INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Skin Lesion Classification: A Deep Learning Approach with Local Interpretable Model-Agnostic Explanations (LIME) for Explainable Artificial Intelligence (XAI)

Sin Yi Hong [a], Lih Poh Lin [b,*]

[a] *Faculty of Engineering and Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia*
[b] *Centre for Multimodal Signal Processing, Biomedical and Bioinformatics Engineering Research Group, Faculty of Engineering and Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia*

*Corresponding author: *linlp@tarc.edu.my*

***Abstract*—The classification of skin cancer is crucial as the chance of survival increases significantly with timely and accurate treatment. Convolution Neural Networks (CNNs) have proven effective in classifying skin cancer. However, CNN models are often regarded as "black boxes", due to the lack of transparency in the decision-making. Therefore, explainable artificial intelligence (XAI) has emerged as a tool for understanding AI decisions. This study employed a CNN model, VGG16, to classify five skin lesion classes. The hyperparameters were adjusted to optimize its classification performance. The best hyperparameter settings were 50 epochs, a 0.1 dropout rate, and the Adam optimizer with a 0.001 learning rate. The VGG16 model demonstrated satisfactory classification performance. The Local Interpretable Model-Agnostic Explanations (LIME) method was implemented as the XAI tool to justify the predictions made by VGG16. The LIME explanation revealed that the correct predictions made by VGG16 were owing to its truthful extraction of the cancer or lesion area, especially for the "vascular lesion" class. Meanwhile, inaccurate classifications were attributed to VGG16 extraction of the background and insignificant parts of the skin as core features. In conclusion, The LIME model allowed visual inspection of the features selected by VGG16, paving the way for improving the CNN model for better feature extraction and classification of skin lesions, offering a promising direction for future research.**

***Keywords*— CNN; deep learning; explainable ai; skin cancer; local interpretable model-agnostic explanations; VGG16; XAI.**

Manuscript received 16 Jun. 2024; revised 21 Aug. 2024; accepted 17 Oct. 2024. Date of publication 30 Nov. 2024.

I. INTRODUCTION

A skin lesion is an abnormal growth of skin cells that could be caused by exposure to ultraviolet light, environmental hazards, or genetic risk factors [1]. Skin lesions such as actinic keratoses, dermatofibroma, and seborrheic Keratoses are benign [2], while melanoma, basal cell carcinoma, and squamous cell carcinoma are usually malignant [3]. About 104,930 new melanoma cancer cases are estimated for the year 2023 in the United States [4]. The 5-year survival rate of skin cancer is as high as 94% when the cancer is still localized but is drastically reduced to 32% after metastasis [4], emphasizing the importance of early detection. The early diagnosis of skin lesions typically involves visual examination before histopathological analysis; nevertheless, classifying skin lesions is challenging due to the high variabilities in the appearance of the skin lesions [5]. The outcomes of diagnostics frequently rely on the dermatologist's skill and expertise, which can sometimes be subjective [6]. The accuracy of skin cancer diagnosis was about 60% (non-dermoscopic image) to 84 % (dermoscopic image) [3]. This shortcoming has motivated the introduction of artificial intelligence (AI) in skin lesion diagnosis. Deep neural networks like Convolutional Neural Networks (CNNs) possess remarkable feature extraction and classification capabilities. As a result, they have witnessed extensive adoption in skin lesion classification in recent years, including application in [7], [8], [9]. Table 1 summarizes several studies that employ CNNs for classifying and detecting skin diseases. Works from the literature suggest that CNNs excel at learning complicated patterns in images, rendering CNNs ideal for analyzing skin lesions where features are sometimes subtle. In other words, CNNs can capture details that may not be easily discernible to the human eye.

TABLE I
REPORTED WORKS OF CNNS IN SKIN DISEASE DETECTION

Method	Dataset	Key Results	Ref
Model used: VGG16, ResNet50, Inception-v3 and Ensemble. Binary cross-entropy loss function	Monkey Pox Skin Lesion Dataset (self-prepared)	[VGG16] Accuracy: 81.48% Precision: 0.85 Recall:0.81 F1 score: 0.83 [ResNet50] Accuracy 82.96 % Precision: 0.87 Recall: 0.83 F1 score: 0.84	[7]
Model used: AlexNet, VGG16, ResNet-18 and fusion CNN models with Support Vector Machine. Fuse the deep features from various layers of CNNs	ISIC image archive	[AlexNet] Avg. AUC: 89.73% [VGG16] Avg. AUC: 88.76% [ResNet-18] Avg. AUC: 88.51%	[10]
Model used: Inception-v3, VGG19, SqueezeNet and ResNet50 Use the .NET framework with the C# language to enable a web service for users	ISIC image archive	[Inception V3] Avg. accuracy: 93% [VGG19] Avg. accuracy: 94.% [ResNet50] Avg. accuracy: 97% [SqueezeNet] Avg. accuracy: 96%	[11]
Model used: ResNet50, Xception and Inception-ResNet-v2, DenseNet121 and Inception-v3 Employ global average pooling followed by a 1x1 convolution Instead of a fully connected layer	Xiangya-Derm (self-prepared)	[ResNet50] Avg. precision: 63% [Inception-v3] Avg. precision:64% [DenseNet121] Avg. precision:69% [Xception] Avg. precision:68% [InceptionResNetv2] Avg. precision:71%	[12]

The implementation of CNNs has proven to be effective in skin lesion classification. However, despite their success, CNN models come with certain limitations, most notably being regarded as "black boxes", which refers to the low level of transparency and interpretability in the decision-making of CNNs [13]. CNNs do not provide explicit explanations or justifications for their classification. When dealing with critical applications such as medical diagnosis, being informed of the logic behind a model's decision is crucial for building the trust and confidence of the medical practitioners in AI [14].

Furthermore, more advanced and robust CNN architectures have been developed in recent years for various tasks. Recent published CNN models continue to grow more complex, with some models already operating with trillions of parameters. As a consequence, the clarity and interpretability of these emerging CNNs were compromised, making it harder to comprehend the prediction process and internal mechanisms [15]. Additionally, In situations where a CNN model makes incorrect predictions, it is often challenging to diagnose the exact cause of the failure [16].

Amid all these limitations, explainable artificial intelligence (XAI) has arisen as a framework and tool for understanding and interpreting AI decisions. XAI generally refers to all techniques and solutions that enable humans to understand AI models. In other words, XAI allows the understanding of the reasons behind the decisions made by the AI [15]. Some common XAI techniques include Gradient-weighted Class Activation Mapping (GRAD-Cam), Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [17], [18], [19], [20]. Nguyen et al. [21] have compared LIME, SHAP, and CAM in describing the ResNet50 CNN in the image classification problem. Considering the hummingbird shown in Fig. 1, which has 49 clusters, all the tested XAI were able to highlight the core feature used by ResNet50 in the classification. As shown in Fig. 2, LIME identifies core features using superpixel regions, SHAP employs weighted color regions, and CAM utilizes heatmaps to highlight essential features.

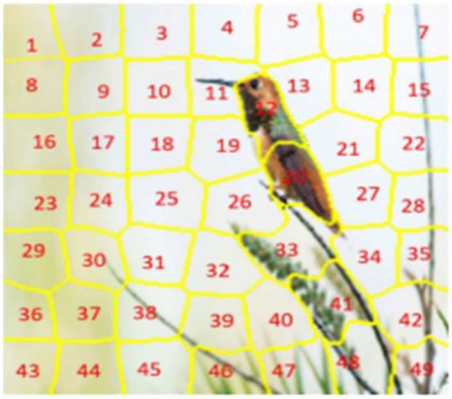


Fig. 1 A hummingbird image with 49 clusters for classification.

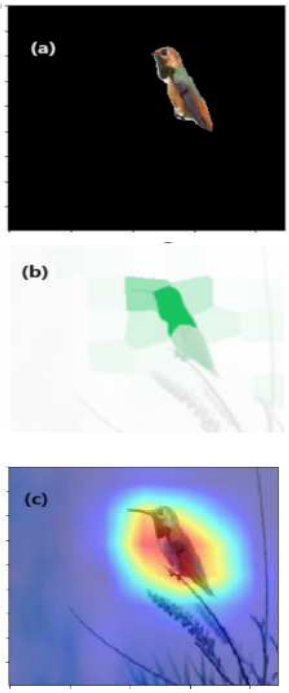


Fig. 2 XAI explanation using (a) LIME, (b) SHAP and (c) CAM [21]

XAI in medicine has been presented in the literature. Yong et al. employed both Grad-CAM and Kernel SHAP techniques on a CNN-based classification of melanoma and benign naevus (moles) [17]. This work's significance lies in using Grad-CAM and Kernel SHAP to perform sanity check experiments, including reproducibility, model dependence,

and sensitivity tests. Yong et al. conducted an extensive analysis, performing 30 model training on 15 randomly selected subclasses of 818 data each. The obtained performance metrics showcase a good 85% mean Area Under Curve (AUC), 1.8% variance, and an 87% recall. The sanity check experiments showed that GradCAM and SHAP were reproducible, model-dependent, and mostly sensitive, but occasionally marked irrelevant features as necessary. The study introduces initial insights connecting accuracy and interpretability in the classification of skin lesions. In the year 2021, the LIME and SHAP models were proposed and compared by Ong et al. to explain COVID-19 diagnoses via X-ray images [18]. A 14-layer SqueezeNet CNN model was first applied to classify the X-rays into trio classes, namely pneumonia, COVID-19, and normal lung, followed by the implementation of XAI based on visual evaluation. Both LIME and SHAP were able to mark the region of interest (ROI) that leads to accurate or inaccurate classification. LIME used superpixel to mark ROI, while SHAP used green and red to mark positive and negative accuracy areas. The author concludes that SHAP is a relatively better XAI for the application as SHAP could always identify the lung region, which is the core feature of an X-ray. The work has provided insight into essential features marked by the CNN network.

Grad-CAM has also been applied to image segmentation models, as demonstrated by Xiao et al. [19]. This study utilized Grad-CAM with several segmentation-deep learning models: MCGU-Net, R2U-Ne, and Double U-Net to segment datasets containing images of colorectal polyps, liver, and skin melanoma. The ROI was presented in the form of heat maps. The heatmaps generated by Grad-CAM revealed distinct insights. For example, in the case of colorectal polyp images, Double U-Net emphasized pixels at the polyp's edge during segmentation, assigning the highest importance to the edge's left and right ends. In contrast, pixels across the whole polyp region were prioritized by R2U-Net, with increasing significance towards the polyp's center. MCGU-Net focused on the center alongside the pixels on the lower edges of the polyp. Regardless, Grad-Cam has effectively identified the image regions that command the primary focus of medical image segmentation models. In short, XAI is emerging in medical image analysis. Table 2 summarizes some reported work of XAI in medical applications.

TABLE II
REPORTED WORKS OF XAI IN MEDICAL APPLICATIONS

Application	Description	XAI	Ref
COVID-19 diagnosis	Dataset from Covidx. Employed SqueezenNet classifier.	LIME and SHAP to give a qualitative visual evaluation	[18]
Melanoma detection	Dataset from HAM10000. Employed Inception classifier	Grad-CAM and KernelSHAP to give qualitative visual inspection and quantitative Structural Similarity Index (SSIM)comparison	[17]
Segmentation of medical images consisting of colorectal polyps, liver CT and melanoma	Datasets: CVC-Clinic, 3Dircadb, Lesion Boundary Segment. Employed MCGU-Net, R2U-Net and Double U-Net segmentation	Grad-CAM to give a qualitative visual inspection	[19]
Adverse Drug Event (ADE) prediction	Dataset from Swedish Health Record Research Bank. Employed RNN: RETAIN and RNN-GRU model	SHAP for qualitative inspection	[20]

Previous studies have established the potential of XAI; nevertheless, its application in skin lesion classification remains limited. This challenge is amplified by the absence of extensive datasets that could fully represent the variations in skin lesions. Considering these gaps, our study developed a CNN model for skin lesion classification. The evaluation of the CNN model's performance is followed by implementing an XAI model LIME to provide justifications for the decisions made by the CNN model. To ensure the robustness of our approach, we have employed a well-established CNN architecture: VGG16. To address the need for sufficient training and validation data, we used the multiple-dataset approach where skin lesion images from 4 different sources were included for this study. LIME presented in this study offers an understanding of the reasoning behind the CNN model's decisions.

II. MATERIALS AND METHOD

A. Dataset and Image Pre-Processing

Four datasets were incorporated into our project. The first dataset was sourced from the Kaggle public dataset 'Skin Cancer ISIC' by Andrey Katanskiy, extracted from The International Skin Imaging Collaboration (ISIC) [22]. The remaining datasets of gold standard lesion diagnosis images, accessible from ISIC archives [23], were derived from the combination of HAM1000 datasets [24], MSK dataset [25] and BCN_20000 Dataset [26]. The combined dataset was then split into train-validation data and test data. The number of test images per class was set at 200. At the same time, the train-validation data went through augmentation (including RandomApply, RandomCrop, RandomRotation, GaussianBlur, RandomAdjustSharpness etc.) or random deletion to reach a uniform 1800 images per class. The train: validation: test ratio of the dataset was 1600:200:200, giving 2000 images per class, as shown in Table 3.

TABLE III
NUMBER OF DATA PER SKIN LESION CLASS

Skin Lesion Class	Train	Valid	Test	Total
Actinic Keratosis (AK)	1600	200	200	2000
Basal Cell Carcinoma (BCC)	1600	200	200	2000
Melanoma (MEL)	1600	200	200	2000
Melanocytic Nevus (NV)	1600	200	200	2000
Vascular Lesion (VASC)	1600	200	200	2000
Total	**8000**	**1000**	**1000**	**10000**

The VGG16 architecture was selected for this study due to its remarkable performance of top-5 test accuracy of 93% on ImageNet, a dataset with over 14 million images distributed among 1000 classes. Furthermore, its robustness has been extensively demonstrated in various detection and classification tasks [27][28][29]. Images were pre-processed into 224 × 224 to suit the VGG16 input requirements. The dataset was normalized as a common pre-processing step in CNNs. The dataset normalization involved scaling the input pixel values to achieve a mean of 0 and a variance of 1. Normalizing data offers several advantages including faster convergence, improved generalization performance, and reduced sensitivity to input changes [30].

B. Hyperparameters Selection and Performance Metrics

Several hyperparameters, including the dropout rate, number of epochs, and type of optimizers, were evaluated to enhance the performance of the CNN models. The dropout rate refers to the ratio of randomly eliminated neurons during training to reduce overfitting. The number of epochs refers to how often the complete dataset is fed through CNN during training. Optimizers adjust the learning rate to minimize the loss function, helping the CNN model to converge. These hyperparameters play critical roles in optimizing CNNs for better performance. The values assessed and the description of the hyperparameters are shown in Table 4.

Quantitative analysis was performed to evaluate the performance of the CNN models. A range of performance metrics, including Accuracy, recall score, Precision score, F1 score, and the Matthews correlation coefficient (MCC) were employed. Descriptions of these metrics are provided in Table 5. Additionally, the loss function of the training and validation sets, which characterize the training process and the potential for overfitting/underfitting, was also examined.

TABLE IV
HYPERPARAMETERS EVALUATED

Hyper-parameter	Description	Value
Epochs	The number of epochs refers to how often the complete dataset is fed through the CNN during training. It affects the training time and the model's performance. The number of epochs has to be sufficient to allow CNN to learn the data features, but an overly high epoch can lead to overfitting	10, 30, 50, 90
Dropout rate	The dropout rate is a regularization technique that randomly drops nodes or neurons in CNN layers to reduce the dependence on specific nodes. Tuning the dropout helps to prevent the overfitting of the CNN models.	0.1, 0.5, 0.7
Type of optimizer	The optimizer is in control of updating the weights of the neural network based on the calculated gradients of the loss function. Different optimizers use various algorithms to perform weight updates which can affect the performance of the CNN and therefore require evaluation.	Adam, SGD, NAdam

TABLE V
PERFORMANCE METRICS

Analysis	Equation and Description	
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	(1)
	To measure the correctly classified data across all data.	
Recall	$\frac{TP}{TP+FN}$	(2)
	To measure total actual positive data detected over total positive data. Likewise understood as true positive rate or sensitivity.	
Precision	$\frac{TP}{TP+FP}$	(3)
	To measure actual predicted positive data out of all predicted positive data.	
F1 score	$\frac{2\ x\ recall\ x\ precision}{recall + precision}$	(4)
	To measure the balance between recall score and precision	
MCC	$\frac{TNxTP - FNxFP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	(5)
	To measure the correlation between the actual and predicted classifications, considering quad results (false/true negatives and false/true positives) to provide a balanced assessment.	

TN = True Negative, FN = False Negative, TP = True Positive, FP = False Positive

C. XAI Application: LIME

LIME was introduced in 2016 by M. T. Ribeiro, S. Singh, and C. Guestrin in their publication titled "Why Should I Trust You? Explaining the Predictions of Any Classifier" [31]. LIME aims to approximate the black-box model locally, making it a post hoc XAI method which does not influence the CNN's training process. Instead. When applying LIME to skin lesion classification, the XAI begins by segmenting the skin lesion image into superpixels, which are pixel clusters with alike characteristics such as colour and

intensity, as shown in Fig. 3(a) [32]. Next, perturbed versions of the original image are generated by randomly masking out a subset of superpixels, creating images with masked regions, as shown in Fig. 3(b). These perturbed images are employed in LIME model training. The superpixels with the highest positive coefficients are considered to have contributed significantly to the prediction of skin lesion type. Local explanations from LIME are accurate within the immediate context or vicinity of the skin lesion image under consideration.

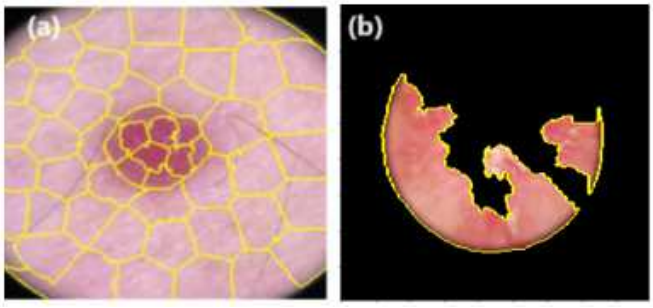


Fig. 3 (a) Skin lesion image segmented into superpixels and (b) Perturbed image to investigate the importance of a specific region

III. Results and Discussion

A. Hyperparameter Selection

Hyperparameter selection is imperative for a CNN as a suboptimal setting could significantly reduce the classification performance [33]. Parameters, including the number of epochs, rate of dropout, and type of optimizer, were evaluated to enhance VGG16's performance in classifying skin cancers. Firstly, dropout rates of 0.1, 0.5, and 0.7 (representing 10%, 50%, and 70% of discarded nodes, respectively) were tested. Fig. 4 illustrates the impact of the dropout rate on accuracy, precision, recall, F1-score, and MCC.

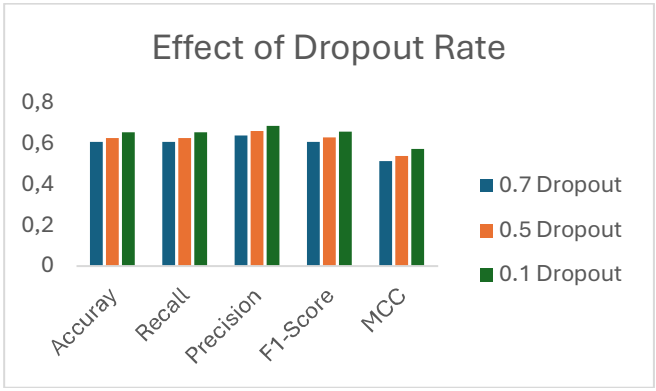


Fig. 4 The effect of dropout rate on classification performance. A 0.1 dropout rate balances preventing overfitting and allowing sufficient learning.

It was observed that the 0.1 dropout rate gave the best performance metrics. Dropouts randomly eliminated neurons and connections during training to prevent rapid adjustment and overfitting. A dropout rate 1.0 indicates that all neurons are dropped and no training occurs, while a rate of 0 means no neuron is dropped, possibly leading to overfitting. For this study, a 0.1 dropout rate appeared to be a reasonable balance where some neurons were eliminated to prevent overfitting while still enabling the model to learn essential features [34]. Fig. 5 demonstrates the effect of epochs on classification performance. Epochs represent one complete pass through the training dataset. A CNN model with low epochs may struggle with underfitting [35], as evidenced by the subpar classification performance observed with only 10 epochs. On the contrary, excessive epochs could lead to overfitting, where CNN memorizes the training sets and cannot generalize on unseen data. For this study, both 50 and 90 epochs reached convergence, but 50 epochs presented the best classification performance and were thus chosen for subsequent analysis.

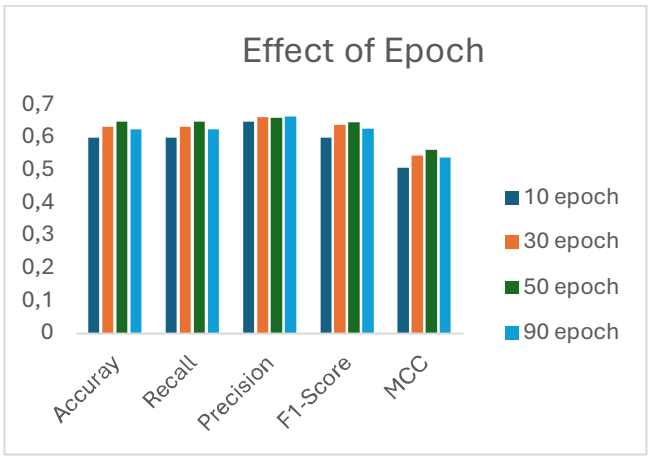


Fig. 5 The impact of epoch on classification performance. Training for 50 epochs yielded optimal convergence and performance.

Optimizers are used to minimize errors between predicted and actual outputs. An optimizer computes the gradient based on the loss function to adjust the weights in the CNN model for better performance. Different optimizers present varying weight adjustments that can substantially affect the classification performances. The optimizers evaluated in this project were Adam, SDG, and Nadam: The Adam optimizer updates the CNN models by looking at past gradients, SGD optimizer uses gradient descent on a random point from the entire dataset for parameter updates, reducing redundant work on large datasets while Nadam optimizer combines Adam and Nesterov's accelerated gradient descent method, boosting learning by considering both past and current gradient trends [36]. A comparison of these optimizers shows that Adam outperformed the other optimizers, as illustrated in Fig. 6.

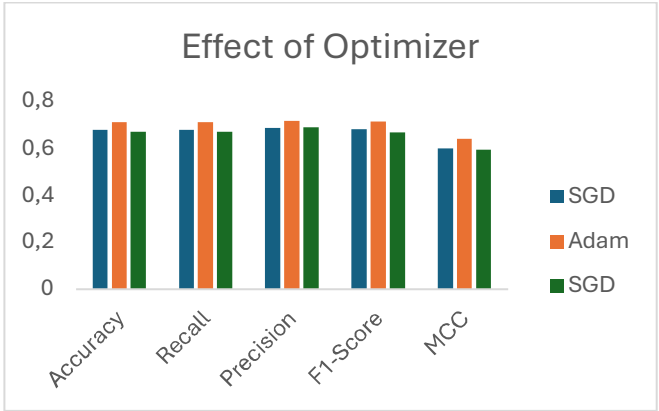


Fig. 6 The influence of optimizer on classification performance. Adam is effective in practice and performs favorably compared to other stochastic optimization methods

This observation is consistent with works reported in the literature, such as that from Kingma et al., which concluded that Adam works well with sparse gradients and is robust for a wide range of non-convex optimization problems using deep learning [36]. Lastly, it is worth noting that the learning rate was consistently set to 0.001 for this study, which is the default value in the PyTorch application of optimizer Adam. This value is commonly used in deep learning frameworks and was observed to be optimal in previous works [37][38].

B. Performance Analysis of CNNs in Skin Lesion Classification

The classification of the CNN models was assessed quantitatively using the individual class classification report and measurements, including accuracy, recall, precision, MCC, and F1 score. According to the results in Fig.7, the VASC class was the easiest to classify, displaying consistently high (> 0.8) precision, recall, and F1-score. This ease of classification is likely due to its distinct blushing and flushing area that a trained VGG16 can easily recognize [39]. On the other hand, the performance of VGG16 in classifying the MEL categories was mediocre, with performance metrics ranging between 0.57 and 0.7. This could be due to the dynamic nature of Melanoma lesions that change over time and have several variants [40]This makes it challenging to capture all stages and variants in a dataset for comprehensive training of VGG16 to classify MEL.

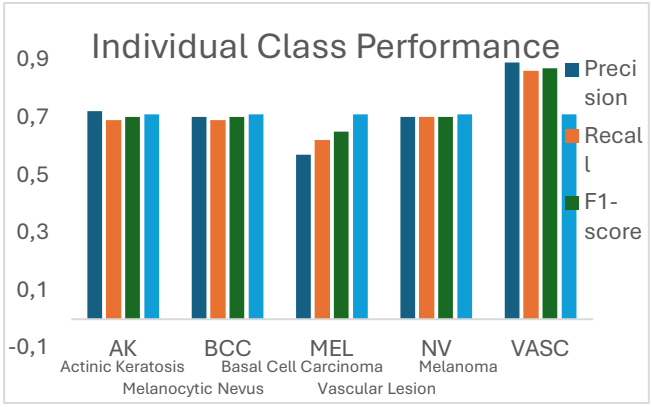


Fig. 7 The performance of the CNN in classifying different skin lesions. VASC class has the highest classification accuracy.

Despite this, it is safe to conclude that the CNN model's classification of skin cancer/lesions is satisfactory, with an overall precision, recall, F1 score, and accuracy of 0.716, 0.712, 0.724, and 0.711, respectively. The MCC was adapted to measure the overall quality of the classification, as high scoring is only produced when the classifier accurately predicts the majority of the negative and positive cases [41]The CNN's MCC was 0.639, indicating a moderately strong level of agreement between the predictions and the actual labels of the skin images. Overall, the CNN model performs reasonably well, with room for improvement.

The training and validation loss were also examined as they visually represent how the VGG16 was learning over time. The decreasing loss as shown in Fig.8 indicates that the VGG16 model was improving its performance on the training data. The validation loss has an overall decreasing trend alongside the training loss, showing that the model was capable of generalizing to new data [42]; The fluctuation observed in the validation loss suggested the CNN model was adjusting its parameters to prevent overfitting and to reach optimal solutions with the unseen data [43].

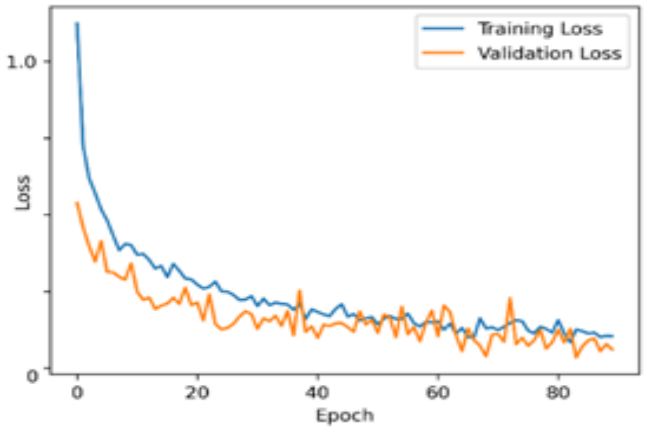


Fig. 8 The loss function of the VGG16 model. Decreasing loss indicates that the CNN model was learning and improving its ability to make accurate predictions

C. The Application of XAI

The VGG16's ability to classify skin cancer/lesion was poorly understood; it was unclear why some images were misclassified. To investigate the features VGG16 used for classification, the test set from the dataset was used for XAI visualization. The saved VGG16 checkpoint was loaded for prediction—the LIME model explained by highlighting segmented areas using superpixel segmentation and feature masking. The segmented regions by LIME made the features used by VGG16 to identify skin cancer classes explainable.

According to the classification report discussed in section B, VGG16 could classify the VASC class well, while it performed poorly on the MEL class. This observation aligns with the LIME segmented explanation, which showed that VGG16 was trained to extract the core area of VASC, as shown in Fig. 9, for accurate classification but struggled with MEL where irrelevant background was extracted, as depicted in Fig. 10. Table 6 provides additional correct and incorrect predictions with remarks. Overall, LIME enabled visual inspection to explain VGG16's decisions. Despite some misclassifications, the CNN model was on the right track in distinguishing between different skin lesion classes. Improvements to the CNN architecture or adding a feature extraction algorithm before CNN classification could enhance classification performance.

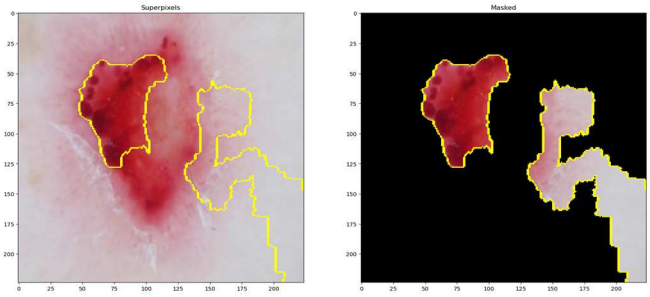


Fig. 9 Superpixels graph and mask graph for VASC correct prediction. LIME shows that VGG16 extracted core features.

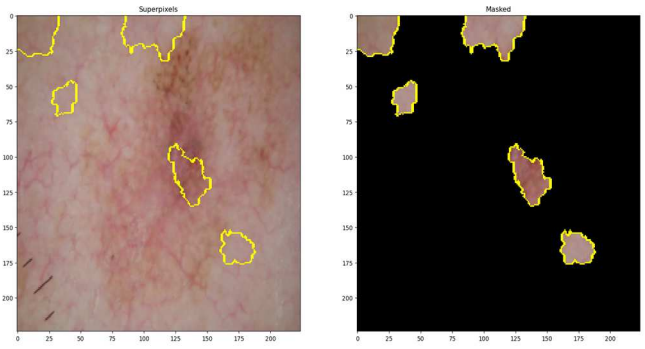
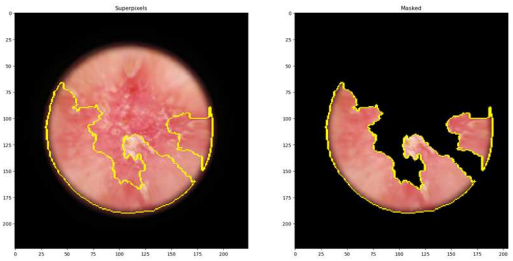
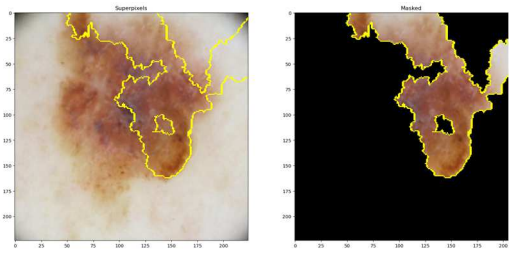
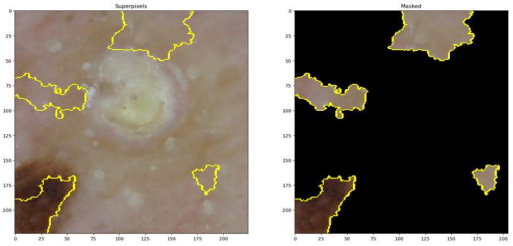
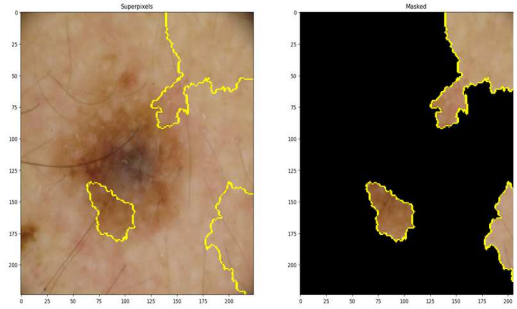


Fig. 10 Superpixels graph and mask graph for MEL incorrect prediction. The image was classified as BCC due to the irrelevant feature extraction by VGG16.

TABLE VI
XAI EXPLANATION OF CNN CLASSIFICATION

Prediction	Superpixels Graph vs Masked Graph
Correct Actual: BCC Predicted: BCC	 Remarks: LIME segmented areas of the image in which VGG16 has utilized as the core feature. Despite not having the entire carcinoma extracted, the feature extracted (probably the redness) was sufficient to allow VGG16 to predict the BCC class correctly.
Correct Actual: MEL Predicted: MEL	 Remarks: LIME segmentation shows that VGG16 has extracted a large area of the skin lesion for the correct classification of MEL.
Incorrect Actual: AK Predicted: NV	 Remark: LIME shows that irrelevant background and a pigmented skin area were extracted leading to incorrect classification. VGG16 may have viewed the pigmented skin (bottom left corner) area as NV.
Incorrect Actual: NV Predicted: BCC	 Remark: LIME shows that irrelevant background was extracted leading to incorrect classification. VGG16 may have viewed the lighter skin area as BCC.

Some improvements can be made to this study. Firstly, different CNN models, such as EfficientNet and DenseNet, could be employed in skin lesion classification to realize better results. Further research should also explore additional hyperparameter tuning to optimize model performance effectively. Bayesian hyperparameter tuning could be explored to efficiently examine the hyperparameter space and automate finding the best hyperparameters with minimal manual intervention. Additionally, segmenting the skin lesion outline before training could enhance training accuracy. Besides, incorporating various XAI models, such as SHAP and Grad-CAM, into CNN models should be considered for future studies. SHAP takes a team-like approach to explain how models make predictions. It helps to explain how each feature contributes to the decisions made, giving AI users valuable insights using colored weighted zones; meanwhile, CAM, being an intrinsic model, analyzes the final convolutional layer of a CNN to understand which parts of the image activate the neurons corresponding to the skin lesion class. It can enhance the interpretability of CNNs by providing heatmaps highlighting essential features.

In addition, quantitative evaluation of XAI models using metrics like fidelity and stability scores should also be explored in future work. The fidelity score measures how closely the explanation matches CNN's decision-making process. In contrast, the stability score measures how consistent the XAI explanation is across data of the same class. Another recommendation for future work is to present XAI explanations using weightage for each feature instead of only visual explanations. While visual explanations can be helpful, they do not always clearly indicate the relative importance of specific features in the decision-making process. Therefore, by explicitly stating the weightage of each feature, the most influential or substantial features affecting the decision of CNNs are made known to the users. In short, exploring different CNN models, implementing Bayesian optimization, comparing different XAI methods, and instigating quantitative XAI explanations can collectively enhance the interpretability of CNN models in skin lesion classification, paving the way for more transparent AI systems in healthcare.

IV. Conclusion

In conclusion, this study has evaluated the classification performance of a VGG16 CNN model in classifying five classes of skin conditions, including three classes of cancerous lesions and two classes of non-cancerous lesions. Upon the optimization of hyperparameters including dropout rate, number of epochs, and optimizer, the VGG16 model achieved satisfactory classification performance, with an average accuracy, precision, recall, F1 score, and MCC of 0.711, 0.716, 0.712, 0.724, and 0.639, respectively. LIME, the post hoc XAI method, was applied to explain the decision made by the VGG16 in skin lesion classification. The LIME explanations use superpixel to demonstrate the features extracted by the VGG16 model visually. LIME showed that accurate predictions by VGG16, particularly the VASC class, were attributed to the truthful extraction of cancer or lesion areas. At the same time, inaccurate classifications were linked to the extraction of background and insignificant parts of the skin as core features. Overall, integrating the LIME with the VGG16 model has allowed visual inspection and justification of the model's predictions, providing a pathway for enhancing the CNN model's performance, particularly in feature extraction. Future work could use quantitative XAI to give specific weight to each feature and highlight its importance in classification.

Acknowledgment

The authors gratefully acknowledge the support provided by Tunku Abdul Rahman University of Management and Technology. This work was not supported by any grant.

References

[1] C. M. Relvas, S. G. Santos, M. J. Oliveira, F. D. Magalhães, and A. M. Pinto, "Nanomaterials for Skin Cancer Photoimmunotherapy," *Biomedicines*, vol. 11, no. 5, 2023, doi:10.3390/biomedicines11051292.

[2] V. Gruber *et al.*, "Common Benign Melanocytic and Non-Melanocytic Skin Tumors among the Elderly: Results of the Graz Study on Health and Aging," *Dermatology*, vol. 239, no. 3, pp. 379–386, 2023, doi: 10.1159/000529219.

[3] M. Tahir, A. Naeem, H. Malik, J. Tanveer, R. A. Naqvi, and S. W. Lee, "DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images," *Cancers (Basel).*, vol. 15, no. 7, 2023, doi: 10.3390/cancers15072179.

[4] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA. Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, 2023, doi: 10.3322/caac.21763.

[5] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[6] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification With Deep Learning: A Systematic Review," *Front. Oncol.*, vol. 12, pp. 1–20, 2022, doi:10.3389/fonc.2022.893972.

[7] A. S. Jaradat *et al.*, "Automated Monkeypox Skin Lesion Detection Using Deep Learning and Transfer Learning Techniques," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, 2023, doi:10.3390/ijerph20054422.

[8] Y. Nie, P. Sommella, M. Carratù, M. O'Nils, and J. Lundgren, "A Deep CNN Transformer Hybrid Model for Skin Lesion Classification of Dermoscopic Images Using Focal Loss," *Diagnostics*, vol. 13, no. 1, pp. 1–18, 2023, doi: 10.3390/diagnostics13010072.

[9] K. M. Hosny and M. A. Kassem, "Refined Residual Deep Convolutional Network for Skin Lesion Classification," *J. Digit. Imaging*, vol. 35, no. 2, pp. 258–280, 2022, doi: 10.1007/s10278-021-00552-0.

[10] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinger, "Skin lesion classification using hybrid deep neural networks," *2019 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1229–1233, 2019, doi:10.1109/icassp.2019.8683352.

[11] A. Khamparia, P. K. Singh, P. Rani, D. Samanta, A. Khanna, and B. Bhushan, "An internet of health things-driven deep learning framework for detection and classification of skin cancer using transfer learning," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 7, pp. 1–11, 2021, doi: 10.1002/ett.3963.

[12] Z. Wu *et al.*, "Studies on Different CNN Algorithms for Face Skin Disease Classification Based on Clinical Images," *IEEE Access*, vol. 7, pp. 66505–66511, 2019, doi: 10.1109/ACCESS.2019.2918221.

[13] C. Linse, H. Alshazly, and T. Martinetz, "A walk in the black-box: 3D visualization of large neural networks in virtual reality," *Neural Comput. Appl.*, vol. 34, no. 23, pp. 21237–21252, 2022, doi:10.1007/s00521-022-07608-4.

[14] Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics*, vol. 12, no. 2, 2022, doi: 10.3390/diagnostics12020237.

[15] K. Borys *et al.*, "Explainable AI in medical imaging: An overview for clinical practitioners - Saliency-based XAI approaches," *Eur. J. Radiol.*, vol. 162, p. 110787, 2023, doi: 10.1016/j.ejrad.2023.110787.

[16] H. Kaur, S. Jindal, and R. Manduchi, "Rethinking Model-Based Gaze Estimation," *Proc ACM Comput Graph Interact Tech*, vol. 5, no. 2, pp. 1–28, 2022, doi: 10.1145/3530797.

[17] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Sally, "Deep Neural Network or Dermatologist?," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, 2019, pp. 48–55. doi: 10.1007/978-3-030-33850-3_6.

[18] J. H. Ong, K. M. Goh, and L. L. Lim, "Comparative Analysis of Explainable Artificial Intelligence for COVID-19 Diagnosis on CXR Image," *Proc. 2021 IEEE Int. Conf. Signal Image Process. Appl. ICSIPA 2021*, pp. 185–190, 2021, doi:10.1109/icsipa52582.2021.9576766.

[19] M. Xiao, L. Zhang, W. Shi, J. Liu, W. He, and Z. Jiang, "A visualization method based on the Grad-CAM for medical image segmentation model," *2021 Int. Conf. Electron. Inf. Eng. Comput. Sci. EIECS 2021*, pp. 242–247, 2021, doi:10.1109/eiecs53707.2021.9587953.

[20] J. Rebane, I. Samsten, P. Pantelidis, and P. Papapetrou, "Assessing the clinical validity of attention-based and SHAP temporal explanations for adverse drug event predictions," *Proc. - IEEE Symp. Comput. Med. Syst.*, pp. 235–240, 2021, doi:10.1109/cbms52027.2021.00025.

[21] H. T. T. Nguyen, Q. C. Hung, K. V. T. Nguyen, and D. K. P. Nguyen, "Evaluation of explainable artificial intelligence: Shap, lime, and cam," *Proc. FPT AI Conf.*, pp. 1–6, 2022, doi: 10.30727/0235-1188-2022-65-1-72-90.

[22] A. Katanskiy, "Skin Cancer ISIC," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic

[23] I. The International Skin Imaging Collaboration, "ISIC Challenge Datasets." [Online]. Available: https://challenge.isic-archive.com/data/#2019

[24] P. Tschandl, C. Rosendahl, and H. Kittler, "Data Descriptor : The HAM 10000 dataset , a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Nat. Publ. Gr.*, pp. 1–9, 2018, doi: 10.1038/sdata.2018.161.

[25] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," *Proceeding 2018 IEEE 15th Int. Symp. Biomed. Imaging*, pp. 168–172, 2018, doi: 10.1109/ISBI.2018.8363547.

[26] M. Combalia *et al.*, "BCN20000: Dermoscopic Lesions in the Wild", 2019." [Online]. Available: https://arxiv.org/abs/1908.02288

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015, doi:10.48550/arXiv.1409.1556.

[28] Z. Liu *et al.*, "Improved Kiwifruit Detection Using Pre-Trained VGG16 with RGB and NIR Information Fusion," *IEEE Access*, vol. 8, pp. 2327–2336, 2020, doi: 10.1109/ACCESS.2019.2962513.

[29] C. Yang and H. Wei, "Plant Species Recognition Using Triangle-Distance Representation," *IEEE Access*, vol. 7, pp. 178108–178120, 2019, doi: 10.1109/access.2019.2958416.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, 2015, doi:10.48550/arXiv.1502.03167.

[31] M. T. Ribeiro and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier Marco," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.

[32] W. Hryniewska, A. Grudzień, and P. Biecek, "LIMEcraft: handcrafted superpixel selection and inspection for Visual eXplanations," *Mach. Learn.*, no. 0123456789, 2022, doi:10.1007/s10994-022-06204-w.

[33] B. K. Law and L. P. Lin, "Development Of A Deep Learning Model To Classify X-Ray Of Covid-19, Normal And Pneumonia-Affected Patients," *Proc. 2021 IEEE Int. Conf. Signal Image Process. Appl. ICSIPA 2021*, pp. 1–6, 2021, doi:10.1109/icsipa52582.2021.9576804.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[35] Y. Sari, Y. F. Arifin, Novitasari, and M. R. Faisal, "The Effect of Batch Size and Epoch on Performance of ShuffleNet-CNN Architecture for Vegetation Density Classification," *ACM Int. Conf. Proceeding Ser.*, pp. 39–46, 2022, doi: 10.1145/3568231.3568239.

[36] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015*, pp. 1–15, 2015, doi: 10.48550/arXiv.1412.6980.

[37] K. S. Wong and L. P. Lin, "A Comparison of Six Convolutional Neural Networks for Weapon Categorization," *Proc. Int. Conf. Electr. Eng. Informatics*, pp. 1–6, 2022, doi:10.1109/iceltics56128.2022.9932092.

[38] Y. Y. Chua and L. P. Lin, "Hyperparameter-Tuned CNN for the Classification of Ten Tomato Plant Diseases," *Proc. - 2022 2nd Int. Conf. Electron. Electr. Eng. Intell. Syst. ICE3IS 2022*, pp. 1–6, 2022, doi: 10.1109/ICE3IS56585.2022.10010269.

[39] P. Zawodny, W. Malec, K. Gill, K. Skonieczna-Żydecka, and J. Sieńko, "Assessment of the Effectiveness of Treatment of Vascular Lesions within the Facial Skin with a Laser with a Wavelength of 532 nm Based on Photographic Diagnostics with the Use of Polarized Light," *Sensors*, vol. 23, p. 1010, 2023, doi: 10.3390/s23021010.

[40] L. E. Davis, S. C. Shalin, and A. J. Tackett, "Current state of melanoma diagnosis and treatment," *Cancer Biol. Ther.*, vol. 20, no. 11, pp. 1366–1379, 2019, doi: 10.1080/15384047.2019.1640032.

[41] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, no. 13, pp. 1–22, 2021, doi: 10.1186/s13040-021-00244-z.

[42] J. Liu and Y. Zhao, "Improved generalization performance of convolutional neural networks with LossDA," *Appl. Intell.*, vol. 53, no. 11, pp. 13852–13866, 2023, doi: 10.1007/s10489-022-04208-6.

[43] C. Wang, J. Sun, W. Xu, and X. Chen, "Depth learning standard deviation loss function," *J. Phys. Conf. Ser.*, vol. 1176, no. 032050, pp. 1–7, 2019, doi: 10.1088/1742-6596/1176/3/032050.