Enhancing Skin Cancer Detection Using Curriculum Learning In Ensemble Deep Learning Context

Riadh Meghatria[1,*,†], Djamel Gaceb[1,*,†], Fayçal Touazi[1,*,†], Tina Boukert[1,†] and Sarah Ben Aidrene[1,†]

[1]*LIMOSE laboratory, CS department, University M'hamed Bougara, Boumerdes, Algeria*

Abstract

Accurate classification of skin lesions, particularly for melanoma detection, remains a critical challenge in medical image analysis. Leveraging recent advances in deep learning, this paper investigates the use of curriculum learning in ensemble deep learning context for melanoma classification. To validate the proposition, three primary strategies are compared: transfer learning of CNNs using VGG16, ResNet50, and EfficientNetB0 models; ensemble learning techniques such as bagging; and curriculum learning that progressively guides training in increasing order of complexity. Experiments conducted on the ISIC 2019 and 2020 dermoscopic image datasets demonstrate that curriculum learning applied to EfficientNetB0 achieves superior classification performance, reaching an F1- score of 90.77%, outperforming conventional fine-tuning and ensemble approaches. These results underscore the potential of integrating curriculum learning in ensemble learning context with state-of-the-art CNN architectures to improve the robustness and accuracy of automated melanoma diagnosis.

Keywords

Skin Lesions, Melanoma Detection, Deep Transfer Learning, Ensemble Deep Learning, Medical Image Analysis, Medical Diagnosis, Curriculum Learning

1. Introduction

The accurate and early diagnosis of skin cancer, particularly melanoma, is a critical public health challenge due to its aggressive nature and potential for metastasis if undetected. Traditional manual interpretation of dermatological images by clinicians is a complex, time-consuming task prone to inter-observer variability, misdiagnosis, or delayed treatment [1]. This inherent difficulty arises from the subtle visual similarities between benign and malignant lesions, significant intra-class variability, and the presence of confounding image artifacts. Consequently, there is an urgent need for robust, automated diagnostic tools to augment clinical decision-making.

Artificial Intelligence (AI), and specifically deep learning (DL), has emerged as a transformative force in medical imaging, offering unprecedented opportunities to enhance diagnostic efficiency and accuracy [2]. Deep Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance across numerous medical image analysis tasks, including detection, segmentation, and classification [3]. Their remarkable ability to automatically learn hierarchical, discriminative features directly from raw image data spares the labor-intensive process of manual feature engineering, making them particularly well-suited for intricate diagnostic problems like skin lesion classification [4]. Furthermore, transfer learning (TL), which involves fine-tuning or reusing features from models pre-trained on large-scale datasets such as ImageNet, has proven highly effective in medical contexts where annotated datasets are typically scarce [5]. Architectures like VGG16, ResNet50, InceptionV3, EfficientNetB0, and Xception have been widely adopted through TL to achieve commendable performance in dermatological applications.

ProfIT AI'25: 5th International Workshop of IT-professionals on Artificial Intelligence, October 15–17, 2025, Liverpool, UK

*Corresponding author.

†These authors contributed equally.

r.meghatria@univ-boumerdes.dz (R. Meghatria); d.gaceb@univ-boumerdes.dz (D. Gaceb); f.touazi@univ-boumerdes.dz (F. Touazi); boukerttina@gmail.com (T. Boukert); sarahbenaidrene@gmail.com (S. B. Aidrene)

0000-0002-4343-4349 (R. Meghatria); 0000-0002-6178-0608 (D. Gaceb); 0000-0001-5949-5421 (F. Touazi); 0009-0008-4849-5897 (T. Boukert); 0009-0001-6814-4225 (S. B. Aidrene)

Despite these advancements, several challenges persist in deploying DL models for reliable clinical diagnosis. Models frequently struggle with pronounced class imbalance in medical datasets, high intra- and inter-class variability of lesions, and the presence of ambiguous or atypical cases that lead to uncertain predictions [6]. While ensemble learning methods [7], which aggregate predictions from multiple models, can mitigate these issues by improving predictive stability and leveraging model diversity [8], they do not address the fundamental learning process on challenging samples. To foster more robust representation learning and address diagnostic uncertainty, curriculum learning (CL) has attracted increasing attention [9]. Inspired by human cognitive development, CL involves structuring the training process by progressively introducing data samples from simple to complex, thereby guiding the model's optimization and enhancing its generalization capabilities [10].

This work systematically investigates advanced deep learning strategies to enhance the automated classification of skin lesions, with a particular emphasis on demonstrating the efficiency of curriculum learning in this challenging domain. Our core objective is to show that a confidence-based curriculum used in ensemble deep learning context, when applied to powerful pre-trained CNNs, can surpass the performance of both conventional TL and ensemble methods for melanoma detection. The main contributions of this paper are summarized as follows:

- A comprehensive empirical evaluation of five pre-trained CNN architectures (VGG16, ResNet50, InceptionV3, EfficientNetB0, and Xception) under two distinct transfer learning modes: feature extraction and fine-tuning.
- The implementation and assessment of a bagging-based ensemble approach applied to the EfficientNetB0 architecture, designed to enhance predictive stability and robustness for skin lesion classification.
- The development of a novel curriculum-learning framework in ensemble deep learning context, where training data is dynamically ordered by diagnostic difficulty (derived from initial classifier confidence), enabling a progressive and more effective model adaptation process.
- A rigorous experimental study conducted on a two-class dataset derived from the ISIC 2019 and ISIC 2020 challenges, providing a robust benchmark for binary classification of melanoma versus nevus.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in automated skin lesion classification. Details of the proposed methods are then provided in section 3, including the transfer learning strategies, the ensemble approach, and the curriculum learning approach applied in ensemble deep learning context. In section 4, the experimental results are presented and discussed, comparing the performance of all implemented approaches. Finally, Section 5 concludes the paper and outlines promising directions for future research.

2. Related Work

In line with numerous research fields and applications influenced by AI revolution over the past decade, the automatic detection and classification of skin cancer from medical images has also witnessed significant advancements. In this context, several automated systems using AI models and techniques have been introduced to facilitate early non-invasive diagnosis. This section provides a review of prior work and existing methodologies for skin cancer detection and classification, spanning from traditional machine learning techniques to state-of-the-art deep learning and ensemble approaches, highlighting their evolution, key contributions, and limitations.

2.1. Machine Learning Approaches

Early efforts in automated skin cancer detection primarily used traditional machine learning algorithms, which typically relied on handcrafted features (such as texture, color, and shape descriptors) extracted from images, followed by classification (like SVM, KNN, and decision trees). While providing preliminary

solutions, their main limitation was the inability to automatically capture the complex discriminative features inherent in dermatoscopic images, often requiring tedious expert intervention.

An early work by Codella et al. (2015) [11] proposed a method that combined deep learning with sparse coding and SVM for melanoma recognition. Aiming to reduce the reliance on extensive annotated data, the authors tried to benefit from unsupervised learning and transferring features from natural photographs, mimicking expert clinical reasoning. Evaluated on 2,624 clinical cases, their method achieved 93.1% precision for distinguishing melanoma from non-melanoma lesions. More recently, Wei et al. (2024) [12] explored the efficacy of classical machine learning algorithms for early skin cancer detection. They applied SVM, KNN, and Decision Tree algorithms on a large ISIC dataset (53,177 images across seven classes). Their results indicated that KNN achieved the highest multiclass precision at 74.93%, while SVM offered better binary classification with an AUC of 0.676. Despite their interpretability and lower computational demands, these methods often require expert-driven feature engineering and struggle to capture complex image patterns, limiting their diagnostic accuracy and scalability. Consequently, these approaches remain valuable as baseline references but are increasingly being outperformed by deep learning models capable of autonomously learning highly discriminative features.

2.2. Deep Learning Approaches

The significant potential of CNNs was underscored by Esteva et al (2017) [4] who demonstrated that these models could achieve classification capabilities comparable to expert dermatologists. Utilizing a dataset of 129,450 clinical and dermoscopic images, their approach employed a GoogLeNet Inception v3 model with transfer learning, reaching 72.1% of precision in disease classification and showing superior performance in specific test cases involving carcinoma, melanoma, and dermoscopy. Following this, Demir et al. (2019) [13] evaluated the performance of two prominent deep CNN architectures, ResNet-101 and Inception-v3, for binary (benign/malignant) skin cancer diagnosis. Their study, conducted on an ISIC-based dataset, reported that Inception-v3 achieved 87.42% precision, slightly outperforming ResNet-101 (84.09%). Expanding on the utility of transfer learning, Laith Alzubaidi et al. (2021)[5] proposed leveraging deep neural network models to enhance skin cancer detection. Their hybrid DCNN model, trained on the HAM10000 dataset, attained a high precision of 98.53% and an F1-score of 89.09%, demonstrating the effectiveness of data augmentation in improving model performance. Further, Aljohani and Turki (2022) [14] conducted a comparative study of eight different CNN architectures, including DenseNet201, MobileNetV2, ResNet50V2, and GoogleNet, for binary melanoma classification using the ISIC 2019 dataset. GoogleNet exhibited the best test performance with a precision of 76.08%, highlighting the varying effectiveness of different established CNN architectures. These CNN-based methods significantly outperform traditional machine learning by capturing hierarchical image features but often require large annotated datasets and substantial computational resources. Moreover, despite these improvements, standard CNNs effectiveness can sometimes be limited by their ability to capture global contextual information and subtle features in complex dermatoscopic images.

2.3. Enhanced and Hybrid Architectures Approaches

To address the limitations of standard CNNs, researchers have introduced some enhanced and hybrid deep learning architectures. These models often combine the strengths of different network types or incorporate task-specific mechanisms. Catal Reis et al. (2022) [15] introduced InSiNet, a CNN adapted for skin cancer classification, achieving up to 94.59% precision on ISIC datasets and outperforming conventional deep learning and traditional machine learning methods. In a comparable way, Shah et al. (2024) [16] integrated explainable AI with Xception for early melanoma detection, attaining 98.5% precision on ISIC 2018 and enhancing interpretability via Grad-CAM and LIME. Building on the idea of specialized components, Kavitha et al. (2024)[3] proposed a CNN-R-CNN hybrid approach for multi-class classification of nine skin cancer types, combining preprocessing and data augmentation to reach 91.32% of precision and 76,92% of F1-score on ISIC images. More advanced hybrid models, such as MetaFormer

introduced by Pacal et al. (2025)[17], merging CNNs with Vision Transformers (ViT) to leverage both local feature extraction and global contextual understanding. MetaFormer, featuring a Focal Self-Attention mechanism, delivered remarkable performance with 92.54% of precision on ISIC 2019 and 95.01% on HAM10000, while maintaining a low parameter count for mobile deployment. Similarly, Ozdemir and Pacal (2025) [10] developed an innovative hybrid deep learning model for multiclass skin cancer classification, integrating ConvNeXtV2 blocks with separable attention mechanisms. Trained on ISIC 2019, this model achieved 93.48% of precision and 91.82% of F1-score, surpassing many existing CNN and ViT-based models. Despite their accuracy gains by capturing complex features, such hybrid architectures, often increase computational demands and pose interpretability challenges.

2.4. Ensemble Methods and Advanced Optimization

To address issues like model bias, variance, and limited generalizability, ensemble methods were explored where predictions of multiple individual models aggregated to produce a more robust and accurate final decision. These techniques frequently outperforming any single model. In fact, numerous studies, in the context of medical image classification, have demonstrated the efficiency of such techniques. For instance, Khaled et al. (2023) [18] proposed a stacking framework that combines multiple CNN architectures to improve classification accuracy across diverse medical imaging modalities and anatomical regions. In a related effort, for brain lesion detection, Laribi et al. (2024) [19] demonstrated that ensembles combining CNNs and ViTs achieved significant improvements. More recently, Khaled et al. (2025) [20] integrated progressive transfer learning with ensemble methods to achieve significant enhancements in mammogram-based breast cancer diagnosis. In the context of skin lesion analysis, ensembles techniques have likewise shown strong potential. Rahman et al. (2021) [21] demonstrated substantial gains, improving recall to 94% by aggregating five diverse CNN architectures on HAM10000 and ISIC datasets for classifying seven types of skin lesions. Bassel et al. (2022) [22] extended this with multi-level stacking of CNN-derived features and traditional classifiers (SVM, Random Forest, KNN, and logistic regression). This yielded 90.9% precision and F1-score, highlighting the benefit of cross-paradigm integration, albeit at increased computational cost. Moreover, subsequent work has focused on enhancing ensemble pipelines through data enrichment and optimization. Chang et al. (2022)[23] combined MELA-CNN image features with clinical metadata (age, sex) and applied K-means SMOTE to address class imbalance, achieving an F1-score of 86.1% and an AUC of 0.970 with XGBoost. Shortly after, Thanka et al. (2023)[24] integrated VGG16-based features with XGBoost and LightGBM classifiers, leveraging GAN-generated synthetic samples to balance the ISIC data and reach 99.1% of precision and 99.4% of recall, though reliance on synthetic data may risk overfitting. Natha et al. (2024)[25] further advanced ensemble strategies by combining Max Voting across pre-trained ensemble models (Random Forest, Gradient Boosting) with Genetic Algorithm-based feature selection, achieving 95.80% of precision and 95.20% of F1-score on ISIC 2018 and HAM10000, while reducing overfitting and improving generalization. Ghosh et al. (2024)[8] extended the feature diversity concept by combining deep learning embeddings from VGG-19, Capsule Network, and ViT with multiple machine learning classifiers through majority voting, achieving 91.6% precision. In addition, Gamil et al. (2024) [26] integrated EfficientNet-B0 features with PCA and an AdaBoost and SVM ensemble, attaining 93% precision on DermIS and 91% on ISIC, indicating strong diagnostic potential despite persistent interpretability challenges. While ensemble methods significantly boost robustness and accuracy, they often come with increased computational complexity and can reduce the overall interpretability of the diagnostic decision.

3. Methodology

In this work, we investigate a range of deep learning strategies for the binary classification of skin lesions, specifically differentiating melanoma from nevus. The methodological framework is designed to enhance both predictive accuracy and model robustness, leveraging advanced CNN architectures and contemporary training paradigms. For this, we structured the study around three approaches: (i)

transfer learning, (ii) ensemble deep learning via bagging, and (iii) curriculum learning used in ensemble deep learning context. A comprehensive description of the dataset is first provided, followed by a detailed explanation of each proposed approach.

3.1. Dataset Description

Our experiments were conducted on the ISIC 2019 and 2020 Melanoma Dataset [27], a fusion of the ISIC 2019 [28, 29, 30] and ISIC 2020 [31] challenges. The ISIC 2019 comprises over 25,000 dermoscopic images across nine lesion categories, including 4,522 melanoma cases, while the ISIC 2020 dataset contains 33,126 images across two classes with only 584 melanoma cases. For our study, only two diagnostic categories of high clinical relevance were considered, namely melanoma (MEL) and nevus (NV) resulting in 11.449 images (5,106 MEL, 6,343 NV). To mitigate class imbalance, we applied random under-sampling of the majority class. Figure 1 presents representative example images from each class.

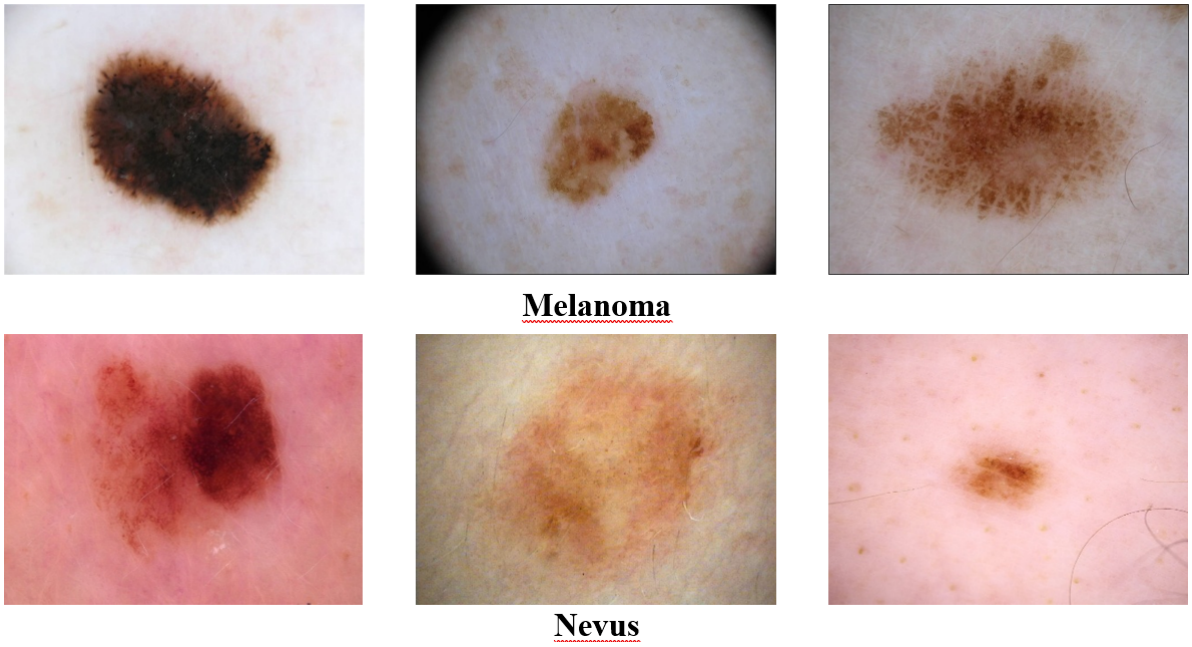

Figure 1: Representative image samples of the two-class dataset.

3.2. Transfer Learning Approach

Given their documented success in medical image analysis, five CNN architectures (VGG16, InceptionV3, Xception, ResNet50, and EfficientNetB0) were employed as backbone models in this approach. To leverage the advantages of transfer learning, each model was initialized with weights pre-trained on the ImageNet dataset. Two transfer learning strategies were explored, resulting in two distinct experimental scenarios for this approach: feature extraction and fine-tuning.

In the feature extraction scenario, as illustrated in Figure 2, the convolutional backbone was frozen, and only the classification layers were trained, thereby utilizing the pretrained weights exclusively for feature representation while adapting the classifier to the target task. Conversely, in the fine-tuning scenario (illustrated in figure 3), a subset of the upper convolutional layers was unfrozen and jointly optimized with the classification head, enabling the adaptation of high-level feature representations to the specific characteristics of the dermatological domain. This approach seeks to achieve an optimal trade-off between the generalization benefits of pretrained representations and the specialization required for accurate skin lesion classification.

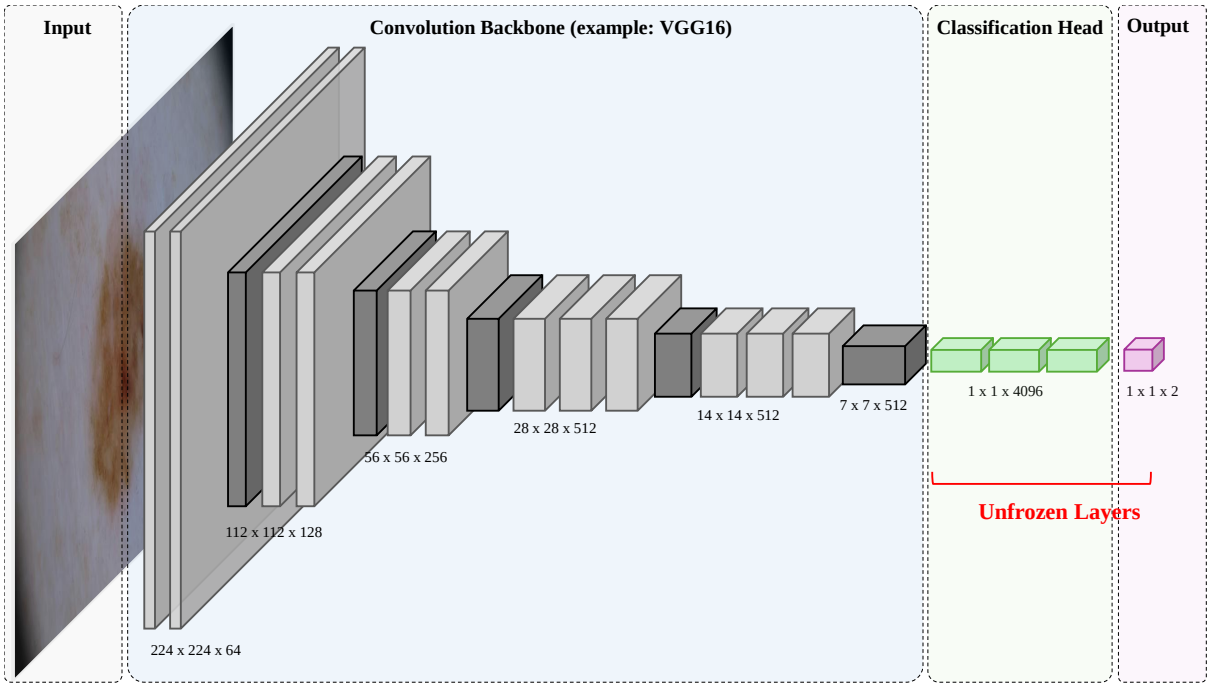

Figure 2: Transfer Learning Approach: Scenario 1 (Feature Extraction).

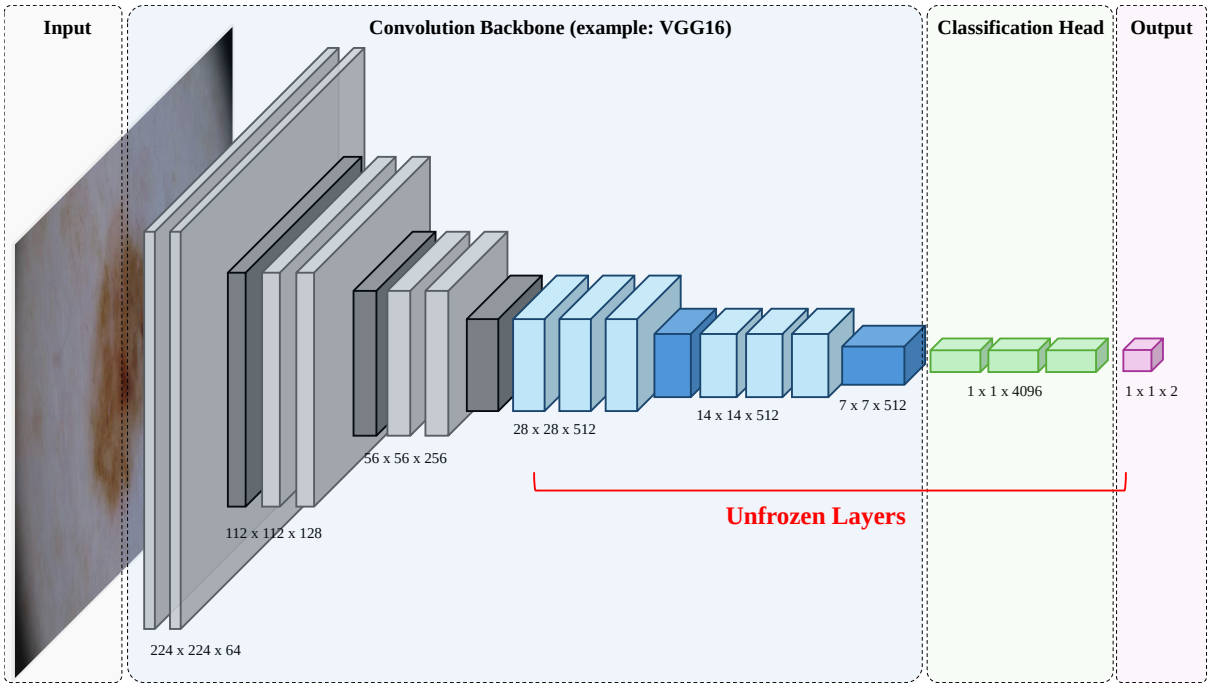

Figure 3: Transfer Learning Approach: Scenario 2 (Fine-tuning).

3.3. Ensemble Learning Approach

In this approach, Ensemble methods were explored. To enhance predictive robustness and reduce variance, we employed an ensemble learning strategy based on bootstrap aggregating (bagging) using multiple EfficientNetB0 models. In this framework, multiple instances of EfficientNetB0 are each trained (fine-tuned) on distinct bootstrap-resampled subsets of the original training set, ensuring diversity among base learners. During inference, predictions from all trained models are combined through majority voting to yield the final classification output. The detailed procedural steps of this approach

are outlined in Algorithm 1.

Algorithm 1 Bagging with EfficientNetB0 Base Learners

Input: Training dataset $D = \{(X_{\text{train}}, y_{\text{train}})\}$,
number of ensemble members T,
base model M (EfficientNetB0)
Output: Final prediction $\hat{y}$
Initialization:
$\mathcal{M} \leftarrow \emptyset$ // set of trained base models
Training phase:
for $t = 1$ to T **do**
  Generate bootstrap sample D_t from D (sampling with replacement)
  Train base learner M_t on D_t
  Update ensemble: $\mathcal{M} \leftarrow \mathcal{M} \cup \{M_t\}$
end for
Prediction phase:
for each new input instance x **do**
  for each $M_t \in \mathcal{M}$ **do**
    Compute prediction: $\hat{y}_t \leftarrow M_t(x)$
  end for
  Aggregate predictions by majority voting:
  $\hat{y} \leftarrow \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T)$
end for
return $\hat{y}$

3.4. Curriculum Learning Approach

Inspired by human pedagogy or learning behavior, the curriculum learning (CL) process begins by relying on simpler samples and progresses to more complex data to improve generalization, reduce sensitivity to noisy or ambiguous samples, and facilitate more stable convergence.

In this approach, we propose a curriculum-learning pipeline that leverages the EfficientNetB0 model as a backbone to be fine-tuned, due to its favorable trade-off between accuracy and computational efficiency. The following steps detail the process of this approach (visualized in figure 4), starting from ranking the samples by difficulty to the final validation phase:

3.4.1. Difficulty-Based Ranking

This stage aims to identify and rank skin lesion images according to their diagnostic complexity. The output of this step is an ordered dataset, starting with the most straightforward cases (characterized by distinct morphological patterns, absence of artifacts, and minimal noise) and progressing to the most challenging cases, involving subtle lesion features, imaging artifacts, or significant anatomical variability. This ordered structure is obtained by leveraging the confidence scores produced by the model's softmax activation, which serve as a quantitative measure of predictive certainty, thereby enabling a systematic and reproducible arrangement of images from easiest to most difficult.

3.4.2. Progressive Curriculum Construction

To allow for a gradual learning, we need to partition the ordered dataset into subsets, each corresponding to a level of difficulty, which will be used to train the model gradually or progressively. Each image is thus associated with a certainty level derived from the predicted probability of its ground-truth label. In our approach, we chose to partition the training set into five such difficulty levels, defined in Table 1:

Table 1
Difficulty-based data stratification in the proposed Curriculum Learning approach

Level	Confidence range	Description
1	$0.8 \leq p < 1.0$	Very certain
2	$0.6 \leq p < 0.8$	Certain
3	$0.4 \leq p < 0.6$	Moderately certain
4	$0.2 \leq p < 0.4$	Uncertain
5	$0.0 \leq p < 0.2$	Highly uncertain

p for ground-truth class.

3.4.3. Training phase

The training process begins with the least complex images, enabling the model to acquire fundamental visual representations and recognize clear diagnostic patterns. Performance is then periodically assessed on a validation set, enabling the refinement of model parameters before progressing to succeeding higher-complexity cases. Subsequently, increasingly challenging images are introduced in a progressive manner by training on added complexity level subsets and validating accordingly. This allows the model to incrementally enhance its feature representations and improve its ability to address atypical or ambiguous cases.

3.4.4. Final consolidation across the entire dataset

After completing or running over all complexity levels, the model is trained on the full training dataset to solidify learned representations and reinforce its generalization capability.

3.4.5. Final validation on an independent test set

Finally, for a rigorous assessment of its robustness and diagnostic accuracy. The model is evaluated on an unseen dataset (test set) that reflects real-world clinical scenarios.

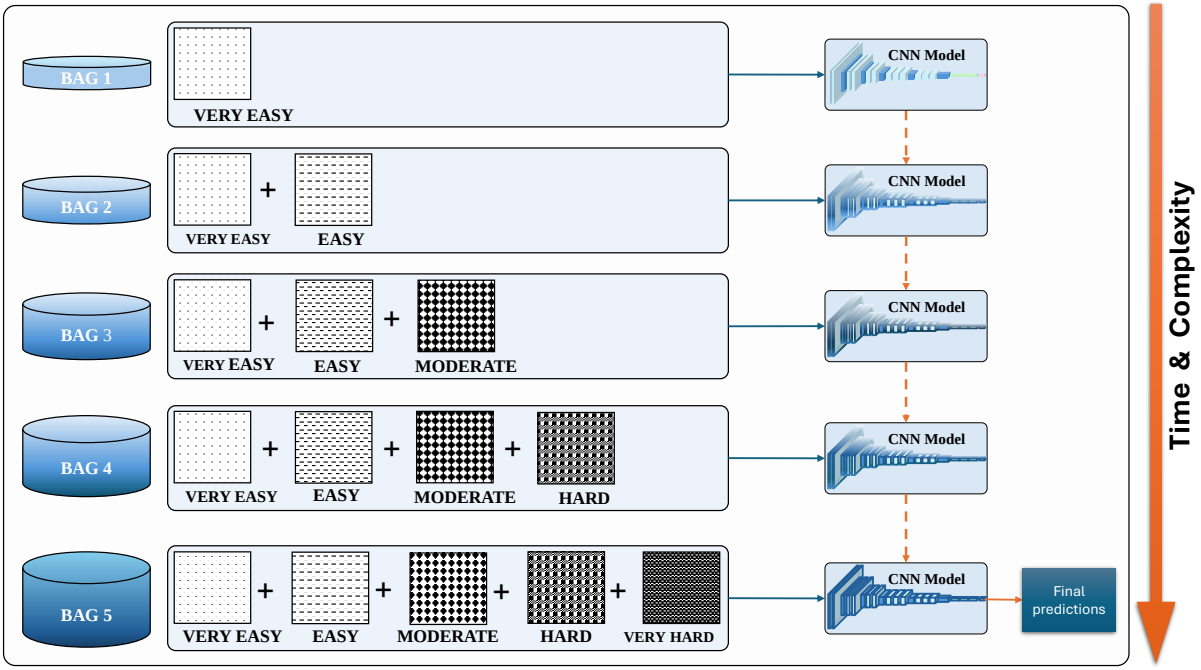

Figure 4: Curriculum Learning Illustration

3.5. Advanced Optimization

Our experiments involved training and evaluating five ImageNet-pretrained CNN backbones (VGG16, InceptionV3, Xception, ResNet50, and EfficientNetB0) using different training strategies on a balanced subset of the ISIC 2019 and ISIC 2020 datasets. To enhance generalization and mitigate overfitting, we applied data augmentation techniques including random flips, brightness/contrast/saturation adjustments, and random crops, alongside stratified train/validation/test splits. Models were trained for 25 epochs with an early stopping patience of 5, and curriculum learning was employed by gradually progressing through five difficulty levels before consolidation on the full dataset. Performance was evaluated using Accuracy, Precision, Recall, and F1-score. Typically, the F1-score is the performance indicator of choice when it comes to comparative evaluations of multiple approaches in medical data analysis, as this clinical context is particularly sensitive to false positives and false negatives. For this reason, the F1-score was adopted as the primary metric in our study.

4. Results

As outlined before, we evaluated three deep learning approaches for the binary classification of skin lesions, distinguishing melanoma from nevus on a combined dataset from ISIC 2019 and ISIC 2020. The results of these evaluations are presented in the following subsections.

4.1. Transfer Learning Approach

In this approach, we trained and evaluated five CNN architecture (InceptionV3, Xception, VGG16, ResNet50, and EfficientNetB0) applied under transfer learning paradigms, both as feature extractors (scenario1) and with fine-tuning (scenario 2). Initial experiments of Scenario 1 where transfer learning was applied with frozen convolutional layers revealed varied performance across the different CNN architectures. ResNet50 and EfficientNetB0 achieved the highest F1-scores among the models, as illustrated in Table 2, with ResNet50 reaching an F1-score of 88.79% and EfficientNetB0 closely following at 88.66% . VGG16 also performed competitively with an F1-score of 87.07%, while InceptionV3 and Xception lagged behind with scores of 81.60% and 84.55%, respectively. These results indicate that the architectures, such as EfficientNetB0 and ResNet50, are better suited for capturing discriminative features in dermoscopic images.

Table 2
Comparative Results of Classification Models Using Transfer Learning in Feature Extraction Mode (Scenario1)

Model	Accuracy	Precision	Rappel	F1-score
InceptionV3	81.72	82.14	81.07	81.60
Xception	85.12	87.89	81.46	84.55
VGG16	87.73	92.01	82.64	87.07
EfficientNetB0	89.03	91.76	85.77	88.66
ResNet50	**89.03**	**90.85**	**86.81**	**88.79**

Table 3 reports the results obtained under the second scenario, in which the five CNN models were fine-tuned with a partial unfreezing of their convolutional layers. As observed, VGG16 achieved the highest performance with an F1-score of 89%, surpassing the best result of Scenario 1 (88.79% with ResNet50). This improvement can be attributed to the fine-tuning, where selectively unfreezing upper convolutional layers enables the model to adapt high-level features to dermatological images.

4.2. Ensemble Learning Strategies Approach

To improve robustness and reduce variance, we implemented ensemble learning using bagging with EfficientNetB0 as the base learner. The ensemble aggregated predictions from multiple independently

Table 3
Comparative Results of Classification Models Using Transfer Learning in Fine-Tuning Mode (Scenario2)

Model	Accuracy	Precision	Rappel	F1-score
InceptionV3	80	84	80	80
EffecienNetB0	84	87	84	84
ResNet50	86	88	86	86
Xception	87	87	87	87
VGG16	**89**	**90**	**89**	**89**

trained instances of EfficientNetB0 models. This approach yielded consistent improvement achieving an F1-score of 88.86% as reported in Table 4.

Table 4
Results of Bagging-Based Classification Using EfficientNetB0

Model	Accuracy	Precision	Rappel	F1-score
EffecienNetB0	89.23	92.03	85.90	88.86

4.3. Curriculum Learning Approach

The most significant performance gains were observed when applying curriculum-learning strategy combined with fine-tuned EfficientNetB0 model. Curriculum learning involved structuring the training process to present images in increasing order of complexity, allowing the model to progressively learn from simpler to more challenging examples. This approach led to a marked improvement in classification metrics, with EfficientNetB0 achieving an F1-score of 90.77%, as listed in Table 5, surpassing both the single-model transfer learning and ensemble bagging approaches.

Table 5
Results of Curriculum Learning Approach Using EfficientNetB0

Model	Accuracy	Precision	Rappel	F1-score
EffecienNetB0	92.03	93.74	87.99	90.77

5. Discussion

The results of this study highlight the significant benefits of applying curriculum learning strategy for the classification of skin lesions using deep convolutional neural networks, particularly EfficientNetB0. Curriculum learning, by organizing training samples from easier to harder based on a confidence-derived difficulty metric, appears to facilitate a more effective learning process. This progressive exposure enables the model to first capture robust feature representations from clear, unambiguous examples before adapting to more challenging, noisy images. Such a training strategy likely contributes to improved convergence stability and ultimately yielded a higher F1-score (90.77%) compared to other approaches. Furthermore, the curriculum learning method not only improved the F1-score but also enhanced precision and recall, reaching 93.74% and 87.99%, respectively, indicating a better balance between false positives and false negatives, which is critical in clinical contexts . This suggests that curriculum learning effectively guides the model to better generalize by focusing on easier samples

first, thereby stabilizing the learning process and reducing misclassification rates. On the other hand, the ensemble learning via bagging revealed complementary strengths compared to curriculum learning, evidenced by an F1-score of 88.86% for EfficientNetB0 ensembles, highlighting the validity of this approach in reducing the variance of the predictions and improving robustness of the model. As well, regarding transfer learning approaches, fine-tuning enhanced performance across most CNN models compared to pure feature extraction, with notable gains observed for VGG16. This aligns with existing literature emphasizing the importance of adapting pretrained weights to domain-specific data for improved representation learning. Nonetheless, curriculum learning surpassed fine-tuning alone, delivering the largest absolute improvement in classification performance within this experimental framework. In summary, the integration of curriculum learning into the training pipeline for skin lesion classification offers a robust and efficient means to improve model accuracy and stability beyond traditional transfer learning and ensemble methods. These findings underscore the value of incorporating data-driven sample difficulty ordering in deep learning workflows, particularly in medical domains where data complexity and variability pose significant challenges. Future work should explore the generalizability of curriculum learning across diverse datasets and investigate its synergy with other advanced techniques and modalities to further enhance clinical applicability.

6. Conclusion

This work presented a comparative study of transfer learning, ensemble deep learning, and curriculum learning strategies for melanoma classification from dermoscopic and clinical images. Our findings confirm the effectiveness of deep CNNs in skin lesion analysis and provide evidence that curriculum learning offers tangible advantages over conventional or even ensemble methods. By gradually exposing models to increasingly difficult samples, curriculum learning used in ensemble deep learning context improves feature refinement and enhances robustness when faced with atypical or ambiguous lesions. Beyond performance gains, curriculum learning contributes to better generalization, an essential property for clinical deployment where unseen cases may deviate significantly from training data. The integration of such progressive training schemes into automated diagnosis systems could therefore support dermatologists with more reliable and efficient decision-making. Future work will focus on expanding curriculum design by incorporating alternative difficulty measures, such as low-level image quality indicators or complexity metrics derived from auxiliary models. Additionally, evaluating the approach across larger and more heterogeneous datasets will be critical to validate its applicability in real-world clinical environments.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim, "Skin cancer detection using deep learning—A review," *Diagnostics*, vol. 13, no. 11, p. 1911, 2023, doi: 10.3390/diagnostics13111911.

[2] S. Wang, G. Cao, Y. Wang, S. Liao, Q. Wang, J. Shi, C. Li, and D. Shen, "Review and prospect: Artificial intelligence in advanced medical imaging," *Front. Radiol.*, vol. 1, art. 781868, Dec. 2021, doi: 10.3389/fradi.2021.781868.

[3] C. Kavitha, S. Priyanka, M. P. Kumar, and V. Kusuma, "Skin cancer detection and classification using deep learning techniques," *Procedia Comput. Sci.*, vol. 235, pp. 2793–2802, 2024, doi: 10.1016/j.procs.2024.04.264.

[4] 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, pp. 115–118, 2017, doi: 10.1038/nature21056.

[5] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al Shamma, and M. A. Fadhel, "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, p. 1590, 2021, doi: 10.3390/cancers13071590.

[6] S. Nahavandi, M. Abdar, M. Samami, S. D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U. R. Acharya, and V. Makarenkov, "Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning," *Comput. Biol. Med.*, vol. 135, p. 104418, 2021, doi: 10.1016/j.compbiomed.2021.104418.

[7] D. Müller, I. Soto-Rey, and F. Kramer, "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks," *IEEE Access*, vol. 10, pp. 66467–66480, 2022, doi: 10.1109/ACCESS.2022.3182399.

[8] S. Ghosh, S. Dhar, R. Yoddha, S. Kumar, A. K. Thakur, and N. D. Jana, "Melanoma skin cancer detection using ensemble of machine learning models considering deep feature embeddings," *Procedia Comput. Sci.*, vol. 235, pp. 3007–3015, 2024, doi: 10.1016/j.procs.2024.04.284.

[9] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4555–4576, Sep. 2022, doi: 10.1109/TPAMI.2021.3069908.

[10] B. Ozdemir and I. Pacal, "A robust deep learning framework for multiclass skin cancer classification," *Sci. Rep.*, vol. 15, art. 4938, 2025, doi: 10.1038/s41598-025-89230-7.

[11] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Proc. Int. Workshop Mach. Learn. Med. Imaging (MLMI)*, vol. 9352, Lecture Notes in Computer Science, L. Zhou, L. Wang, Q. Wang, and Y. Shi, Eds. Cham, Switzerland: Springer, 2015, pp. 118–126, doi: 10.1007/978-3-319-24888-2_15.

[12] Y. Wei, D. Zhang, M. Gao, A. Mulati, C. Zheng, and B. Huang, "Skin cancer detection based on machine learning," *J. Knowl. Learn. Sci. Technol.*, vol. 3, no. 2, pp. 72–86, 2024, doi: 10.60087/jklst.vol3.n2.p86.

[13] A. Demir, F. Yilmaz, and O. Kose, "Early detection of skin cancer using deep learning architectures: ResNet-101 and Inception-v3," in *Proc. Med. Technol. Congr. (TIPTEKNO)*, Izmir, Turkey, 2019, pp. 1–4, doi: 10.1109/TIPTEKNO47231.2019.8972045.

[14] K. Aljohani and T. Turki, "Automatic classification of melanoma skin cancer with deep convolutional neural networks," *AI*, vol. 3, no. 2, pp. 512–525, 2022, doi: 10.3390/ai3020029.

[15] H. C. Reis, V. Turk, K. Khoshelham, and S. Kaya, "InSiNet: A deep convolutional approach to skin cancer detection and segmentation," *Med. Biol. Eng. Comput.*, vol. 60, no. 3, pp. 643–662, Mar. 2022, doi: 10.1007/s11517-021-02473-0.

[16] S. A. H. Shah, S. T. H. Shah, R. Khaled, A. Buccoliero, S. B. H. Shah, A. Di Terlizzi, G. Di Benedetto, and M. A. Deriu, "Explainable AI-based skin cancer detection using CNN, particle swarm optimization and machine learning," *J. Imaging*, vol. 10, no. 12, p. 332, 2024, doi: 10.3390/jimaging10120332.

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

[18] K. Mamar, D. Gaceb, F. Touazi, C. A. Aouchiche, Y. Bellouche, and A. Titoun, "New CNN stacking model for classification of medical imaging modalities and anatomical organs on medical images," in *Proc. Int. Conf. Intell. Data Process. Appl. Med. Imaging (IDDM)*, 2023, pp. 174–188.

[19] N. Laribi, D. Gaceb, F. Touazi, A. Rezoug, A. Sahad, and M. O. Reggai, "Ensemble deep learning of CNN vs. Vision Transformers for brain lesion classification on MRI images," in *Proc. Int. Conf. Intell. Data Process. Appl. Med. Imaging (IDDM)*, 2024, pp. 203–219.

[20] M. Khaled, F. Touazi, and D. Gaceb, "Improving breast cancer diagnosis in mammograms with progressive transfer learning and ensemble deep learning," *Arab. J. Sci. Eng.*, vol. 50, pp. 7697–7720, 2025, doi: 10.1007/s13369-024-09428-1.

[21] Z. Rahman, M. S. Hossain, M. R. Islam, M. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin lesion classification based on ensemble learning," *Informat. Med. Unlocked*, vol. 25, p. 100659, 2021, doi: 10.1016/j.imu.2021.100659.

[22] A. Bassel, A. B. Abdulkareem, Z. A. A. Alyasseri, N. S. Sani, and H. J. Mohammed, "Automatic malignant and benign skin cancer classification using a hybrid deep learning approach," *Diagnostics*,

vol. 12, no. 10, p. 2472, Oct. 2022, doi: 10.3390/diagnostics12102472.

[23] C.-C. Chang, Y.-Z. Li, H.-C. Wu, and M.-H. Tseng, "Melanoma detection using XGB classifier combined with feature extraction and K-means SMOTE techniques," *Diagnostics*, vol. 12, no. 7, p. 1747, 2022, doi: 10.3390/diagnostics12071747.

[24] M. R. Thanka, E. B. Edwin, V. Ebenezer, K. M. Sagayam, B. J. Reddy, H. Günerhan, and H. Emadifar, "A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning," *Comput. Methods Programs Biomed. Update*, vol. 3, p. 100103, 2023, doi: 10.1016/j.cmpbup.2023.100103.

[25] P. Natha and P. RajaRajeswari, "Advancing skin cancer prediction using ensemble models," *Computers*, vol. 13, no. 7, p. 157, 2024, doi: 10.3390/computers13070157.

[26] S. Gamil, F. Zeng, M. Alrifaey, M. Asim, and N. Ahmad, "An efficient AdaBoost algorithm for enhancing skin cancer detection and classification," *Algorithms*, vol. 17, no. 8, p. 353, 2024, doi: 10.3390/a17080353.

[27] Q. Deng, J. Beltran, and D. Lee, "Assessment of segmentation impact on melanoma classification using convolutional neural networks," *J. Comput. Sci. Eng.*, vol. 15, no. 3, pp. 115–124, 2021, doi: 10.5626/JCSE.2021.15.3.115.

[28] 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. 180161, 2018, doi: 10.1038/sdata.2018.161.

[29] N. Codella, D. Gutman, M. E. Celebi, B. Helba, M. Marchetti, S. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern, "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)," *arXiv preprint arXiv:1710.05006*, 2017. doi: 10.48550/arXiv.1710.05006.

[30] 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 preprint arXiv:1908.02288*, 2019. Available: arxiv.org/abs/1908.02288.

[31] V. Rotemberg *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, p. 34, 2021, doi: 10.1038/s41597-021-00815-z.