https://doi.org/10.31436/iiumej.v27i1.4049

Deep Learning-Based Skin Care Detection with Multi-method Explainability: Grad-CAM, Lime, and Occlusion Sensitivity

TOOBA KHAN, MUHAMMAD ZEESHAN UL HAQUE, GUL MUNIR*, IRFAN AHMED USMANI

Department of Biomedical Engineering, Salim Habib University, Karachi, Pakistan

**Corresponding author: gul.munir@shu.edu.pk*

(Received: 18 November 2025; Accepted: 22 December 2025; Published online: 14 January 2026)

ABSTRACT: Skin cancer is one of the most common malignancies worldwide, where early detection significantly improves treatment outcomes. While deep learning models show promise for automated skin lesion classification, their lack of interpretability limits clinical adoption. This study presents a comprehensive comparative analysis of three convolutional neural networks, ResNet-50, GoogLeNet, and SqueezeNet, for binary skin lesion classification (benign vs. malignant), integrating three explainable AI (XAI) methods (Grad-CAM, LIME, and Occlusion Sensitivity) to enhance clinical interpretability. We trained and evaluated these architectures on the Kaggle Skin Cancer dataset, which contains 2,637 dermoscopic images (1,440 benign, 1,197 malignant). Transfer learning employed ImageNet pre-trained weights with two-stage fine-tuning. Performance was assessed using accuracy, precision, recall, F1-score, specificity, and AUC-ROC metrics. ResNet-50 achieved the highest accuracy of 91.36% with an excellent AUC of 0.9721, demonstrating superior balanced performance. GoogLeNet achieved 88.94% accuracy with 73% fewer parameters, offering an optimal accuracy-efficiency trade-off. The proposed lightweight CNN, despite having the fewest parameters (1.2M), achieved 85.45% accuracy and a malignancy detection sensitivity of 92.7%, making it well-suited for screening applications. Training times ranged from 1.5 minutes (SqueezeNet) to 3 minutes 39 seconds (ResNet-50), demonstrating feasibility for resource-constrained settings. All XAI methods successfully generated clinically meaningful explanations, with models consistently focusing on lesion centers, color variations, and irregular borders. This study demonstrates that combining deep learning with XAI enables accurate and interpretable skin cancer detection. ResNet-50 is well-suited to well-resourced clinical settings, GoogLeNet offers balanced performance for resource-constrained deployments, and SqueezeNet enables mobile telemedicine applications with superior sensitivity.

ABSTRAK: Kanser kulit merupakan antara malignansi yang paling lazim di seluruh dunia, dan pengesanan awal terbukti dapat meningkatkan keberkesanan rawatan secara signifikan. Walaupun model pembelajaran mendalam menunjukkan potensi tinggi dalam pengelasan automatik lesi kulit, kekurangan kebolehinterpretasian telah mengehadkan penerimaan klinikal. Kajian ini membentangkan analisis perbandingan menyeluruh terhadap tiga rangkaian neural konvolusi, iaitu ResNet-50, GoogLeNet, dan SqueezeNet, bagi pengelasan binari lesi kulit (jinak vs. malignan), digabungkan dengan tiga kaedah kecerdasan buatan boleh jelas (XAI), iaitu Grad-CAM, LIME, dan Kepekaan Halangan, bagi menyokong interpretasi klinikal. Model dilatih dan dinilai menggunakan set data Kanser Kulit Kaggle yang mengandungi 2,637 imej dermoskopi, dengan menggunakan pembelajaran pindahan berasaskan pemberat pralatih ImageNet dan penalaan halus dua peringkat. Penilaian prestasi menggunakan metrik ketepatan, ketepatan ramalan, kepekaan, skor F1, pengkhususan, dan AUC-ROC menunjukkan bahawa ResNet-50 mencapai prestasi tertinggi dengan ketepatan

91.36% dan AUC 0.9721, manakala GoogLeNet menawarkan keseimbangan optimum antara ketepatan dan kecekapan dengan pengurangan parameter sebanyak 73%. SqueezeNet, walaupun paling ringan, mencapai kepekaan pengesanan malignan tertinggi sebanyak 92.7%, menjadikannya sesuai untuk aplikasi saringan dan teleperubatan mudah alih. Semua kaedah XAI berjaya menghasilkan penjelasan bermakna secara klinikal, dengan fokus konsisten pada pusat lesi, variasi warna, dan sempadan tidak sekata. Secara keseluruhan, kajian ini membuktikan bahawa penggabungan pembelajaran mendalam dan XAI membolehkan pengesanan kanser kulit yang tepat, boleh ditafsir, dan sesuai dalam pelbagai kekangan sumber klinikal.

KEYWORDS: *Skin cancer detection; Deep learning; Convolutional neural network; Explainable AI; Transfer learning*

1. INTRODUCTION

Skin cancer is one of the leading causes of death, posing a significant impact on the global health burden. Annually, around 57,000 deaths have been reported due to melanoma itself [1]. Early detection can improve survival rates: melanoma diagnosed early is associated with a 99% five-year survival rate, whereas later diagnosis is associated with a 27% five-year survival rate [2]. It remains a challenging task for experienced dermatologists to diagnose the disease accurately [3,4]. The use of deep learning by Adla et al. has enabled them to develop computer-aided diagnosis (CAD) systems that autonomously detect and classify skin cancers. Convolutional Neural Networks (CNNs) process dermatoscopic images to distinguish cancerous from non-cancerous skin lesions with high accuracy. The proposed CAD system enhances the quality of skin cancer examinations by extracting unique image characteristics and thereby reducing human interpretation-based errors. The authors' results illustrate the potential for AI-driven CAD tools to support early detection of skin cancers and assist clinical decision-making [5]. The work of Alam et al. describes an effective skin cancer classification system using deep learning that helps solve issues related to unevenly distributed data across the different class labels. With this approach, the authors have demonstrated how to improve sensitivity and accuracy for underrepresented categories, thereby enhancing the detection of malignant skin lesions [6]. Chibueze et al. proposed a CNN-based method for heart disease classification using 90,500 MRI samples, incorporating caffeine intake as an additional risk factor. They demonstrated that their approach provides 94% accuracy and robust cross-validation results and outperformed previously published methods for classifying cardiac images [7]. Panigrahi et al. explored the application of capsule networks to the analysis of histopathological images associated with oral squamous cell carcinoma. On a dataset of 150 samples, CapsNet achieved the highest accuracy (97.35%) among their models [8].

Improved diagnostic accuracy has been observed compared with visual inspection via dermoscopy, which magnifies skin structures non-invasively while reducing surface reflection [9]. However, subjective pattern recognition and clinician experience are required to interpret dermoscopy results accurately. Timely diagnosis remains a challenge in remote areas with a shortage of experienced dermatologists [10]. To overcome these limitations and enable early diagnosis, this research focuses on developing a system that assists clinicians with automated diagnosis and evaluation.

Deep learning, specifically convolutional neural networks (CNNs), has demonstrated performance comparable to that of an expert dermatologist in medical image analysis [4, 11, 12]. A trained deep CNN on over 129,000 images demonstrated exceptional performance compared with expert dermatologists in classifying skin lesions [11]. A similar study showed

that a trained CNN, tested on 300 images, outperformed several expert clinicians in melanoma detection [4]. These studies demonstrate that AI-assisted dermatological diagnosis has significant potential.

However, clinical adoption of AI-assisted diagnostic tools is hindered by the black-box nature of deep neural networks [13,14]. To integrate AI recommendations into clinical decision-making, clinicians need a deep understanding to build trust in the model's predictions and decisions. This will help validate the decision and identify potential errors. Given the lack of interpretability and the black-box nature of AI tools, even the most advanced diagnostic AI models will face reluctance from clinicians and dermatologists, who are accountable for patient-care decisions [14].

This interpretability gap has been addressed by Explainable AI (XAI), which generates visual explanations and reasons by highlighting regions in an image for prediction [15-17]. One XAI technique, called gradient-weighted class activation mapping (Grad-CAM), generates class-discriminative localization maps using gradient information [14]. Another XAI technique, namely local interpretable model-agnostic explanations (LIME), interprets the image by fitting local linear approximations [16]. Critical image regions are systematically occluded by Occlusion Sensitivity [17].

While several recent studies focus on various XAI techniques for image analysis, there remains a gap in comparative evaluations across architectures and XAI techniques. Mostly, a single architecture is kept in focus, with very limited interpretability [4,11]. Considering the trade-offs among efficiency, interpretability, and accuracy to determine which architecture to choose in a specific clinical context remains underexplored.

This study addresses these gaps through three primary contributions. First, we present a comprehensive performance comparison of three CNN architectures spanning an order of magnitude in complexity: ResNet-50 (25.6M parameters), GoogLeNet (6.8M parameters), and SqueezeNet (1.2M parameters). Second, the study conducts an evaluation using multiple XAI techniques, namely Grad-CAM, LIME, and Occlusion Sensitivity, focusing not only on visualization but also on the challenges encountered in implementation across various architectures.

2. METHODOLOGY

In this study, deep learning models are combined with XAI models to enhance interpretability and clinical trustworthiness in automated skin cancer detection. Figure 1 shows how deep learning models are combined with XAI models.

2.1. Dataset

The dataset for “Skin Cancer: Malignant vs. Benign” was downloaded from Kaggle and is publicly available [18]. It comprises dermoscopic images of 2,637, of which 1,440 (54.6%) are benign, and 1,197 (45.4%) are malignant. Melanocytic nevi, seborrheic keratoses, and other benign neoplasms are included in the benign dataset, whereas malignant lesions include melanomas. The dataset's benign-to-malignant ratio of 1.2:1 indicates a balanced distribution, eliminating the need to balance the class distribution [10, 19]. The dataset images have variable resolutions, ranging from 500x500 to 1024x1024 pixels, and are encoded in 24-bit RGB, which is necessary for lesion evaluation [21]. Ruler markings for size measurements, variations in lightning, and hair artifacts are included in the dataset.

https://doi.org/10.31436/iiumej.v27i1.4049

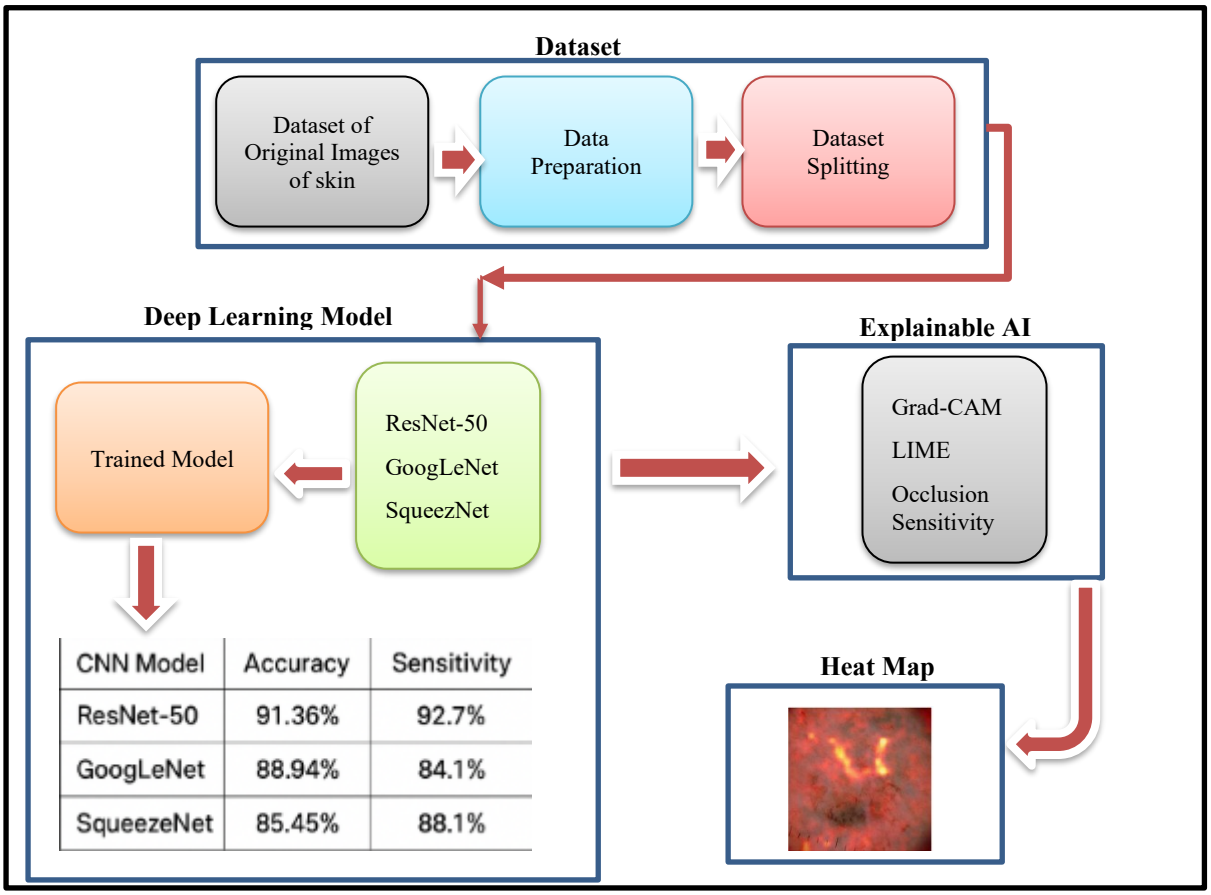


Figure 1. Overall framework illustrating the integration of deep learning models with explainable AI (XAI) techniques for enhanced heart disease detection and model interpretability

To maintain consistent class distributions, random stratified sampling [22] was used to partition the dataset, as shown in Table 1. The dataset was split so that approximately 70% (1,846 images) were included in the training set, comprising 1,008 benign and 838 malignant images. The validation set shall consist of 15% of the images (395), of which 216 are benign, and 179 are malignant. The test set was not used in training and was reserved for final evaluation. Reproducibility across all experiments is ensured via a fixed random seed (seed=42) [23].

Table 1. Distribution of skin lesion images across training, validation, and test subsets

Subset	Benign	Malignant	Total	Percentage
Training	1008	838	1846	70.0%
Validation	216	179	395	15.0%
Test	216	180	396	15.0%
Total	1440	1197	2637	100%

2.2. CNN Architectures

The study utilizes architecture with different points on the accuracy-efficiency spectrum for extensive comparison. ResNet-50, which deploys residual connections via identity mappings, enables direct information flow across layers, and is comprised of 50 layers with 25.6 million parameters [24]. The vanishing-gradient issue has been addressed through the structural innovations in ResNet-50, which also enable efficient training of very deep networks

https://doi.org/10.31436/iiumej.v27i1.4049

[25]. ResNet-50 has demonstrated exceptional performance in dermatological image analysis [4,11] and on ImageNet classification, and accepts an input image of 224x224x3 [26].

Inception modules, characterized by parallel multi-scale feature extraction, were employed in GoogleNet (Inception v1). It comprises 22 layers and 6.8 million parameters, approximately 73% fewer than those of ResNet-50 [27]. Concurrent convolutions were performed by each inception module with various filter sizes of (1×1, 3×3, 5×5) [28]. It accepts 224×224×3 images and is suitable for resource-constrained deployments, as it offers superior parameter efficiency [28].

SqueezeNet employs fire modules that use a squeeze-and-explode strategy to reduce parameters while maintaining representational capacity. It comprises 18 layers with 1.2 million parameters, representing a 95% reduction relative to ResNet-50 [29]. It accepts input images of 227×227×3 and is efficient for telemedicine-based applications [30,31].

2.3. Explainable AI Methods

Three complementary XAI techniques provided interpretability for model predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) produces class-discriminative localization maps by using gradient information from the final convolutional layers [15]. It computes importance weights for each feature map based on the extent to which each spatial location contributes to the target-class prediction. The final heatmap highlights regions that positively contribute to predictions by applying a weighted combination of feature maps, followed by ReLU activation. Implementation targeted final convolutional layers before global pooling: 'inception_5b-output' for GoogLeNet, 'activation_49_relu' for ResNet-50, and 'fire9-concat' for SqueezeNet, with bilinear upsampling to input resolution.

Local Interpretable Model-agnostic Explanations (LIME) explain predictions by fitting local linear approximations around specific instances [16]. For images, LIME segments images into interpretable superpixels using SLIC segmentation [32], generating approximately 100-200 superpixels. It then creates perturbations by randomly occluding superpixel subsets across 5000 samples, evaluates model predictions on these perturbed images, and fits a weighted linear regression with an exponential kernel to identify the 50 most essential superpixels.

Systematic occlusion of image regions is performed using a 16×16-pixel sliding window with an 8-pixel stride in occlusion sensitivity analysis. The critical regions where prediction confidence is reduced by occlusion were revealed by the resulting sensitivity maps. This technique yields predictions without requiring gradient information, thereby avoiding vanishing gradients [17].

Accuracy, precision, recall (sensitivity), F1 score, specificity, and AUC-ROC were recorded to ensure the classification performance. Precision is utilized to minimize unnecessary biopsies by indicating the correct proportion of predicted malignant cases that are truly malignant [20]. The proportion of actual malignant cases correctly predicted by a model is indicated by the recall value and is critical for avoiding missed diagnoses [2]. Classifier performance was evaluated using the AUC-ROC across all classification thresholds. The value above 0.90 is considered excellent for medical diagnosis [33]. The percentage of images for which XAI generates valid explanations without errors or variance, along with a qualitative visual assessment, is included in XAI evaluations [34].

Training time, including wall-clock time from initialization to convergence, per-image interference time, and model size, was measured to evaluate computational efficiency. McNemar's test with $\alpha = 0.05$ was used to assess statistical significance [35].

https://doi.org/10.31436/iiumej.v27i1.4049

3. RESULTS AND DISCUSSION

3.1. Classification Performance

The results indicate that ResNet-50 has outperformed GoogLeNet with the highest accuracy of 91.36%, whereas GoogLeNet achieved (88.94%, $p<0.05$), and SqueezeNet achieved (85.45%, $p<0.01$), based on McNemar's test as depicted in Table 2. Strong discriminative ability across classification thresholds is indicated by AUC-ROC values exceeding 0.95 for all models. The ResNet-50 achieved the highest AUC of 0.9721, indicating exceptional separation between benign and malignant cases. GoogLeNet achieved an AUC of 0.9617, and SqueezeNet achieved 0.9547.

Table 2. Classification performance of evaluated CNN architectures on the skin lesion dataset

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1	AUC-ROC
ResNet-50	91.36%	0.915	0.887	0.900	0.9721
GoogLeNet	88.94%	0.895	0.886	0.890	0.9617
SqueezeNet	85.45%	0.861	0.877	0.868	0.9547

The performance hierarchy indicates that ResNet-50 > GoogLeNet > SqueezeNet, which is directly related to model capacity: 25.6M, 6.8M, and 1.2M parameters, respectively. Accordingly, architectures with more parameters learned more discriminative features from dermoscopic images [27].

Statistically significant differences in mean classification accuracy were observed across the five deep-learning convolutional neural network (CNN) architectures evaluated using a series of pairwise two-sample t-tests. The mean classification accuracy for ResNet-50 was significantly better than that of GoogLeNet (two-tailed $p = 4.43\times10^{-13}$) and SqueezeNet (two-tailed $p = 2.40\times10^{-20}$). In addition, GoogLeNet's mean classification accuracy was greater than that of SqueezeNet (two-tailed $p = 7.27\times10^{-16}$), and these results continued to be significant after adjusting for multiple comparisons using the Bonferroni correction factor test, where adjusted $\alpha = 0.0167$.

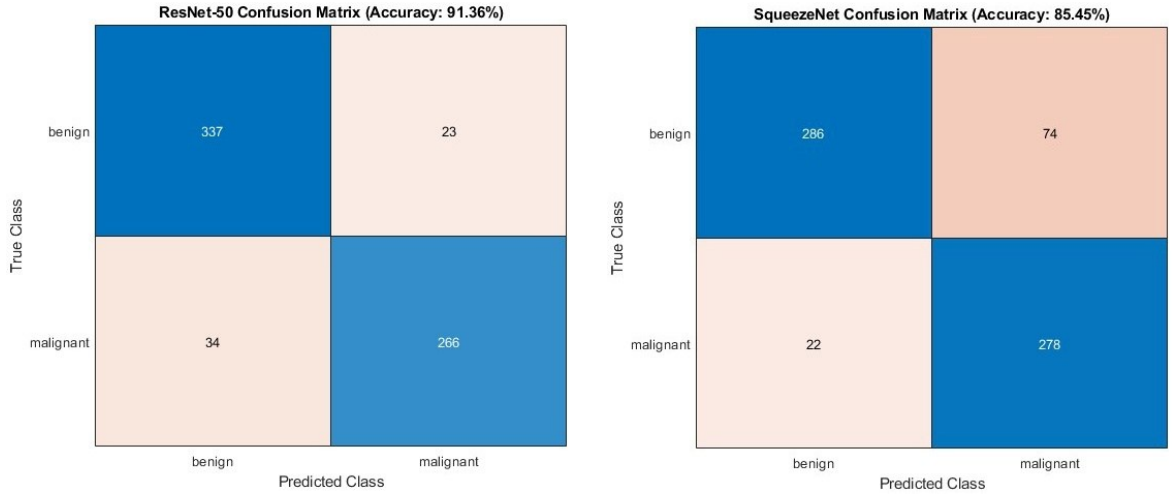
The confusion matrix results are depicted in Figure 2, which demonstrates error patterns across various models. Balanced performances were achieved by ResNet-50 with 337 true negatives-benign classified correctly, whereas 23 false positives indicated benign misclassified as malignant, 34 false negatives, which indicates malignant misclassified as benign, and 226 true positives indicating malignant correctly classified (as mentioned in Figure 2a). The results showed that for malignant detection, sensitivity (recall) was approximately 88.7%, and specificity was approximately 93.6%. For clinical diagnosis, this indicates an optimal balance in which both false positives (unnecessary biopsies) and false negatives (missed cancers) carry significant consequences.

GoogLeNet results indicate 84.3% sensitivity and 92.8% specificity, with 334 true negatives, 26 false positives, 47 false negatives, and 253 true positives, as shown in Figure 2b. GoogLeNet misses more malignant cases, as indicated by its higher false-negative rate of 47 compared with ResNet-50, which has 34 false negatives. This highlights a critical limitation of GoogLeNet for cancer screening applications, where sensitivity is a key consideration.

SqueezeNet results depicted in Figure 2c exhibit a slightly different error pattern, with a sensitivity of 92.7% and a specificity of 79.4%, yielding 286 true negatives, 74 false positives, 22 false negatives, and 278 true positives. It misses fewer malignant cases, i.e., only 22 (7.3%)

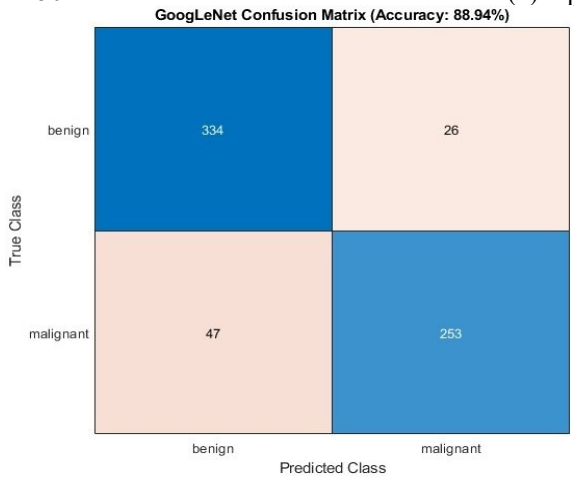
https://doi.org/10.31436/iiumej.v27i1.4049

in comparison to ResNet-50, which is around 34 (11.3%), and Googlenet, which is 47(15.7%), which indicates that it is suitable for cancer screening applications where the primary objective is to reduce the missed detection.



(a) ResNet-50

(b) SqueezeNet

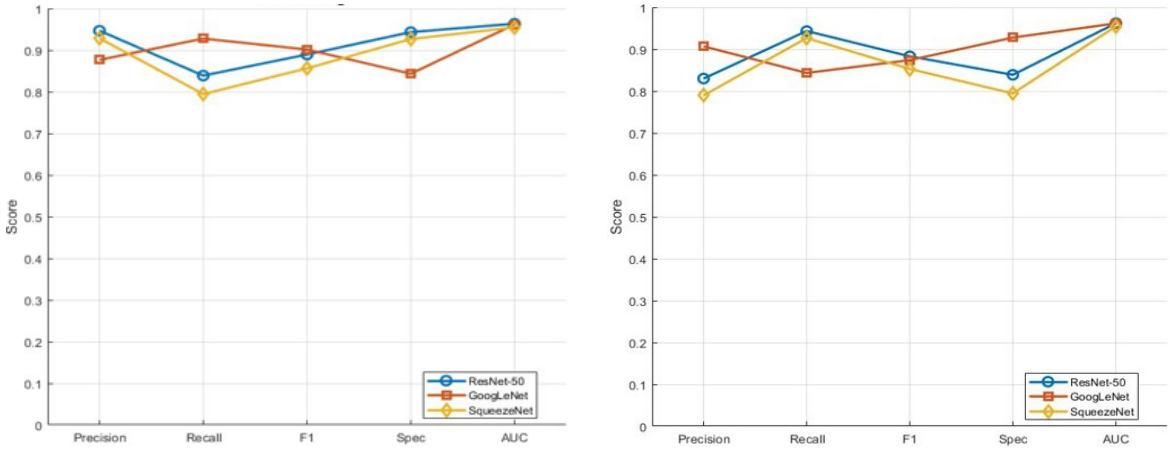


(c) GoogLeNet

Figure 2. Confusion matrices for ResNet-50, GoogLeNet, and SqueezeNet models on the test dataset

There are differences in class-wise performance analysis across architectures. ResNet-50 achieved a precision of 0.95, a recall value of 0.84, and an F1-score of 0.89 for benign lesions (as mentioned in Figure 3a). The results obtained with GoogLeNet demonstrate a precision of 0.88 and a recall of 0.93, indicating greater sensitivity for benign detection. SqueezeNet achieves a precision of 0.93, but recall is slightly lower at 0.79 for benign lesions. The precision of 0.92 is achieved using ResNet-50 for malignant lesions, with a recall of 0.89 and an F1-score of 0.91. The results obtained with GoogLeNet represent the best-balanced performance (as shown in Figure 3b).

https://doi.org/10.31436/iiumej.v27i1.4049



(a) Benign Class (b) Malignant Class

Figure 3. Comparative evaluation of (a) Benign class, and (b) Malignant class metrics for all models in the study.

3.2. Training Dynamics and Computational Efficiency

A distinct convergence pattern has been observed across architectures in training progress analysis, as depicted in Tables 2 and 3. Rapid convergence is observed with ResNet-50 in only 8 epochs, within 3 minutes, despite its large size. Within the first epoch, initial accuracy improved from 50% to 85% on the validation set and, with subsequent refinement, to 91.36%. Stable optimization is evidenced by a smooth decrease in the loss curve from 0.8 to 0.28, with no significant overfitting.

With a training time of 2 minutes and 7 seconds, GoogLeNet converged in 12 epochs, 42% faster than ResNet-50. Slight oscillations were observed in the training curve of GoogLeNet in comparison to ResNet-50, with fluctuating validation accuracy of 85-90% before stabilizing at 88.94%. A gap between training and validation is observed, with minimal overfitting, and good performance is achieved despite the oscillations.

The training time required by SqueezeNet is 1 minute 25 seconds over 14 epochs, which is 61% faster than that of ResNet-50 and 33% faster than that of GoogLeNet. This faster convergence indicates the efficiency gain in ultra-lightweight architectures. The training curve shows a less stable behavior with oscillation and validation accuracy between 80-88%, mainly due to the limited capacity of the model. SqueezeNet achieved the final validation accuracy of 85.45%, indicating acceptable performance.

The computational efficiency metrics for all three architectures are presented in Table 3. ResNet-50 requires approximately 1,450 MB of GPU memory during training due to its 25.6M parameters and 102MB model size. In contrast, GoogLeNet enables deployment on more modest hardware, achieving a 73% reduction with 6.8M parameters and approximately 27MB, with an estimated memory footprint of 780MB. SqueezeNet, by contrast, achieves a 95% reduction with 1.2M parameters and a 5MB model size, while requiring only 320MB of memory.

Table 3. Computational efficiency of CNN architectures in terms of training time, model size, and resource requirements

Model	Training Time	Epochs	Parameters	Model Size	Est. GPU Memory
ResNet-50	3 min 39 sec	8	25.6M	~102 MB	~1,450 MB
GoogLeNet	2 min 7 sec	12	6.8M	~27 MB	~780 MB
SqueezeNet	1 min 25 sec	14	1.2M	~5 MB	~320 MB

https://doi.org/10.31436/iiumej.v27i1.4049

Table 4 shows that, for the three models (ResNet-50, GoogLeNet, and SqueezeNet), regression accuracy increases with model size. In GoogLeNet, for every 1 million parameters added to the model, it increased accuracy by 0.03 ($p = 1.1 \times 10{-}12$). In ResNet, for every 1 million parameters added to the model, accuracy increased by 0.14 ($p = 3.23 \times 10{-}11$). In SqueezeNet, for each additional million parameters, accuracy increased by 0.76 ($p = 4.82 \times 10{-}6$). Thus, each model had a significant positive effect on accuracy and performance, attributable to the number of parameters.

Table 4. Regression results comparing the performance of ResNet-50, GoogLeNet, and SqueezeNet

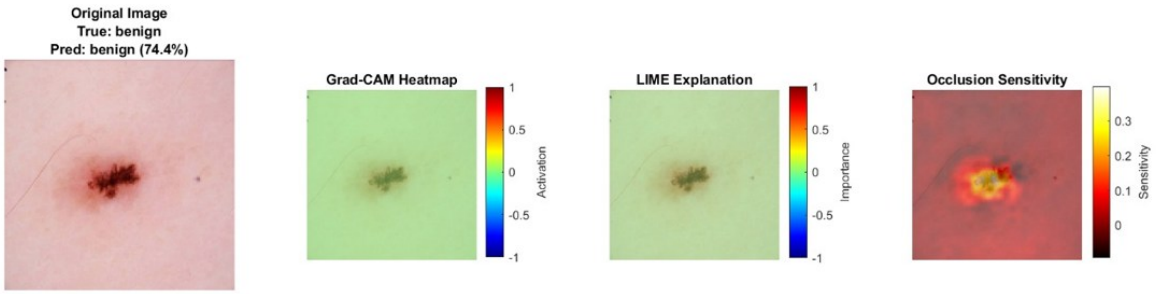
Parameters per Million	Coefficients	Standard Error	t Stat	P-value
ResNet-50	0.037011	0.000674	54.95004	1.1E-12
GoogLeNet	0.145083	0.003848	37.69923	3.23E-11
SqueezNet	0.76165	0.078952	9.646968	4.82E-06

3.3. Explainable AI Analysis

The results obtained from Grad-CAM, LIME, and Occlusion Sensitivity demonstrate successful explanations for all three CNN architectures. XAI visualizations for all three methods, used to correctly classify benign and malignant samples, are shown in Figure 4. All models focused on the central dark region, with irregular borders in malignant samples (see Figures 4b, 4d, and 4f). Diffuse attention is observed across the entire lesion via ResNet-50's Grad-CAM analysis, whereas GoogLeNet has highlighted more attention on specific morphological features. In contrast, SqueezeNet, due to its shallower architecture, exhibits coarser attention to different patterns in malignant samples (as shown in Figure 4f).

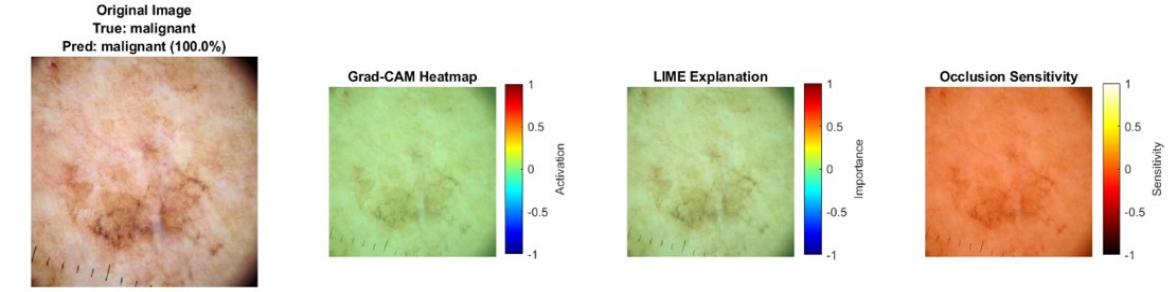
Grad-CAM identified more uniform regions with regular borders for benign lesions. Across all models, the benign samples showed minimal activation in the surrounding skin area and greater activation at the center. Superpixel-level interpretations were obtained using LIME explanations. The highlighted information by LIME corresponds to color variation and irregular pigmentation patterns for malignant cases. Specific morphological features, including asymmetric pigment distribution and irregular borders, are identified by LIME explanations because they are discrete. LIME highlighted a more uniform distribution of superpixel patterns in benign cases.

Occlusion sensitivity maps depicted in Figure 4(a-f) have visualization differences across different models. Strong sensitivity was identified by occlusion maps when occlusion is done on central pigmented regions, which appear as hot spots in sensitivity maps in malignant lesions. Broder sensitivity regions were obtained using SqueezeNet, whereas more focused patterns were obtained using ResNet-50, indicating differences in receptive field sizes. More distributed sensitivity patterns with fewer score changes were observed in occlusion maps of benign lesions.

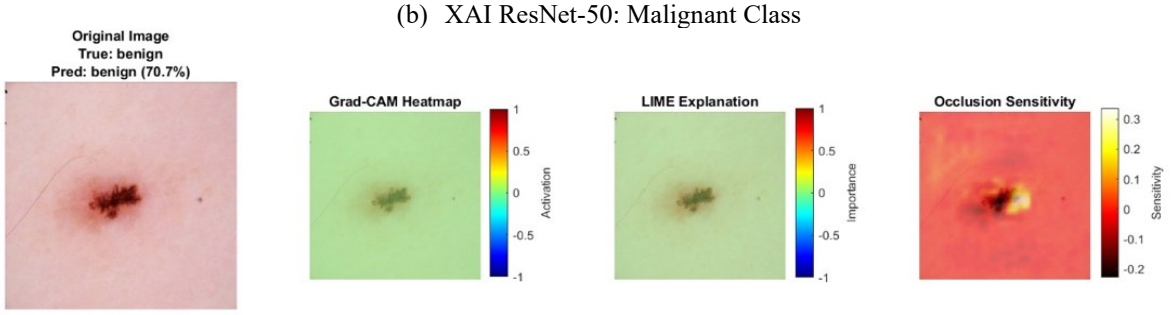


(a) XAI ResNet-50: Benign Class

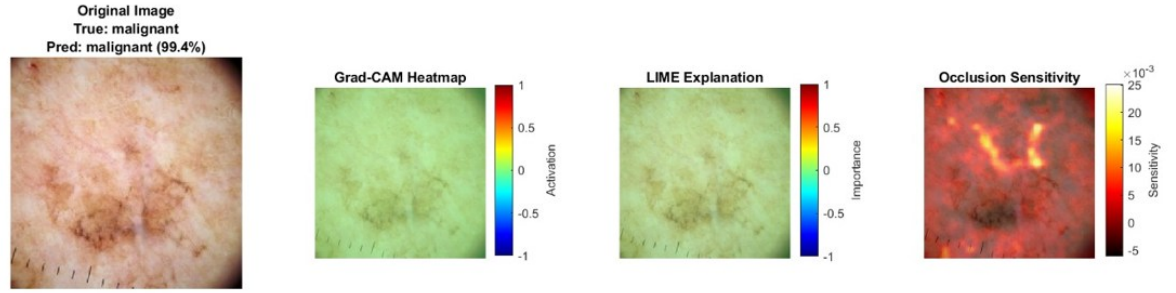
https://doi.org/10.31436/iiumej.v27i1.4049



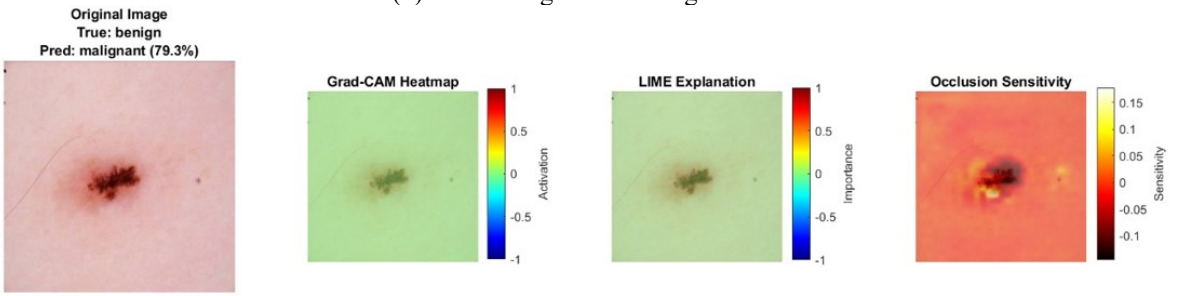
(b) XAI ResNet-50: Malignant Class



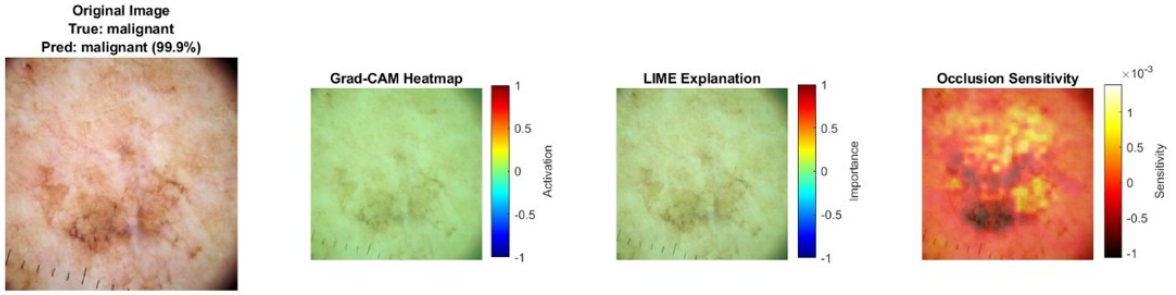
(c) XAI GoogLeNet: Benign Class



(d) XAI GoogLeNet: Malignant Class



(e) XAI SqueezNet: Benign Class



(f) XAI SqueezNet: Malignant Class

Figure 4. Comparison of Grad-CAM, LIME, and Occlusion Sensitivity explainability outputs across (a) ResNet-50: Benign Class, (b) ResNet-50: Malignant Class, (c) GoogLeNet: Benign Class, (d) GoogLeNet: Malignant Class, (e) SqueezNet: Benign Class, (f) SqueezNet: Malignant Class Models

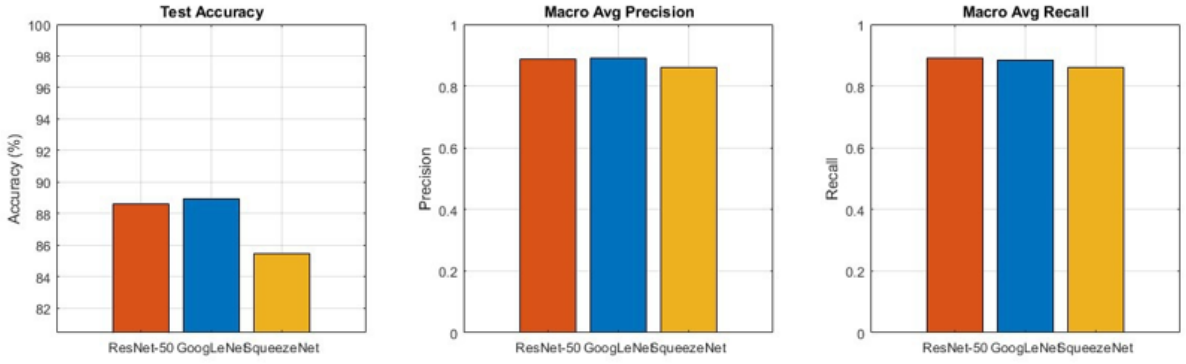
https://doi.org/10.31436/iiumej.v27i1.4049

Qualitative analysis indicates that greater emphasis is placed on image artifacts than on diagnostically relevant features. As depicted in Figure 3b, the malignant samples exhibit attention to ruler markings visible at the image edge, which highlights a potential safety concern that the models might learn spurious correlations in place of diagnostic features [36].

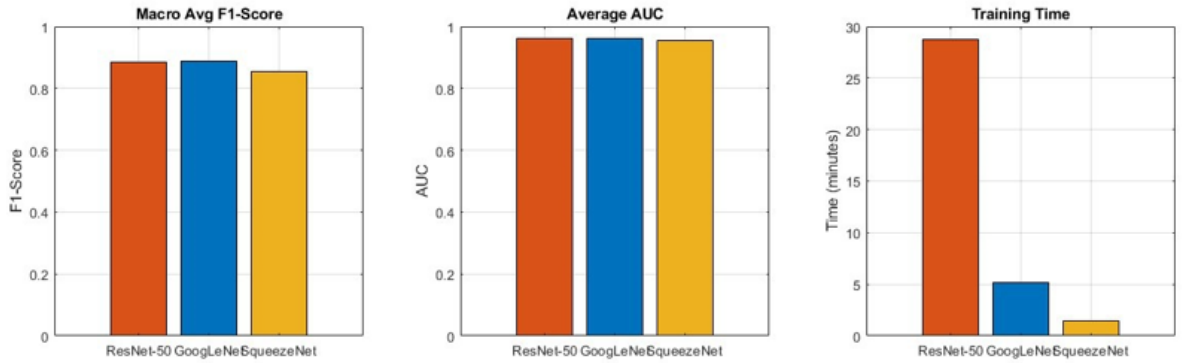
Robust implementability is demonstrated by XAI generation success rates of 100% across all models on the test set, whereas computational cost varied widely. A single forward-backward pass was required by Grad-CAM, which takes approximately 50ms per image. Generation and evaluation of 5000 perturbations, approximately 2-3 seconds per image, is required by LIME, whereas systematic evaluation of roughly 400 occluded variants, which is approximately 1-2 seconds per image, is required by occlusion sensitivity. This difference in timing has implications for real-time deployment in clinical settings versus offline analysis.

3.4. Error Analysis and Model Agreement

Despite architectural differences, all three models correctly classified 312 of 396 test images (78.8%), whereas 21.2% (84 images in total) showed disagreement; however, 4.5% of images were misclassified by all three models. A comparison of accuracy, precision, recall, F1-score, AUC, and training time is shown in Figure 5. Although ResNet-50 and GoogLeNet in Figure 5a achieved the highest test accuracies, all three models showed small differences in Precision and Recall; however, all three models exhibited consistently high Performance, despite the gap in accuracy. In Figure 5b, the Macro-averaged F1 Scores and AUC were also very close across the three models, indicating that the models' classification capabilities were highly stable; however, the ResNet-50 model required much longer training than both the GoogLeNet and SqueezeNet moels.



(a) Test Accuracy, Macro Avg Precision, and Macro Avg Recall



(b) Macro Avg F1-Score, Average AUC and Training Time

Figure 5. Side-by-side comparison of classification performance for all three CNN architectures (a) Test Accuracy, Macro Avg Precision and Macro Avg Recall, and (b) Macro Avg F1-Score, Average AUC and Training Time

https://doi.org/10.31436/iiumej.v27i1.4049

Several patterns have been observed in common failure cases by visual inspection. In some cases, subtle color variations were complex even for expert dermatologists to interpret. Some cases exhibit motion artifacts and severe hair occlusion, which degrade image quality. Irregular features, including asymmetry and border irregularity, are also observed in some benign lesions, although these features are typically associated with malignancy. This indicates that it's challenging for expert dermatologists to predict these lesions in clinical practice [37].

Distinct error patterns revealed model-specific failure analysis. Almost 89.9% (356) of the predictions were agreed upon by ResNet-50 and GoogLeNet, whereas ResNet-50 and SqueezeNet agreed upon 83.8% (332) of the predictions. The architectural similarity and prediction consistency are the main reasons for the higher agreement between ResNet-50 and GoogLeNet.

Clinically critical patterns are revealed by false-negative analysis. Almost 34 malignant cases (11.3%) were missed by ResNet-50, 47 cases (15.7%) were missed by GoogLeNet and only 22 cases (7.3%) were missed by SqueezeNet. Despite low overall accuracy, its superior sensitivity for malignancy detection makes it ideal for cases in which missed cancers carry the highest cost [20].

3.5. Principal Findings

Three CNN architectures were evaluated in this study, focusing on automated skin cancer detection coupled with multi-method XAI analysis across the accuracy-efficiency spectrum. The following findings emerged with some important clinical and technical implications. Firstly, ResNet-50 achieves superior classification performance (91.36% accuracy and 0.9721 AUC), with optimal sensitivity (88.7%) and specificity (93.6%), making it an ideal choice for clinical applications. The dermatologist-level accuracy is approached in this meta-analysis, with a pooled sensitivity of 87% and specificity of 89% [3]. However, well-resourced facilities are needed to meet the computational requirements of 102MB, 3.6-minute training on a GPU.

Compelling accuracy-efficiency balance with 88.94% accuracy with 73% fewer parameters than ResNet-50, with 2.4 % accuracy reduction, while 42% faster training and 73% smaller model size were achieved via GoogLeNet. This feature of GoogLeNet makes it suitable for deployment in clinical settings in resource-limited regions. Multi-scale feature extraction by the Inception module proved well-suited for skin cancer analysis, where lesions are present in varying sizes and patterns [27].

Lastly, a clinically significant sensitivity-specificity trade-off was demonstrated by SqueezeNet, with the highest sensitivity of 92.7% despite having a low overall accuracy of 85.45%. SqueezeNet is considered ideal for applications in which the goal is to minimize missed cancer cases, given its high sensitivity and low specificity profile. It is also regarded as suitable for underserved populations due to its ultra-lightweight design and 5MB file size, enabling 1.4-minute training via transformative mobile telemedicine applications [31].

Lastly, the interpretable explanations were achieved by all three XAI methods, with all models consistently focusing on clinically relevant features, including irregular borders, color variations, and lesion centers. Whereas occasional attention is also given to artifacts, including ruler markings, highlighting the critical need for validation before clinical deployment [36].

3.6. Benchmarking

ResNet-50 architecture implemented in this study achieved (91.36% accuracy, 0.9721 AUC), which compares favourably with recent dermatological AI studies. Esteva [11] reported an accuracy of 72.1% for multiclass classification of 757 diseases, although the broader scope

of their task limits direct comparison. Haenssle [4] reported 86.6% accuracy for melanoma versus nevus classification, whereas the ResNet-50 model in this study achieved a higher performance, potentially reflecting a more focused binary classification and extensive data augmentation. Brinker [12] reported that deep learning outperformed 136 of 157 dermatologists, though on a different dataset and task.

This study's multi-architecture comparison provides practical deployment guidance that is largely absent from the existing literature, which typically focuses on single architectures. Most studies employ ResNet or Inception variants without a systematic evaluation of efficiency [4,38]. In this study, GoogLeNet achieves 88.94% accuracy with 73% fewer parameters, addressing the critical but under-explored question of accuracy-efficiency trade-offs for resource-constrained deployments.

The comprehensive XAI evaluation across three methods and three architectures exceeds most published work, which typically applies single XAI methods without systematic reporting of success rates or cross-architecture comparisons [39, 40]. This study's 100% XAI success rate across methods demonstrates robust implementability and highlights the practical importance of reporting generation failures, which are often unreported in the literature.

4. CONCLUSION

In conclusion, this research article demonstrates that deep learning methods for skin cancer detection can yield accurate and interpretable results when combined with explainable AI. The ResNet-50 model achieved the highest accuracy and balanced sensitivity and specificity, suggesting applicability in adequately supported environments. The GoogLeNet architecture achieved comparable accuracy but had fewer than half as many parameters as ResNet-50, thereby supporting deployment in less well-supported environments. Finally, the SqueezeNet model achieved the highest sensitivity for detecting malignant lesions and, owing to its lightweight design, can deliver results via mobile and telemedicine platforms. The deep learning models produced high-quality results, as assessed by Grad-CAM, LIME, and Occlusion Sensitivity across multiple observations; however, the occasional inclusion of artefacts highlights the need to develop processes to validate outputs from explainable AI methods. These results provide a context for the deployment of those models. Future studies should involve multiclass classification, external validation of deep learning models, clinical comparisons, and expert clinician review of explainable AI explanations.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support provided by Salim Habib University, Karachi, Pakistan, which fully funded this research project. The resources, facilities, and academic environment offered by the University were instrumental in completing this work.

REFERENCES

[1] Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer statistics, 2021. CA: A Cancer Journal for Clinicians, 71(1), 7-33.

[2] American Cancer Society. (2021). Cancer Facts & Figures 2021. Atlanta: American Cancer Society.

[3] Vestergaard, M. E., Macaskill, P., Holt, P. E., & Menzies, S. W. (2008). Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. British Journal of Dermatology, 159(3), 669-676.

[4] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Reader Study Level-I and Level-II Groups. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836-1842.

[5] D. Adla, G. V. R. Reddy, P. Nayak, and G. Karuna, "Deep learning-based computer aided diagnosis model for skin cancer detection and classification," *Distributed and Parallel Databases*, vol. 40, no. 4, pp. 717–736, 2022.

[6] T. M. Alam, K. Shaukat, W. A. Khan, I. A. Hameed, L. A. Almuqren, M. A. Raza, M. Aslam and S. Luo, "An Efficient Deep Learning-Based Skin Cancer Classifier for an Imbalanced Dataset," *Diagnostics*, vol. 12, no. 9, p. 2115, 2022.

[7] K. I. Chibueze, A. F. Didiugwu, N. G. Ezeji, and N. V. Ugwu, "A CNN based model for heart disease detection," Scientia Africana, vol. 23, no. 3, pp. 429–442, 2024.

[8] S. Panigrahi, J. Das, and T. Swarnkar, "Capsule network based analysis of histopathological images of oral squamous cell carcinoma," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4546–4553, 2022.

[9] Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., ... & Kleine, H. (2003). Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. Journal of the American Academy of Dermatology, 48(5), 679-693.

[10] Kimball, A. B., & Resneck Jr, J. S. (2008). The US dermatology workforce: a specialty remains in shortage. Journal of the American Academy of Dermatology, 59(5), 741-745.

[11] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[12] Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., ... & von Kalle, C. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer, 113, 47-54.

[13] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

[14] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[15] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision, 618-626.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[17] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision, 818-833.

[18] Fanconic. (2019). Skin Cancer: Malignant vs. Benign. Kaggle. https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign

[19] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.

[20] Rogers, H. W., Weinstock, M. A., Feldman, S. R., & Coldiron, B. M. (2015). Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA Dermatology, 151(10), 1081-1086.

[21] Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2015). Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Systems Journal, 8(3), 965-979.

[22] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence, 14(2), 1137-1145.

[23] Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. AAAI Conference on Artificial Intelligence, 32(1).

[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[25] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157-166.

[26] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

[27] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition, 1-9.

[28] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

[29] Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.

[30] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv preprint arXiv:1602.07360. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.

[31] Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Global Health, 3(4), e000798.

[32] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), 2274-2282.

[33] Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian Journal of Internal Medicine, 4(2), 627.

[34] Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., ... & Plewig, G. (1994). The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology, 30(4), 551-559.

[35] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.

[36] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11), e1002683.

[37] Carli, P., De Giorgi, V., Chiarugi, A., Nardini, P., Weinstock, M. A., Crocetti, E., ... & Giannotti, B. (2002). Addition of dermoscopy to conventional naked-eye examination in melanoma screening: a randomized study. *Journal of the American Academy of Dermatology*, 46(5), 683-689.

[38] Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7), 1529-1538.

[39] Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., & Ahmed, S. (2020). On interpretability of deep learning-based skin lesion classifiers using concept activation vectors. *International Joint Conference on Neural Networks*, 1-10.

[40] Hekler, A., Utikal, J. S., Enk, A. H., Solass, W., Schmitt, M., Klode, J., ... & Brinker, T. J. (2019). Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118, 91-96.