, E. Z. Chen, and H. Jiang, "From deep learning towards finding skin lesion biomarkers," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 2797–2800.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[8] D. Thanh and S. Dvoenko, "A denoising of biomedical images," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, no. 5, p. 73, 2015.

[9] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020.

[10] A. H. Shahin, A. Kamal, and M. A. Elattar, "Deep ensemble learning for skin lesion classification from dermoscopic images," in *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*. IEEE, 2018, pp. 150–153.

[11] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018.

[12] J. A. A. Salido and C. Ruiz, "Using deep learning to detect melanoma in dermoscopy images," *Int. J. Mach. Learn. Comput*, vol. 8, no. 1, pp. 61–68, 2018.

[13] F. Sherif, W. A. Mohamed, and A. Mohra, "Skin lesion analysis toward melanoma detection using deep learning techniques," *International Journal of Electronics and Telecommunications*, vol. 65, no. 4, pp. 597–602, 2019.

[14] T. A. Kumar, R. Rajmohan, M. Pavithra, S. A. Ajagbe, R. Hodhod, and T. Gaber, "Automatic face mask detection system in public transportation in smart cities using iot and deep learning," *Electronics*, vol. 11, no. 6, p. 904, 2022.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014002135

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[17] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European Journal of Cancer*, vol. 113, pp. 47–54, 2019.

[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] T. Chowdhury, A. R. Bajwa, T. Chakraborti, J. Rittscher, and U. Pal, "Exploring the correlation between deep learned and clinical features in melanoma detection," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2021, pp. 3–17.

[21] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[22] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1996–2000.

[23] F. Nunnari, M. A. Kadir, and D. Sonntag, "On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2021, pp. 241–253.

[24] M. Sadeghi, P. K. Chilana, and M. S. Atkins, "How users perceive content-based image retrieval for identifying skin images," in *Understanding and interpreting machine learning in medical image computing applications*. Springer, 2018, pp. 141–148.

[25] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2482–2493, 2020.

[26] J. Yang, F. Xie, H. Fan, Z. Jiang, and J. Liu, "Classification for dermoscopy images using convolutional neural networks based on region average pooling," *IEEE Access*, vol. 6, pp. 65 130–65 138, 2018.

[27] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" in *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support*. Springer, 2019, pp. 48–55.

[28] H. Zunair and A. B. Hamza, "Melanoma detection using adversarial training and deep transfer learning," *Physics in Medicine &amp Biology*, vol. 65, no. 13, p. 135005, jul 2020. [Online]. Available: https://doi.org/10.1088/1361-6560/ab86d3

[29] S. H. Kassani and P. H. Kassani, "A comparative study of deep learning architectures on melanoma detection," *Tissue and Cell*, vol. 58, pp. 76–83, 2019.

[30] H. M. Ünver and E. Ayan, "Skin lesion segmentation in dermoscopic images with combination of yolo and grabcut algorithm," *Diagnostics*, vol. 9, no. 3, p. 72, 2019.

[31] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, p. 102470, 2022.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[33] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[34] M. Manandhar, S. Hawkes, K. Buse, E. Nosrati, and V. Magar, "Gender, health and the 2030 agenda for sustainable development," *Bulletin of the World Health Organization*, vol. 96, no. 9, p. 644, 2018.

[35] R. Erol, *Skin Cancer Malignancy Classification with Transfer Learning*. University of Central Arkansas, 2018.

[36] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[39] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

[40] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[43] V. N. Gudivada and V. V. Raghavan, "Content based image retrieval systems," *Computer*, vol. 28, no. 9, pp. 18–22, 1995.

[44] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[45] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[47] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

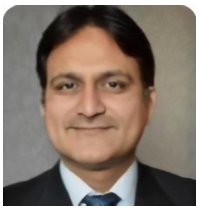
[48] Y. H. Bhosale and K. Sridhar Patnaik, "Iot deployable lightweight deep learning application for covid-19 detection with lung diseases using raspberrypi," in *2022 International Conference on IoT and Blockchain Technology (ICIBT)*, 2022, pp. 1–6.

[49] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection." *MAICS*, vol. 710, no. 1, pp. 120–127, 2011.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

NATASHA NIGAR received her PhD degree from the School of Computer Science, University of Birmingham, United Kingdom, in 2021. She is currently working as Assistant Professor with the Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan. From 2008 to 2009, she was a Lab Engineer at National University of Computer and Emerging Sciences, Lahore, Pakistan. She worked as Software Engineer at Palmchip Pvt. Limited, Lahore, Pakistan from 2009 to 2011; and Senior Software Quality Assurance Engineer at Netsol Technologies, Lahore, Pakistan from 2011 to 2013. She was awarded Faculty Development Program Scholarship to pursue her PhD studies. Her research interests include computational intelligence, optimization in dynamic and uncertain environments, and machine learning.

MUHAMMAD UMAR received his M.Sc degree from the Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan, in 2020. He is currently working as a Software Developer at TEK Headquarters, Irvine California, USA. From 2018 to 2020, he was working as a Senior Software Engineer at NETSOL Technologies Inc., Lahore, Pakistan. He worked as Software Engineer at B.I.S.E, Lahore, Pakistan from 2017 to 2018; and at Ebryx Pvt Ltd., Lahore, Pakistan from 2016 to 2017. He has been awarded Outstanding Performance Award and certificate of Appreciations during his professional career. His research interests include artificial intelligence, machine learning and cloud computing.



MUHAMMAD KASHIF SHAHZAD recieved his Bachelor's degree in engineering from UET, Lahore, Pakistan in 2000 and Master's and PhD degrees in industrial systems engineering from University of Grenoble, France in 2008 and 2012. He is presently working as Chief Technical Officer (CTO) at Power Information Technology Company (PITC), Ministry of Energy, Power Division, Govt. of Pakistan. His research interests include data models interoperability, advanced software engineering, technology, smart grid solutions, and engineering data management. He has vast experience of working in large scale European R&D projects IMPROVE and INTEGRATE. Dr. Shahzad specializes in designing and delivering technology driven smart grid solutions and is working with USAID in developing solutions to improve Pakistans power secror in deregulated market. Dr. Shahzad has 38 publications and 2 book chapters. Moreover, he has 20+ years of professional experience designing, business process re-engineering and managing large scale software development projects.

SHAHID ISLAM received the B.S. and M.S. degrees in computer science from University of Engineering and Technology, Lahore, Pakistan, in 2003 and 2008, respectively. He is currently working as Assistant Professor at Rachna College of University of Engineering and Technology, Lahore, Pakistan. His research interests include cloud computing, machine learning, semantic web, m-learning, and intelligent agent applications.

DOUHADJI ABALO received the PhD degree in mathematics from University of Lomé in 2021. Currently, he is working as assistant professor in Mathematics Harmonic Analyse at University of Lomé. His research interests include Topology, Pure Mathematics, Matrix Theory, and Harmonic Analysis.