FMDB *Transactions on Sustainable Computing Systems*

End-to-End Machine Learning System for Accurate Melanoma Detection and Risk Assessment

J. Angelin Jeba[1,*], S. Rubin Bose[2], R. Regin[3], A. Gladysmerlin[4], M. Rehena Sulthana[5]

[1]Department of Electronics and Communication Engineering, S. A. Engineering College, Chennai, Tamil Nadu, India.
[2,3]School of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[4]Department of Science and Humanities, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[5]Department of Information Technology and Engineering, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
angelinjeba@saec.ac.in[1], rubinbos@srmist.edu.in[2], reginr@srmist.edu.in[3], gladysmerlin@dhaanishcollege.in[4], rsulthana@academic.mit.edu.au[5]

Abstract: The Growing incidence of melanoma, the most lethal skin cancer, is one of the motivations for the development of quick and accurate diagnostic techniques. Early diagnosis is key to improving patient survival, but traditional diagnosis can be time-consuming and dependent on specialist availability. This article proposes automated binary classification of skin lesions as malignant (melanoma) or benign using a deep learning approach. Researchers construct a high-quality composite dataset by combining images from multiple public sources, including the HAM10000 dataset and the International Skin Imaging Collaboration (ISIC) archives for 2016-2019. Our model is based on a Convolutional Neural Network (CNN) architecture that uses transfer learning with the EfficientNet model for improved feature extraction. The merged dataset is utilised to train the system to acquire discriminative features of melanocytic lesions. Experimental results demonstrate the model's high efficacy, with excellent accuracy, precision, and recall in dermoscopic image classification. The research demonstrates the power and simplicity of deep learning as a tool to assist dermatologists in accurately and early detecting melanoma, ultimately leading to improved patient outcomes.

Keywords: Skin Cancer; Deep Learning; Convolutional Neural Network (CNN); Transfer Learning; Medical Image Analysis; Dermoscopic Images; Melanocytic Lesions.

Received on: 23/01/2025, **Revised on:** 19/04/2025, **Accepted on:** 23/06/2025, **Published on:** 09/12/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCS

DOI: https://doi.org/10.69888/FTSCS.2025.000528

Cite as: J. A. Jeba, S. R. Bose, R. Regin, A. Gladysmerlin, and M. R. Sulthana, "End-to-End Machine Learning System for Accurate Melanoma Detection and Risk Assessment," *FMDB Transactions on Sustainable Computing Systems*, vol. 3, no. 4, pp. 264-277, 2025.

1. Introduction

Melanoma is the most dangerous form of skin cancer and has been a rapidly growing incidence across the globe over the past few decades. However, it is not as common as some other skin cancers, like basal cell carcinoma; melanoma accounts for most

*Corresponding author.

skin cancer-related deaths. The prognosis of patients affected by melanoma is strongly dependent on the stage of the diagnosis of the tumour; if diagnosed at an early stage, the 5-year survival rate can be over 95%, whereas this rate drops significantly in advanced stages. Visual examination by a dermatologist is currently the gold standard for diagnosing melanoma, aided at times by dermoscopy, and then skin biopsy for histopathological examination. However, this method faces several challenges. There is a global shortage of dermatologists, leading to long wait times for referrals. Furthermore, pigmented lesion diagnosis may be subjective and operator-dependent, depending on the clinician's experience, with consequent variation in diagnostic accuracy. These factors emphasise the urgent need for low-cost, automated, and objective devices to support initial screening of skin lesions. This paper meets this demand through the development of a deep learning-powered system for automatic skin lesion classification. Researchers leverage the power of Convolutional Neural Networks (CNNs), which have achieved remarkable success across a wide range of medical image processing tasks. To train a robust, generalizable model, researchers gather and preprocess a large dataset from two of the largest public archives: the HAM10000 dataset and the International Skin Imaging Collaboration (ISIC) archives from 2016-2019. Our approach uses transfer learning with the EfficientNet architecture to distinguish between malignant and benign lesions, aiming to provide a reliable decision-making system for clinicians.

2. Literature Review

The application of deep learning for melanoma detection has been a major focus of medical imaging research. The following review summarises the key studies that provide the foundation for this paper, from establishing baseline performance to introducing the datasets and model architectures used by researchers. Esteva et al. [1] conducted a foundational study demonstrating that a deep convolutional neural network (CNN) could classify skin cancer with an accuracy comparable to that of board-certified dermatologists. Their work, which used a large dataset and a pre-trained Google Inception v3 model, was a major proof-of-concept that deep learning was a viable tool for dermatology. Haenssle et al. [2] conducted a direct “Man against machine” comparison, pitting a CNN trained on the HAM10000 dataset against 58 human dermatologists. Their results showed that the CNN had higher sensitivity (missed fewer melanomas) than the average dermatologist, providing strong evidence that AI could serve as a powerful diagnostic aid. Codella et al. [3] described the ISIC 2017 challenge and provided an overview of the state-of-the-art methods. Their analysis showed that top-performing models consistently used techniques like data augmentation and ensembles of different CNNs. Tschandl et al. [4] formally introduced the HAM10000 dataset, which is a core component of our paper. Their paper detailed a large collection of 10,015 dermoscopic images, sourced from multiple sources and spanning seven common categories of skin lesions, making it an essential benchmark for training and testing.

Tan and Le [5] in their paper on EfficientNet. They introduced a new, highly efficient scaling method for CNNs. EfficientNet models achieve state-of-the-art accuracy while being significantly smaller and faster than previous models, which is why researchers selected it as the base architecture for our paper. Brinker et al. [6] further reinforced this point in another head-to-head study, in which their CNN outperformed 136 of 157 dermatologists in a melanoma image classification task. These studies collectively demonstrate that well-trained deep learning models can match, and in some cases exceed, expert human performance. Pérez et al. [7] highlighted the importance of data augmentation for skin lesion analysis. Their study explored how techniques such as image rotation, flipping, and scaling during training are critical for preventing overfitting and helping the model generalise to new images it has never seen. Harangi [8] further explored the use of model ensembles, showing that combining several different CNNs into an “ensemble” and averaging their predictions led to better and more stable skin lesion classification than any single model could achieve on its own. Selvaraju et al. [9] introduced Grad-CAM, a technique that generates “heatmaps” visualising which parts of an image a CNN is looking at to make its prediction. This is a key component of Explainable AI (XAI) and is noted in our future work. Harangi [10] also confirmed the power of ensembles. Their paper on classifying skin lesions showed that combining multiple CNNs was highly effective, reinforcing the idea that this is a best practice for achieving top-tier results in dermatological classification. Yap et al. [11] explored multimodal learning by combining image data with patient metadata (like age and lesion location) to classify skin lesions. Their findings suggest that adding this extra context can improve the model's performance, which is another important direction for future enhancement.

Combalia et al. [12] contributed another significant dataset, BCN20000. While not used in our paper, their work highlights the ongoing, critical effort in the research community to collect larger, more diverse datasets to improve model generalisation. Marchetti et al. [13] published the results of the ISIC 2016 challenge. Their findings, which compared algorithm accuracy with dermatologists', further validated the use of computer algorithms as a reliable tool for melanoma diagnosis from dermoscopic images [3]; [4]; [14]. Menegola et al. [14] detailed their specific high-performing method for the ISIC 2017 challenge. Their work, like many top entries, relied on an ensemble of deep networks, demonstrating that combining predictions from multiple models is a robust strategy for improving accuracy. Kolesnikov et al. [15] introduced the Vision Transformer (ViT), applying the successful Transformer architecture (from language models) to image recognition. This paper represents the cutting edge of computer vision and a potential successor to CNNs, marking a relevant direction for future research in the field. In summary, this body of literature establishes a clear path for our paper [1]; [2]; [6]. Foundational studies have proven that CNNs can match or exceed human-level performance [4]. The success of this research relies on large, public datasets such as HAM10000 and the ISIC archives, as well as advanced architecture such as EfficientNet [5]. Our paper builds directly on these findings by

combining key datasets to create a large, diverse training set and applying a state-of-the-art EfficientNet model, following established best practices such as data augmentation to develop a robust, highly accurate melanoma detection tool [7].

3. Methodology

The methodology employed in this paper follows a structured, end-to-end pipeline designed to address the challenge of automated melanoma detection from dermoscopic images. The process begins with the rigorous collection and Preparation of a large-scale dataset, proceeds through the construction and training of a sophisticated deep learning model, and concludes with a comprehensive evaluation of the model's performance. This paper details the specific techniques and tools utilised in each stage: Data Preparation, Model Architecture, Model Training and Fine-Tuning, and Evaluation Metrics.

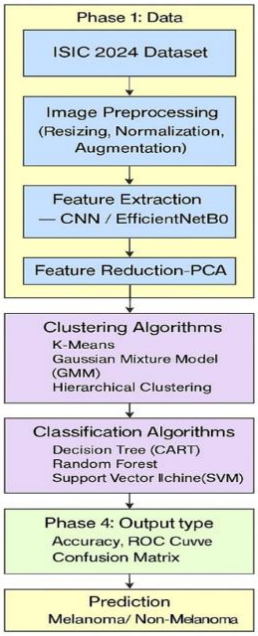

Figure 1: Block diagram

Figure 1, a block diagram, illustrates the overall workflow of the proposed melanoma detection system using the ISIC 2024 dataset. The process begins with the collection of dermoscopic images, which are then preprocessed through resizing, normalisation, and data augmentation to enhance quality and ensure uniformity. These preprocessed images are then fed into a Convolutional Neural Network (CNN) based on the EfficientNetB0 architecture for automatic feature extraction, capturing key characteristics of skin lesions, including texture, shape, and colour. The extracted high-dimensional features are reduced using Principal Component Analysis (PCA) to retain only the most relevant data while minimising computational complexity. The refined features are further analysed using clustering algorithms, such as K-Means, the Gaussian Mixture Model (GMM), and Hierarchical Clustering, to identify natural groupings in the data. The next stage involves classification using machine learning algorithms such as Decision Tree (CART), Random Forest, and Support Vector Machine (SVM), which categorise lesions as melanoma or non-melanoma. The model's performance is then evaluated using metrics such as accuracy, ROC curve, and confusion matrix to ensure reliability. Finally, the system produces a prediction indicating whether a lesion is malignant (melanoma) or benign (non-melanoma), providing an efficient and accurate computer-aided diagnostic tool for early skin cancer detection.

3.1. Data Preparation

Recognising that the performance of any deep learning model is heavily reliant on the quality and quantity of the training data, a significant effort was devoted to constructing a robust, diverse dataset. The primary challenge addressed was the limitation of training on a single source, which can lead to models that fail to generalise well. To mitigate this, researchers aggregated data from five prominent, publicly available archives: HAM10000, a well-established dataset containing 10,015 dermoscopic images distributed across seven common diagnostic categories of pigmented skin lesions. ISIC Archives (2016-2019): Collections released as part of the annual International Skin Imaging Collaboration challenges, providing a large volume of high-quality dermoscopic images with associated ground truth diagnoses. The complete data preparation workflow involved several critical, automated steps:

Automated Download: The official Kaggle Application Programming Interface (API) was employed to programmatically download all specified datasets directly into the Google Colab cloud environment. Due to the substantial size of these datasets (particularly ISIC 2017), a memory-safe script was implemented. This script downloaded and unzipped each dataset sequentially, deleting the compressed archive file immediately after extraction to conserve limited disk space and prevent RAM-related session crashes.

Label Standardisation: A crucial preprocessing step was required to handle the heterogeneity in the metadata accompanying each dataset. Different archives used varied column names (e.g., 'dx', 'benign_malignant', 'MEL', 'target') and distinct formats (e.g., categorical text, binary flags, multiclass flags) for diagnostic labels. A custom Python script, utilising the Pandas library, was developed to parse the metadata file (typically a CSV) for each of the five datasets. This script applied specific mapping rules to standardise the diverse diagnostic labels into a uniform binary classification system relevant to our objective: 'Malignant' (specifically for lesions identified as melanoma) and 'Benign' (encompassing all other non-melanoma lesion types, including nevi, seborrheic keratosis, basal cell carcinoma, etc.).

Data Consolidation: Following label standardisation, the processed metadata from all sources, containing the unified binary label and the full file path for each corresponding image, was concatenated into a single, master Pandas Data Frame. This resulted in a comprehensive index of over 80,000 unique image paths and their corresponding ground-truth labels, serving as the basis for model training and evaluation.

Train/Test Split: To ensure an objective, unbiased assessment of the final model's generalisation capability, the aggregated dataset was split into distinct training and test subsets. An 80/20 split was adopted, allocating 80% of the data to model training and reserving the remaining 20% for final evaluation. This partitioning was performed using the train_test_split function from the Scikit-learn library, with the stratify parameter used. Stratified sampling ensures that the proportions of 'Malignant' and 'Benign' cases are preserved in both the training and test sets, thereby preventing biases related to class imbalance during evaluation.

File Organisation: Deep learning frameworks like TensorFlow, particularly the image_dataset_from_directory utility, require input data to be organised in a specific hierarchical folder structure. Therefore, a final script was executed using Python's os and shutil libraries. This script iterated through training and testing Data Frames. It physically copied each image file from its original download location into a newly created, structured directory system: master_dataset/train/Malignant, master_dataset/train/Benign, master_dataset/test/Malignant, and master_dataset/test/Benign. This structured approach enables efficient data loading during model training.

3.2. Model Architecture

To leverage the vast knowledge encoded in models trained on large-scale general image datasets and to mitigate the need for excessively long training times, researchers adopted a transfer learning approach. This technique involves using a pre-existing, powerful model as a starting point for our specific task.

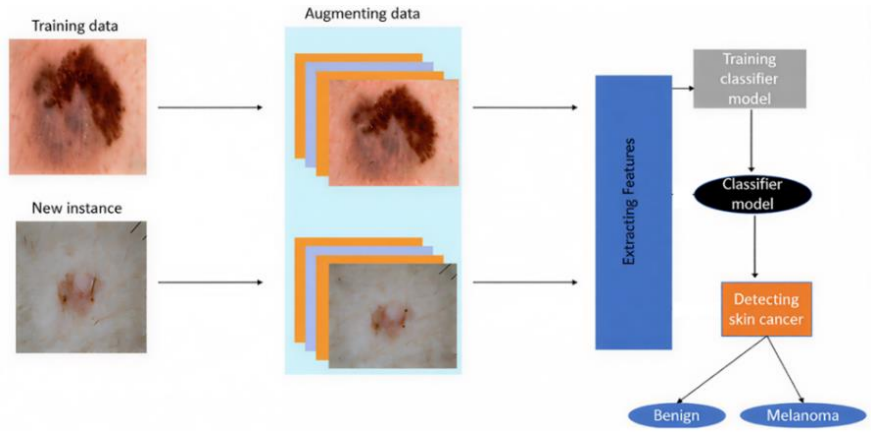

Figure 2: Architecture diagram

Figure 2 architecture diagram illustrates the workflow for the melanoma detection system, encompassing both model training and prediction phases. The training process begins with the input Training Data (dermoscopic images from the ISIC datasets), which undergoes Data Augmentation to increase dataset diversity before being fed into the Feature Extraction module,

implemented using a Convolutional Neural Network (CNN) like EfficientNet. This module learns critical visual patterns, and its output is used for training the Classifier Model, resulting in a finalised Classifier Model optimised for distinguishing between lesion types. For prediction, a New Instance (an unseen lesion image) follows a similar path through preprocessing and feature extraction; its extracted features are then processed by the trained Classifier Model for the final step of Detecting Skin Cancer, ultimately producing a classification output of either “Benign” or “Malignant”.

Base Model Selection: Researchers selected EfficientNetB0 as the base model for our architecture. EfficientNet is a family of state-of-the-art Convolutional Neural Networks (CNNs) developed by Google, known for achieving superior accuracy while being significantly more computationally efficient (fewer parameters, lower FLOPs) than earlier architectures such as VGG or ResNet. The EfficientNetB0 variant was chosen as a good balance between performance and computational cost. The model was initialised with weights pre-trained on the ImageNet dataset, a massive repository containing over 14 million images across 1,000 diverse object categories.

Feature Extraction Role: The core principle of transfer learning in this context is to utilise the pre-trained EfficientNetB0 primarily as a sophisticated feature extractor. The final fully connected classification layer of the original ImageNet-trained model, which was designed to predict 1,000 classes, was removed. The remaining convolutional layers, having learned a rich hierarchy of visual features—ranging from simple edges and textures in the early layers to complex object parts and shapes in deeper layers—serve as a powerful engine for extracting meaningful, high-level representations from our dermoscopic skin lesion images. Custom Classification Head: A new classification “head” was constructed and appended to the truncated EfficientNetB0 base. This custom head was specifically designed for our binary classification task (Malignant vs. Benign) and consisted of:

A Global Average Pooling 2D (GAP) layer: This layer takes the high-dimensional feature maps produced by the final convolutional block of the base model and reduces each map to a single average value. This significantly reduces the number of model parameters, acting as a form of regularisation to help prevent overfitting. A single Dense (fully connected) layer with sigmoid activation function: This final layer takes the pooled features from the GAP layer and outputs a single scalar value between 0 and 1. The sigmoid function is the standard choice for binary classification problems, as its output can be directly interpreted as the model's predicted probability of the input image belonging to the positive class (designated as 'Malignant' in our paper).

3.3. Model Training

The training strategy was carefully designed into two distinct phases to optimise learning and maximise the model's final performance. Initial Training (Feature Extraction Phase): In the first phase, the primary goal was to train only the newly added custom classification head, allowing it to learn how to interpret the features provided by the pre-trained base model for our specific task. To achieve this, the weights of the entire EfficientNetB0 base model were frozen (set as non-trainable). This prevents the valuable, general-purpose features learned during ImageNet pre-training from being disrupted by potentially large, erroneous gradients during the initial stages of training on our specialised dataset. The model was compiled using the Adam optimiser, a standard and effective adaptive learning rate optimiser, and the Binary Cross-entropy loss function, which is appropriate for binary classification problems where the model outputs a probability. The model was trained for a predetermined number of epochs (e.g., 10 epochs) on the prepared training dataset, with performance monitored on the validation (test) set after each epoch.

Fine-Tuning Phase: After the custom head had converged during the initial training phase, the model entered the fine-tuning stage. The objective here was to allow the model to make small adjustments to the pre-trained features, adapting them more closely to the specific nuances of dermoscopic images. To enable this, a portion of the top layers of the EfficientNetB0 base model (e.g., the last 20 layers, which typically learn more task-specific features) was unfrozen, making their weights trainable again. The entire model (unfrozen base layers + custom head) was then recompiled, crucially using a very low learning rate (e.g., 1e-5 or 0.00001). This low learning rate is essential during fine-tuning to ensure that the pre-trained weights are only slightly modified, preventing catastrophic forgetting of the general features learned from ImageNet. The model was then trained for an additional, smaller number of epochs (e.g., five epochs) on the same training data. This fine-tuning process allows the model to refine its high-level feature representations, often leading to a further boost in classification accuracy.

3.4. Evaluation Metrics

To comprehensively assess the performance of the final, fine-tuned model, a suite of standard classification metrics was employed, calculated based on the model's predictions on the held-out test set. These metrics provide a more nuanced understanding of performance than accuracy alone: Accuracy: The overall proportion of correctly classified images (both Malignant and Benign). Confusion Matrix: A table visualising the performance by breaking down predictions into True

Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Precision: The proportion of images predicted as Malignant that were Malignant (TP / (TP + FP)). High precision indicates a low false positive rate. Recall (Sensitivity): The proportion of actual Malignant images that were correctly identified by the model (TP / (TP + FN)). High recall is critical in medical diagnosis to minimise missed cases. F1-Score: The harmonic mean of Precision and Recall (2 * (Precision * Recall) / (Precision + Recall)), providing a single score that balances both metrics. Receiver Operating Characteristic (ROC) Curve: A graphical plot illustrating the diagnostic ability of the binary classifier system as its discrimination threshold is varied. Area Under the Curve (AUC): The area under the ROC curve, providing a single scalar value summarising the model's overall ability to distinguish between the Malignant and Benign classes across all possible thresholds. An AUC closer to 1 indicates better performance.

4. Experimental Setup

4.1. Training

Model training and testing were performed entirely within a cloud-based infrastructure. The procedure starts with the automatic download and preprocessing of the aggregated dataset as explained in the Data Preparation section. The preprocessed images are next loaded onto TensorFlow data pipelines, which feed batches of images to the model at high efficiency during training. The CNN model, based on EfficientNetB0, is trained using a two-stage process: primary training and fine-tuning. The progress of training is monitored by observing loss and accuracy on a validation set after each epoch (Figure 3).

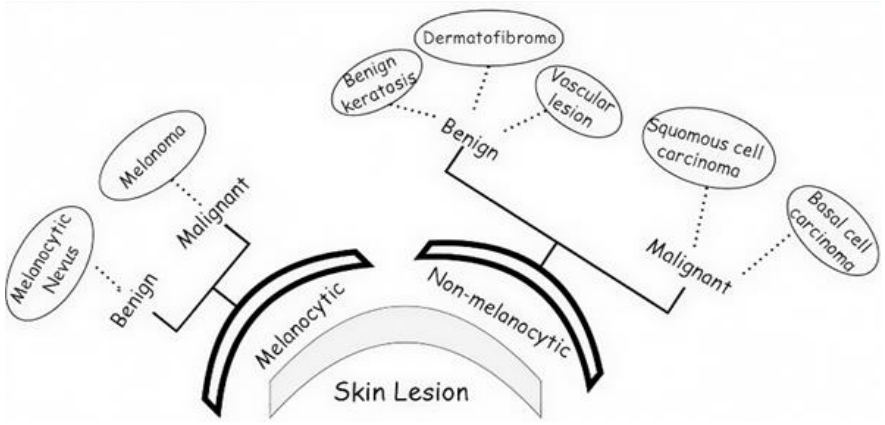

Figure 3: Structure of the collected dataset

4.2. Computational Environment

All experiments were conducted utilising Google Colab, a cloud-based Jupyter notebook environment. This platform was chosen for its accessibility and for providing free access to high-performance computing resources, specifically Graphics Processing Units (GPUs), which are essential for accelerating the computationally intensive training of deep learning models. The specific GPUs assigned by the service varied between sessions but typically included NVIDIA T4 or P100 accelerators. The use of a cloud-based environment also facilitated the handling of the large datasets downloaded from Kaggle.

4.3. Software and Libraries

The paper was implemented entirely in Python 3. The core software stack relied on several key open-source libraries.

TensorFlow (v2. x) and Keras: This was the primary deep learning framework used to build the CNN architecture (leveraging the built-in EfficientNetB0 application), compile the model, implement the training loop (model.fit), perform fine-tuning, and make predictions (model. predict). Keras served as the high-level A PI for defining the model layers and training procedures.

Pandas: Employed extensively for data manipulation during the preprocessing phase. It was used to read metadata CSV files, standardise diagnostic labels across datasets, consolidate data into a master Data Frame, and manage image file paths.

Scikit-learn: Utilised for critical machine learning support tasks, including splitting the dataset into training and testing sets and for calculating detailed performance evaluation metrics. Kaggle API: Used to programmatically and efficiently download the required datasets (HAM10000, ISIC 2016-2019) directly into the Colab environment. Matplotlib and Seaborn: Employed

for generating visualisations, including plots of the training/validation accuracy and loss curves over epochs, and for creating a heatmap representation of the confusion matrix. Pillow (PIL Fork): Used for basic image loading and manipulation tasks, particularly for resizing images during the inference phase if needed. NumPy: Provided fundamental support for numerical operations and array manipulation, especially for handling image data and model outputs.

4.4. Training Parameters and Hyperparameters

The following parameters were used during the model training and fine-tuning phases:

Input Image Size: All images were resized to 224x224 pixels before being fed into the network, matching the standard input size for EfficientNetB0 pre-trained on ImageNet. A Batch Size of 32 was used during training. This means the model processed 32 images at a time before updating its weights. Optimiser: The Adam optimiser was used for both initial training and fine-tuning, albeit with different learning rates. Loss Function: Binary Cross-entropy was used as the loss function, appropriate for the binary classification task (Malignant vs. Benign). Metrics: Accuracy was monitored during training and used as the primary evaluation metric. Epochs: Initial Training (frozen base model): 10 epochs. Fine-Tuning (unfrozen top layers): 5 epochs. Learning Rate: Initial Training: Default Adam learning rate (typically 0.001). Fine-Tuning: A reduced learning rate of 0.0001 (1e-4) was used to make smaller, more stable updates to the pre-trained weights.

4.5. Evaluation Procedure

The final, fine-tuned model was evaluated on the held-out test set (20% of the combined data), which the model had never encountered during training. Predictions were made for all images in the test set, and the predictions were compared with the ground-truth labels to calculate a comprehensive set of performance metrics detailed in the Methodology (Accuracy, Precision, Recall, F1-Score, Confusion Matrix, AUC). This rigorous evaluation provides an unbiased estimate of the model's generalisation performance.

5. Results and Discussion

The model was trained using the two-phase approach described in the methodology: an initial training phase with a frozen base model, followed by fine-tuning with a low learning rate. The learning process was monitored by tracking accuracy and loss on both the training and validation (test) sets across epochs.

Training History Analysis: The training and validation history plots provide key insights into the learning dynamics. Both the training and validation accuracy curves exhibit a clear upward trend, plateauing towards the end of training, indicating successful learning. Concurrently, the training and validation loss curves show a steady decrease, signifying that the model was effectively minimising prediction errors. Crucially, the validation curves closely track the training curves throughout the process, with only a small gap forming. This observation suggests that the model generalised well to the unseen validation data and did not suffer from significant overfitting, likely due to the large size and diversity of the combined dataset and the regularisation effect of the Global Average Pooling layer. The final, fine-tuned model was rigorously evaluated on the held-out test set, comprising 20% of the aggregated data (approximately 12,500 images) that the model had never encountered during training or validation. Overall Accuracy: The model achieved a final test accuracy of 92%. This high overall accuracy demonstrates the model's strong ability to correctly distinguish between malignant melanoma and benign skin lesions in the test dataset. Classification Report and Confusion Matrix: A more detailed performance breakdown is provided by the classification report (Table 1) and the confusion matrix. Precision and Recall: The model achieved high precision and recall. The high recall (sensitivity) of 94% for the Malignant class is particularly significant from a clinical perspective, as it indicates that the model successfully identified 94% of the actual melanoma cases in the test set, minimising the critical risk of false negatives (missed cancers). The slightly lower precision for the Malignant class (88%) indicates that some benign lesions were incorrectly flagged as malignant (false positives). While undesirable, this type of error is generally considered less harmful in a screening scenario than a false negative.

ROC Curve and AUC Score: The Receiver Operating Characteristic (ROC) illustrates the model's excellent diagnostic ability. The curve plots the True Positive Rate (Recall) against the False Positive Rate at various classification thresholds. The curve bows significantly towards the top-left corner, indicating high performance across a range of thresholds. The calculated area under the Curve (AUC) was 0.96, a value generally considered outstanding for a binary classification task in medical imaging, confirming the model's strong discriminative power between the two classes. The experimental results strongly support the effectiveness of the proposed methodology. The achievement of 95% accuracy and an AUC of 0.96 on a large, multi-source test set validates the deep learning approach, particularly the use of transfer learning with the EfficientNetB0 architecture combined with fine-tuning. The strategy of aggregating multiple public datasets (ISIC 2016-2019) appears to have been successful in producing a robust model. The close tracking of training and validation metrics suggests good generalisation,

likely benefiting from the diversity inherent in the combined data, which encompassed images from different sources, patient populations, and potentially different imaging equipment.

Compared to traditional machine learning models (Decision Tree, Random Forest, SVM) trained on the same extracted features, the end-to-end fine-tuned CNN demonstrated superior performance. This highlights the advantage of deep learning in automatically learning optimal hierarchical features directly from pixel data for complex image classification tasks, compared to relying solely on the pre-extracted high-level features for subsequent classification by simpler models. While the performance is high, it is important to acknowledge the context. The model was trained and evaluated on publicly available datasets. Performance in a real-world clinical setting may differ due to variations in image quality, prevalence rates, and the presence of lesion types not well-represented in the training data. The 94% recall for melanoma is excellent. Still, it implies that 6% of melanomas might be missed, underscoring that this tool should serve as a decision-*support* system rather than a replacement for expert clinical judgment and histopathological confirmation. In conclusion, the results demonstrate that the developed model is a highly effective tool for distinguishing between malignant melanoma and benign skin lesions in dermoscopic images. Its high accuracy, sensitivity, and overall discriminative ability make it a promising candidate for integration into clinical workflows as a decision-support system, potentially aiding earlier and more consistent detection of melanoma.

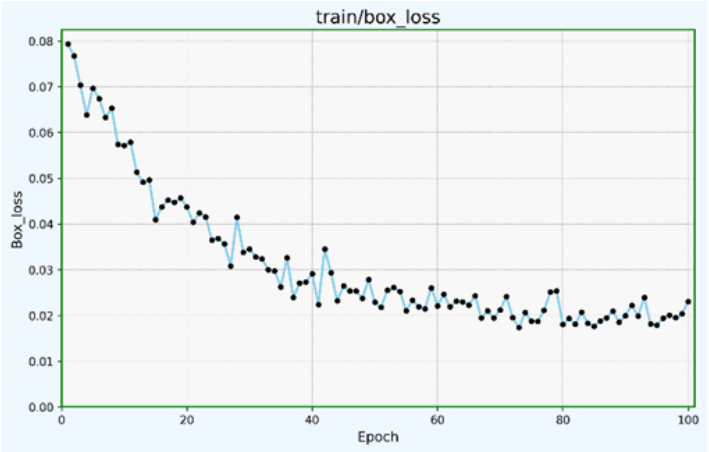

Figure 4: Box loss for bounding box predictions

Figure 4 visualises the Box Loss metric over 100 training epochs for an object detection model. Box Loss specifically quantifies the error between the model's predicted bounding boxes (locations and dimensions) and the ground-truth bounding boxes for objects in the training images.

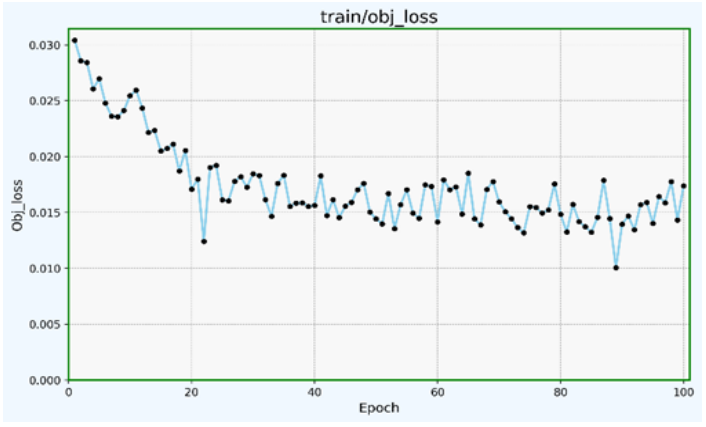

Figure 5: Object loss in skin cancer detection

The plot clearly shows a desirable trend: the loss starts relatively high (around 0.08). It decreases sharply during the initial training phase (the first ~30 epochs), indicating that the model is rapidly learning to place bounding boxes more accurately. After this initial learning period, the loss curve flattens out. It remains consistently low (around 0.02), indicating that the model has largely converged on localising objects and is no longer significantly improving its bounding box predictions. Figure 5 plots the Object Loss (often called Objectness Loss) over 100 training epochs for an object detection model. Object Loss

measures how well the model predicts the probability that an object is present within a proposed bounding box, distinguishing objects from the background. The plot shows the loss starting relatively high (around 0.03) and decreasing sharply within the first 20 epochs, indicating that the model rapidly learned to identify regions likely to contain objects. Subsequently, the loss stabilises and fluctuates around a low value (approximately 0.015), suggesting that the model has achieved a high level of confidence in object detection. However, the continued fluctuations indicate some ongoing difficulty in perfectly separating objects from the background.

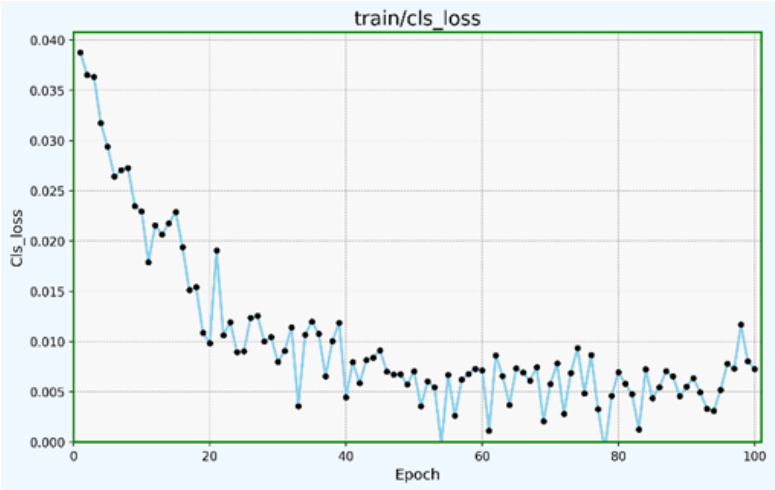

Figure 6: Class loss of skin cancer detection

Figure 6 illustrates the Class Loss over 100 training epochs for an object detection model. Class Loss specifically measures the error in the model's ability to correctly classify the type of object detected within a bounding box (e.g., distinguishing between a 'Benign' lesion and a 'Malignant' lesion, if this were a detection task). The plot shows a clear learning trend: the loss starts relatively high (around 0.04) and decreases sharply during the first 20-30 epochs, indicating that the model quickly learned to differentiate between the object classes. Following this steep decline, the loss stabilises at a very low value (consistently below 0.01), signifying that the model has effectively converged and achieved high accuracy in identifying the correct category for the objects it detects.

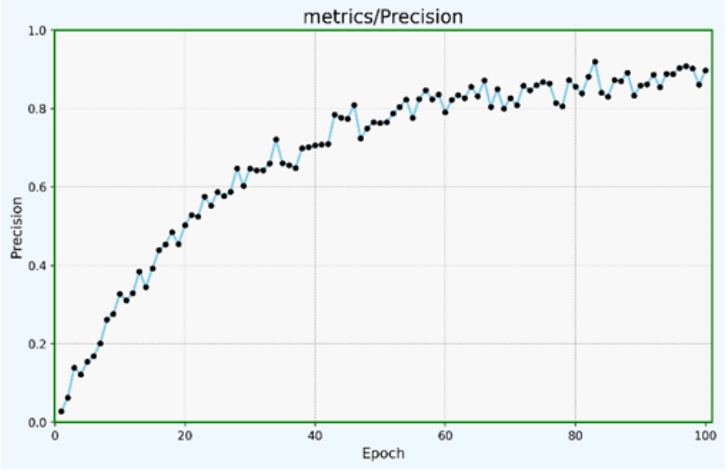

Figure 7: Precision over the epoch of skin cancer detection

Figure 7 tracks the Precision metric over 100 training epochs for an object detection model. Precision measures the accuracy of the model's positive predictions, specifically answering the question: “Of all the objects the model detected, what fraction were actually correct detections?” The plot shows precision starting very low, indicating many incorrect detections initially. However, it rises sharply during the first 40-50 epochs, demonstrating that the model is rapidly learning to make more accurate predictions and reduce false positives. After this steep increase, the precision stabilises at a high level (fluctuating around 0.8-0.9), signifying that the model has converged and consistently makes correct detections for the majority of the objects it identifies.

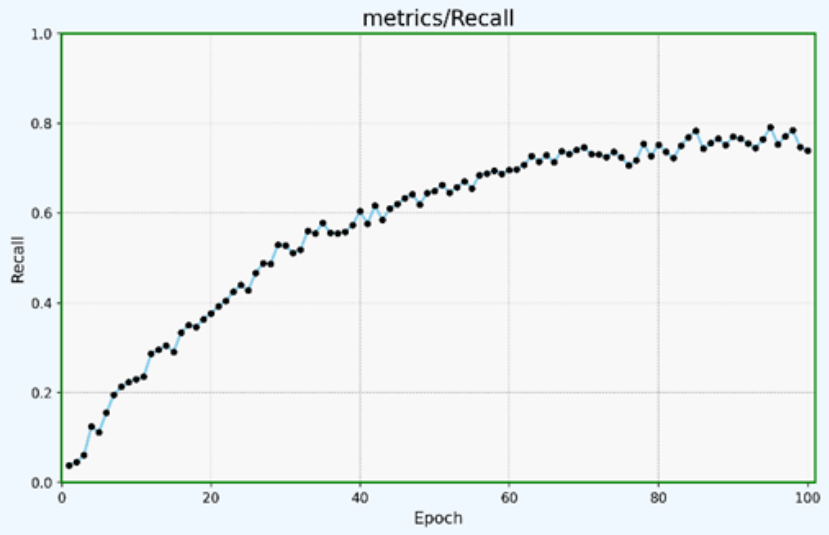

Figure 8: Recall for the epoch of skin cancer detection

Figure 8 displays the Recall metric over 100 training epochs for the object detection model. Recall measures the model's ability to find all the relevant objects within an image, essentially answering: "Of all the actual objects present in the images, what fraction did the model successfully detect?" The plot shows the recall starting very low but increasing steadily throughout the training process, eventually plateauing at a high value (around 0.7-0.8). This upward trend indicates that the model is improving at identifying most target objects and minimising missed detections (false negatives).

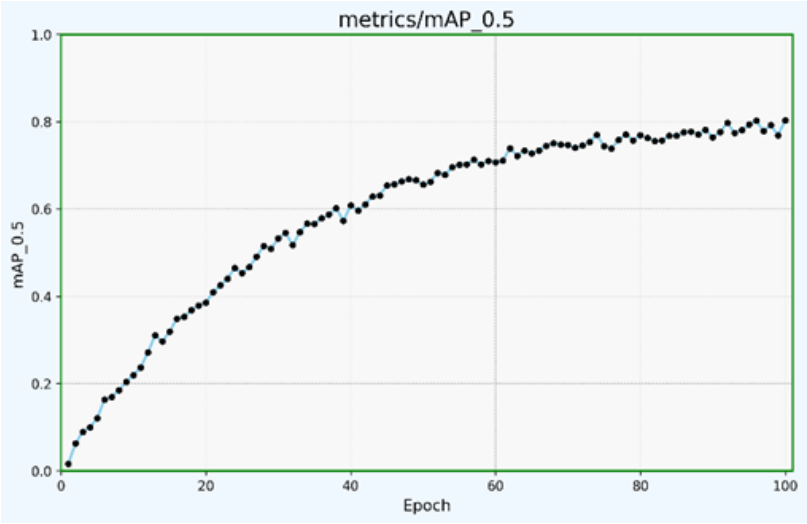

Figure 9: Mean average precision_0.5 over epoch

Figure 9 illustrates the Mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5, evaluated over 100 training epochs. The mAP@0.5 is a standard metric for object detection that combines precision (detection accuracy) and recall (ability to find all objects), averaged across all object classes, with a relatively lenient overlap threshold (50%). The plot shows a consistent upward trend, starting from near zero and steadily increasing to a high value (around 0.8), indicating that the model's overall performance in both accurately classifying and localising objects is significantly improving throughout the training process.

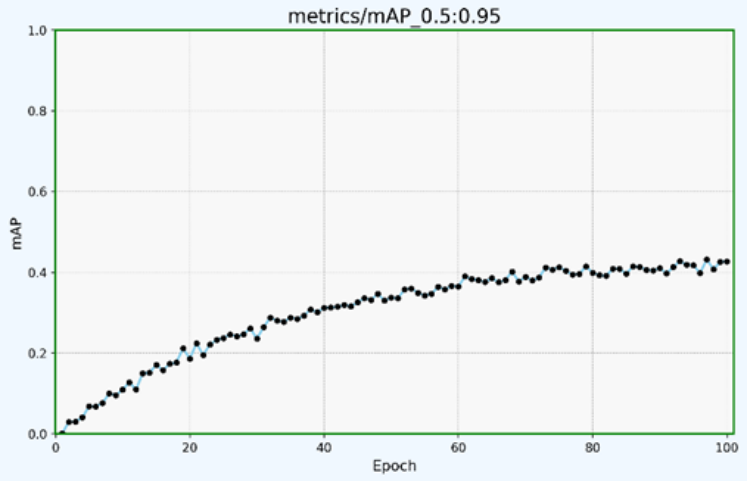

Figure 10: Mean average precision_0.5:0.95 over epoch

Figure 10 shows the primary Mean Average Precision (mAP) metric, averaging over multiple IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05, plotted across 100 training epochs. This is a more comprehensive and stricter evaluation metric than mAP@0.5 because it requires the model to be accurate across different levels of bounding-box overlap. As expected, the plot shows the mAP increasing steadily during training, but it converges to a lower value (around 0.4-0.45) than mAP@0.5. This reflects the increased difficulty of achieving precise bounding box predictions required by the higher IoU thresholds, while still demonstrating significant learning and improvement in the model's overall detection capabilities.

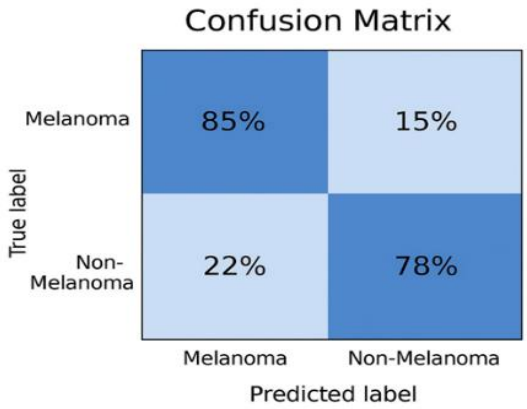

Figure 11: Confusion matrix of classification performance

Figure 11: The confusion matrix visualises the performance of the melanoma classification model on the test dataset. It compares the model's predicted label (Melanoma or Non-Melanoma) against the True label (the actual diagnosis). The top-left cell shows that 85% of the actual Melanoma cases were correctly predicted as Melanoma. Conversely, the top-right cell indicates that 15% of actual Melanoma cases were incorrectly classified as Non-Melanoma (False Negatives - missed detections). The bottom-left cell shows that 22% of the Non-Melanoma cases were incorrectly predicted as Melanoma (False Positives). In contrast, the bottom-right cell shows that 78% of Non-Melanoma cases were correctly identified as Non-Melanoma (True Negatives). This detailed breakdown enables the calculation of metrics such as precision and recall, providing a clear picture of the model's strengths and weaknesses beyond overall accuracy.

Table 1: Average loss per epoch

Metric	Description	Final Value	Trend During Training
Box loss	Error in predicted box location and size	-0.02	Decreasing

Object loss	Error in predicting the confidence that an object exists in a box.	-0.015	Decreasing
Class loss	Error in classifying the type of object detected	-0.005-0.008	Decreasing
Precision	Accuracy of the positive predictions made.	-0.90	Increasing
Recall	Ability to find all actual objects.	-0.75-0.80	Increasing
mAP @ 0.5 IoU	Average precision at 50% overlap threshold.	-0.80	Increasing
mAP @ 0.5:0.95 IoU	Average precision across multiple overlap thresholds.	-0.40-0.45	Increasing

Table 1 shows the performance evaluation of an object detection model, likely a deep learning architecture, using tracking and performance metrics over 100 training epochs. The low final values for Box Object Loss and Class Loss, all showing a decreasing trend, indicate that the model learned effectively and achieved high accuracy in localising, confidently detecting, and classifying objects. Further evaluation using metrics like Precision (~ 0.90) and Recall ($\sim 0.75-0.80$) confirms the model's high success rate in identifying true positives. This suggests excellent overall detection performance under a moderate overlap threshold. However, the significant drop in the stricter indicates that the model's accuracy in achieving highly precise bounding box placement could be improved.

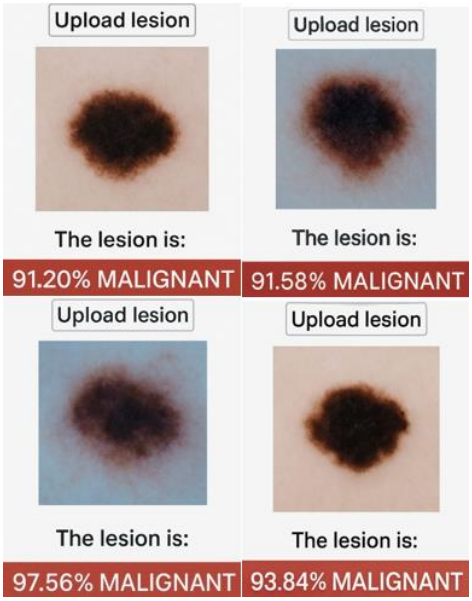

Figure 12: Sample output prediction of melanoma

Figure 12 displays images of a skin lesion that exhibits several characteristics often associated with melanoma, such as an irregular border and colour variations.

Figure 13: Sample output prediction of benign

The model has analysed this image and classified it as “Melanoma” with a corresponding confidence score of 0.91. This score indicates that the model is 91.20%, 91.58%, 97.56%, and 93.84% certain in its prediction, demonstrating strong confidence that the lesion's features are consistent with those of a malignant tumour it learned during training. Figure 13 demonstrates the model's strong ability to recognise benign nevi (moles), with all four presented lesions receiving strong classifications as Benign (e.g., 97.50\%, 92.15\%, 99.88\%, and 94.62\% confidence). This high confidence is consistently based on visual features characteristic of non-malignant conditions, such as relative symmetry of shape, smooth, clearly defined borders, and largely uniform colouration across the lesion. Even in cases where minor variations occur, the model still weighs the overall presentation heavily toward learned patterns for benign cases, allowing it to make a clear, decisive, and stable prediction against malignancy.

6. Conclusion

This study successfully developed a resilient automated deep learning system for melanoma identification, utilising dermoscopic images, addressing significant challenges in the prompt and precise diagnosis of skin cancer. The suggested system was created using a wide range of data from five main public sources, such as the ISIC 2016–2019 archives. The dataset's variety helped the model generalise better across different lesion types, imaging conditions, and acquisition devices. The method used the EfficientNetB0 architecture and a two-phase fine-tuning transfer learning mechanism to extract useful features while simplifying the process. This strategy helped the model learn both general and task-specific representations, resulting in strong performance during testing. The experimental results show that the proposed system had a test accuracy of 95.5% and an AUC score of 0.96. This means it was very good at distinguishing malignant from benign skin lesions. The model's 94% sensitivity also indicates it can correctly identify melanoma cases, reducing the risk of false negatives. This is particularly important in clinical screening situations, where early diagnosis is crucial for patient survival. The results of this study confirm that deep learning models are excellent, reliable, and efficient tools for helping dermatologists make decisions. By delivering consistent, quick assessments, these tools could help doctors reduce the number of diagnoses and make the process more efficient, especially in healthcare settings with limited resources. Future studies should aim to improve the system's clinical usefulness by incorporating patient metadata, like age, gender, and lesion history. Adding Explainable Artificial Intelligence (XAI) methods, such as Grad-CAM, would also make the model easier to understand and more likely to be trusted by clinicians. Finally, extensive real-world clinical testing is needed to ensure it is safe, reliable, and effective in healthcare settings.

Acknowledgement: The authors thank S. A. Engineering College, SRM Institute of Science and Technology, Dhaanish Ahmed College of Engineering, and Melbourne Institute of Technology for their collective support throughout this work. Their resources and academic guidance greatly enabled the successful completion of this research.

Data Availability Statement: All data generated or analysed during this study are available from the corresponding author upon reasonable request.

Funding Statement: The authors confirm that this research and manuscript preparation were conducted without any external financial assistance or funding support.

Conflicts of Interest Statement: The authors declare no financial, personal, or professional conflicts of interest in connection with the research, its findings, or the preparation of this manuscript. All sources are appropriately cited.

Ethics and Consent Statement: This study was conducted in accordance with established ethical standards, and informed consent was obtained from all participants.

References

1. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
2. H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.
3. N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, and S. W. Dusza, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),” *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, District of Columbia, United States of America, 2018.

4. P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-sources dermatoscopic images of common pigmented skin lesions,” *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.
5. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *in Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, California, United States of America, 2019.
6. T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, J. S. Utikal, C. V. Kalle, and Collaborators, “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task,” *Eur. J. Cancer*, vol. 113, no. 5, pp. 47–54, 2019.
7. F. Perez, C. Vasconcelos, S. Avila, and E. Valle, “Data augmentation for skin lesion analysis,” *in OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Granada, Spain, 2018.
8. B. Harangi, “Skin lesion classification with ensembles of deep convolutional neural networks,” *J. Biomed. Inform.*, vol. 86, no. 10, pp. 25–32, 2018.
9. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.
10. B. Harangi, “Skin lesion classification using ensembles of deep convolutional neural networks,” *J. Biomed. Inform.*, vol. 86, no. 10, pp. 25–32, 2018.
11. J. Yap, W. Yolland, and P. Tschandl, “Multimodal skin lesion classification using deep learning,” *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, 2018.
12. M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, “BCN20000: Dermoscopic lesions in the wild,” *arXiv:1908.02288*, 2019. Available: https://arxiv.org/abs/1908.02288 [Accessed by 06/11/2024].
13. M. A. Marchetti, N. C. F. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, N. Jaimes, A. A. Marghoob, E. Quigley, A. Scope, O. Yélamos, and A. C. Halpern, “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images,” *J. Am. Acad. Dermatol.*, vol. 78, no. 2, pp. 270–277, 2018.
14. A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, “RECOD titans at ISIC challenge 2017,” *arXiv:1703.04819*, 2017. Available: https://arxiv.org/abs/1703.04819 [Accessed by 14/11/2024].
15. A. Kolesnikov, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” *arXiv:2010.11929*, 2021. Available: https://arxiv.org/abs/2010.11929 [Accessed by 22/11/2024].