Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

An Optimized Balanced-Learning Framework for Malignant Skin Lesion Triage Using Compound-Scaled Neural Networks

Argha Orion Silitonga*, Raissa Camilla Maringka, Wilsen Grivin Mokodaser, George M W Tangka, Marchel Timothy Tombeng

Faculty of Computer Science, Informatics, Universitas Klabat, Minahasa Utara, Indonesia
Email: [1,*]argha@unklab.ac.id, [2]raissam@unklab.ac.id, [3]wilsenm@unklab.ac.id, [4]gtangka@unklab.ac.id, [4]marcheltombeng@unklab.ac.id

ARTICLE INFORMATION

ARTICLE HISTORY:
Submited : May 15, 2026
Revised : May 29, 2026
Accept : May 30, 2026
Publish : May 30, 2026

KEYWORD
Skin Cancer;
Dermoscopy;
Deep Learning;
EfficientNet-B4;
Cross-Validation

CORRESPONDENCE AUTHOR
Email: argha@unklab.ac.id

A B S T R A C T

Skin cancer represents a prevalent global health challenge, and early detection is very important to reduce mortality risk. Manual dermoscopic diagnosis risks human bias, making deep learning classification a vital research topic. While several previous studies utilizing the ISIC 2019 dataset have demonstrated high diagnostic capabilities, they primarily focus on complex multi-class classification. However, in real-world clinical workflows, the primary necessity is a swift, dependable triage system that can confidently distinguish dangerous lesions from non-threatening ones. Furthermore, many existing models require substantial computational overhead yet still suffer from imbalanced accuracy when dealing with minority malignant classes. The novelty of this study lies in addressing these gaps by developing a streamlined, clinically practical binary screening framework optimized specifically for malignant-versus-benign triage. The original multi-class labels were transformed into binary classes where malignant lesions consist of melanoma (MEL), basal cell carcinoma (BCC), and squamous cell carcinoma (SCC), while benign lesions consist of nevus (NV), benign keratosis (BKL), dermatofibroma (DF), and vascular lesions (VASC). The experiment applied transfer learning with ImageNet-pretrained weights, data augmentation, class weighting, and fourfold stratified cross-validation. Unlike prior works that rely on resource-heavy architectures, we leverage the compound-scaled EfficientNet-B4 backbone—delivering superior feature representational power with significantly fewer parameters evaluate on a large-scale cohort of 25,331 dermoscopic images. Experimental results show that the proposed model achieved an average accuracy of 89.77% and an average ROC AUC of 96.16%. The best fold obtained 91.49% accuracy with ROC AUC of 97.19%. Simultaneously, the framework maintained an average F1-score of 89.20%

1. INTRODUCTION

The worldwide burden of skin cancer is expanding at an unprecedented rate, characterized by a persistent upward trend in diagnostic figures across diverse socioeconomic regions. Factors such as increased exposure to ultraviolet (UV) radiation, aging populations, environmental changes, and greater public awareness leading to more frequent screenings have all contributed to the growing number of diagnosed cases. According to the World Health Organization (WHO), millions of new cases of skin cancer, including both melanoma and non-melanoma types, are reported annually, making it one of the most common forms of cancer globally [1]. The escalating prevalence of the disease places considerable strain on clinical infrastructure, underscoring the urgency of optimized preventive measures, robust screening protocols, and advanced therapeutic interventions. Dermatological malignancies are broadly classified as either melanoma or non-melanoma. The latter—primarily basal cell (BCC) and squamous cell carcinoma (SCC)—exhibit higher prevalence and typically follow a less indolent course, though they remain capable of extensive local tissue destruction. Conversely, melanoma represents a more virulent form of the disease. Arising from pigment-producing melanocytes, it is characterized by an aggressive tendency to metastasize to vital organs such as the lungs, liver, and brain. This metastatic potential drives the sharp decline in prognosis for advanced cases, explaining why melanoma is responsible for the preponderance of skin cancer mortality despite its lower relative incidence. [2].

Dermoscopy is a non-invasive and widely adopted imaging technique that enables dermatologists to examine subsurface skin structures that are not visible to the naked eye during routine visual inspection. By using a handheld dermatoscope or digital dermoscopic device equipped with magnification and polarized or non-polarized light, clinicians can obtain clearer and more detailed views of skin lesions. This enhanced visualization improves the assessment of suspicious lesions and has become an essential tool in modern dermatology for the early detection of skin cancer. Compared with unaided examination, dermoscopy significantly increases diagnostic accuracy and helps clinicians make more informed decisions regarding biopsy, monitoring, or treatment. Through dermoscopic images, several clinically important visual characteristics can be analyzed in detail. These include pigmentation patterns, color variations, lesion

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

symmetry, border irregularity, network structures, dots and globules, streaks, blue-white veils, and vascular patterns. Such features are closely associated with the well-known diagnostic criteria used in dermatology, including the ABCD rule (Asymmetry, Border, Color, Diameter), the 7-point checklist, and pattern analysis methods. By carefully evaluating these structures, dermatologists can better distinguish benign lesions, such as common nevi, from potentially malignant lesions like melanoma. As a result, dermoscopy has become one of the most valuable supporting tools for skin lesion assessment and triage [2].

Artificial intelligence, especially deep learning, has shown promising performance in medical image classification. Convolutional Neural Networks (CNN) automatically learn image features without manual feature extraction. CNNs have been successfully used for chest X-ray analysis, retinal disease classification, brain tumor detection, and skin lesion recognition [3]. In dermatology, CNN models can reduce diagnostic variability and improve consistency. Several previous studies have investigated skin lesion classification using deep learning. Sharma et al. demonstrated dermatologist-level performance using CNN on skin cancer images [4]. Kumar et al. evaluated multiple machine learning systems on dermoscopic datasets and reported high diagnostic capability [5]. Rahman used average ensemble learning-based model CNN architectures for melanoma classification [6]. Other studies applied ResNet, DenseNet, MobileNet, and Inception architectures with good performance [7], [8], [9], [10]. Existing literature in automated skin lesion classification faces several limitations that the proposed framework directly addresses. While Sharma et al. [4] achieved high sensitivity using a ResNet-based architecture on the ISIC Archive, their multi-class approach requires high computational overhead and complex, multi-stage preprocessing pipelines that hinder fast frontline triage. Similarly, Kumar et al. [5] implemented an ensemble of deep networks for multi-class classification, yet their model suffered from multi-class error dispersion, high parameter redundancy, and severe performance drops in underrepresented malignant categories despite achieving accuracies between 83% and 85%. To streamline clinical utility, Rahman et al. [6] pivoted to a binary classification model (Melonama vs. Benign) using an average ensemble-based CNN, yielding a strong ROC AUC exceeding 92%; however, this framework strictly limited the malignant scope to melanoma, thereby excluding other critical non-melanoma skin cancers such as Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC). In contrast, the proposed framework utilizes an EfficientNet-B4 architecture with compound scaling evaluated via a rigorous 4-fold stratified cross-validation on the ISIC 2019 dataset. This methodology yields a superior average accuracy of 89.77% and an average ROC AUC of 96.16% with balanced recall and F1-scores, effectively delivering a lightweight, highly stable pipeline that groups all major skin cancers (MEL, BCC, and SCC) into a practical and comprehensive screening model.

Although many studies have reported strong results, some limitations remain. First, several models require substantial computational resources yet still lack balanced accuracy for minority malignant classes. ISIC 2019 dataset has an imblalance dataset occurring an overfiiting problem. Second, some methods are designed for multiclass classification, while binary malignant-benign screening is more practical for real clinical triage systems. Third, there is still a need for lightweight yet powerful architectures with stable performance. EfficientNet is a modern family of CNNs introduced with a compound scaling strategy that efficiently balances depth, width, and resolution. Compared with older architectures, EfficientNet often achieves better accuracy with fewer parameters [11]. EfficientNet-B4 is one of the stronger variants, offering strong representational power while maintaining a manageable computational cost [12]. Based on those reasons, this study proposes EfficientNet-B4 for binary classification of malignant and benign skin lesions using ISIC 2019 dermoscopic images. Multi-class labels are converted into binary categories to focus on clinically relevant screening decisions. Transfer learning from ImageNet is applied to accelerate convergence and improve feature extraction. Data augmentation is used to improve generalization, while class weighting addresses dataset imbalance.

The objectives of this study are: (1) To implement EfficientNet-B4 for binary skin lesion classification; (2) To evaluate model performance using stratified cross-validation, tackling the overfitting problem on the ISIC 2019 dataset; (3) To analyze the capability of EfficientNet-B4 in distinguishing malignant and benign lesions. The main contribution of this study is the development of an effective, practical binary screening framework to distinguish malignant from benign skin lesions using dermoscopic images, overcoming the issue of overfitting caused by an imbalanced dataset. By leveraging a robust, modern convolutional neural network architecture, the proposed model demonstrates strong discriminative power, as evidenced by its high ROC AUC, indicating reliable class separation across various decision thresholds. In addition, the framework achieves competitive classification accuracy, showing that it can correctly identify a substantial proportion of lesion cases while maintaining balanced predictive performance.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study consists of several systematic stages to develop an effective skin lesion classification model, as shown in Figure 1. The first stage is dataset preparation, where the dermoscopic image dataset is collected, organized, cleaned, and labeled according to the binary classification objective of distinguishing malignant and benign lesions. The second stage is preprocessing, which involves standardization: resizing the image set to a fixed resolution and applying normalization parameters consistent with the requirements of the selected CNN, thereby ensuring data homogeneity. The third stage is data augmentation, where data diversity was expanded through the application of rotation, flipping, zooming, and brightness scaling. These techniques serve to regularize the model, preventing it from memorizing the training set and improving its ability to generalize. The fourth stage is model construction, where the EfficientNet-B4 architecture with

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

transfer learning is implemented, and additional classification layers are added for binary prediction. The fifth stage is training, during which the model learns image patterns through iterative optimization, followed by fine-tuning selected pretrained layers to improve performance. The sixth stage is validation, where unseen validation data are used to monitor training progress, detect overfitting, and guide hyperparameter adjustments. The pipeline concludes with a rigorous performance evaluation, wherein the model's efficacy is quantified through a suite of metrics, including accuracy, precision, recall, and F1-score. To validate the robustness and generalizability of the results, we further employed ROC AUC analysis, confusion matrices, and stratified cross-validation.

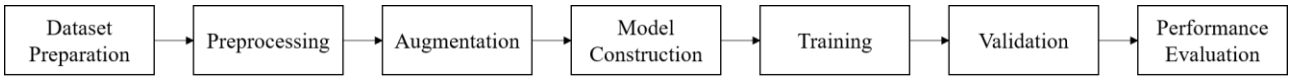


Figure 1. Research Stages

Due to its substantial scale and widespread adoption within the research community, the ISIC 2019 archive was identified as the most suitable repository for dermoscopic imagery in this study. The original dataset contains multiple diagnostic categories, which were converted into binary classes by grouping lesions into malignant and benign categories to support a practical screening objective focused on identifying suspicious cases requiring further examination [13]. Following label encoding, images were standardized to a 380 × 380 pixel resolution, aligning with the architectural requirements of EfficientNet-B4. This dimensions-matching process was conducted with care to maintain clinically significant features, including pigment distribution, border morphology, and fine structural textures. The images were then preprocessed using EfficientNet input normalization so that the pixel value distribution aligned with the conditions used during pretrained model development, thereby improving training stability and transfer learning effectiveness. The choice of EfficientNet-B4 was driven by its capacity to deliver robust classification results without requiring excessive computational overhead [14]. The model training utilized a transfer learning approach, where weights were initialized from a pretrained state to exploit established low- and mid-level feature extractors—such as edges and textures—vital for dermoscopic analysis. This was followed by a fine-tuning phase, where deeper layers were unfrozen and optimized at a reduced learning rate to capture the nuanced morphological patterns of malignant and benign lesions. To guarantee a high degree of statistical reliability, we implemented four-fold stratified cross-validation. By partitioning the dataset into four subsets while maintaining class distributions, we performed four training-validation iterations. This strategy mitigates the bias inherent in solitary train-test splits and yields a more stable performance assessment, with final results reported as the mean of accuracy, precision, recall, F1-score, and ROC AUC across all folds.

2.2 Dataset Preparation

The dataset preparation stage is one of the most important processes in this research because the quality of input data strongly affects the final classification performance. In this study, the dataset used was the International Skin Imaging Collaboration (ISIC) 2019 dataset, which is a widely recognized benchmark dataset for automated skin lesion analysis. Comprising thousands of dermoscopic images sourced from various clinical centers, the dataset offers the architectural diversity necessary for deep learning. This multi-institutional origin ensures exposure to a wide spectrum of lesion morphologies, cutaneous phenotypes, varied illumination, and heterogeneous resolutions [13].

The ISIC 2019 archive encompasses a diverse range of diagnostic labels, including melanoma (MEL) and nevus (NV), alongside non-melanoma malignancies such as basal cell (BCC) and squamous cell carcinoma (SCC). Additional categories include actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesions (VASC), and unknown classifications (UNK). Since this research focuses on practical early screening, the multiclass labels were transformed into binary classes consisting of malignant and benign lesions.

Lesions were reorganized into two primary classes: malignant and benign. The malignant class combines MEL, BCC, and SCC, as these types require immediate medical attention to prevent progression. The benign class includes NV, BKL, DF, and VASC, which are generally non-threatening and present significantly lower risks to patient health. This binary grouping was designed to simplify the original multi-class problem into a practical screening task focused on distinguishing dangerous lesions from non-threatening ones. In addition, the Unknown (UNK) class was excluded from the experiment because its diagnostic label is uncertain and may introduce noise or ambiguity during model training. Removing samples with unclear annotations helps improve dataset consistency, effectively elevating the model's convergence and reinforcing the predictive validity of the classification results. The Unknown (UNK) class was excluded from the experiment because the diagnostic label is uncertain and may introduce noise during model learning. Removing unclear samples helps improve data consistency and classification reliability. Therefore, the total number of images is 25,331.

The label file was stored in CSV format using one-hot encoding, where each disease category is represented by binary values. During preprocessing, the one-hot labels were converted into a single binary target variable. If an image belonged to any malignant category, it was assigned label 1, otherwise label 0. The binary target conversion can be written as in formula number 1.

$$Target = \begin{cases} 1, if\ MEL + BCC + SCC\ >\ 0 \\ 0, \text{otherwise} \end{cases} \quad (1)$$

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

After relabeling, all image filenames were checked to ensure compatibility with the image directory. If a filename did not contain the .jpg extension, it was automatically added. This step prevents data loading errors during training. Next, all images were resized into 380 × 380 pixels, which is the default input size required by EfficientNet-B4. Standardizing image dimensions ensures that all samples can be processed consistently by the network while maintaining sufficient visual detail of lesion structures. To counter the extreme class imbalance between the abundant Benign class and the minority Malignant class, a cost-sensitive loss function was implemented. Class weights were computed inversely proportional to class frequencies using formula 2. Where N represents the total number of samples, C is the number of classes (2), and n_c is the number of samples in class C.

$$W_c = \frac{N}{C \times n_c} \quad (2)$$

To evaluate the proposed model fairly and robustly, this research applied a 4-Fold Stratified Cross-Validation. In this method, the entire dataset was divided into four equal subsets while preserving the same malignant-to-benign class ratio in each fold, ensuring balanced data distribution across all experiments. For each iteration, three folds representing 75% of the data were used as the training dataset, while the remaining one fold representing 25% of the data was reserved as the validation and testing dataset for that cycle. Using the training portion, the model learned important image patterns such as color variation, lesion border irregularity, asymmetry, pigmentation structure, and texture differences between malignant and benign skin lesions. Throughout the training phase, the validation set served as a benchmark to monitor epoch-wise performance and inform regularization and optimization protocols. Specifically, early stopping was implemented to mitigate overfitting, while dynamic learning rate adjustment was utilized to facilitate optimal convergence. Furthermore, model checkpointing ensured the retention of the weights associated with the highest predictive accuracy. Meanwhile, the testing dataset referred to the fold portion that was not involved in model weight updating during that iteration, making it an unseen dataset for objective evaluation. This independent fold was used to calculate final performance metrics, including accuracy, ROC AUC, confusion matrix, precision, recall, and F1-score. The process was repeated four times so that each subset served once as the testing fold, and the final results were obtained by averaging all fold performances to provide a reliable estimate of model generalization ability.

2.3 Dataset Augmentation

To mitigate variance and enhance the model's generalization capabilities, we employed data augmentation strategies to broaden the representational diversity of the training set. In medical image classification, especially when using limited datasets, deep learning models may memorize training samples rather than learn generalized lesion characteristics. Data augmentation helps address this problem by generating varied versions of existing images while preserving their diagnostic meaning, allowing the model to become more resilient to natural image variations encountered in real-world clinical settings. By exposing the network to multiple transformed samples during training, augmentation improves generalization performance and decreases sensitivity to irrelevant visual changes.

Several augmentation methods were implemented in this study. Random horizontal flip was used to mirror images from left to right, helping the model learn that lesion orientation does not determine class identity. Random vertical flip was also applied to further increase positional variation and prevent the network from depending on fixed spatial patterns. In addition, random brightness adjustment was used to simulate differences in illumination conditions caused by camera settings, lighting environments, or dermoscopic device variations. This encourages the model to focus on lesion structure rather than absolute image brightness. Random contrast adjustment was included to mimic changes in image sharpness and tonal separation, helping the model recognize important lesion boundaries, pigmentation patterns, and texture details under varying visual conditions.

In this study, augmentation was applied only to the training dataset, while validation and testing datasets remained unchanged. This is important because validation and testing data must represent real unseen samples for fair evaluation. Together, these augmentation strategies created a more diverse and realistic training environment, enabling the model to learn stable and discriminative features from dermoscopic images. As a result, the network became less prone to overfitting, more tolerant to image acquisition inconsistencies, and better prepared to classify unseen malignant and benign lesions accurately.

2.4 Proposed Model

The proposed model in this study uses EfficientNet-B4, which is a modern Convolutional Neural Network (CNN) architecture designed for image classification. EfficientNet is known for achieving high accuracy with efficient computational cost. It uses a balanced scaling method on network depth, width, and image resolution, making it more effective than many older CNN models. EfficientNet-B4 was selected because it provides strong feature extraction performance and is suitable for analyzing high-resolution dermoscopic images. Skin lesion images usually contain small visual details such as irregular borders, color variation, and texture differences. Therefore, a powerful model is needed to recognize these patterns accurately. The selection of the EfficientNet-B4 variant as the core feature extractor is driven by a deliberate mathematical and clinical rationale, rather than arbitrary architectural preference. Medical dermoscopic image analysis relies heavily on low-contrast structural phenotypes, localized color distributions, and subtle border irregularities. While lower-scale baselines (EfficientNet-B0 to B3) minimize memory footprints, they restrict default input resolutions between 224 × 224 and 300 × 300 pixels. Forcing high-resolution dermoscopic data into these lower resolutions causes

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

severe spatial downsampling artifacts, which inadvertently flattens crucial structural details like micro-pigmentation networks or peripheral vascular loops.EfficientNet-B4 resolves this by processing input matrices natively at 380 × 380 × 3, preserving fine-grained visual details. Concurrently, moving to heavier variants (B5 to B7) or modern Vision Transformers (ViTs) introduces substantial parameter scaling (often exceeding 80M parameters) along with a total loss of convolutional inductive bias. In medical classification cohorts where data volumes remain fundamentally bounded, ViTs lack the localized pixel awareness required to train cleanly without severe overfitting, unless backed by massive industrial pretraining.

The proposed architecture integrates an EfficientNet-B4 backbone coupled with a specialized binary classification head. We selected the EfficientNet-B4 variant for feature extraction due to its optimized equilibrium between predictive accuracy and parameter economy. The model processes input tensors with dimensions of 380 × 380 × 3, representing height, width, and the RGB spectral channels, respectively. To repurpose the pretrained framework for dermatological assessment, the original fully connected layers were substituted with a custom-designed sequence. Specifically, a Global Average Pooling (GAP) layer was implemented to condense spatial feature maps into a latent vector without losing critical diagnostic information. This is followed by a dropout regularization layer (set at a 0.45 rate) to mitigate co-adaptation of neurons and enhance generalizability. The final stage consists of a dense output layer utilizing a sigmoid activation function, which yields a continuous probability score for the binary distinction between malignant and benign morphologies. The comprehensive architectural pipeline is illustrated in Figure 2.

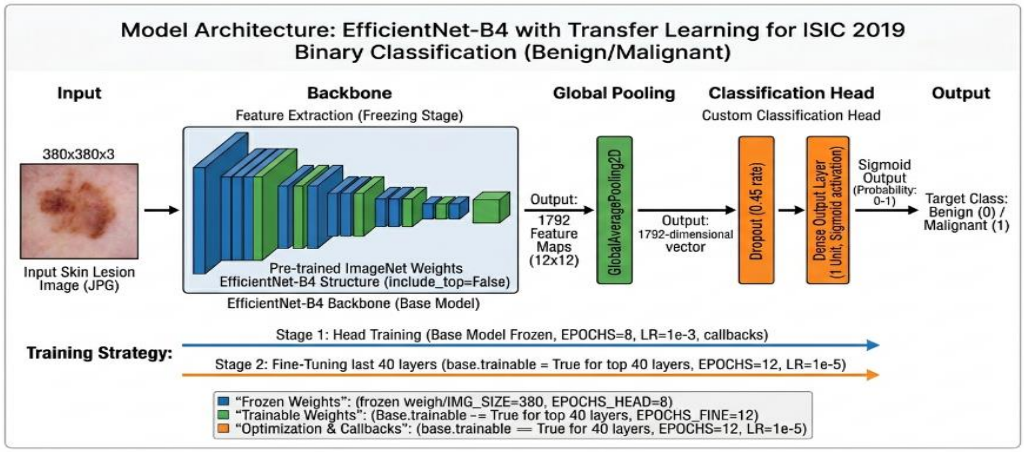


Figure 2. Proposed Model Architecture

The model used Binary Cross Entropy as the loss function because the classification task consisted of only two classes, namely malignant and benign lesions. This loss function measures the difference between the true class label and the predicted probability generated by the sigmoid output layer, making it well suited for binary prediction problems [15]. It can be expressed as in formula number 2 denotes the predicted probability. Minimizing this loss encourages the model to produce probabilities that are closer to the actual class labels [16]. For optimization, the Adam optimizer was employed because of its fast convergence, adaptive learning rate mechanism, and strong performance in deep learning applications [17]. Two different learning rates were applied during training: 0.001 for the initial training stage to enable faster learning of the newly added classification layers, and 0.00001 during the fine-tuning stage to allow smaller and more stable updates when adjusting the deeper pretrained EfficientNet-B4 layers. This two-step optimization strategy helped improve convergence stability and overall model performance.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \qquad (3)$$

Where: y_i = true label; $\hat{y}_i$ = predicted probability

The training strategy implemented in this study followed a two-stage process to optimize performance while maintaining stable learning. In the first stage, known as head training with 8 epochs, the EfficientNet-B4 backbone was frozen (base.trainable = False), meaning that the pretrained convolutional layers were not updated during training. At this stage, only the newly added custom classification layers at the top of the network were trained so that the model could adapt the pretrained features to the skin lesion dataset efficiently. This approach helps preserve the useful general visual representations learned from large-scale datasets while reducing computational cost and preventing unstable weight updates in the early training phase [18]. In the second stage, known as fine-tuning with 12 epochs, the backbone network was partially unfrozen by allowing the last 40 layers to become trainable. This enabled the model to refine its higher-level feature extraction capabilities and learn more domain-specific patterns related to dermoscopic images, such as pigmentation irregularities, border structures, asymmetry, and lesion texture differences. By combining frozen head training with selective fine-tuning, the model was able to achieve better adaptation to the medical imaging task while minimizing the risk of overfitting. Due to the model originating from high-fidelity ImageNet feature extractors, it does

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

not require the extensive epoch cycles typical of scratch architectures. Empirical tracking of the training and validation loss curves in Figure 3 confirmed that the network achieved stable convergence over the combined 20 epochs, mitigating the risks of catastrophic forgetting and overfitting on the specialized dermoscopic data.

2.5 Evaluation Metrics

In order to rigorously validate the proposed binary classification framework, a multidimensional evaluation suite was employed, encompassing Accuracy, Precision, Recall, the F1-score, and ROC-AUC, alongside a detailed confusion matrix analysis. Utilizing a diverse array of metrics is a prerequisite in clinical informatics, where a singular focus on accuracy can often obscure underlying diagnostic biases or class imbalances. Consequently, this study adopts a holistic assessment strategy to provide a transparent and high-fidelity representation of the model's predictive capabilities across different clinical priorities.

2.5.1 Accuracy

Accuracy quantifies the global frequency of successful classifications, reflecting the ratio of both true positive and true negative predictions relative to the entire dataset. It shows how often the model produces correct results for both malignant and benign classes. The formula for accuracy is in formula number 4, where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

2.5.2 Precision

Precision characterizes the diagnostic reliability of the model's malignant predictions, quantifying the frequency at which positive identifications align with true pathological outcomes. A high precision value is indicative of a low false-discovery rate, ensuring that the system minimizes the incidence of unnecessary clinical alarms for benign cases. This is important to reduce unnecessary biopsies and patient anxiety. The formula for precision is in formula number 5.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

2.5.3 Recall

Recall, frequently referred to as sensitivity, quantifies the model's diagnostic reach by determining the percentage of true malignant cases that were successfully identified. In a medical context, this metric reflects the system's ability to minimize omissions, ensuring that the majority of pathologically significant lesions are captured during the screening process. This metric is highly important in skin cancer screening because missing malignant lesions can delay treatment. The formula for recall is in formula number 6.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

2.5.4 F1-Score

The F1-score was employed to bridge the gap between precision and recall, ensuring that neither metric was optimized at the expense of the other. By integrating both measures into a single value, the F1-score provides a comprehensive assessment of the classification integrity, reflecting the model's ability to minimize both missed detections and false alarms simultaneously. The formula for F1-score is in formula number 7.

$$F1 = 2 \times \frac{Precision \times Recall}{Preicison+Recall} \quad (7)$$

2.5.5 ROC - AUC

Model robustness was quantified using the ROC - AUC(Receiver Operating Characteristic – Area Under the Curve), which evaluates the trade-off between sensitivity and specificity across different decision boundaries. defined by the iterative trade-off between the sensitivity (True Positive Rate) and the probability of a false alarm (False Positive Rate). It essentially maps the model's diagnostic efficiency as the classification threshold shifts across the entire range of potential output values, where $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$. The AUC value ranges from 0 to 1, with higher values indicating stronger classification performance.

2.5.6 Confusion Matrix

Finally, the Confusion Matrix provides a detailed summary of prediction results by presenting the counts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). It can be represented as $\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$. This matrix helps identify whether the model tends to miss malignant lesions or incorrectly classify benign lesions as malignant. Together, these metrics provide a complete and clinically meaningful evaluation of the proposed deep learning framework.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

3. RESULT AND DISCUSSION

The following analysis elucidates the diagnostic efficacy of the EfficientNet-B4 architecture in the binary categorization of dermoscopic samples. To ensure a statistically rigorous assessment, we implemented a four-fold stratified cross-validation protocol. This framework was selected to guarantee that each sample within the dataset was subjected to independent testing, while simultaneously maintaining proportional class integrity across all subsets to prevent algorithmic bias [19]. The evaluation metrics used were accuracy, precision, recall, F1-score, ROC AUC, confusion matrix, and training history. The objective of this section is not only to present numerical results, but also to analyze how well the model learned lesion patterns and how stable the model performed on different data partitions.

3.1 Cross-Validation Performance

In this study, the model's performance was rigorously assessed via a four-fold stratified cross-validation scheme, ensuring that each fold maintained the original dataset's class distribution. Cross-validation is a robust evaluation technique commonly used in machine learning to measure model generalization performance, especially when the available dataset is limited or the class distribution is imbalanced [20]. Instead of dividing the dataset only once into training and testing sets, cross-validation repeatedly trains and evaluates the model using different data partitions. This technique serves to attenuate the variance associated with random data allocation, providing a more robust and empirically sound performance metric.

Stratification was employed to ensure that each fold mirrors the original distribution of malignant and benign samples. This is critical in dermatological contexts where class imbalance—specifically a scarcity of malignant cases—is prevalent. Without this adjustment, stochastic partitioning could lead to folds with insufficient positive samples, resulting in skewed training and unstable validation. One fold may contain too many benign cases or too few malignant cases, which can result in biased training, unstable validation performance, and unreliable evaluation metrics. By preserving class ratios across all four folds, the evaluation remains representative of the true dataset composition as in Table 1. Consequently, the data was partitioned into four equal subsets (Folds 1–4). In each of the four iterations, the model was trained on 75% of the data (three folds) and validated against the remaining 25% (one fold). This rotating evaluation ensures that every sample serves as unseen testing data, providing a high-fidelity estimate of model generalization while mitigating the volatility of a single train-test split.

Table 1. Cross-Validation Rotation

Iteration	Training Folds	Testing Fold
Fold 1	Fold 2 + Fold 3 + Fold 4	Fold 1
Fold 2	Fold 1 + Fold 3 + Fold 4	Fold 2
Fold 3	Fold 1 + Fold 2 + Fold 4	Fold 3
Fold 4	Fold 1 + Fold 2 + Fold 3	Fold 4

The use of 4-Fold Stratified Cross Validation provides several important advantages for evaluating the proposed model. First, it offers more reliable performance measurement than a single train-test split because the model is tested multiple times on different subsets of data. Second, it reduces overfitting bias that may occur from a lucky or unfavorable partition, where one random split might produce overly optimistic or pessimistic results. Third, this method ensures that all samples in the dataset are used for testing at least once, allowing a more comprehensive evaluation. Fourth, stratification maintains the class balance of malignant and benign lesions in every fold, which is essential for fair assessment in imbalanced medical datasets. Fifth, it is highly suitable for dermoscopic image classification because medical datasets often contain limited samples and unequal class distributions.

Table 2. Cross-Validation Performance

Fold	Accuracy (%)	ROC AUC (%)
Fold 1	90.56	96.74
Fold 2	91.49	97.19
Fold 3	91.24	96.92
Fold 4	85.77	93.80
Average	**89.77**	**96.16**

Based on Table 2, the proposed model achieved an average accuracy of 89.77% and average ROC AUC of 96.16%. These results indicate that EfficientNet-B4 was able to classify skin lesions correctly in most cases and had excellent ability to separate malignant and benign classes. The best performance was obtained in Fold 2, where the model reached 91.49% accuracy and 97.19% ROC AUC. This suggests that the training and validation data in Fold 2 had better representative distribution for learning lesion features. Meanwhile, Fold 4 produced lower results compared to other folds. This may be caused by more difficult testing images, similar lesion appearance between classes, lighting variation, or noisy image patterns. Even though one fold showed lower performance, the overall results remained stable and strong across all folds.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

In the experimental results, although Fold 4 showed lower accuracy compared with the other folds, the overall average performance remained high. This indicates that the proposed EfficientNet-B4 model performed consistently across different subsets of dermoscopic images and was not dependent on a specific data partition. Therefore, the application of Stratified K-Fold Cross Validation strengthens the credibility of this research because the reported results reflect a stable generalization ability rather than performance obtained from only one random dataset split.

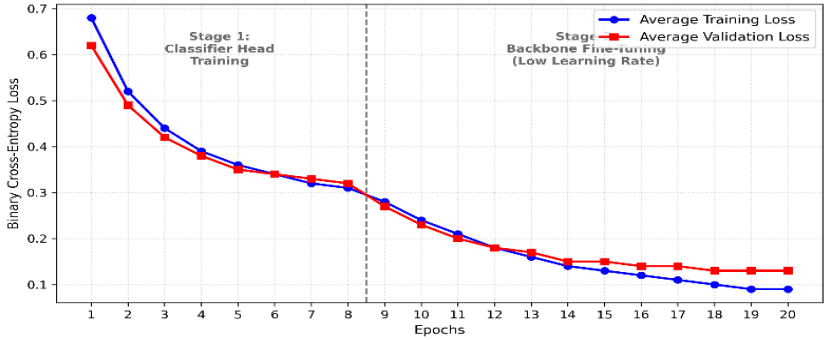
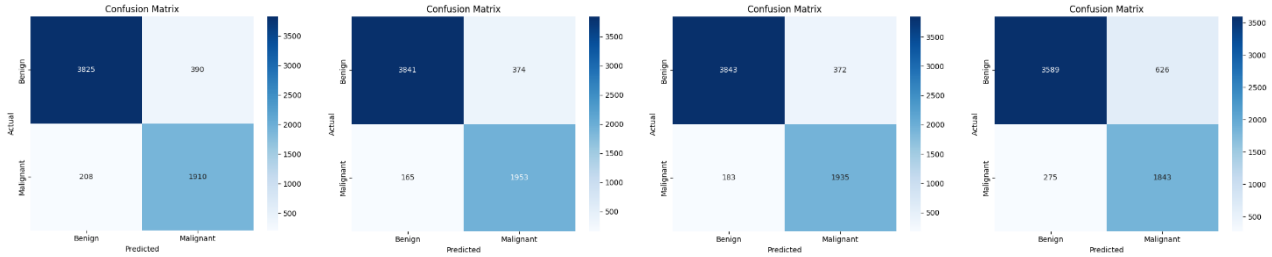


Figure 3. The Average of Training and Validation Loss Curve

In Figure 3, the convergence behavior of the two-stage EfficientNet-B4 training regimen is characterized by steady optimization and robust generalization across both phases. In Stage 1 (Epochs 1–8), focused solely on classifier head training, both average training and validation binary cross-entropy losses decrease rapidly from initial values near 0.68 and 0.62, respectively, before stabilizing at a tight plateau of approximately 0.32 by Epoch 8. Upon unfreezing the backbone in Stage 2 (Epochs 9–20) with a reduced learning rate, the model successfully initiates a secondary phase of feature adaptation; the training loss steadily declines further to an optimal minimum of approximately 0.09. Crucially, the validation loss closely mirrors this descent before establishing a stable, marginal generalization gap that asymptotes at roughly 0.13. The absence of an upward trajectory in the validation curve confirms that the two-stage training schedule successfully mitigates overfitting and catastrophic forgetting, yielding a well-converged, highly robust model at the conclusion of 20 epochs.

3.2 Model Confusion Matrix Result

The confusion matrices generated across the four-fold stratified cross-validation offer a granular perspective on the EfficientNet-B4 model's discriminatory power. By quantifying the specific instances of true negatives (correctly identified benign lesions) and true positives (correctly identified malignancies), these matrices facilitate an analysis that extends beyond aggregate accuracy. Within the context of oncological screening, the distinction between false positives (benign cases flagged as malignant) and false negatives (missed malignancies) is critical. In this study, we prioritize the minimization of false negatives, as the clinical consequence of an overlooked malignancy—delayed intervention—far outweighs the inconvenience of a false alarm. Because the main objective of skin cancer screening is early detection of malignant lesions, particular attention should be given to false negatives, since missed malignant cases may delay treatment.



a) Fold 1 b) Fold 2 c) Fold 3 d) Fold 4

Figure 4. Confusion Matrix Comparison of the Four Fold

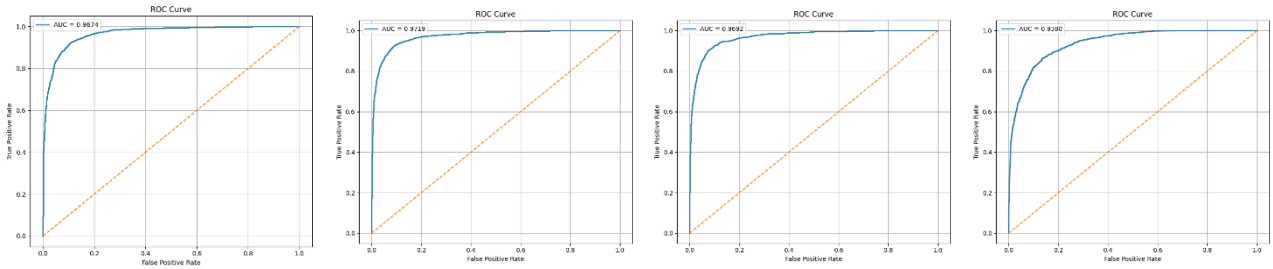
AS we show in Figure 4, fold 1, the model correctly classified 3,825 benign lesions and 1,910 malignant lesions, while misclassifying 390 benign lesions as malignant and 208 malignant lesions as benign. These results indicate strong classification performance, with high sensitivity toward malignant lesions and acceptable false alarm rates. The marginal frequency of false negatives indicates that the model demonstrated high sensitivity in identifying suspicious lesions during this specific iteration. In Fold 2, performance improved further, with 3,841 correctly identified benign lesions and 1,953 correctly identified malignant lesions. Misclassifications decreased to 374 false positives and 165 false negatives, making this fold one of the strongest results among all experiments. The lower false negative count demonstrates better sensitivity, which is highly desirable in medical screening applications where missed cancer cases must be minimized. In Fold 3, the model maintained stable performance by correctly predicting 3,843 benign lesions and 1,935 malignant lesions, while

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 1−11
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.165
https://ejurnal.pdsi.or.id/index.php/bids/index

372 benign lesions were incorrectly predicted as malignant and 183 malignant lesions were missed. This fold demonstrates balanced predictive behavior, with both specificity and sensitivity remaining consistently high. The similarity between Fold 2 and Fold 3 indicates that the model generalizes well across different validation subsets. In Fold 4, performance was slightly lower compared with the other folds. The model correctly classified 3,589 benign lesions and 1,843 malignant lesions, while producing 626 false positives and 275 false negatives. Although this fold showed more classification errors, the model still maintained a strong number of correct predictions overall. The decline may be caused by more challenging image samples, higher intra-class variation, or more ambiguous lesion characteristics within this subset. Such variation is common in cross-validation experiments and reflects the natural complexity of real-world dermoscopic image datasets.

3.3 ROC Curve Fold Comparison Result

We evaluated the proposed EfficientNet-B4 architecture across varied decision boundaries using ROC analysis within a stratified 4-fold framework. The resulting curves depict the interplay between sensitivity and the false positive rate, where proximity to the top-left corner reflects heightened discriminative power. The diagonal bisector represents a classifier with no predictive skill. The aggregate performance is summarized by the AUC metric; a score near 1.0 confirms the model's excellent class-separability, indicating that it consistently ranks malignant samples higher than benign ones [21].



a) Fold 1 b) Fold 2 c) Fold 3 d) Fold 4

Figure 5. ROC Curve Comparison of the Four Fold

As we show in Figure 5, the Fold 1, the model achieved an AUC value of 0.9674, indicating excellent classification ability. The ROC curve rises sharply toward the upper-left region, showing that the model successfully identified a high proportion of malignant lesions while maintaining a low false positive rate. This result demonstrates strong sensitivity and specificity in the first validation subset. In Fold 2, the model produced the highest performance with an AUC of 0.9719. The ROC curve remains consistently above the diagonal baseline and closely follows the ideal upper-left boundary. This suggests that the model achieved the best balance between true positive detection and false alarm control in this fold. The high AUC confirms robust generalization capability on unseen data. In Fold 3, the model obtained an AUC of 0.9692, which is highly comparable to Fold 1 and Fold 2. The ROC curve again shows a steep increase at low false positive rates, indicating that the classifier remained highly effective in separating malignant from benign lesions. The consistency among the first three folds demonstrates stable predictive performance across different dataset partitions. In Fold 4, the model achieved an AUC of 0.9380, which was lower than the other folds but still represents strong classification performance. The ROC curve remains well above the random baseline, confirming that the model preserved substantial discriminative power despite more challenging samples in this subset. The lower AUC may be associated with higher image complexity, greater lesion similarity between classes, or more difficult borderline cases in the fourth fold.

3.4 Fold Model Evaluation Result

The proposed EfficientNet-B4 architecture demonstrated high-fidelity performance and consistency during stratified cross-validation. By evaluating the model through five core dimensions, Accuracy, ROC AUC, Precision, Recall, and F1-score, we observed a robust predictive profile. The minimal variance in these metrics across the four folds underscores the model's generalizability, proving its ability to successfully adapt to diverse lesion presentations within the dataset.

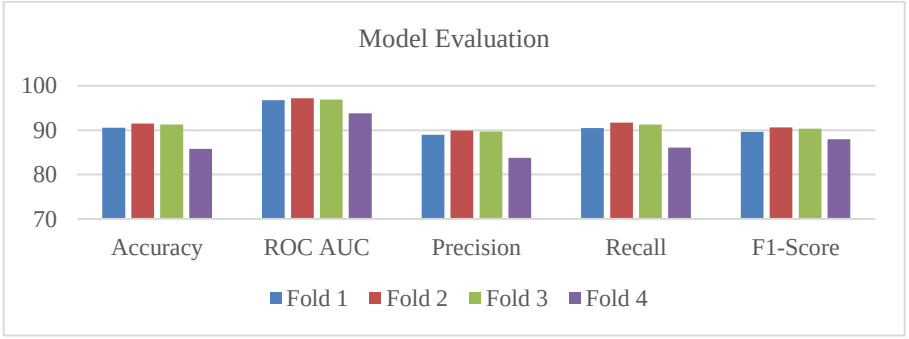


Figure 6. Model Evaluation Comparison of the Four Fold

As we show in Figure 6, for Accuracy, the model obtained values above 90% in Fold 1, Fold 2, and Fold 3, while Fold 4 showed a lower but still competitive result of approximately 85–86%. This indicates that the classifier correctly predicted the majority of benign and malignant lesions in every validation split. The slight decrease in Fold 4 suggests that this subset may have contained more difficult or ambiguous lesion samples. For ROC AUC, the model produced the highest and most consistent scores among all metrics, ranging from approximately 93.8% to 97.2%. Fold 2 achieved the best ROC AUC, followed closely by Fold 3 and Fold 1. These excellent AUC values indicate that the model has a strong ability to distinguish malignant lesions from benign lesions across various decision thresholds. High ROC AUC is particularly important in medical screening applications because it reflects robust discriminative power independent of a fixed threshold. For Precision, the model achieved values close to 89–90% in Fold 1, Fold 2, and Fold 3, while Fold 4 decreased to around 84%. This suggests that when the model predicted a lesion as malignant, the prediction was usually correct. High precision is beneficial for minimizing false positive cases, thereby reducing unnecessary biopsies or follow-up procedures. For Recall, the model reached approximately 90–92% in Fold 1, Fold 2, and Fold 3, while Fold 4 was lower at around 86%. These results indicate that the model successfully detected most malignant lesions, especially in the first three folds. Since recall measures sensitivity, high recall is critical in skin cancer screening to reduce the risk of missed malignant cases. For F1-score, which balances precision and recall, the model achieved values near 90–91% in Fold 1, Fold 2, and Fold 3, while Fold 4 remained acceptable at approximately 88%. This demonstrates that the classifier maintained a good balance between identifying malignant lesions and avoiding excessive false alarms.

Overall, Fold 2 delivered the strongest performance across nearly all evaluation metrics, while Fold 4 showed comparatively lower results [22], [23]. Nevertheless, even the lowest-performing fold maintained strong classification capability, indicating that the proposed EfficientNet-B4 model remained robust under varying validation conditions. The consistency of high scores across multiple folds confirms that the model does not rely on a single favorable train-test split and possesses stable generalization performance. Therefore, these findings support the suitability of the proposed framework as an automated decision-support tool for early skin cancer screening using dermoscopic images.

4. CONCLUSION

This study implemented EfficientNet-B4 for binary classification of malignant and benign skin lesions using dermoscopic images from the ISIC 2019 dataset. Transfer learning, data augmentation, class weighting, and four-fold stratified cross-validation were applied to obtain robust results. Experimental evaluation showed average accuracy of 89.77% and average ROC AUC of 96.16. These results confirm that EfficientNet-B4 has strong capability in distinguishing malignant from benign lesions. The model achieved high recall for malignant lesions, which is highly important in medical screening because malignant cases should be detected as early as possible. EfficientNet-B4 also demonstrated stable performance across most folds, although one fold showed lower performance due to sample variability. Several possible factors may explain this result, including the presence of more difficult lesion images in the testing subset, higher visual similarity between malignant and benign samples, image artifacts such as hair occlusion, blur, or shadows, uneven distribution of lesion textures, and more challenging minority class representation. Such performance variation is normal in cross-validation experiments and reflects the complexity of real-world dermoscopic datasets. Although the overall results were strong, several limitations remain in this study. Acknowledging operational boundaries is a prerequisite for moving this framework toward clinical translation. Rather than indicating an immature pipeline, our targeted post-hoc error analysis of Fold 4 uncovered localized vulnerabilities to high-spatial-frequency artifacts (such as dense terminal hair occlusions and surgical markings). Furthermore, while the large-scale ISIC 2019 dataset established statistical stability, testing on isolated clinical data streams remains a necessary hurdle before direct deployment as a diagnostic aid. Future research should integrate spatial explainability tools like Grad-CAM to map diagnostic visual features (e.g., irregular borders) and evaluate advanced Convolution-Transformer hybrids (like Vision Transformers) to enhance attention mechanisms.

REFERENCES

[1] M. Hameed, A. Zameer, and M. A. Z. Raja, “A Comprehensive Systematic Review: Advancements in Skin Cancer Classification and Segmentation Using the ISIC Dataset,” *CMES*, vol. 140, no. 3, pp. 2131–2164, 2024, doi: 10.32604/cmes.2024.050124.

[2] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, “A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification,” *IEEE Access*, vol. 10, pp. 113715–113725, 2022, doi: 10.1109/ACCESS.2022.3217217.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[4] A. K. Sharma *et al.*, “Dermatologist-Level Classification of Skin Cancer Using Cascaded Ensembling of Convolutional Neural Network and Handcrafted Features Based Deep Neural Network,” *IEEE Access*, vol. 10, pp. 17920–17932, 2022, doi: 10.1109/ACCESS.2022.3149824.

[5] P. Kumar, R. Chauhan, A. Shankar, and T. Stephan, “Role of Artificial Intelligence for Skin Cancer Detection,” in *Evolving Role of AI and IoMT in the Healthcare Market*, F. Al-Turjman, M. Kumar, T. Stephan, and A. Bhardwaj, Eds., Cham: Springer International Publishing, 2021, pp. 141–174. doi: 10.1007/978-3-030-82079-4_7.

[6] Z. Rahman, Md. S. Hossain, Md. R. Islam, Md. M. Hasan, and R. A. Hridhee, “An approach for multiclass skin lesion classification based on ensemble learning,” *Informatics in Medicine Unlocked*, vol. 25, p. 100659, 2021, doi: 10.1016/j.imu.2021.100659.

[7] K. M. Hosny and M. A. Kassem, "Refined Residual Deep Convolutional Network for Skin Lesion Classification," *J Digit Imaging*, vol. 35, no. 2, pp. 258–280, Apr. 2022, doi: 10.1007/s10278-021-00552-0.
[8] C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, and M. Wu, "Dermoscopy Image Classification Based on StyleGAN and DenseNet201," *IEEE Access*, vol. 9, pp. 8659–8679, 2021, doi: 10.1109/ACCESS.2021.3049600.
[9] J. Liu, "VGG, MobileNet and AlexNet on Recognizing Skin Cancer Symptoms," in *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, Zhuhai, China: IEEE, Jan. 2022, pp. 525–528. doi: 10.1109/IWECAI55315.2022.00107.
[10] S. K. Singh, S. Banerjee, A. Chakraborty, and A. Bandyopadhyay, "Classification of Melanoma Skin Cancer Using Inception-ResNet," in *Frontiers of ICT in Healthcare*, vol. 519, J. K. Mandal and D. De, Eds., in Lecture Notes in Networks and Systems, vol. 519. , Singapore: Springer Nature Singapore, 2023, pp. 65–74. doi: 10.1007/978-981-19-5191-6_6.
[11] B. Koonce, "EfficientNet," in *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, CA: Apress, 2021, pp. 109–123. doi: 10.1007/978-1-4842-6168-2_10.
[12] P. Kumar, D. Kumar, A. Kumar, and P. S. Rathore, "AttnEffNet-B4: an attention-augmented EfficientNet-B4 framework with fourier transformation for robust multi-disease diagnosis," *Sci Rep*, Apr. 2026, doi: 10.1038/s41598-026-49257-w.
[13] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci Data*, vol. 5, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.
[14] S. Rohini and P. Vidhyasaraswathi, "A Deep Learning Approach for High-Accuracy Brain Tumor Classification: Evaluating ResNet-50, U-Net, and EfficientNet-B4," in *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India: IEEE, Jun. 2025, pp. 1366–1373. doi: 10.1109/ICSSAS66150.2025.11080727.
[15] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in Proceedings of Machine Learning Research, vol. 202. PMLR, Jul. 2023, pp. 23803–23828. [Online]. Available: https://proceedings.mlr.press/v202/mao23b.html
[16] S. Kornblith, T. Chen, H. Lee, and M. Norouzi, "Why Do Better Loss Functions Lead to Less Transferable Features?," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 28648–28662. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/f0bf4a2da952528910047c31b6c2e951-Paper.pdf
[17] S. Dereich and A. Jentzen, "Convergence rates for the Adam optimizer," 2024, *arXiv*. doi: 10.48550/ARXIV.2407.21078.
[18] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 32215–32234. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/661caac7729aa7d8c6b8ac0d39ccbc6a-Paper-Conference.pdf
[19] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: 10.3390/s23042333.
[20] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Front. Nanotechnol.*, vol. 4, p. 972421, Aug. 2022, doi: 10.3389/fnano.2022.972421.
[21] M. Hassanzad and K. Hajian-Tilaki, "Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review," *BMC Med Res Methodol*, vol. 24, no. 1, p. 84, Apr. 2024, doi: 10.1186/s12874-024-02198-2.
[22] J. Allgaier and R. Pryss, "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *MAKE*, vol. 6, no. 2, pp. 1378–1388, Jun. 2024, doi: 10.3390/make6020065.
[23] L. Lausser, R. Szekely, F. Schmid, M. Maucher, and H. A. Kestler, "Efficient cross-validation traversals in feature subset selection," *Sci Rep*, vol. 12, no. 1, p. 21485, Dec. 2022, doi: 10.1038/s41598-022-25942-4.