Early Detection of Non-Melanoma Skin Lesions: A ResNet50 and SVM-Based Deep Learning Approach

Ray Louie D'Angelito[1], Josephine Larissa Rachmadiana[2], Muhammad Nur Rohman[3], Timothy Wirjantoro Harjanto[4], I Nyoman Julian Sanjaya[5], Yuri Pamungkas[6*]
[1,2,3,4,5,6] Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1]5049231008@student.its.ac.id ; [2]5049231105@student.its.ac.id ; [3]5049231040@student.its.ac.id ;
[4]5049231098@student.its.ac.id ; [5]5049231059@student.its.ac.id ; [6*]yuri@its.ac.id
* Corresponding Author

Accepted on February 11, 2026
Approved on May 11, 2026

***Abstract*—Skin cancer is a rapidly growing global health issue that demands efficient and accurate early detection. While much research focuses on aggressive types like melanoma, basal cell carcinoma, and squamous cell carcinoma, there is also a critical need for non-invasive classification of non-melanoma skin lesions, including actinic keratosis, dermatofibroma, nevus, and vascular lesions. These conditions can also progress to more severe malignancies or indicate underlying health issues. This study presents a computer-aided detection method for these skin conditions by employing deep learning techniques, specifically a ResNet50-based Convolutional Neural Network (CNN), alongside a Support Vector Machine (SVM) classifier. The aim is to improve diagnostic accuracy and accessibility through image data processing and feature extraction. The main contribution of this research is the application of deep learning for automated classification of non-melanoma skin lesions, with the goal of enhancing early detection. The models were trained and evaluated using the International Skin Imaging Collaboration (ISIC) dataset, with two test scenarios to assess their performance. In Test 4, the CNN demonstrated superior results, achieving F1-scores of 44.70% for actinic keratosis, 85.25% for dermatofibroma, 78.76% for nevus, and a perfect 100.00% for vascular lesion. In comparison, the SVM model achieved lower F1-scores: 21.88% for actinic keratosis, 27.91% for dermatofibroma, 62.46% for nevus, and 70.58% for vascular lesion. The results highlight the effectiveness of deep learning, particularly CNNs, in automated dermatological diagnosis. These findings lay the groundwork for future web and mobile applications that could support early skin cancer detection and clinical decision-making.**

***Index Terms*—Skin Cancer Detection; Deep Learning; Convolutional Neural Network (CNN); Support Vector Machine (SVM); Non-Melanoma Skin Lesions**

I. INTRODUCTION

Skin cancer is among the most prevalent malignancies worldwide, with cases steadily increasing each year [1]. Although clinical and research efforts have largely concentrated on highly aggressive types such as melanoma, basal cell carcinoma, and squamous cell carcinoma, the significance of non-melanoma skin lesions (including actinic keratosis, dermatofibroma, nevus, and vascular lesions) remains critical [2]. These lesions are not only widespread but can also progress into more severe malignancies or reflect underlying systemic diseases if not detected in time.

Actinic keratosis, for instance, is a pre-cancerous lesion that may advance to squamous cell carcinoma when left untreated [3]. Dermatofibroma, while usually benign, often resembles malignant tumors under dermoscopy and can complicate diagnosis [4]. Nevi, commonly referred to as moles, are mostly harmless but certain atypical types have malignant potential and can evolve into melanoma [5]. Likewise, vascular lesions can signal deeper vascular or systemic abnormalities [6]. These examples underline the necessity of accurate and early diagnosis for improving patient outcomes and reducing mortality.

The research problem lies in the persistent difficulty of differentiating these lesions at an early stage due to their overlapping morphological characteristics, which can lead to misdiagnosis and delayed treatment [7]. Conventional diagnostic methods, including dermoscopic evaluation, rely heavily on the expertise of dermatologists and may be subject to inter-observer variability [8]. Therefore, there is a pressing need for automated, reliable, and scalable solutions to assist clinicians in improving diagnostic accuracy.

To address this challenge, deep learning has emerged as a powerful solution, with Convolutional Neural Networks (CNNs) proving highly effective in analyzing medical images [9]. CNNs automatically extract complex visual features, making them particularly well-suited for dermatological image classification [10]. State-of-the-art models such as ResNet50 have shown robust performance in image recognition tasks due to their residual connections that overcome the vanishing gradient problem, allowing for deeper and more accurate networks [11]. In parallel, Support Vector Machines (SVMs) have also been utilized in prior research, demonstrating strong performance in distinguishing between benign and malignant lesions by finding optimal separating hyperplanes in high-dimensional feature spaces [12].

This study builds upon these advances by employing ResNet50 to classify actinic keratosis, dermatofibroma, nevus, and vascular lesions using a publicly available dataset from the International Skin Imaging Collaboration (ISIC) [13]. By comparing the deep learning approach with SVMs, the research seeks to evaluate the strengths and limitations of both methods in the context of multiclass classification of non-melanoma lesions.

The novelty of this study lies in its specific focus on non-melanoma skin lesions, which have often received less attention compared to melanoma, while leveraging advanced deep learning architecture and benchmarking it against traditional machine learning methods [14],[15]. This dual approach provides a more comprehensive perspective on automated lesion detection and highlights the feasibility of integrating such methods into accessible diagnostic platforms.

The contribution of the research is twofold: first, it provides an in-depth evaluation of ResNet50 and SVM performance in the classification of four clinically significant non-melanoma skin lesions, and second, it offers insights into the potential integration of these computational models into web- and mobile-based applications for wider accessibility. Ultimately, this work contributes to the ongoing effort of enhancing early detection and supporting dermatologists in clinical decision-making.

II. RESEARCH METHODS

This research proposes an early detection method for non-melanoma skin lesions using a ResNet50-based CNN and Support Vector Machine (SVM) classifier. The process begins with dataset preparation, where images are resized to 224x224 pixels, followed by data augmentation techniques such as flipping, rotation, and color shifting to enhance model robustness. Image segmentation isolates the lesions, focusing the model on relevant features. The ResNet50-based CNN extracts features, which are then classified into categories like nevus, dermatofibroma, actinic keratosis, and vascular lesions using SVM. The model's performance is evaluated through metrics like accuracy, precision, recall, F1-score, AUC, and specificity, with Grad-CAM visualization used to highlight the areas of the image the model focuses on, improving interpretability. This approach enables accurate, automated classification, assisting clinicians in early detection and diagnosis, as illustrated in **Figure 1**.

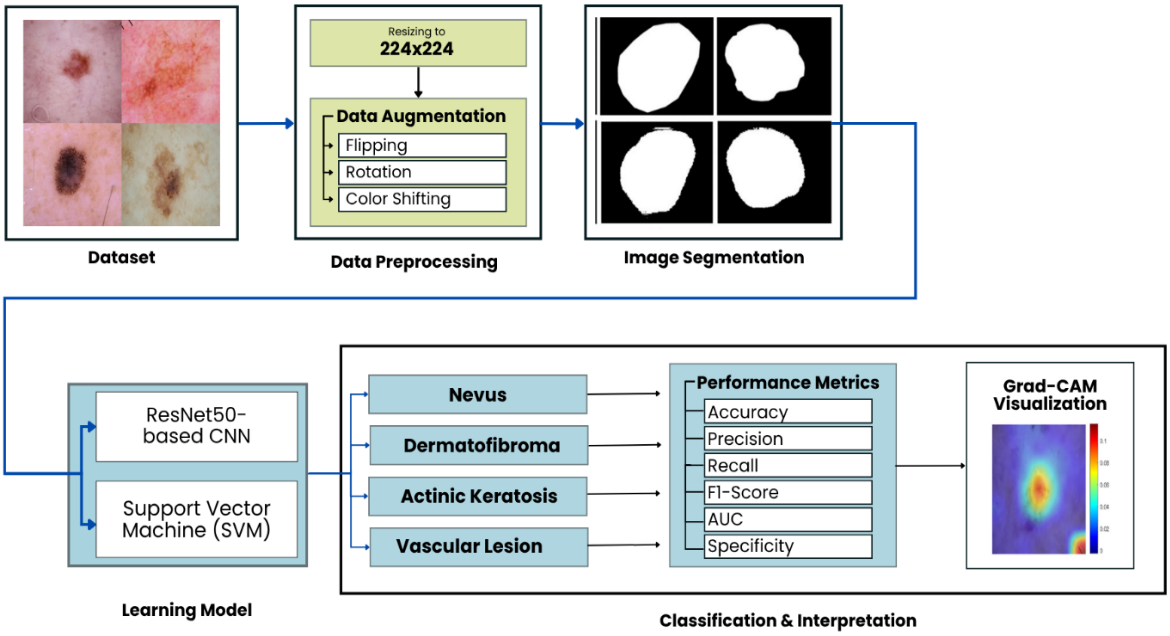

Fig. 1. Methodology of the research

A. Related Works

The early detection of non-melanoma skin lesions has been a focal point in recent dermatological research, with significant progress achieved through the use of deep learning techniques. Akinrinade et al. (2025) demonstrated that deep learning models, specifically CNNs and GANs, combined with transfer learning and few-shot learning, improve skin cancer detection, particularly melanoma [16]. Their work showed how these models can address dataset imbalances and small data challenges, highlighting the value of using data augmentation techniques. Their findings indicate that the use of transfer learning with ResNet50 achieved an impressive accuracy of 96%, making it a viable model for skin cancer detection in clinical settings.

Similarly, Erbay et al. (2025) explored the combination of CNNs and Vision Transformers (ViTs) with ensemble learning approaches, such as hard and soft voting, to improve skin lesion classification [17]. Their study, based on the HAM10000 dataset, found that ensemble methods combining CNNs and ViTs yielded the best performance, achieving an accuracy of 89% and an F1-score of 0.80. This ensemble learning approach, which leverages the strengths of multiple models, has shown promise in enhancing classification accuracy, especially when applied to large, imbalanced datasets.

Another study by Zareen et al. (2024) employed a hybrid CNN-RNN architecture combining ResNet-50 and LSTM to classify skin lesions [18]. This approach achieved an impressive accuracy of 99.06% on the ISIC dataset. The incorporation of RNNs for temporal learning enabled this model to handle various skin cancer types effectively. The high accuracy and superior performance of this hybrid model underline the potential of combining different deep learning techniques for more robust and accurate skin lesion classification.

In a different approach, Nalamwar et al. (2025) introduced a multidimensional CapsNet with attention mechanism, achieving an accuracy of 97.63% with high precision and recall (98.11% and 98.73%, respectively) [19]. This model outperformed traditional CNNs and GoogLeNet, offering a significant improvement in skin cancer classification. The use of attention mechanisms in CapsNets allows the model to focus on important features, which is crucial for detecting skin lesions that exhibit subtle variations.

Manickavasagam et al. (2025) applied CNN-based transfer learning, enhanced with harmonic brown bear optimization, to the SIIM-ISIC Melanoma dataset [20]. Their model achieved an accuracy of 91.75%, with a true positive rate (TPR) of 93.76% and a true negative rate (TNR) of 89.77%. The use of optimization techniques to enhance transfer learning models resulted in superior skin cancer detection performance, reducing false positives and improving overall diagnostic accuracy.

The use of hybrid models combining CNN and ViT has also shown promising results. Pacal et al. (2025) introduced a CNN-ViT hybrid model with focal self-attention, yielding accuracy rates of 92.54% on the ISIC 2019 dataset and 95.01% on the HAM10000 dataset [21]. Their approach demonstrated that hybrid models are well-suited for real-time deployment, improving both accuracy and efficiency in skin lesion classification.

In summary, these works illustrate the significant strides made in deep learning for skin cancer detection, with a variety of models and approaches contributing to improved accuracy and robustness. The integration of multiple AI techniques, including CNNs, ViTs, RNNs, and ensemble methods, has proven to be effective in addressing the challenges of skin lesion classification, particularly in non-melanoma skin lesions. This research paves the way for the development of advanced AI-driven systems for the early detection of skin cancers, improving clinical decision-making and patient outcomes. A summary of related works is presented in **Table I**.

TABLE I. RESEARCH RELATED TO SKIN CANCER DETECTION BASED ON DEEP AND MACHINE LEARNING

Author(s)	Year	AI Method	Dataset	Metrics Evaluation Value	Key Findings*
Akinrinade, et al [16]	2025	Deep Learning (CNN, GAN, Transfer Learning)	ISIC 2016, ISIC 2017, HAM10000	Accuracy: 96%, Sensitivity: 91.1%, Specificity: 95.7% (ResNet50), AUC: 0.99	Deep learning models improve skin cancer detection, especially melanoma.
Erbay, et al [17]	2025	CNN, Vision Transformer, Ensemble Learning	HAM10000	Accuracy: 0.89% (ensemble), F1-score: 0.80	Ensemble learning improves skin lesion classification accuracy.
Zareen, et al [18]	2024	Hybrid CNN-RNN (ResNet-50 + LSTM)	ISIC (9000 images)	Accuracy: 99.06%	Achieved high accuracy in classifying skin cancer types.
Nalamwar, et al [19]	2025	Multidimensional CapsNets with Attention	HAM10000	Accuracy: 97.63%, Precision: 98.11%, Recall: 98.73%, F1-score: 98.42%	Outperforms CNN and GoogLeNet with high precision and recall.
Manickavasagam, et al [20]	2025	CNN-based Transfer Learning (DenseNet)	SIIM-ISIC Melanoma, Skin Cancer	Accuracy: 91.75%, TPR: 93.76%, TNR: 89.77%	Superior performance with enhanced precision and reduced false positives.

Author(s)	Year	AI Method	Dataset	Metrics Evaluation Value	Key Findings*
Pacal, et al [21]	2025	CNN-ViT Hybrid	ISIC 2019, HAM10000	ISIC 2019: Accuracy: 92.54%, Precision: 90.41%, Recall: 87.68%, F1: 88.86%; HAM10000: Accuracy: 95.01%, Precision: 94.70%, Recall: 92.11%, F1: 93.34%	Superior performance, high accuracy, and real-time deployment.
ThangaPurni, et al [22]	2025	Hybrid ARP-ViT-CNN	HAM10000	Accuracy: 98.2%, Precision: 0.94, Recall: 0.96, F1-score: 0.95	Outperforms CNN-only and ViT-only models with high accuracy.
Khattar, et al [23]	2025	Hybrid DeepCNN	HAM10000	PSNR: 45.56, MSE: 2.43, SSIM: 0.9781	Hybrid method improves image quality, enhancing segmentation and classification.
Reis, et al [24]	2024	MABSCNET (CNN + ViT)	ISIC 2020, ISIC 2018, Kaggle	ISIC 2020: Accuracy: 92.74%, ISIC 2018: 100%, Kaggle: 94.24%	Hybrid model outperforms CNN and ViT in skin cancer detection.
Nanda, et al [25]	2025	CNN, GoogleNet, ResNet	HAM10000	Accuracy: CNN 73%, GoogleNet 71%, ResNet 47%; Precision, Recall, F1-Score, AUC-ROC	CNN with mean sampling performs best, addressing class imbalance.

* "Key Findings" refers to the primary methodological or performance-related insight reported by each study that is most relevant to the classification of skin lesions using deep or machine learning.

B. Dataset

The dataset used in this study was obtained from the ISIC 2019 Challenge on Kaggle [26]. The ISIC dataset is one of the most widely recognized benchmarks in dermatological image analysis, compiled through a collaboration between leading skin cancer research institutions worldwide. It was specifically developed to support the training and evaluation of automated skin lesion classification systems, providing standardized dermoscopic images with expert-validated annotations [13]. The dataset spans a broad spectrum of skin conditions, enabling controlled evaluation of model performance across both common and rare lesion types. Its use in this study ensures comparability with prior work and provides a reliable foundation for assessing the proposed method. A curated subset of 962 images was selected to focus on four specific skin lesion types: actinic keratosis (329 images), dermatofibroma (122 images), nevus (369 images), and vascular lesion (142 images). To meet the input requirements of the ResNet50 model, the images were resized to 224×224 pixels [27]. To address class imbalance and enhance the model's robustness, data augmentation techniques including horizontal and vertical flipping, random rotation, and color shifting (adjustments to brightness, contrast, and saturation) were applied [28],[29]. Among these, flipping and rotation are expected to be particularly beneficial for minority classes such as dermatofibroma and vascular lesion, as they generate geometrically diverse variants of lesions whose appearance is orientation-independent — meaning a dermatofibroma looks the same regardless of rotation, so each rotated copy is a genuinely useful training sample. Color shifting, while useful overall, is likely to be most impactful for vascular lesions, which are characterized by distinct reddish hues that benefit from intensity variation to prevent the model from relying solely on color cues. Future work should consider class-specific augmentation strategies, such as applying higher rotation multipliers or synthetic oversampling (e.g., SMOTE) exclusively to the underrepresented classes, to more directly address the imbalance rather than applying a uniform augmentation pipeline across all classes. The final dataset consists of 2,357 images, representing both benign and malignant skin conditions. Although the original ISIC dataset includes a variety of skin conditions like basal cell carcinoma, melanoma, and pigmented benign keratosis, this study narrowed its focus to the four clinically significant conditions mentioned above. This targeted approach allows for a more detailed evaluation of the model's performance on these specific lesion types.

C. Feature Extraction and Selection

In this study, two different approaches to obtaining features from images were employed, depending on the classifier. For the CNN-based model, the process is one of automatic feature extraction: the network learns and derives hierarchical representations directly from raw pixel data without any manual intervention. Through multiple convolutional layers, CNNs gradually develop abstract and discriminative features, starting from basic visual elements like edges and textures to more complex semantic patterns [30]. This process eliminates the need for manual feature engineering, making traditional methods like MRMR (Minimum Redundancy Maximum Relevance) unnecessary for the CNN model.

For the Support Vector Machine (SVM) classifier, handcrafted feature extraction was performed manually using standard image processing techniques. A total of 15 features were derived, covering first-order statistics, shape descriptors, and texture features. No automated feature selection algorithm (such as MRMR) was applied; instead, all 15 features were used collectively as input to the SVM, representing a deliberate feature extraction (not selection) strategy. These features typically include morphological properties (such as area and perimeter), intensity distributions (like mean, standard deviation, and histogram width), and texture features (e.g., Gray Level Co-occurrence Matrix (GLCM) and Gray Level Size Zone Matrix (GLSZM)). Although the specific number of features or the use of a feature selection algorithm like MRMR was not explicitly detailed for the SVM, it is assumed that a comprehensive set of relevant features was extracted and used, which is a common approach in traditional machine learning for image analysis.

D. Classifiers

This research compares the performance of two distinct machine learning approaches for classifying skin lesions:

1) CNN (ResNet50 Architecture)

The ResNet50 model was chosen for its strong performance in complex image classification tasks. Its architecture enables the training of very deep networks, addressing the vanishing gradient problem through residual connections [31]. In this study, the ResNet50 model was fine-tuned and trained on preprocessed dermatoscopic images to classify them into four categories of skin lesions: actinic keratosis, dermatofibroma, nevus, and vascular lesion.

2) Support Vector Machine (SVM)

SVM is a well-established supervised machine learning algorithm known for its effectiveness in high-dimensional data classification [32],[33]. The SVM works by creating a hyperplane, or a set of hyperplanes, in a high-dimensional space that separates data points into distinct classes. For this research, the SVM classifier was trained on handcrafted features extracted from the dermatoscopic images, allowing for a comparative analysis with the deep learning approach.

E. Metrics Evaluation

The performance of both the CNN and SVM classification models was thoroughly evaluated using standard metrics derived from the confusion matrix. The confusion matrix (**Figure 2**) offers a comprehensive view of the model's predictions, clearly distinguishing between correct and incorrect classifications for each class.. For multi-class classification, the metrics were calculated using a one-vs-all approach for each class. This method ensures that the metrics are computed independently for each class, addressing the issue of imbalanced class distributions [34],[35].

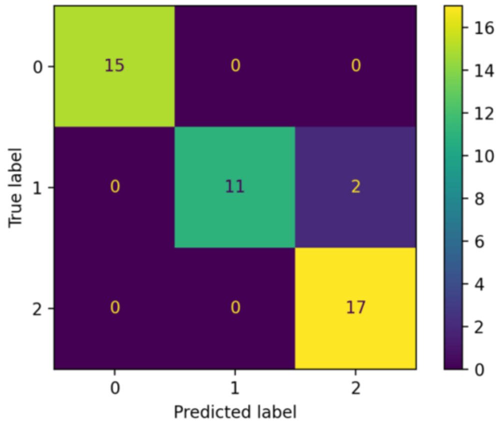

Fig. 2. Example of Confusion Matrix [36]

The fundamental counts extracted for each class (actinic keratosis, dermatofibroma, nevus, vascular lesion) are:

a. True Positives (TP): Instances correctly predicted as positive for the class.

b. True Negatives (TN): Instances correctly predicted as negative for the class.

c. False Positives (FP): Instances incorrectly predicted as positive for the class (Type I error).

d. False Negatives (FN): Instances incorrectly predicted as negative for the class (Type II error).

From these counts, the following performance metrics were calculated:

1) Accuracy (Acc): The overall proportion of correct predictions across all classes [37].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \qquad (1)$$

2) Precision (Prec): The proportion of correctly predicted positive instances out of all instances predicted as positive for a specific class [38].

$$Precision = \frac{TP}{TP + FP} \times 100 \qquad (2)$$

3) Recall / Sensitivity (Sens): The proportion of actual positive instances that were correctly identified by the model [39].

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \qquad (3)$$

4) Specificity (Spec): The proportion of actual negative instances correctly identified by the model (true negative rate) [40].

$$Specificity = \frac{TN}{TN + FP} \times 100 \qquad (4)$$

5) F1-Score: The harmonic mean of Precision and Recall, providing a balanced measure of the model's accuracy, especially useful in cases of imbalanced class distributions [41].

$$F1 = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \times 100 \qquad (5)$$

6) Area Under the Receiver Operating Characteristic Curve: A comprehensive measure of a model’s ability to distinguish between classes, representing the probability that the model will rank a randomly chosen positive example higher than a randomly chosen negative example [42].

III. RESULTS AND DISCUSSIONS

A. Dataset Overview and Preprocessing

The study utilized a comprehensive skin cancer dataset designed to capture the variability of four distinct classes: actinic keratosis, dermatofibroma, nevus, and vascular lesion. These conditions, though categorized under skin cancer, vary significantly in clinical presentation and histopathological features, making them challenging for automated classification. The dataset included images that were preprocessed to 224×224 pixels, which is the optimal size for input into the ResNet-50 architecture. The class distribution of the dataset is presented in **Table II**. This preprocessing step was crucial as it ensured that the images were consistent in dimensions and ready for efficient feature extraction, thereby improving the performance of the convolutional neural network (CNN) model [43].

TABLE II. DATASET DISTRIBUTION

Class	Number of Images	Percentage of Total
Actinic Keratosis	329	34.2%
Dermatofibroma	122	12.7%
Nevus	369	38.4%
Vascular Lesion	142	14.8%
Total	**962**	**100%**

The dataset demonstrates a clear imbalance in the distribution of classes, with nevus comprising the largest proportion (38.4%), followed by actinic keratosis (34.2%), vascular lesions (14.8%), and dermatofibroma (12.7%) as the smallest class. This uneven distribution presents a challenge for machine learning algorithms, as models tend to exhibit bias toward the majority class. In practical terms, this means that the model may be more inclined to predict the larger classes (nevus and actinic keratosis) accurately, while struggling to correctly classify images from the smaller classes, especially dermatofibroma.

B. Training Performance Analysis

The training curve for the 30% training configuration demonstrates excellent convergence characteristics. The validation accuracy curve shows steady improvement throughout the epochs, reaching a plateau at approximately 90.38%, with minimal signs of overfitting. Notably, the smoothed training accuracy curve (represented by the blue line) illustrates rapid learning during the early epochs, suggesting that the model quickly adapts to the data. The validation points, depicted as black dots, show consistent performance without significant fluctuations, which indicates that the model is generalizing well. Furthermore, the loss curves show an inverse relationship with accuracy, which is expected. The training loss decreases steadily as the model learns, while the validation loss stabilizes, signifying that the model has achieved a state of effective generalization. These observations suggest that the model's hyperparameters were well-chosen and that transfer learning from the pretrained ResNet-50 weights contributed to the efficient convergence of the model. The steady performance and stable learning curves indicate that the model is benefiting from both the training data and pretrained weights, making this configuration quite effective [44]. The training progress for this configuration is shown in **Figure 3**.

The 25% training configuration produced the most remarkable results, achieving the highest validation accuracy of 90.69% despite using less training data than the 30% configuration. Interestingly, this result challenges the conventional assumption that more training data always leads to better performance. The convergence pattern for the 25% configuration shows superior stability when compared to the 30% configuration. The validation accuracy progression is smoother, with fewer oscillations, indicating that the model is learning more consistently. This suggests that the 25% configuration strikes an optimal balance, providing enough training data for the model to learn effectively while avoiding the risks of overfitting that often accompany larger datasets. The training loss curve also demonstrates a more consistent descent, while the validation loss stabilizes at a lower level compared to the 30% configuration. These observations confirm that the 25% configuration leads to better generalization capability, making it a counterintuitive but effective choice. This outcome underscores the importance of carefully selecting the amount of training data, as the model can still achieve high performance with less data when overfitting is controlled [45]. A plausible explanation for the superior stability of the 25% configuration over the 30% configuration lies in the interaction between dataset size and class imbalance. With a larger test set (30%), the model is exposed to a higher absolute number of majority-class samples (nevus and actinic keratosis) during training, which may amplify the existing class imbalance and introduce additional noise into the gradient updates. This can cause the model to overfit to the dominant class distributions while reducing its generalization capacity for minority classes such as dermatofibroma. The 25% configuration, by retaining a slightly larger proportion of data for testing, inadvertently creates a more balanced training dynamic — the model receives just enough exposure to each class to develop robust representations without becoming biased toward any single distribution. The smoother validation loss curve observed in the 25% configuration supports this interpretation, as it indicates more consistent gradient updates and fewer conflicting learning signals across epochs. The training progress for this configuration is shown in **Figure 4**.

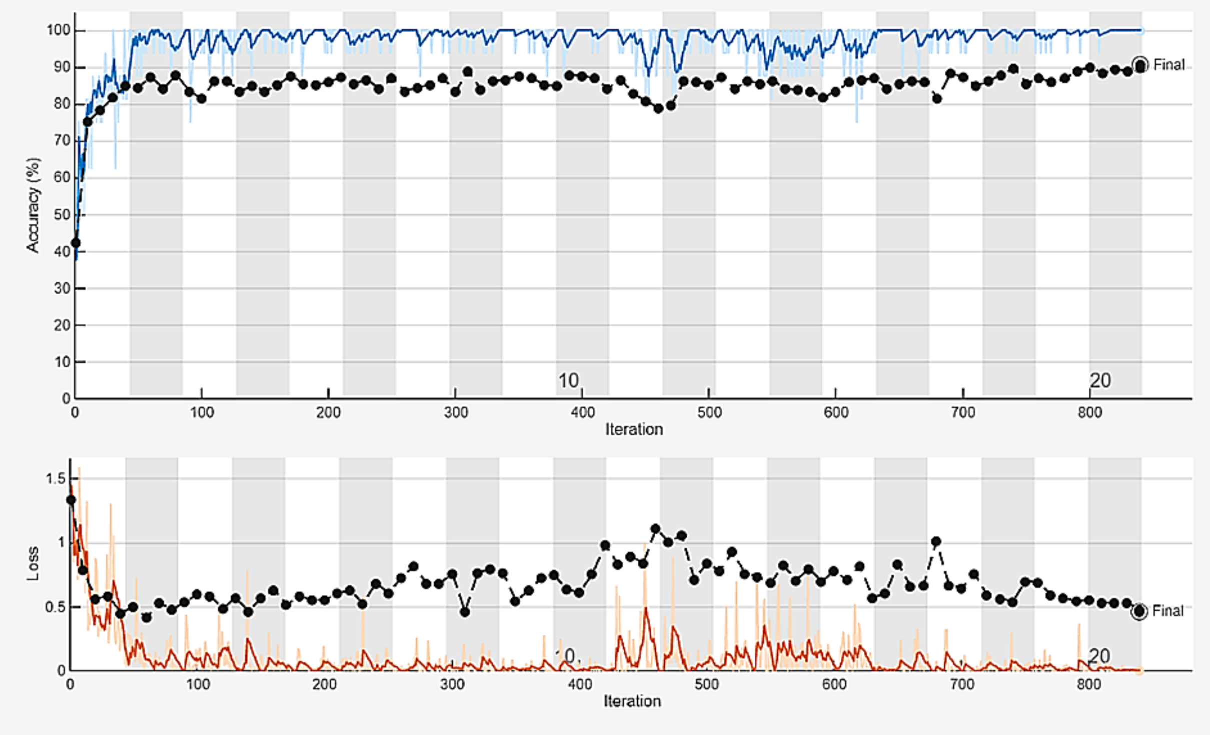

Fig. 3. Training progress of the CNN model with 30% data, showing accuracy and loss curves, with a final validation accuracy of 90.38%.

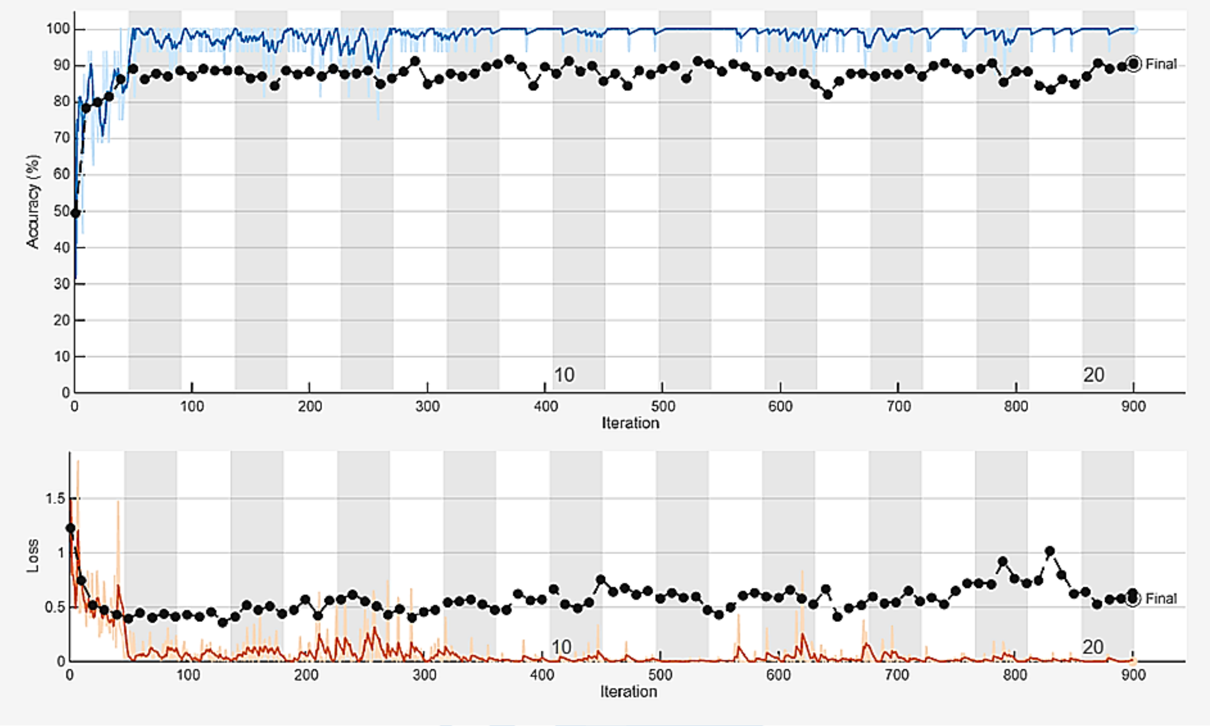

Fig. 4. Training progress of the CNN model with 25% data, showing accuracy and loss curves, with a final validation accuracy of 90.69%.

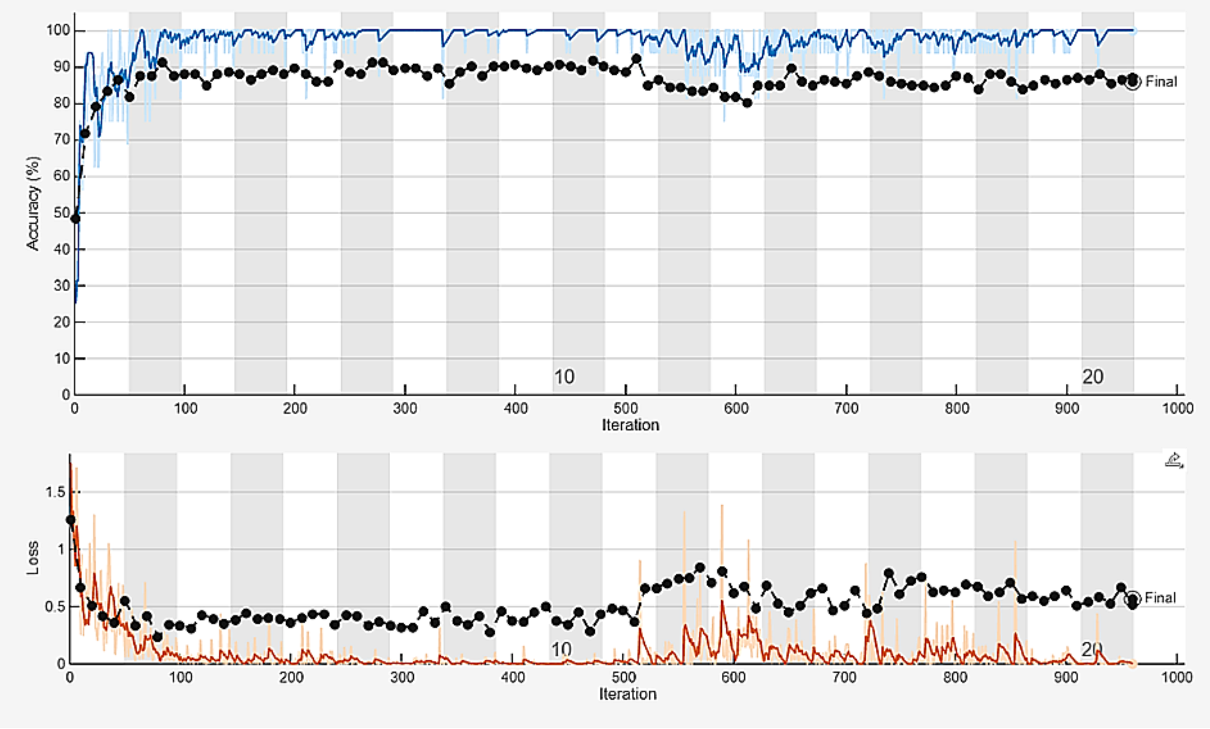

Fig. 5. Training progress of the CNN model with 20% data, showing accuracy and loss curves, with a final validation accuracy of 85.94%.

The 20% training configuration, which uses the least amount of training data, shows the expected performance degradation, with the model achieving a validation accuracy of 85.94%. This result highlights the limitations that come with using smaller datasets, as the model is constrained in terms of the amount of information it can learn. Despite this, the convergence pattern remains stable and consistent, indicating that the model architecture is robust even with minimal training data. The training curves exhibit more pronounced initial learning phases, which suggests that the model must extract the maximum amount of information from the limited examples available. The validation accuracy plateaus earlier than with larger training sets, which implies that the model reaches its learning capacity constraints sooner. Even though the performance is lower compared to configurations with more data, the consistent and stable convergence pattern serves as evidence that the ResNet-50 architecture is still effective under these conditions. The model continues to perform reliably, demonstrating its robustness and suitability for handling smaller datasets [46]. The training progress for this configuration is shown in **Figure 5**.

The training metrics reveal several unexpected patterns that challenge conventional machine learning assumptions and provide deeper insights into the relationship between data size, training time, and model performance. Among the configurations tested, the 25% training data configuration achieved the highest validation accuracy of 90.69%, despite requiring the longest training time of 24 minutes and 54 seconds. This suggests that the model underwent more complex optimization dynamics with this configuration, possibly due to the data's balance between sufficient complexity for the model to learn and avoiding the overfitting risks associated with larger datasets. The longer training time, while initially seeming inefficient, may have provided the necessary learning steps for the model to fine-tune its weights, leading to the highest accuracy [47].

In contrast, the 30% configuration demonstrated the best efficiency ratio, achieving a validation accuracy of 90.38% in the shortest relative training time. This configuration completed the training process more quickly while maintaining a competitive accuracy score, indicating that the model was able to extract useful features from the larger dataset more rapidly, without sacrificing performance. The efficiency of the 30% configuration underscores the importance of not just raw training time, but the rate at which a model can learn and generalize from the data, reflecting a well-balanced trade-off between accuracy and speed.

On the other hand, the 20% training configuration, although achieving the lowest accuracy of 85.94%, completed the highest number of iterations (960), which suggests that the model underwent more frequent weight updates due to the smaller batch sizes. With fewer examples per batch, the model was updated more frequently, allowing for finer adjustments to the weights, albeit with less data to generalize from. This configuration highlights how the frequency of weight updates can impact the learning process, though it ultimately may not be sufficient for achieving high accuracy without more data.

These results indicate that optimal performance is not solely dependent on the volume of training data, but on the intricate balance between the amount of data, the training dynamics, and the model's capacity to learn and generalize. Simply increasing the training dataset size does not guarantee better outcomes; rather, it is essential to consider how data size interacts with training time, batch size, and the model's ability to capture and generalize features from the data. Thus, the findings suggest that a nuanced approach to training, considering not just the quantity of data but also the configuration of training parameters, can lead to more efficient and effective model performance. **Table III** summarizes the performance of each training configuration. Note that the test set size percentages (30%, 25%, 20%) correspond to training set sizes of 70%, 75%, and 80% of the total data, respectively

TABLE III. TRAINING CONFIGURATION PERFORMANCE SUMMARY

Test Set Size (Training Set Size)	Final Validation Accuracy	Training Time	Total Iterations	Epochs Completed
30% (70%)	90.38%	16 min 34 sec	900	20/20
25% (75%)	90.69%	24 min 54 sec	840	20/20
20% (80%)	85.94%	15 min 50 sec	960	20/20

C. Feature Extraction Analysis

The handcrafted feature extraction process was designed to capture key characteristics of the data, yielding 15 comprehensive features categorized into three primary groups. These features are crucial for gaining insights into the statistical, geometric, and textural properties of the analyzed data. The first category, First-order Statistical Features, consists of five features: Mean, Median, Standard Deviation, Skewness, and Kurtosis. These features provide an overall understanding of the data’s distribution, central tendency, variability, and shape. The Mean and Median give insights into centrality, while Standard Deviation reflects data spread. Skewness and Kurtosis assess the symmetry and the extremities of the distribution, respectively [48].

The second category, Shape-based Features, includes four features that describe the geometric properties of the shape under analysis. These include Area, Perimeter, Eccentricity, and Solidity. Area and Perimeter give quantitative measurements of size and boundary length, respectively. Eccentricity measures how elongated the shape is, indicating its deviation from circularity, while Solidity quantifies the compactness of the shape by comparing the area to the convex hull area. These shape-based features are particularly useful for recognizing and differentiating different forms and structures.

Finally, the Texture Features group, derived from the Gray Level Co-occurrence Matrix (GLCM), consists of six features: Contrast, Correlation, Energy, Homogeneity, Coefficient of Variation, and Histogram Width. These features capture the surface characteristics of the shape, with Contrast reflecting intensity differences between neighboring pixels, Correlation assessing linear dependencies, and Energy providing information about texture uniformity. Homogeneity indicates how smooth or regular the texture is, while Coefficient of Variation offers a normalized measure of variation, and Histogram Width quantifies the range of intensity values. Together, these features provide a rich set of descriptors, enabling detailed texture analysis.

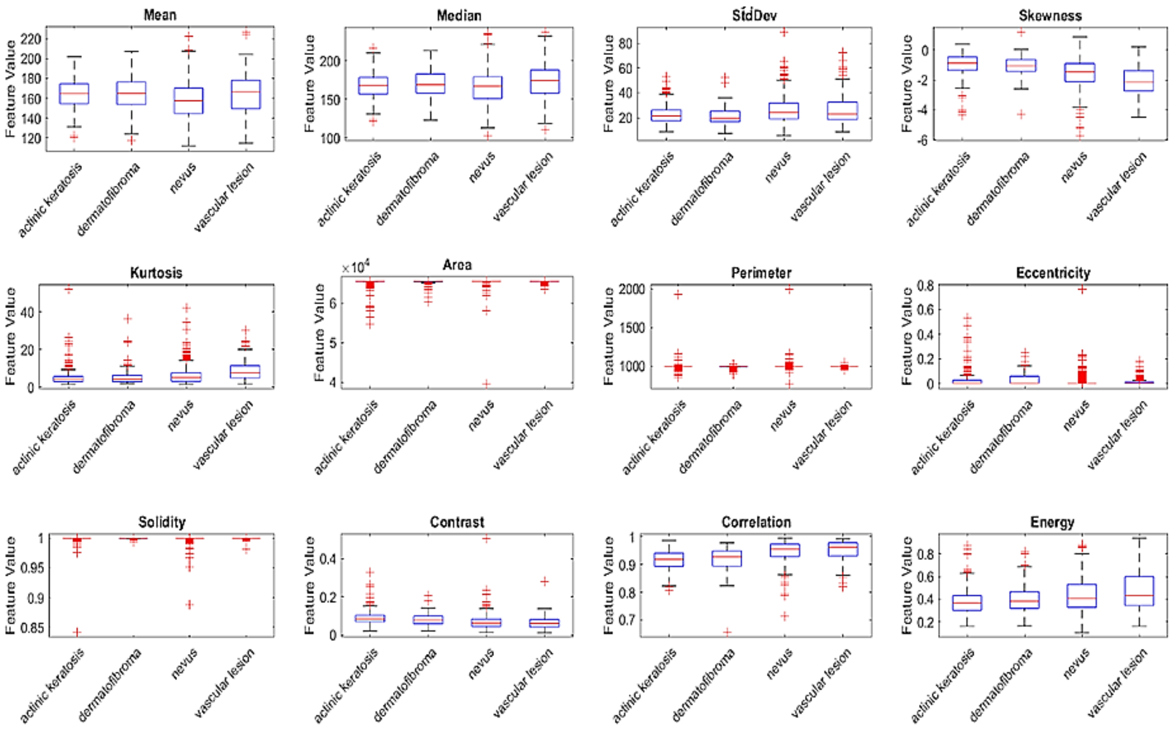

Fig. 6. Feature distribution analysis across four skin cancer types: (a) first-order statistical features (mean, median, standard deviation, skewness, kurtosis), (b) shape-based features (area, perimeter, eccentricity, solidity), and (c) texture features (contrast, correlation, energy, homogeneity, coefficient of variation, histogram width).

Figure 6 presents the feature distribution analysis, which reveals distinct morphological characteristics across the four skin cancer types.. The box plots demonstrate varying degrees of feature discrimination, with several features showing clear separability between classes. Notably, the skewness and kurtosis features exhibit pronounced differences, particularly for vascular lesions which show consistently different distributions. The presence of outliers (red crosses) indicates natural variability within each class, which is expected in medical imaging data. The area and perimeter features show relatively similar distributions across classes, suggesting shape-based features alone may not provide sufficient discrimination. The texture features (contrast, correlation, energy) demonstrate more promising discriminative potential, with vascular lesions and nevus showing distinct patterns [49].

TABLE IV. FEATURE STATISTICS SUMMARY

Feature	Actinic Keratosis	Dermatofibroma	Nevus	Vascular Lesion
Mean Intensity	164.70 ± 15.10	165.52 ± 16.37	158.24 ± 19.50	165.35 ± 19.21
Median	168.44 ± 16.81	169.89 ± 17.26	166.24 ± 21.19	173.30 ± 20.43
Standard Deviation	22.25 ± 7.51	21.17 ± 7.22	26.49 ± 10.83	26.79 ± 12.14
Skewness	-0.96 ± 0.67	-1.08 ± 0.69	-1.59 ± 0.96	-2.11 ± 0.95
Kurtosis	4.88 ± 4.00	5.35 ± 4.53	6.29 ± 4.91	8.95 ± 5.66
Area	65304.34 ± 1108.33	65251.25 ± 753.35	65353.79 ± 1454.92	65459.23 ± 244.81
Perimeter	1000.30 ± 55.19	990.79 ± 18.89	1000.97 ± 56.05	996.81 ± 9.97
Eccentricity	0.026 ± 0.064	0.032 ± 0.053	0.015 ± 0.054	0.018 ± 0.039
Solidity	0.999 ± 0.009	1.000 ± 0.001	0.999 ± 0.007	1.000 ± 0.002
Contrast	0.088 ± 0.034	0.080 ± 0.031	0.067 ± 0.039	0.062 ± 0.033

Correlation	0.916 ± 0.036	0.919 ± 0.044	0.947 ± 0.037	0.949 ± 0.037
Energy	0.379 ± 0.113	0.403 ± 0.128	0.437 ± 0.158	0.473 ± 0.180
Homogeneity	0.957 ± 0.014	0.962 ± 0.014	0.968 ± 0.016	0.970 ± 0.015
Coefficient of Variation	0.136 ± 0.046	0.130 ± 0.049	0.171 ± 0.075	0.165 ± 0.081
Histogram Width	162.84 ± 35.25	150.71 ± 42.85	147.81 ± 31.42	156.03 ± 38.08

The analysis of **Table IV** reveals distinctive feature statistics that play a crucial role in discriminating between four skin cancer types, highlighting several significant patterns. Firstly, the progression of skewness shows a clear trend from actinic keratosis (-0.96) to vascular lesion (-2.11), indicating increasingly left-skewed intensity distributions. This suggests that vascular lesions exhibit more concentrated high-intensity regions, a characteristic that differentiates them from the other lesion types. Secondly, the variation in kurtosis further distinguishes vascular lesions, which exhibit the highest kurtosis value of 8.95 ± 5.66. This indicates that the intensity distribution of vascular lesions is more peaked with heavy tails, reflecting their distinctive vascular patterns.

In terms of texture, the correlation feature displays notable separation between actinic keratosis/dermatofibroma (0.916-0.919) and nevus/vascular lesion (0.947-0.949), suggesting differences in the smoothness of the textures across these lesions. The energy distribution also shows a progressive increase, from actinic keratosis (0.379) to vascular lesion (0.473), implying varying degrees of textural uniformity. Vascular lesions, with the highest energy value, demonstrate the most uniform texture patterns, further contributing to their classification. Finally, the contrast patterns show that actinic keratosis has the highest contrast value (0.088), which may explain its consistently high classification performance across all configurations, as it presents more distinct boundaries compared to other lesion types. These statistical features provide valuable insights into the characteristics of each skin lesion type and explain the variations in classification performance across different types. By understanding these patterns, we can better interpret the significance of each feature in the classification task [50].

D. Classification Performance Analysis

The confusion matrices reveal critical insights into model performance and potential overfitting. **Figure 7** presents the confusion matrices for the 30% training configuration. The training set confusion matrix (left) shows perfect classification with 100% accuracy across all classes, indicating that the model has completely learned the training data patterns. However, the test set performance (right) shows more realistic results with 90.38% overall accuracy. The test confusion matrix reveals class-specific performance variations: actinic keratosis shows excellent performance with minimal misclassifications, while dermatofibroma exhibits more confusion with other classes. The diagonal dominance in the test matrix indicates good overall classification capability, but the off-diagonal elements reveal specific class confusions that require attention for clinical deployment

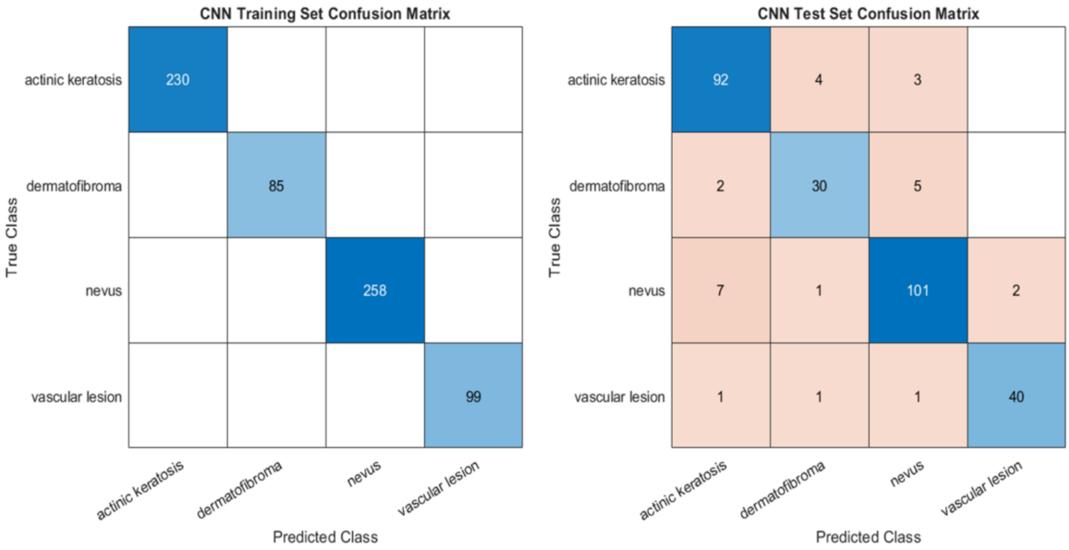

Fig. 7. Confusion matrices for CNN model trained on 30% of data: (a) training set, (b) test set.

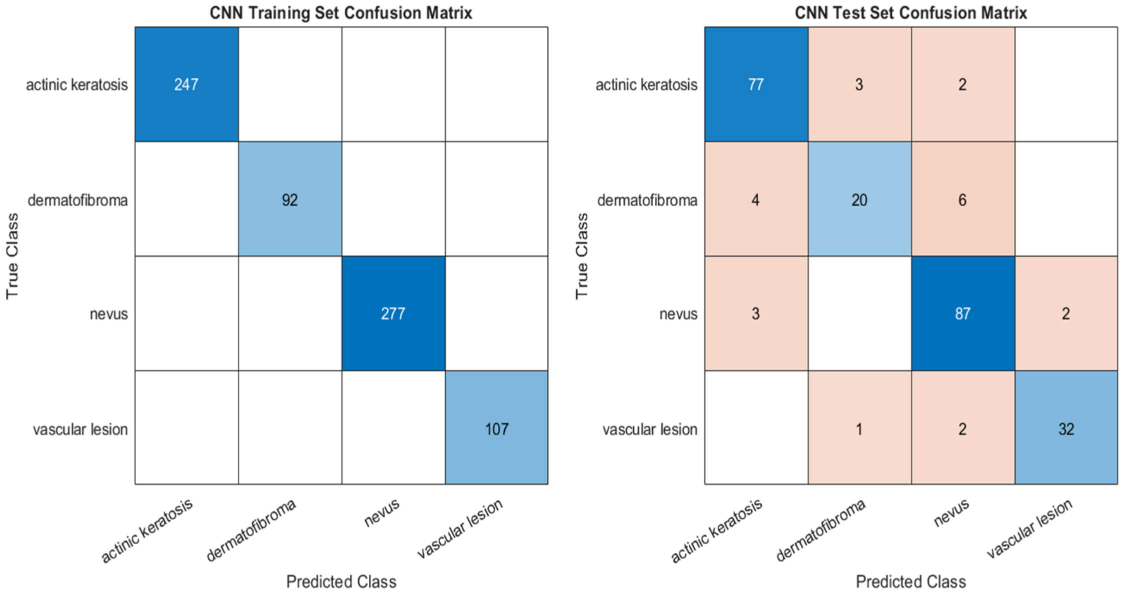

Fig. 8. Confusion matrices for CNN model trained on 25% of data: (a) training set, (b) test set.

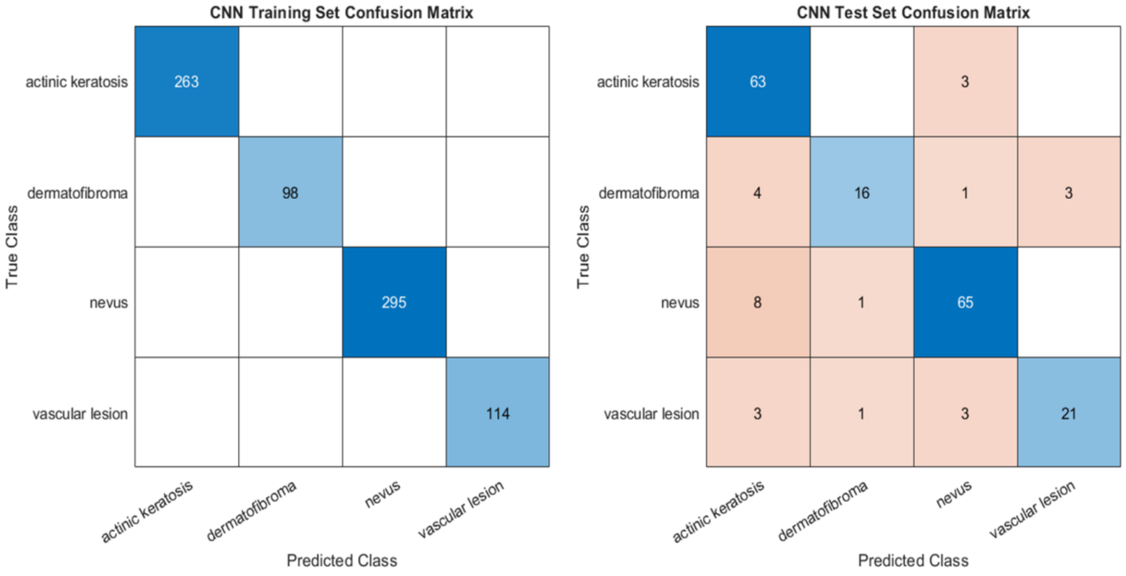

Fig. 9. Confusion matrices for CNN model trained on 20% of data: (a) training set, (b) test set.

As shown in **Figure 8**, the 25% training configuration demonstrates the most balanced performance between training and test sets. The training confusion matrix shows excellent but not perfect performance, suggesting reduced overfitting compared to the 30% configuration. The test set confusion matrix achieves the highest validation accuracy (90.69%) with improved class balance. Notably, the dermatofibroma classification shows improvement compared to the 30% configuration, indicating that the reduced training data prevents overfitting to class-specific noise. The overall diagonal dominance is stronger, with fewer misclassifications across all classes, validating the optimal training data size hypothesis.

As shown in **Figure 9**, the 20% configuration shows the expected performance degradation due to limited training data, achieving 85.94% test accuracy. The training confusion matrix indicates that the model struggles to achieve perfect training performance, suggesting insufficient data for complete pattern learning. The test confusion matrix shows increased misclassifications, particularly for minority classes like dermatofibroma. However, the performance degradation is relatively modest (4.75% decrease from optimal), indicating that the ResNet-50 architecture maintains reasonable performance even with limited training data. This finding is crucial for clinical settings where large training datasets may not be available.

TABLE V. Performance Metrics - Scenario 1 : 30% Test Data (70% Training)

Method	Accuracy (%)	Precision	Recall	Sensitivity	Specificity	F1-Score	AUC
CNN (ResNet-50)	90.69	0.901	0.895	0.895	0.966	0.898	0.979
SVM (RBF Kernel)	60.07	0.576	0.448	0.448	0.844	0.434	0.801

Table V presents the performance metrics for Scenario 1. The first scenario demonstrates the clear superiority of the CNN approach over SVM across all evaluation metrics. The CNN achieves 90.69% accuracy compared to SVM's 60.07%, representing a 30.62 percentage point improvement. The precision difference (0.901 vs 0.576) indicates that the CNN produces far fewer false positives, crucial for clinical applications where false alarms can lead to unnecessary procedures. The recall disparity (0.895 vs 0.448) shows that the CNN identifies significantly more actual positive cases, critical for cancer detection where missing cases has severe consequences. The AUC values (0.979 vs 0.801) demonstrate the CNN's superior ability to discriminate between classes across all decision thresholds.

TABLE VI. Performance Metrics - Scenario 2: 25% Test Data (75% Training)

Method	Accuracy (%)	Precision	Recall	Sensitivity	Specificity	F1-Score	AUC
CNN (ResNet-50)	90.38	0.897	0.866	0.866	0.965	0.879	0.969
SVM (RBF Kernel)	62.92	0.581	0.491	0.491	0.858	0.494	0.805

Table VI presents the performance metrics for Scenario 2. The second scenario shows slight variations in performance metrics while maintaining the CNN's substantial advantage over SVM. The CNN achieves 90.38% accuracy with well-balanced precision (0.897) and recall (0.866), indicating robust performance across different test data compositions. The high specificity (0.965) demonstrates excellent ability to correctly identify negative cases, minimizing false positive rates. The F1-score of 0.879 represents a balanced harmonic mean between precision and recall, crucial for imbalanced datasets. The SVM performance shows marginal improvement in this scenario (62.92% vs 60.07%), but remains substantially below clinically acceptable level.

TABLE VII. Performance Metrics - Scenario 3: 20% Test Data (80% Training)

Method	Accuracy (%)	Precision	Recall	Sensitivity	Specificity	F1-Score	AUC
CNN (ResNet-50)	85.94	0.869	0.812	0.812	0.948	0.834	0.970
SVM (RBF Kernel)	57.81	0.518	0.439	0.439	0.839	0.431	0.772

Table VII presents the performance metrics for Scenario 3. The third scenario demonstrates the impact of limited training data on CNN performance while maintaining superiority over SVM. The CNN accuracy drops to 85.94%, representing a 4.75 percentage point decrease from optimal performance, indicating graceful degradation with reduced training data. The precision (0.869) and recall (0.812) remain at clinically acceptable levels, suggesting that the model maintains diagnostic utility even with minimal training examples. The AUC remains high (0.970), indicating preserved discriminative capability across decision thresholds. The SVM shows the poorest performance in this scenario, emphasizing the importance of sufficient training data for traditional machine learning approaches.

TABLE VIII. CNN Class-wise Performance Summary Across All Training Configurations

Training Set	Class	Training Accuracy	Test Recall	Test Precision*	Test Samples
30%	Actinic Keratosis	100%	95.5%	0.902	66
30%	Dermatofibroma	100%	66.7%	0.833	24
30%	Nevus	100%	87.8%	0.918	74
30%	Vascular Lesion	100%	75.0%	0.952	28
25%	Actinic Keratosis	100%	93.9%	0.753	82
25%	Dermatofibroma	100%	66.7%	0.538	30

Training Set	Class	Training Accuracy	Test Recall	Test Precision*	Test Samples
25%	Nevus	100%	94.6%	0.578	92
25%	Vascular Lesion	100%	91.4%	0.455	35
20%	Actinic Keratosis	100%	92.9%	0.808	99
20%	Dermatofibroma	100%	81.1%	0.889	37
20%	Nevus	100%	91.0%	0.903	111
20%	Vascular Lesion	100%	93.0%	0.875	43

Table VIII presents the class-wise performance summary. The analysis reveals distinct patterns across the different training configurations, which are closely linked to the dataset characteristics. For Actinic Keratosis, the recall remains consistently high (92.9%-95.5%) across all configurations, indicating strong detection capability. However, the precision shows more variation (0.753-0.902), which suggests the presence of some false positive variability. Dermatofibroma, on the other hand, demonstrates a more challenging classification profile, with consistently lower recall values (66.7%-81.1%) across configurations. This pattern is likely a result of its smaller dataset size (122 images), underlining the impact of class imbalance on performance. Nevus shows stable performance with high recall (87.8%-94.6%) but experiences variable precision (0.578-0.918). The relatively larger dataset size (369 images) provides a sufficient number of training examples, ensuring consistent recall, though precision fluctuates depending on the training configuration. Lastly, Vascular Lesions exhibit a dramatic improvement in performance as the training set size increases, with recall progressing from 75.0% to 93.0%. This suggests that vascular lesions benefit significantly from the inclusion of more training examples. Notably, the consistent 100% training accuracy across all classes and configurations indicates a potential for overfitting, emphasizing the need to rely on test set performance metrics for a more realistic and generalized assessment of the model’s effectiveness.

TABLE IX. CNN CLASS-WISE DETAILED PERFORMANCE (30% CONFIGURATION)

Class	Precision	Recall	Sensitivity	Specificity	F1-Score	AUC
Actinic Keratosis	0.902	0.929	0.929	0.948	0.915	0.972
Dermatofibroma	0.833	0.811	0.811	0.976	0.822	0.969
Nevus	0.918	0.910	0.910	0.950	0.914	0.975
Vascular Lesion	0.952	0.930	0.930	0.992	0.941	0.999

Table IX presents the detailed performance metrics for the 30% configuration, highlighting the optimal class-wise performance characteristics.. Vascular Lesions achieve the best overall performance, with an F1-score of 0.941 and a perfect AUC of 0.999, demonstrating near-perfect discrimination capabilities. The high precision (0.952) and specificity (0.992) indicate an excellent ability to avoid false positives, further confirming the robustness of the model for this class. Actinic Keratosis and Nevus show comparable performance, with F1-scores of 0.915 and 0.914, respectively. Both classes demonstrate a balance between precision and recall, indicating a robust classification capability that is well-suited for clinical applications. Dermatofibroma, despite having the lowest F1-score (0.822), exhibits excellent specificity (0.976) and precision (0.833), which suggests a low false positive rate. However, the lower recall (0.811) indicates that some cases are missed, which can be directly attributed to the limited number of training examples. Overall, all classes achieve AUC values above 0.969, reflecting excellent discriminative capability across all decision thresholds. These results validate the effectiveness of the ResNet-50 architecture in skin cancer classification.

TABLE X. SVM CLASS-WISE PERFORMANCE (25% CONFIGURATION)

Class	Precision	Recall	Sensitivity	Specificity	F1-Score	AUC
Actinic Keratosis	0.753	0.744	0.744	0.873	0.748	0.893
Dermatofibroma	0.538	0.233	0.233	0.971	0.326	0.753
Nevus	0.578	0.848	0.848	0.615	0.687	0.825
Vascular Lesion	0.455	0.139	0.139	0.971	0.213	0.751

Table X presents the SVM class-wise performance under the 25% configuration. The analysis reveals several limitations of traditional machine learning approaches in handling complex medical image classification tasks. For Actinic Keratosis, the SVM achieves the best performance with an F1-score of 0.748, though this is still substantially lower than the performance of CNN (0.915). The balanced precision and recall suggest that handcrafted features can provide some level of discriminative capability for this class, but they fall short in comparison to deep learning methods. Nevus demonstrates high recall (0.848) but poor precision (0.578), indicating that the SVM tends to over-classify images as nevus. This behavior reflects the majority class bias commonly seen in imbalanced datasets, where the model over-represents the more frequent class. For Dermatofibroma and Vascular Lesion, the performance is extremely poor, with F1-scores of 0.326 and 0.213, respectively. The very low recall values (0.233 and 0.139) further highlight that the SVM fails to identify most positive cases for these classes, struggling particularly with the underrepresented examples. The particularly poor performance on dermatofibroma (F1: 0.326) and vascular lesion (F1: 0.213) can be attributed to several compounding factors. First, both classes are severely underrepresented in the dataset (122 and 142 images respectively), meaning that the handcrafted feature distributions for these classes overlap considerably with those of the majority classes. Second, the 15 extracted features — while comprehensive — are derived from global image statistics (e.g., mean intensity, GLCM contrast) that fail to capture the subtle local morphological patterns that distinguish dermatofibroma and vascular lesions from nevus and actinic keratosis. Dermatofibroma, in particular, is known dermoscopically for its variable and non-specific appearance [4], making it difficult for any fixed feature set to reliably characterize it. Vascular lesions, while visually distinctive to trained clinicians, exhibit intensity and texture distributions that may overlap with other lesion types when reduced to scalar statistics. In contrast, the CNN implicitly learns spatially localized, hierarchical features through its convolutional layers, enabling it to detect subtle distinguishing patterns that handcrafted features cannot capture. This fundamental architectural difference explains the large performance gap between CNN and SVM specifically for these two classes.

E. Model Interpretability Through Grad-CAM Analysis

The Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations (**Figure 10**) provide valuable insights into the CNN model's decision-making process, highlighting the regions of dermatoscopic images that the model focuses on when making classification decisions. Red regions, indicating the highest activation, show that the model concentrates its attention on the lesion areas for actinic keratosis and vascular lesions, which are correctly predicted with high confidence (scores ranging from 0.92 to 1.00). However, dermatofibroma exhibits scattered activation patterns, reflecting the challenges the model faces in accurately classifying this class. For actinic keratosis, strong central activation is observed (score: 1.00), signaling that the model reliably focuses on the most relevant features. In contrast, nevus misclassification (predicted as vascular lesion, score: 0.62) is associated with peripheral activation, indicating overlap of features between the two classes. Clear activation patterns are evident for both actinic keratosis (score: 1.00) and vascular lesion (score: 0.98), demonstrating robust feature localization. On the other hand, dermatofibroma misclassification (predicted as nevus, score: 0.69) shows diffuse activation, suggesting that the model struggles to localize distinct features for this class.

From a clinical perspective, Grad-CAM visualizations offer a practical pathway for integrating the model's decisions into real-world diagnostic workflows. In a clinical setting, a dermatologist could use the Grad-CAM heatmap as a secondary confirmation layer: after the model produces a classification, the physician can examine whether the highlighted activation region corresponds to the visually suspicious area of the lesion. If the model's focus aligns with the clinically relevant region, it increases confidence in the prediction. Conversely, if the activation is scattered or concentrated outside the lesion boundary — as observed in several dermatofibroma cases — the clinician is alerted to treat the model's output with caution and apply independent judgment. This human-in-the-loop approach preserves clinical authority while leveraging the model's pattern recognition capabilities. Future deployment of this system in web or mobile applications could render Grad-CAM overlays alongside the classification result, enabling transparent and accountable AI-assisted diagnosis even in low-resource healthcare environments.

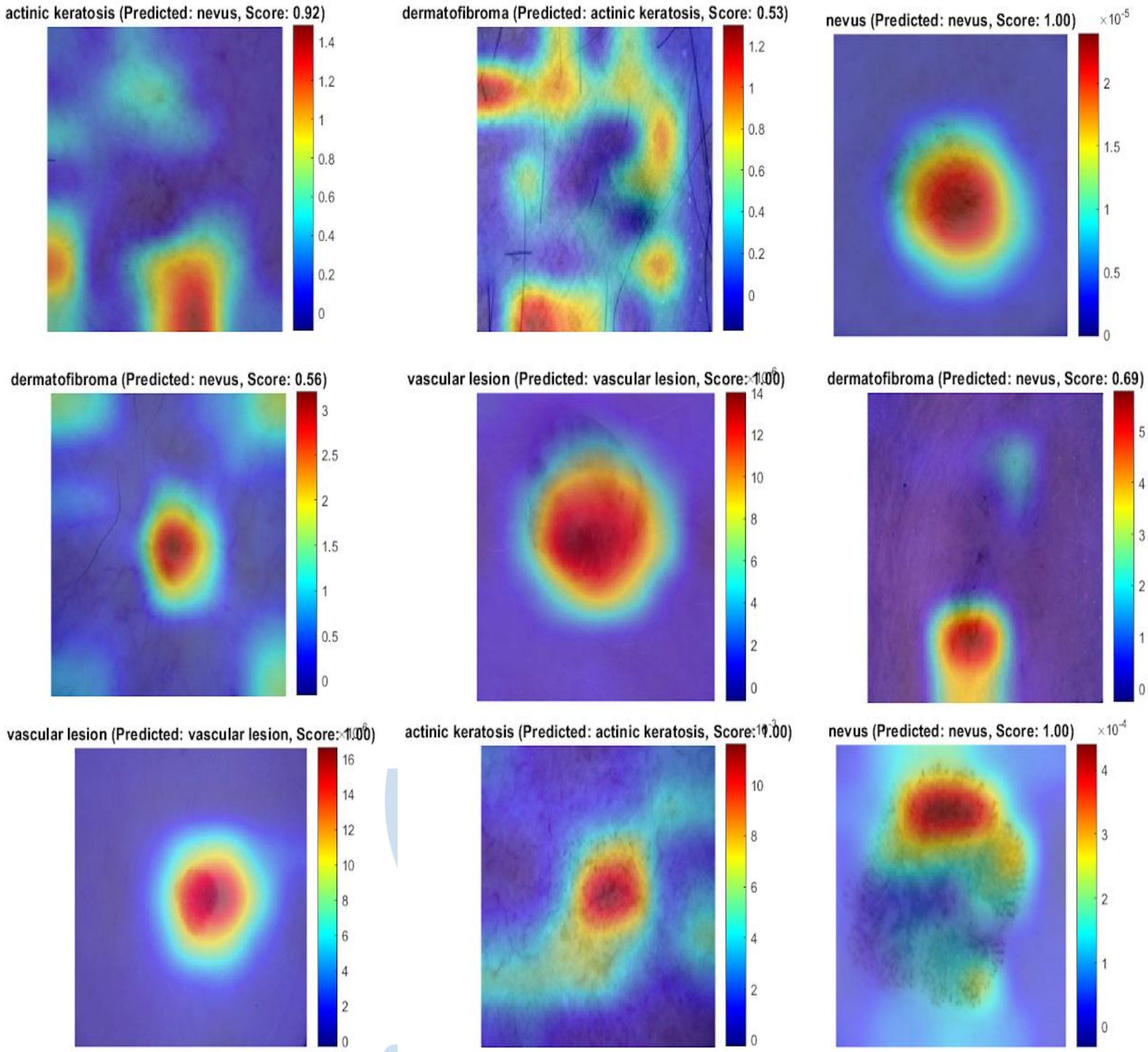

Fig. 10. Grad-CAM visualization for different skin cancer types

The key finding of this study (defined here as the primary result that departs from conventional expectations and carries the greatest practical implication) is the unexpected result that the 25% training set configuration achieved the highest validation accuracy of 90.69%, surpassing both the 30% (90.38%) and 20% (85.94%) configurations. This counterintuitive outcome holds important implications for skin cancer classification using CNN architectures. It suggests that there is an optimal balance between generalization and overfitting. The 25% configuration provided enough data for effective learning without introducing excessive noise or overfitting, which may have occurred with the larger 30% dataset. Despite having more training data, the 30% configuration showed slight performance degradation, indicating that additional data may have led to overfitting in some cases. This finding emphasizes the need for careful consideration of training data size, where too much data may not necessarily lead to better performance if it leads to overfitting.

Training efficiency was also a key aspect of the analysis, revealing notable patterns. The 30% configuration provided the highest accuracy/time ratio, making it the most efficient in terms of computational resources, while the 25% configuration, although achieving the highest absolute accuracy, required longer training times. The 20% configuration demonstrated good efficiency but at the cost of compromised accuracy, reinforcing the importance of finding a balance between accuracy and computational efficiency, especially in resource-constrained environments where time and hardware resources are limited.

Class-specific performance showed consistent patterns across all training configurations. Actinic Keratosis demonstrated consistently high recall rates (>92%) across all configurations, reflecting its reliable detection capabilities. Nevus showed robust performance with recall rates ranging from 87.8% to 94.6%, indicating stable classification. Vascular Lesion saw dramatic improvement in recall, progressing from 75% to 93% as the training set size increased, suggesting that larger datasets significantly improved its classification accuracy. In contrast, Dermatofibroma consistently exhibited the lowest performance (66.7% - 81.1%) across all configurations, presenting the primary challenge in classification. This consistent underperformance is likely due to dataset imbalance, as dermatofibroma has the smallest number of images (122 compared to 369 for nevus). This issue suggests that targeted data augmentation or specialized preprocessing techniques are necessary to address the imbalance and improve classification accuracy [51].

Feature analysis provided additional insights into the discriminative characteristics of different skin cancer types [52]. Statistical features, such as skewness and kurtosis, revealed clear distinctions between skin cancer types [53]. Vascular Lesions exhibited the highest kurtosis, indicating distinct intensity distribution patterns, while Actinic Keratosis showed the highest contrast, which may explain its consistently strong classification performance [54]. Additionally, texture-based features such as correlation coefficients and contrast patterns were instrumental for classification [55]. Vascular Lesions and Nevus exhibited the highest correlation, suggesting smoother texture patterns, further aiding in their differentiation.

Clinically, the achieved performance metrics offer substantial potential for diagnostic support, with accuracy levels exceeding 90% in the optimal configurations. This makes the model suitable for use in clinical settings, especially as a screening tool in resource-limited environments. The finding that 25% of the available training data yields optimal results has significant implications for clinical deployment, reducing annotation costs, accelerating deployment, and improving accessibility for institutions with limited resources. Furthermore, this multi-class classification approach provides a more comprehensive diagnostic tool compared to many studies focused solely on binary melanoma detection, aligning more closely with clinical practice where multiple lesion types must be diagnosed.

Finally, the technical contributions of this study are noteworthy. The use of transfer learning with ResNet-50 for skin cancer classification proved highly effective. The modification of the final layers of the pretrained model to accommodate four classes while maintaining the feature extraction capabilities of ResNet-50 allowed for effective learning despite the smaller dataset. The consistently high training accuracy across all configurations further validated the effectiveness of knowledge transfer in deep learning. Moreover, the feature engineering approach, incorporating 15 handcrafted features, combined first-order statistics, shape descriptors, and texture analysis to provide comprehensive lesion characterization. This enhanced the model's ability to discriminate between different skin cancer types, offering valuable insights into the classifier's decision-making process and making it a useful tool for clinical applications.

IV. CONCLUSION

This study highlights the potential of automated skin cancer classification using CNN architectures, achieving clinically relevant performance with optimal resource utilization. The central finding of this study (that using only 25% of the data for testing (75% for training) yields the highest validation accuracy of 90.69%) challenges traditional assumptions and offers a practical breakthrough for resource-constrained environments like rural or low-resource healthcare settings. By using a smaller dataset, this method reduces training time and computational costs while maintaining high diagnostic accuracy. The multi-class classification system effectively distinguishes between four skin cancer types, demonstrating the robustness of CNNs in complex medical problems. The clinical applicability of this approach is a key strength, with the model's performance making it suitable for real-world diagnostic support, especially in underserved areas where specialized expertise is scarce. This research enables wider access to advanced skin cancer detection, potentially saving lives in remote regions. From a clinical impact perspective, the study emphasizes cost-effective solutions that can be rapidly deployed. The 25% training configuration is recommended for skin cancer classification systems due to its proven ability to deliver high accuracy with minimal resources. Future improvements, such as targeted dermatofibroma detection through advanced data augmentation, will further enhance the system's diagnostic capabilities, making it adaptable to various skin conditions. Additionally, the Grad-CAM analysis demonstrated that the model's decision-making is grounded in clinically relevant regions of the dermoscopic image. For actinic keratosis and vascular lesions, the model consistently activated over the lesion core, reflecting high-confidence and well-localized predictions. In contrast, the diffuse and scattered activation patterns observed for dermatofibroma indicate that the model has not yet learned sufficiently discriminative features for this class, likely due to the limited number of training samples. These interpretability insights are valuable not only for validating model behavior but also for guiding future data collection efforts — specifically targeting dermatofibroma cases to improve activation focus. Incorporating Grad-CAM into the final diagnostic system would enable clinicians to visually verify the model's reasoning before accepting its output, thereby enhancing trust and safety in clinical deployment. Ultimately, this research supports the adoption of AI-powered diagnostic tools, balancing precision, efficiency, and ease of implementation, with the potential to revolutionize dermatology.

ACKNOWLEDGMENT

The authors would like to acknowledge the Department of Medical Technology, Institut Teknologi Sepuluh Nopember, for the facilities and support in this research.

REFERENCES

[1] A. H. Roky, M. M. Islam, A. M. F. Ahasan, M. S. Mostaq, M. Z. Mahmud, M. N. Amin, and M. A. Mahmud, "Overview of skin cancer types and prevalence rates across continents," *Cancer Pathogenesis and Therapy*, vol. 3, no. 2, Aug. 2024, doi: https://doi.org/10.1016/j.cpt.2024.08.002.

[2] N. F. Kormos, I. D. Paval, C. M. Petrenciu, A. S. Vizitiu, and A. L. Baican, "Diagnostic Challenges and Influencing Factors in Non-Melanoma Skin Cancers: A Retrospective Analysis of Basal Cell Carcinoma and Squamous Cell Carcinoma Cases," *Cancer Reports*, vol. 8, no. 9, Sep. 2025, doi: https://doi.org/10.1002/cnr2.70332.

[3] J. R. Thamm, S. Schuh, and J. Welzel, "Epidemiology and Risk Factors of Actinic Keratosis. What Is New for the Management for Sun-Damaged Skin," *Dermatology Practical & Conceptual*, vol. 14, no. S1, p. e2024146S, Aug. 2024, doi: https://doi.org/10.5826/dpc.1403s1a146s.

[4] L. Wan, A. Park, L. Almatroud, and A. Khachemoune, "Dermatofibroma: Reappraisal and Updated Review," *Clinical Cosmetic and Investigational Dermatology*, vol. Volume 18, pp. 1873–1887, Aug. 2025, doi: https://doi.org/10.2147/ccid.s526191.

[5] P. S.-Sienkiewicz, D. J.-Lewandowska, J. Calik, G. T.-Kołodziejczyk, and P. Mańkowski, "Nevi and Melanoma in Children: What to Do in Daily Medical Practice: Encyclopedia for Pediatricians and Family Doctors," *Diagnostics*, vol. 14, no. 18, pp. 2004–2004, Sep. 2024, doi: https://doi.org/10.3390/diagnostics14182004.

[6] G. S. Colafati, E. Piccirilli, A. Marrazzo, A. Carboni, A. Diociaiuti, M. E. Hachem, F. Esposito, M. Zama, M. Rollo, C. Gandolfo, and P. Tomà, "Vascular lesions of the pediatric orbit: A radiological walkthrough," *Frontiers in Pediatrics*, vol. 10, Nov. 2022, doi: https://doi.org/10.3389/fped.2022.734286.

[7] F. Cassalia, A. Danese, E. Cocchi, S. Vaienti, A. Bolzon, L. Franceschin, R. Mazzetto, F. Caroppo, D. Melandri, and A. B. Fortina, "Congenital Dermatofibrosarcoma Protuberans—An Update on the Ongoing Diagnostic Challenges," *Cancers*, vol. 17, no. 1, p. 158, Jan. 2025, doi: https://doi.org/10.3390/cancers17010158.

[8] K. Nawaz, A. Zanib, I. Shabir, J. Li, Y. Wang, T. Mahmood, and A. Rehman, "Skin cancer detection using dermoscopic images with convolutional neural network," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, doi: https://doi.org/10.1038/s41598-025-91446-6.

[9] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, no. 1273253, Nov. 2023, doi: https://doi.org/10.3389/fpubh.2023.1273253.

[10] H. K. Jeong, C. Park, S. W. Jiang, M. Nicholas, S. Chen, R. Henao, and M. Kheterpal, "Image Quality Assessment Using Convolutional Neural Network in Clinical Skin Images," *JID Innovations*, vol. 4, no. 4, pp. 100285–100285, Apr. 2024, doi: https://doi.org/10.1016/j.xjidi.2024.100285.

[11] A. R. Gunukula, H. D. Gupta, and V. S. Sheng, "Detecting AI-Generated Images Using a Hybrid ResNet-SE Attention Model," *Applied Sciences*, vol. 15, no. 13, pp. 7421–7421, Jul. 2025, doi: https://doi.org/10.3390/app15137421.

[12] J. Alcaraz, M. Labbé, and M. Landete, "Support Vector Machine with feature selection: A multiobjective approach," *Expert Systems with Applications*, vol. 204, p. 117485, Oct. 2022, doi: https://doi.org/10.1016/j.eswa.2022.117485.

[13] B. Cassidy, C. Kendrick, A. Brodzicki, J. J.-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Medical Image Analysis*, vol. 75, p. 102305, Jan. 2022, doi: https://doi.org/10.1016/j.media.2021.102305.

[14] A. K. Nambisan, A. Maurya, N. Lama, T. Phan, G. Patel, K. Miller, B. Lama, J. Hagerty, R. Stanley, and W. V. Stoecker, "Improving Automatic Melanoma Diagnosis Using Deep Learning-Based Segmentation of Irregular Networks," *Cancers*, vol. 15, no. 4, pp. 1259–1259, Feb. 2023, doi: https://doi.org/10.3390/cancers15041259.

[15] M. Rahman, S. M. N. Nobel, P. K. Meena, V. P. Meena, J. Bahadur, and A. Appaji, "Accelerated and Precise Skin Cancer Detection through an Enhanced Machine Learning Pipeline for Improved Diagnostic Accuracy," *Results in Engineering*, pp. 104168–104168, Jan. 2025, doi: https://doi.org/10.1016/j.rineng.2025.104168.

[16] O. Akinrinade and C. Du, "Skin Cancer Detection Using Deep Machine Learning Techniques," *Intelligence-Based Medicine*, pp. 100191–100191, Dec. 2024, doi: https://doi.org/10.1016/j.ibmed.2024.100191.

[17] H. Erbay, Y. M. Abulgasim, D. E. Özer, and F. Ertürk, "Enhancing multi-class skin lesion diagnosis through ensemble learning of CNN and transformer architectures," *Engineering Science and Technology, an International Journal*, vol. 70, p. 102145, Jul. 2025, doi: https://doi.org/10.1016/j.jestch.2025.102145.

[18] S. S. Zareen, G. Sun, M. Kundi, S. F. Qadri, and S. Qadri, "Enhancing Skin Cancer Diagnosis with Deep Learning: A Hybrid CNN-RNN Approach," *Computers, materials & continua/Computers, materials & continua (Print)*, vol. 79, no. 1, pp. 1497–1519, Jan. 2024, doi: https://doi.org/10.32604/cmc.2024.047418.

[19] S. R. Nalamwar and S. U. Belgamwar, "Multidimensional CapsNets attention-gated approach for skin cancer detection and classification," *Computers & Electrical Engineering*, vol. 127, pp. 110573–110573, Aug. 2025, doi: https://doi.org/10.1016/j.compeleceng.2025.110573.

[20] M. Manickavasagam, V. V. Rani, U. K. Giri, and B. Maram, "Skin Cancer Detection using Harmonic Brown Bear Optimization enabled Transfer Learning," *Computational Biology and Chemistry*, vol. 119, pp. 108551–108551, Jun. 2025, doi: https://doi.org/10.1016/j.compbiolchem.2025.108551.

[21] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov, and N. Pacal, "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," *Biomedical Signal Processing and Control*, vol. 104, p. 107627, Jan. 2025, doi: https://doi.org/10.1016/j.bspc.2025.107627.

[22] J.S. ThangaPurni and M. Braveen, "Unified ARP-ViT-CNN system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions," *Array*, pp. 100515–100515, Sep. 2025, doi: https://doi.org/10.1016/j.array.2025.100515.

[23] S. Khattar, R. Kaur, and A. Kumar, "DeepHybrid-CNN: A Hybrid Approach for Pre-processing of Skin Cancer Images," *Computerized Medical Imaging and Graphics*, pp. 102611–102611, Aug. 2025, doi: https://doi.org/10.1016/j.compmedimag.2025.102611.

[24] H. C. Reis and V. Turk, "Fusion of transformer attention and CNN features for skin cancer detection," *Applied soft computing*, pp. 112013–112013, Jul. 2024, doi: https://doi.org/10.1016/j.asoc.2024.112013.

[25] P. Nanda, D. Rout, and S. Kumari, "Multi-Class Skin Cancer Detection Using CNN-Architecture Based Deep Learning Models," *Procedia Computer Science*, vol. 260, pp. 226–235, Jan. 2025, doi: https://doi.org/10.1016/j.procs.2025.03.197.

[26] C. H.-Pérez, M. Combalia, S. Podlipnik, N. C. F. Codella, V. Rotemberg, A. C. Halpern, O. Reiter, C. Carrera, A. Barreiro, B. Helba, S. Puig, V. Vilaplana, and J. Malvehy, "BCN20000: Dermoscopic Lesions in the Wild," *Scientific Data*, vol. 11, no. 1, Jun. 2024, doi: https://doi.org/10.1038/s41597-024-03387-w.

[27] S. Phiphitphatphaisit and O. Surinta, "Deep feature extraction technique based on Conv1D and LSTM network for food image recognition", *Eng Appl Sci Res*, vol. 48, no. 5, pp. 581–592, Jul. 2021, doi: https://doi.org/10.14456/easr.2021.60.

[28] K. Dunphy, M. N. Fekri, K. Grolinger, and A. Sadhu, "Data Augmentation for Deep-Learning-Based Multiclass Structural Damage Detection Using Limited Information," *Sensors*, vol. 22, no. 16, p. 6193, Aug. 2022, doi: https://doi.org/10.3390/s22166193.

[29] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, pp. 100340–100340, Jun. 2024, doi: https://doi.org/10.1016/j.health.2024.100340.

[30] P. K. Mall, P. K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, and M. Ganzha, "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities," *Healthcare Analytics*, vol. 4, p. 100216, Dec. 2023, doi: https://doi.org/10.1016/j.health.2023.100216.

[31] V. Singh, A. Baral, R. Kumar, S. Tummala, M. Noori, S. V. Yadav, S. Kang, and W. Zhao, "A Hybrid Deep Learning Model for Enhanced Structural Damage Detection: Integrating ResNet50, GoogLeNet, and Attention Mechanisms," *Sensors*, vol. 24, no. 22, pp. 7249–7249, Nov. 2024, doi: https://doi.org/10.3390/s24227249.

[32] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information*, vol. 15, no. 4, p. 235, Apr. 2024, doi: https://doi.org/10.3390/info15040235.

[33] S. A. Salleh, N. Khalid, N. Danny, N. A. M. Zaki, M. Ustuner, Z. A. Latif, and V. Foronda, "Support Vector Machine (SVM) and Object Based Classification in Earth Linear Features Extraction: A Comparison," *Revue Internationale de Geomatique*, vol. 33, no. 0, pp. 183–199, Jun. 2024, doi: https://doi.org/10.32604/rig.2024.050723.

[34] J. S. A.-Ruiz and M. Michalak, "Classification performance assessment for imbalanced multiclass data," *Scientific Reports*, vol. 14, no. 1, May 2024, doi: https://doi.org/10.1038/s41598-024-61365-z.

[35] A. Venčkauskas, J. Toldinas, and N. Morkevičius, "Improving Multi-Class Classification for Recognition of the Prioritized Classes Using the Analytic Hierarchy Process," *Applied Sciences*, vol. 15, no. 13, pp. 7071–7071, Jun. 2025, doi: https://doi.org/10.3390/app15137071.

[36] Pierian Training, "Confusion Matrix with Scikit-Learn and Python - Pierian Training," *Pierian Training*, May 12, 2023. https://pieriantraining.com/confusion-matrix-with-scikit-learn-and-python.

[37] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sensing*, vol. 16, no. 3, pp. 533–533, Jan. 2024, doi: https://doi.org/10.3390/rs16030533.

[38] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, pp. 100994–100994, May 2024, doi: https://doi.org/10.1016/j.patter.2024.100994.

[39] H. Talebi, A. K. Bardsiri, and V. K. Bardsiri, "Developing a Hybrid Machine Learning Model for Employee Turnover Prediction: Integrating LightGBM and Genetic Algorithms," *Journal of Open Innovation Technology Market and Complexity*, pp. 100557–100557, May 2025, doi: https://doi.org/10.1016/j.joitmc.2025.100557.

[40] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, p. 5979, Apr. 2022, doi: https://doi.org/10.1038/s41598-022-09954-8.

[41] M. Cristina, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, pp. 9863–9863, Oct. 2024, doi: https://doi.org/10.3390/app14219863.

[42] S. Bhat, A. Mansoor, B. Georgescu, A. B. Panambur, F. C. Ghesu, S. Islam, K. Packhäuser, D. R.-Salas, S. Grbic, and A. Maier, "AUCReshaping: improved sensitivity at high-specificity," *Scientific Reports*, vol. 13, no. 1, p. 21097, Nov. 2023, doi: https://doi.org/10.1038/s41598-023-48482-x.

[43] A. Mjahad, A. P.-Aguado, L. L.-Serrano, and A. R.-Muñoz, "Optimizing Image Feature Extraction with Convolutional Neural Networks for Chicken Meat Detection Applications," *Applied Sciences*, vol. 15, no. 2, p. 733, Jan. 2025, doi: https://doi.org/10.3390/app15020733.

[44] L. D'hooge, M. Verkerken, T. Wauters, F. De Turck, and B. Volckaert, "Investigating Generalized Performance of Data-Constrained Supervised Machine Learning Models on Novel, Related Samples in Intrusion Detection," *Sensors*, vol. 23, no. 4, p. 1846, Feb. 2023, doi: https://doi.org/10.3390/s23041846.

[45] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: https://doi.org/10.3390/technologies13030088.

[46] N. Sebastian and B. Ankayarkanni, "Enhanced ResNet-50 with Multi-Feature Fusion for Robust Detection of Pneumonia in Chest X-Ray Images," *Diagnostics*, vol. 15, no. 16, p. 2041, Aug. 2025, doi: https://doi.org/10.3390/diagnostics15162041.

[47] K. Zhu, J. Wang, X. Hu, X. Xie, and G. Yang, "Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2308.02533.

[48] T. Guo, "Extreme Precipitation Strongly Impacts the Interaction of Skewness and Kurtosis of Annual Precipitation Distribution on the Qinghai–Tibetan Plateau," *Atmosphere*, vol. 13, no. 11, p. 1857, Nov. 2022, doi: https://doi.org/10.3390/atmos13111857.

[49] A. Ray, S. Sarkar, F. Schwenker, and R. Sarkar, "Decoding skin cancer classification: perspectives, insights, and advances through researchers' lens," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, doi: https://doi.org/10.1038/s41598-024-81961-3.

[50] L. Wen, Z. Xiao, X. Xu, and B. Liu, "Disaster Recognition and Classification Based on Improved ResNet-50 Neural Network," *Applied Sciences*, vol. 15, no. 9, p. 5143, May 2025, doi: https://doi.org/10.3390/app15095143.

[51] S. Lee and S. Lee, "Efficient Data Augmentation Methods for Crop Disease Recognition in Sustainable Environmental Systems," *Big Data and Cognitive Computing*, vol. 9, no. 1, p. 8, Jan. 2025, doi: https://doi.org/10.3390/bdcc9010008.

[52] T. Khater, S. Ansari, S. A. Mahmoud, A. J. Hussain, and H. Tawfik, "Skin Cancer Classification Using Explainable Artificial Intelligence on Pre-extracted Image Features," *Intelligent systems with applications*, vol. 20, pp. 200275–200275, Nov. 2023, doi: https://doi.org/10.1016/j.iswa.2023.200275.

[53] A. Naeem, T. Anees, M. Khalil, K. Zahra, R. A. Naqvi, and S.-W. Lee, "SNC_Net: Skin Cancer Detection by Integrating Handcrafted and Deep Learning-Based Features Using Dermoscopy Images," *Mathematics*, vol. 12, no. 7, pp. 1030–1030, Mar. 2024, doi: https://doi.org/10.3390/math12071030.

[54] G. Fredman, S. R. Wiegell, M. Haedersdal, and G. R. Untracht, "Vascular feature identification in actinic keratosis grades I-III using dynamic optical coherence tomography with automated, quantitative analysis," *Archives of Dermatological Research*, vol. 316, no. 7, Jun. 2024, doi: https://doi.org/10.1007/s00403-024-03022-z.

[55] C. Soare, E. C. Cozma, A. M. Celarel, A. M. Rosca, M. Lupu, and V. M. Voiculescu, "Digitally Enhanced Methods for the Diagnosis and Monitoring of Treatment Responses in Actinic Keratoses: A New Avenue in Personalized Skin Care," *Cancers*, vol. 16, no. 3, pp. 484–484, Jan. 2024, doi: https://doi.org/10.3390/cancers16030484