Methodology for Deep Ensemble Learning and Explainable AI in Imbalanced Melanoma Classification

Puneetha B H [1], Rubini P [2]

[1] Postdoctoral Research Fellow, Lincoln University College, Malaysia.

[2] Professor ,CSE , School of Engineering and Technology, CMR University , Bengaluru.

[a]Puneethgowdar.bh@gmail.com, [b] prubini.cs@gmail.com

Abstract: Automated analysis of dermoscopic images has the potential to support the early identification of melanoma, but reliable classification remains challenging because of class imbalance, variations in image quality, limited labelled data, and the difficulty of interpreting deep learning decisions. This study presents a methodological framework for developing an ensemble-based melanoma classification system with an emphasis on systematic model selection, prediction fusion, and explainability. The framework uses dermoscopic images from the ISIC 2019 and ISIC 2020 datasets and follows a sequence of data preparation, image preprocessing, transfer-learning model development, candidate model evaluation, ensemble construction, and explanation generation. Class imbalance is addressed through majority-class downsampling, while image enhancement, centre cropping, and normalization are applied before model training. Five pretrained convolutional neural networks—VGG-19, ResNet-50, ResNet-101, DenseNet-121, and InceptionV3—are considered as candidate learners. Based on their classification performance, suitable models are selected for ensemble construction. Several prediction-fusion strategies are then examined, including hard voting, probability averaging, maximum-rule fusion, and weighted probability averaging. The weighting mechanism considers multiple evaluation measures rather than relying only on accuracy. Finally, SHAP is incorporated to examine the image regions contributing to individual predictions. This methodology provides a structured basis for evaluating whether combining complementary deep learning models can improve melanoma classification while making model decisions more interpretable

Keywords: Melanoma Classification, Dermoscopic Images, Transfer Learning, Deep Ensemble Learning, Class Imbalance, Explainable AI, SHAP

1. Introduction

Melanoma is a serious form of skin cancer in which early detection plays an important role in improving patient outcomes. Dermoscopic imaging provides a non-invasive way to examine skin lesions and has become an important source of images for automated melanoma classification[1]. However, the visual characteristics of dermoscopic images can vary considerably because of differences in image quality, lighting, focus, lesion appearance, and surrounding skin. These variations make automated classification a challenging task[2].

Deep learning has been increasingly applied to melanoma classification because of its ability to learn complex visual patterns from medical images. Convolutional neural networks (CNNs) can automatically extract discriminative features from dermoscopic images and classify lesions into malignant and benign categories. However, relying on a single deep learning model may not provide consistent performance across different image characteristics. The use of multiple deep learning models can provide complementary predictions and offers an alternative approach for improving classification robustness[3].

Another major challenge is the imbalance between malignant and benign samples in melanoma datasets. For example, the ISIC 2020 dataset contains substantially more benign images than malignant

images, with malignant cases representing only 1.8% of the dataset with known diagnoses. The ISIC 2019 dataset also shows considerable imbalance, with malignant cases representing 17.8% of the dataset. Such imbalance can influence the learning process and affect the ability of a classifier to identify malignant cases effectively[4].

In addition to predictive performance, interpretability is an important consideration when applying deep learning to healthcare. Deep neural networks generally provide limited information about the reasoning behind their predictions[5]. This lack of transparency can make it difficult to assess whether a model is focusing on meaningful characteristics of a skin lesion or on unrelated image features. Explainable Artificial Intelligence (XAI) methods can therefore be used to investigate the factors contributing to individual predictions[6].

To address these challenges, the study develops a methodological framework that combines data balancing, image preprocessing, transfer learning, deep ensemble learning, and explainable AI. Multiple pretrained CNN architectures are first evaluated as candidate models[7-9]. The stronger models are then selected for ensemble construction, where different prediction-fusion mechanisms are compared. The framework also incorporates SHAP to examine the contribution of image regions to individual classification decisions[10].

The main aim of this work is therefore to establish a systematic methodology for melanoma classification that not only evaluates classification performance but also investigates the reliability and interpretability of the resulting predictions. The subsequent sections describe the dataset preparation, preprocessing procedure, transfer-learning models, ensemble design, evaluation strategy, and explainability approach used in the study.

2. Review Methodology

Deep learning has been widely explored for automated melanoma classification using dermoscopic images. Convolutional neural networks (CNNs) have demonstrated the ability to learn visual characteristics of skin lesions directly from images, reducing the dependence on manually designed features[1-3]. However, variations in lesion appearance and the imbalance between benign and malignant samples remain important challenges in developing reliable classification systems[4].

Transfer learning has been used to address the difficulty of training deep networks when medical datasets contain limited labelled examples. Pretrained CNN architectures can be adapted to dermoscopic image classification by fine-tuning their learned representations. The paper considered in this study follows this approach by evaluating several pretrained architectures, including VGG-19, ResNet-50, ResNet-101, DenseNet-121, and InceptionV3, before selecting suitable models for ensemble learning[5-6].

Another direction is the use of ensemble learning, where predictions from multiple deep learning models are combined instead of depending on a single classifier. This approach is useful when different CNN architectures learn complementary image characteristics. Previous work has also investigated ensemble and stacking approaches for melanoma classification, including combinations of transfer-learning CNNs and other machine learning models. For example, an interpretable melanoma diagnosis approach combined several deep learning models and used ensemble strategies to improve classification performance[7-9].

The handling of class imbalance is particularly important in melanoma classification. Dermoscopic datasets can contain considerably fewer malignant cases than benign cases. The study considered here addresses this issue by applying downsampling to the majority class before combining the datasets. This creates a balanced set of benign and malignant samples for subsequent experimentation[10-12].

Several ensemble fusion strategies have also been investigated, including hard voting, soft voting, maximum-rule fusion, and weighted probability averaging. Rather than assuming that every model should contribute equally, performance-based weighting provides a way to give greater influence to stronger classifiers. The considered study evaluates these alternatives and uses Precision, Recall, F1-score, and ROC-AUC when determining model weights[13].

While classification performance is important, interpretability is another major requirement for medical AI systems. Explainable AI methods can provide information about the image regions that influence a model's prediction. SHAP is used in the considered methodology to generate explanations for individual predictions and examine the contribution of image features[14].

Recent research continues to explore the combination of deep learning, ensemble methods, and explainability for melanoma and skin-lesion classification. For example, recent work has investigated explainable ensemble approaches using SHAP and Grad-CAM, while other approaches integrate segmentation information with classification to produce explanations that are more closely related to clinically relevant lesion regions[15].

Overall, the literature indicates that melanoma classification can benefit from combining class-imbalance handling, transfer learning, multiple CNN architectures, ensemble prediction, and explainability. Based on these observations, the present study focuses on developing a systematic methodology in which candidate CNN models are evaluated first, suitable models are selected for ensemble construction, alternative fusion strategies are compared, and SHAP is subsequently used to examine the basis of model predictions. This provides the foundation for the methodology and experimental design presented in the following sections.

3. Research Problem and Objectives

Although deep learning models have shown potential for melanoma classification, developing a reliable classification system remains challenging because dermoscopic images can vary in quality and appearance, while available datasets may contain a large difference between benign and malignant samples. In addition, using a single deep learning model may not provide sufficiently consistent predictions across different image characteristics[2][8]. Another important concern is that the reasoning behind a model's prediction is difficult to understand when the model operates as a black box. Based on these issues, this study considers a methodology that combines transfer learning, ensemble learning, performance-based model weighting, and Explainable AI (XAI). The methodology first evaluates several pretrained CNN architectures and then selects suitable models for ensemble construction. Different prediction-combination strategies are compared to determine an effective approach for melanoma classification[10]. Finally, SHAP-based explanations are used to examine the image regions contributing to the predictions.

The main objectives are:

1. **To prepare a balanced dermoscopic image dataset** by addressing the class imbalance present in the ISIC datasets.
2. **To evaluate multiple transfer-learning CNN architectures** for malignant and benign skin-lesion classification.
3. **To identify suitable base learners** for constructing a deep ensemble.
4. **To compare different ensemble strategies**, including hard voting, soft voting, maximum-rule fusion, and weighted probability averaging.

5. **To develop a performance-based weighting approach** using classification metrics for combining the selected models.
6. **To incorporate SHAP-based explanations** to investigate which image regions contribute to model predictions.
7. **To evaluate the complete methodology** using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The above objectives provide the basis for the experimental design and methodology described in the following section.

4. Proposed Methodology

The proposed methodology is designed as a sequence of stages, beginning with dataset preparation and ending with model explanation. The overall process includes data collection, class balancing, image preprocessing, transfer-learning model development, model selection, ensemble construction, performance-based weighting, and explainability analysis. This structure allows each stage to be evaluated before moving to the next stage. Figure 1 illustrates the overall methodology. The input dermoscopic image is provided to three selected deep learning models, namely ResNet-101, DenseNet-121, and InceptionV3.

4.1 Dataset Preparation

The study uses dermoscopic images from the ISIC 2019 and ISIC 2020 datasets. ISIC 2020 contains 33,126 training images, while ISIC 2019 contains 25,331 training images. The images are classified into two categories: benign and malignant. A major issue in the datasets is class imbalance. In ISIC 2020, the number of benign images is much higher than the number of malignant images. Therefore, images with unknown diagnosis are first removed, followed by downsampling of the majority class. The same balancing procedure is applied to the ISIC 2019 data. The resulting datasets are combined to obtain 10,212 images, consisting of 5,106 benign and 5,106 malignant images.

4.2 Image Preprocessing

The images are preprocessed to reduce variations that may affect model learning. Image enhancement is performed using PIL/OpenCV operations. Centre cropping is applied to reduce differences in lesion position, followed by normalization of pixel values. The preprocessing settings include colour enhancement, sharpness adjustment, brightness and contrast modification, centre cropping to 75% of the image dimensions, and pixel normalization by dividing the pixel values by 255.

4.3 Transfer Learning Models

Five pretrained CNN architectures are considered as candidate models: **VGG-19, ResNet-50, ResNet-101, DenseNet-121, and InceptionV3**. These models are pretrained using ImageNet and adapted for the binary classification task. The original top classification layers are removed and replaced with task-specific fully connected layers. ReLU activation and L1/L2 regularization are incorporated into the network. The input images are resized to **224 × 224 × 3** before being provided to the models. For training, the methodology uses the Adam optimizer, categorical cross-entropy loss, a learning rate of **0.0001**, batch size of **64**, and a maximum of **1000 epochs**. Early stopping is applied with a patience of 100 epochs. Online augmentation is also used through flipping, rotation, zooming, shearing, and image shifting.

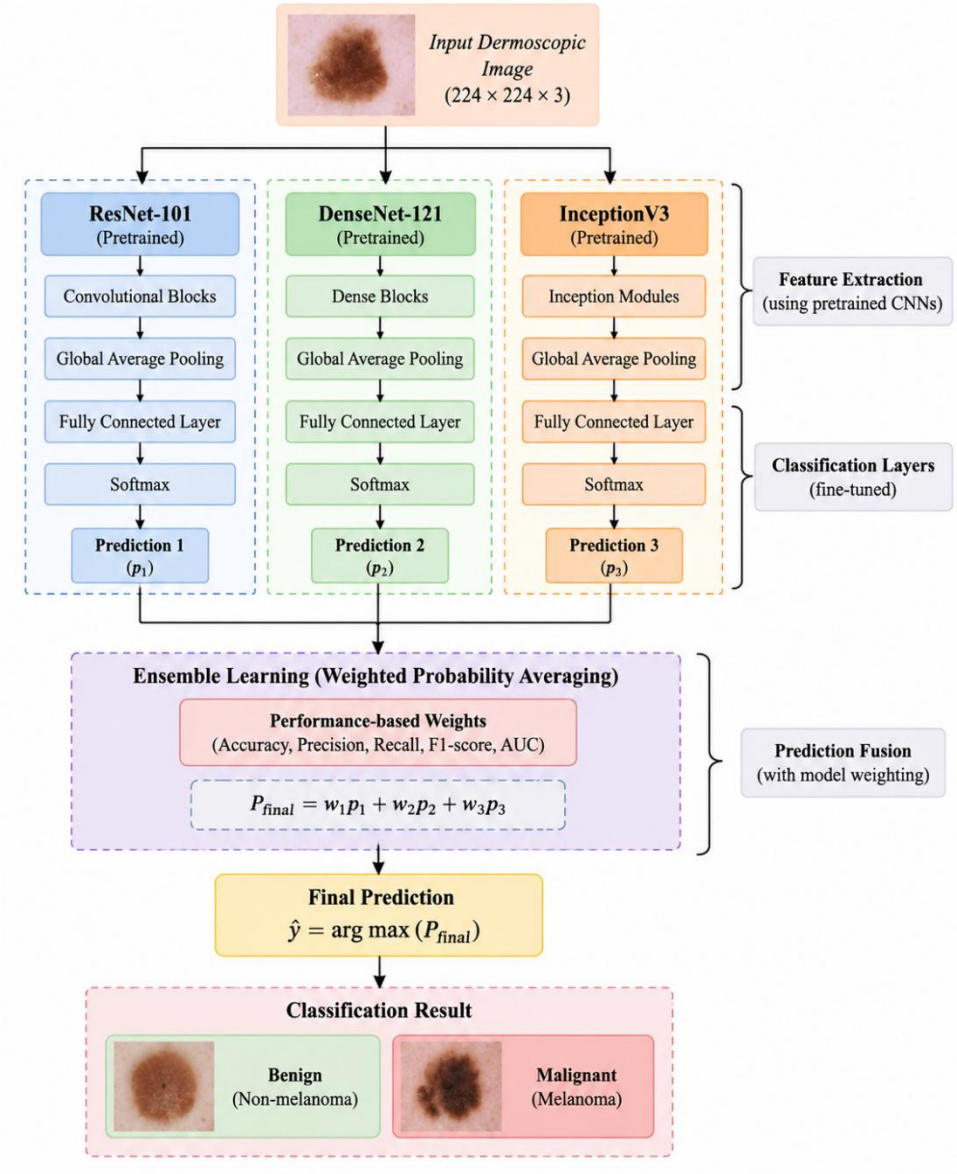

Figure 7: Proposed Ensemble learning framework

4.4 Model Selection

The candidate models are evaluated using **Accuracy, Precision, Recall, F1-score, and ROC-AUC**. Four-fold cross-validation is used during the evaluation of the candidate models. Based on the experimental evaluation, **ResNet-101, DenseNet-121, and InceptionV3** are selected as the base learners for the ensemble. The selected models provide different learned representations of the dermoscopic images, which makes them suitable candidates for combining their predictions in the ensemble stage.

4.5 Ensemble Construction

The predictions of the three selected models are combined using four ensemble strategies:

- Hard majority voting
- Soft majority voting
- Maximum-rule fusion
- Weighted probability averaging

The weighted approach assigns different contributions to the base learners according to their performance. This avoids treating all models as equally reliable and forms the basis for the final ensemble methodology. This methodology provides a clear experimental pipeline in which the effect of **model selection and prediction fusion** can be evaluated systematically.

5. Experimental Design

The experimental design evaluates the proposed methodology in three main stages. First, five transfer-learning models—VGG-19, ResNet-50, ResNet-101, DenseNet-121, and InceptionV3—are evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Four-fold cross-validation is used for model evaluation. Next, ResNet-101, DenseNet-121, and InceptionV3 are selected as base models. Their predictions are combined using hard voting, soft voting, maximum-rule fusion, and weighted probability averaging. Finally, different weighting approaches are compared using Precision, Recall, F1-score, and ROC-AUC. The best-performing ensemble is then examined using SHAP to understand the image regions influencing the predictions. This design provides a systematic comparison of individual models, ensemble strategies, model weighting, and explainability.

6. Explainable AI Analysis

To improve the interpretability of the classification system, **SHAP (SHapley Additive exPlanations)** is used to examine the contribution of image regions to the model's predictions. The Gradient Explainer is applied to the ensemble framework and the feature representations generated by the selected CNN models. The SHAP results indicate which regions contribute positively or negatively to a prediction. This provides an additional way to examine whether the models are focusing on relevant lesion regions rather than relying mainly on surrounding skin or image artefacts. This explainability stage complements the classification experiments by providing information about **why a particular prediction was made**, rather than evaluating the model only through numerical performance measures.

7. Conclusion and Future Work

This study presents a methodological framework for melanoma classification using transfer learning, ensemble learning, and Explainable AI. Multiple pretrained CNN models are evaluated, suitable models are selected, and their predictions are combined using different ensemble strategies. The experimental design also incorporates performance-based weighting and SHAP analysis to improve the interpretability of the classification results. The reported findings indicate that the weighted ensemble provides better overall performance than the individual models, while SHAP helps identify image regions contributing to the predictions. However, the analysis also indicates that image artefacts such as hair and circular structures can influence model decisions. Future work will focus on improving sensitivity, reducing the effect of image artefacts, and investigating lesion segmentation and colour calibration. Further validation of the generated explanations against clinical dermoscopic features and expert dermatologist assessment can also strengthen the reliability of the methodology

Acknowledgment

The authors sincerely thank their respective institutions and departments for providing the necessary support, facilities, and encouragement to carry out this research work successfully.

References

1. American Cancer Society. (2022). Cancer Facts and Figures 2023. American Cancer Society.
2. Lakhtakia, R., Mehta, A., & Nema, S. (2009). Melanoma: A frequently missed diagnosis. Medical Journal Armed Forces India, 65(3), 292–294. https://doi.org/10.1016/S0377-1237(09)80036-1
3. Li, Y., & Shen, L. (2018). Skin lesion analysis towards melanoma detection using deep learning network. Sensors, 18(2), 556. https://doi.org/10.3390/s18020556
4. Harangi, B., Baran, A., & Hajdu, A. (2018). Classification of skin lesions using an ensemble of deep neural networks. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). https://doi.org/10.1109/EMBC.2018.8512800
5. Bisla, D., Choromanska, A., Berman, R. S., Stein, J. A., & Polsky, D. (2019). Towards automated melanoma detection with deep learning: Data purification and augmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/CVPRW.2019.00330
6. Ali, R., Hardie, R. C., Narayanan, B., & De Silva, S. (2019). Deep learning ensemble methods for skin lesion analysis towards melanoma detection. 2019 IEEE National Aerospace and Electronics Conference (NAECON). https://doi.org/10.1109/NAECON46414.2019.9058245
7. Adegun, A. A., & Viriri, S. (2020). Deep learning-based system for automatic melanoma detection. IEEE Access, 8, 7160–7172. https://doi.org/10.1109/ACCESS.2019.2962812
8. Wei, L., Ding, K., & Hu, H. (2020). Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network. IEEE Access, 8, 99633–99647. https://doi.org/10.1109/ACCESS.2020.2997710
9. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata. MethodsX, 7, 100864. https://doi.org/10.1016/j.mex.2020.100864
10. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR).
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90
12. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2017.243
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.308

14. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287
15. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4768–4777.