STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2025, pp 0–8.
Published online in International Academic Press (www.IAPress.org)

A Hybrid Fusion Approach for Skin Cancer Detection using Deep Learning on Clinical Images and Machine Learning on Patient Metadata

Aya Saber Abd El Aziz[1*], Mohamed Mohamed El-gazzar[2], Mary Monir Saeid[1]

[1] *Faculty of Computers and Artificial Intelligence, Department of Information Systems, Fayoum University, Egypt*
[2] *Faculty of Computers and Artificial Intelligence, Modern University for Technology Information, Egypt*

Abstract
Skin cancer continues to be a global health issue, with early detection being critical for improving treatment outcomes. While deep learning models like convolutional neural networks (CNNs) have proven highly efficient in skin cancer classification from dermatological images, they often disregard the valuable patient metadata that can contribute to better diagnostic accuracy. In the present study, we introduce a multimodal late fusion framework that integrates both skin cancer images and patient metadata. The approach leverages the Inception-ResNet-v2 (IRv2) model to extract image features, and a stacking ensemble model consisting of Extra Trees and Random Forest classifiers for patient metadata preprocessing. Then, a final voting classifier applying a soft voting strategy, which aggregates class probabilities from Logistic Regression and Random Forest as base voters, is employed for the final fusion methodology. This leads to an accuracy of 95.9% on the HAM10000 dataset. Our results highlight the potential of multimodal approaches in healthcare applications.

Keywords Skin Cancer Classification, Machine Learning, Multimodal Fusion, skin cancer, Patient Metadata, Medical Diagnostics.

DOI: 10.19139/soic-2310-5070-2811

1. Introduction

The prevalence of skin malignancies, including melanoma and non-melanoma, continues to rise [1]. Globally, 132,000 instances of melanoma and 2 to 3 million of non-melanoma cancers are detected yearly [2]. To address the challenge of early melanoma detection, the use of image analysis in medical diagnosis increased dramatically in recent years [3]. Innovations in diagnostic technologies, particularly in artificial intelligence, have shown promising results in supporting dermatologists [4]. However, many existing methods only analyze dermatological images [5, 6]. Given the complexity of skin cancer, incorporating additional patient data for instance age, gender, and lesion location, which are known to influence skin cancer risk can significantly enhance classification accuracy [7]. To address this, we propose a multimodal fusion method that integrates dermatological images with patient-specific data. By leveraging complementary information from both visual and non-visual data sources, our objective is to improve the model's diagnostic performance.

- Putting forward a hybrid fusion strategy that combines machine learning on patient metadata with deep learning on clinical images.
- Enhancing skin cancer detection accuracy by integrating multimodal features.
- Conducting comprehensive experiments on a publicly available dataset (HAM10000)
- Providing a detailed comparison with existing state-of-the-art methods.

Following the introduction, Section two reviews existing research on skin cancer detection and multimodal fusion. Section three describes the datasets and the proposed fusion model. Section four presents the experimental

ISSN 2310-5070 (online) ISSN 2311-004X (print)

setup, evaluation measures, and performance comparisons. Section five discusses the limitations of the proposed approach. Section six concludes the study by highlighting key findings, and Section seven outlines future research directions.

2. Literature Review

This literature review examines recent advancements in multimodal approaches for skin cancer classification, summarized in Table 1. A multimodal fusion framework based on deep learning was presented by [8] for skin cancer classification, combining intra-and inter-modality self-attention cross-attention to process both images and metadata collected via cellphones. Applied to the PAD-UPES-20 dataset, the model uses ResNet-50 as an image encoder and a multi-layer perceptron (MLP) for metadata processing. The fusion framework accomplished an average accuracy of 76.8%.

The authors of [9] introduced a combined artificial neural network framework for early skin cancer diagnosis through color space transformations. The suggested model combines EfficientNet-B1, a Convolutional Neural Network (CNN) for the identification of features in dermoscopic images, employing an MLP to handle patient metadata and manually calculated mean and median color space values. The hybrid approach, which fuses the two models using a common dense layer, outperformed the standalone EfficientNet-B1 model, achieving 86% accuracy across seven skin lesion classes on the HAM10000 dataset.

A multimodal data fusion network was proposed by [10] to enhance skin cancer diagnosis using both patient skin images and corresponding clinical information. The network consists of three main components: feature extraction, an attention mechanism, and a multimodal feature fusion module. Feature extraction was performed using neural networks such as VGGNet-19, ResNet-50, DenseNet-121, and Inception-V3, alongside a fully connected layer for patient clinical data. Attention mechanisms, such as channel attention, spatial attention, and mixed attention, were examined in this study, in addition to fusion strategies like feature and model fusion. An accuracy of 80% was attained by the model on the PAD-UPES-20 dataset, outperforming single-modality approaches.

As presented by [11], a neural network combining multiple modalities with a modified cross-entropy loss function featuring multiple pre-trained CNN architectures was described. DenseNet-161, Inception-v4, and ResNeXt-50 were utilized to extract features from input images, while an MLP processed patient metadata sourced from the International Skin Imaging Collaboration (ISIC) Archive. The extracted features were merged by means of a fully connected layer to integrate visual and metadata information. The DenseNet-161-based model accomplished an accuracy of 85.2%.

Several CNN architectures were used by [12], including DarkNet-53, DenseNet-201, GoogLeNet, Inception-V3, InceptionResNet-V2, ResNet-50, ResNet-101, and Xception. Utilizing a weighted sum approach, the predictions of these models were combined, yielding the best accuracy of 89% with DarkNet-53 and 82% with GoogLeNet on the ISIC 2020 dataset.

The authors of [13] introduced a novel Multi-Model Late Feature Fusion Network (MLFF-Net) that combines DenseNet-121 and Vision Transformers (ViTs) to enhance feature extraction and classification accuracy. DenseNet-121 captures local features through dense connections, while ViTs capture global context and long-range dependencies. Features from both models are fused using a Multi-Receptive Field Feature Fusion Block (FFB) to gather multi-scale information, followed by a classification layer. This approach achieved 86% accuracy on the HAM10000 dataset.

The multimodal approach put forth by the authors of [14] uses pre-trained CNNs to extract complex patterns and features from the HAM10000 dataset and an extra 100 Squamous Cell Carcinoma (SCC) images. This method utilizes machine learning techniques and incorporates data from several sources. In particular, this work captures temporal dependencies in textual metadata by combining CNNs for image processing with LSTMs for text handling. The models were trained and evaluated using the merged dataset, reaching a 92% accuracy rate. The model performed well in classifying skin cancer through the usage of dense and convolutional layers.

3. Materials and Methods

In this work, we propose a multimodal skin cancer classification model based on a late fusion strategy that integrates dermoscopic skin images with patient metadata. The proposed system is composed of three main components. The first component is an image classification model based on the Inception-ResNet-v2 (IRV2) architecture. The second component is a metadata classification model constructed using a stacking ensemble approach, which incorporates Extra Trees and Random Forest classifiers to improve predictive performance. Finally, the third component is a fusion mechanism that combines predictions from both the image and metadata models using Logistic Regression and Random Forest as base-level classifiers. The final prediction is obtained through a consensus decision by averaging the class probabilities from both modalities, allowing the fusion model to consider the confidence levels of each network. The proposed approach is evaluated using the publicly available HAM10000 dataset [15], which contains 10,015 dermoscopic images categorized into seven skin lesion classes: vascular lesions, squamous cell carcinoma, basal cell carcinoma, melanoma, melanocytic nevi, dermatofibroma, and keratosis.Figure 1 illustrates the overall architecture of our proposed multimodal framework, highlighting the interaction between the three core modules and emphasizing the integration of visual and metadata features.

3.1. Image Preprocessing and Classification

In the image classification task, we used an Inception-ResNet-v2 deep learning model, which is a state-of-the-art architecture of convolutional neural networks that leverages the benefits of Inception networks and residual networks [16]. The model is widely used for complex image classification tasks and is effective for dermoscopic images as it retains fine details.

3.1.1 Data Preparation

To prepare the seven-class dataset, a stratified train-test split was applied with 85% of the data saved for training and 15% for testing. The goal of the stratification is to guarantee that each class has equal representation in the testing and training sets. Each of the training and testing sets was organized into separate directories based on their respective classes. This directory structure facilitates efficient image handling during the model training and evaluation process. The images corresponding to each class label were moved into their respective class subdirectories within the train and test directories.

3.1.2 Data Augmentation

Data augmentation was employed before training to increase the training set's diversity to improve the model's robustness and prevent overfitting. Images were rotated randomly with angles ranging from 0 to 180 degrees. Zooming (up to 10%), and random flips in both directions were among the augmentation techniques used. The `ImageDataGenerator` from the TensorFlow Keras library was used to implement these augmentations, which generated augmented images dynamically during the training process. The dataset's diversity was greatly expanded by the augmentation, which also enhanced its generalizability. Table 2 presents the distribution of images in the training and testing sets.

3.1.3 Model Architecture

The Inception-ResNet-V2 model was customized by removing the top layer and adding a new layer appropriate for the classification task. Specifically, a global average pooling layer was added after the convolutional base, followed by a fully connected dense layer with ReLU activation. A dropout layer was included before the output layer to prevent overfitting. The final layer is a SoftMax layer to produce class probabilities for the seven categories.

Class weights were assigned for each class to address any potential class imbalances, with a weight of 5.0 for melanoma and 1.0 for the rest of the classes. This weighting aimed to ensure that the model treats all classes appropriately during training, particularly emphasizing melanoma's importance for its clinical significance.

3.1.4 Model Compilation and Training

We employed the Adam optimizer with a learning rate of 0.01 and an epsilon value of 0.1 for model compilation. The loss function used was categorical cross-entropy, and accuracy was monitored as the performance metric. This learning rate was chosen due to multiple experiments with different values to balance the training stability.

The model training process included the use of early stopping and model checkpoint to enhance generalization and prevent overfitting. A validation split of 10% from the training data was used during training to monitor performance and trigger early stopping when improvements plateaued. The training was first conducted over 5 epochs, then later continued for a total of 38 epochs, with the number of steps per epoch calculated based on the size of the training batches.

3.2. Patient Metadata Preprocessing and Classification

The patient metadata classification task was performed using a stacking ensemble method to benefit from the strengths of several classifiers and overcome their individual weaknesses [17]. Random Forest and Extra Trees classifiers were used as base models in the ensemble, with a Logistic Regression model serving as the final estimator.

3.2.1 Data Preparation

A stratified train-test split was applied to ensure proportional representation of all classes in the training and testing sets during the metadata preparation phase. The dataset was split into 85% for training and 15% for testing. To ensure proper alignment between the metadata and the images during the late fusion step, the `image_id` column in the metadata was used to match each metadata entry to its corresponding test image.

A custom Python script was used to extract additional features from the dermoscopic images to enhance the metadata model. These features included color (mean and standard deviation of RGB channels), texture (using Gray Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP)), and shape (lesion area and perimeter from grayscale images). Texture features included contrast, dissimilarity, homogeneity, energy, correlation, and angular second moment (ASM). All extracted features were saved alongside the original metadata, including lesion ID, image ID, diagnosis, patient age, sex, and lesion localization. Rows with missing values were removed to maintain data integrity. Table 3 summarizes the newly extracted metadata features.

3.2.2 Data Preprocessing

The enhanced metadata underwent several preprocessing steps. Categorical features (e.g., sex and localization) were encoded using Label Encoding and One-Hot Encoding. Missing values in numerical fields were imputed using the median. All numerical features were standardized to have zero mean and unit variance to ensure fair treatment across features.

To address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training data. Additionally, class weights were computed and incorporated into the classifiers to further mitigate imbalance effects. This comprehensive preprocessing ensured the metadata was clean and well-prepared for model training, thereby improving predictive performance.

3.2.3 Stacking Ensemble Model

A stacking ensemble model was applied to combine the predictions of multiple classifiers, enhancing the overall performance [17, 18]. The Random Forest and Extra Trees classifiers were used as base models. Both were trained with class weighting to account for imbalance. Their predictions were then input into a Logistic Regression model as the meta-learner.The stacking approach leverages the strengths of diverse classifiers, allowing the model to learn complex patterns more effectively. This technique improves both accuracy and generalization capability of the metadata classification pipeline.

3.3. Late Fusion: An Ensemble Approach for Image and Metadata Models

A late fusion technique employing soft voting [18] was used to merge the predictions from the metadata classification model (stacking ensemble) with the image classification model (Inception-ResNet-v2). By merging

the estimated probability of both modalities into a single decision-making framework, this methodology aims to capitalize on the advantages of both image-based features and patient metadata.

3.3.1 Generating and Combining Probabilistic Predictions from Multiple Modalities

The first step in the fusion process was to generate class probabilities from both models. Probabilities for the seven classes are produced by the trained Inception-ResNet-v2 model, indicating its level of confidence in each class label for the test images. The stacking ensemble generates its own class probabilities based on the preprocessed information by combining Random Forest and Extra Trees classifiers with a logistic regression meta-estimator.

Next, the probabilities from both models are combined to create a unified feature set, which serves as the soft voting classifier's input. The soft voting strategy can take into account the relative confidence levels of each model by using probabilities instead of hard class predictions, which improves the final prediction.

3.3.2 Soft Voting Classifier Design and Training

The soft voting classifier was developed by combining two meta-models: Logistic Regression and Random Forest. Logistic Regression was selected due to its ability to process probabilistic inputs and its effectiveness in establishing a linear decision boundary, which helps assess the confidence of predictions from base models [18]. In contrast, Random Forest was chosen for its robustness in managing complex feature interactions and capturing non-linear patterns in the data [19].

The soft voting mechanism averaged the class probabilities predicted by each model to produce the final output, allowing for more refined predictions by incorporating the confidence levels of both models. A stratified train-test split was used during training to maintain the distribution of class labels in both sets. The final model was implemented using a soft voting classifier that combined the outputs of Inception-ResNet-v2 and the stacking ensemble, optimizing the interactions and weights of the predictions.

To formalize this, the final predicted probability for a given class c is computed as:

$$P_{\text{final}}(c) = \frac{1}{n} \sum_{i=1}^{n} P_i(c) \tag{1}$$

where $P_i(c)$ is the predicted probability of class c from the i-th model. This formulation ensures that each model contributes equally to the final prediction, enabling a robust ensemble decision.

4. RESULTS AND DISCUSSION

In this study, a multi-class classification fusion model was developed and evaluated to classify skin lesion images into seven categories: vascular lesions, squamous cell carcinoma, basal cell carcinoma, melanocytic nevi, dermatofibroma, and keratosis. A variety of metrics, including accuracy, precision, recall, F1 score, Matthews Correlation Coefficient (MCC), Cohen's Kappa, and confusion matrices, were used to evaluate the performance of the image model, stacking model, and fusion model (Soft Voting Classifier). All models were tested on the same dataset, consisting of 828 samples representing seven different skin lesion categories.

4.1. Evaluation Metrics and Formulas

The following metrics were used to evaluate the models:

- **Accuracy**: The proportion of correct predictions among all cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

  where:

  – TP = True Positives

  - *TN* = True Negatives
  - *FP* = False Positives
  - *FN* = False Negatives

- **Precision**: The proportion of positive predictions that were accurate.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

- **Recall**: The proportion of actual positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

- **F1 Score**: The harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

- **Matthews Correlation Coefficient (MCC)**: An indicator of the quality of binary classifications.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

- **Cohen's Kappa**: A statistical measure of the degree of inter-rater agreement between categorical items.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{7}$$

  where P_o is the observed agreement, and P_e is the expected agreement. "'

4.2. Image Model Results

The image model, based on the Inception-ResNet-v2 architecture, achieved an accuracy of 90%. The classification report for this model is detailed in Table 4, while the corresponding confusion matrix is presented in Figure 2, which provides a more detailed view of how well the model discriminates between the different classes, showing the true positives, false positives, and false negatives. The model performed particularly well for the `nv` (melanocytic nevi) class, achieving a high F1-score of 0.96, indicating strong and balanced performance in both precision and recall for this class.

However, the recall and F1-score were lower for the `df` class (0.33 and 0.36 respectively), suggesting that the model struggled to identify this class reliably. The model's highest F1-score of 0.96 was for the `nv` class, demonstrating a good balance between precision and recall. The `mel` (melanoma) class, on the other hand, had an F1 score of 0.50, indicating room for improvement in both precision and recall for this particular class. Additionally, the ROC curve for the image model is included in Figure 3 to demonstrate the model's performance as a means of differentiating classes. The curve highlights the model's strong discriminative ability, particularly for the `nv` class, but also shows areas where further improvements, especially in rare classes like `akiec` and `mel`, could enhance performance.

4.3. Stacking Model Results

The stacking ensemble model, which combined the Random Forest and Extra Trees classifiers with a Logistic Regression meta-model, achieved an accuracy of 81%. This model effectively leveraged the diversity of base classifiers to produce a robust final prediction. The classification report for the stacking model is shown in Table 5,

and the corresponding confusion matrix is presented in Figure 4. The model demonstrated strong performance in the 5 class (precision: 0.86, recall: 0.97, F1-score: 0.91), which contained many samples, but struggled with classes such as 0, 1, and 6 where the precision, recall, and F1-scores were relatively low. Specifically, the 0 class had a precision and recall of 0, indicating that the model did not identify any instances of this class correctly. Similarly, the 6 class had low values across all metrics, suggesting the need for further model improvement or better class balancing.

4.4. Fusion Model Results (Soft Voting Classifier)

The Soft Voting Classifier, which combined the predictions of the image model and the stacking model, achieved an impressive accuracy of 95.9%. This classifier demonstrated robust performance across multiple evaluation metrics presented in Table 6, and the confusion matrix is visualized in Figure 5. Including a macro-average precision of 0.99, a macro-average recall of 0.79, and a macro-average F1-score of 0.87. These results indicate the model's ability to balance precision and recall across all classes effectively.

4.5. Comparison of Model Performance

Table 7 below summarizes the accuracy and key metrics achieved by each model evaluated in this study. Among the models tested, the Soft Voting Classifier demonstrated the best overall performance, excelling across all evaluation metrics. This model effectively leveraged the strengths of both the image-based model and the metadata-based model, resulting in a significant improvement in accuracy and recall for underperforming classes.

Notably, the Soft Voting Classifier enhanced the classification performance for challenging classes such as class 0, where other models struggled to achieve consistent results. Furthermore, the classifier maintained excellent performance for class 5, achieving a perfect recall of 1.00, indicating that all samples in this class were correctly identified.

The fusion model also exhibited a strong overall performance, achieving an Average Precision (AUC-PR) of 0.9899, which demonstrates the classifier's high confidence and reliability in its predictions. Additionally, metrics such as Matthews Correlation Coefficient (MCC = 0.8781), Cohen's Kappa (0.8707), and an overall ROC-AUC score of 0.9992 further highlight the robustness and generalization ability of the model.

4.6. Comparison with Other Multimodal Approaches on the HAM10000 Dataset

In comparison to other multimodal approaches on the HAM10000 dataset, the Fusion Model (Soft Voting Classifier) outperforms the models proposed in previous works. For instance, the model introduced in [9] achieves an accuracy of 86%, combining EfficientNet-B1 with an MLP to process both image features and patient metadata. Similarly, the model in [13], utilizing DenseNet-121 and Vision Transformers, reports an accuracy of 86%. Additionally, the multimodal approach proposed in [14] employs pre-trained CNNs to extract complex patterns and features from the HAM10000 metadata and an extra 100 Squamous Cell Carcinoma (SCC) images. This approach combines CNNs for image data processing with Long Short-Term Memory Networks (LSTMs) to capture temporal connections in textual metadata, achieving a notable accuracy of 92%.

However, our Fusion Model achieves a significantly higher accuracy of 95.9%, demonstrating its superior performance by 9.9% compared to [9] and [13]and by 3.9% compared to [14]. Table 8 below summarizes the accuracy achieved by each model evaluated in this study and provides a comparison with models from previous work. By skillfully integrating image characteristics with patient information, the fusion model tackled disparities

and subtleties within the dataset, resulting in enhanced performance over standalone models and other multimodal models.

5. Limitations

While the proposed multimodal model demonstrates strong performance, several limitations must be acknowledged. First, the model was trained and evaluated solely on the HAM10000 dataset, which may limit its generalizability to broader clinical populations or other imaging modalities. Second, the model shows relatively lower performance in classifying underrepresented classes such as dermatofibroma (df), which affects the robustness across all lesion types. Third, the metadata features although useful, may carry inherent demographic or sampling biases that were not explicitly corrected in this study.

Despite these limitations, the proposed approach establishes a solid foundation for future research in multimodal skin lesion classification. Future work could address these issues by evaluating the model on diverse datasets, exploring advanced imbalance-handling techniques (e.g., focal loss), and incorporating fairness-aware modeling to reduce metadata bias.

6. Conclusion

The present work introduces and evaluates a novel multimodal Fusion Model for skin cancer classification using the HAM10000 dataset. By integrating both image-based features, extracted with Inception-ResNet-v2, and patient metadata, processed using a Soft Voting Classifier, the model achieved an impressive accuracy of 95.9%. Additionally, it demonstrated excellent performance across various evaluation metrics, including ROC-AUC (0.9950), AUC-PR (0.9996), Matthews Correlation Coefficient (MCC = 0.8781), and Cohen's Kappa (0.8707), showcasing its robustness and reliability in classifying skin lesions.In comparison to previous multimodal approaches, such as [9],[13], and [14], the present work significantly outperforms these models. The results highlight the effectiveness of combining image and metadata features, demonstrating superior classification accuracy and handling of challenging classes. The Fusion Model proves to be a promising approach for accurate and reliable skin cancer classification, contributing to the ongoing efforts to improve early diagnosis and treatment of skin cancer.

7. Future Work

Future work could focus on exploring additional techniques to handle the existing class imbalance within the dataset. This could involve the application of advanced data augmentation, resampling methods, or more sophisticated loss functions to better balance the classes.

Introducing a dedicated validation set during training could help mitigate overfitting and improve model generalization, leading to more robust results on unseen data. Furthermore, investigating additional ensemble methods, such as boosting or hybrid approaches, may further enhance classification performance.

Finally, expanding the dataset by incorporating additional data modalities or clinical features could provide even richer insights into skin cancer classification, further strengthening the model's potential for real-world clinical applications.

REFERENCES

1. M. Ciażyńska, G. Kamińska-Winciorek, D. Lange, B. Lewandowski, A. Reich, M. Sławińska, M. Pabianek, K. Szczepaniak, A. Hankiewicz, M. Ułańska, and J. Morawiec, "The incidence and clinical analysis of non-melanoma skin cancer," *Scientific Reports*, vol. 11, no. 1, p. 4337, Feb. 2021.

2. C. M. Abdalla, J. A. Sanches, and R. R. Munhoz, *Oncodermatology: An Evidence-Based, Multidisciplinary Approach to Best Practices*. Springer Nature, Jul. 2023.
3. J. Wang, H. Zhu, S. H. Wang, and Y. D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351–380, Feb. 2021.
4. H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, Jul. 2021.
5. W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of skin cancer based on skin lesion images using deep learning," *Healthcare*, vol. 10, no. 7, p. 1183, Jun. 2022.
6. Z. Lan, S. Cai, X. He, and X. Wen, "Fixcaps: An improved capsules network for diagnosis of skin cancer," *IEEE Access*, vol. 10, pp. 76 261–76 267, Jun. 2022.
7. J. Höhn *et al.*, "Integrating patient data into skin cancer classification using convolutional neural networks: systematic review," *Journal of Medical Internet Research*, vol. 23, no. 7, p. e20708, Jul. 2021.
8. C. Ou *et al.*, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, p. 1029991, Oct. 2022.
9. S. Tajjour, S. Garg, S. S. Chandel, and D. Sharma, "A novel hybrid artificial neural network technique for the early skin cancer diagnosis using color space conversions of original images," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 276–286, Jan. 2023.
10. Q. Chen, M. Li, C. Chen, P. Zhou, X. Lv, and C. Chen, "Mdfnet: application of multimodal fusion method based on skin image and clinical data to skin cancer classification," *Journal of Cancer Research and Clinical Oncology*, vol. 149, no. 7, pp. 3287–3299, Jul. 2023.
11. P. A. Lyakhov, U. A. Lyakhova, and D. I. Kalita, "Multimodal analysis of unbalanced dermatological data for skin cancer recognition," *IEEE Access*, vol. 11, pp. 131 487–131 507, Nov. 2023.
12. H. El-khatib, A. M. Ştefan, and D. Popescu, "Performance improvement of melanoma detection using a multi-network system based on decision fusion," *Applied Sciences*, vol. 13, no. 18, p. 10536, Sep. 2023.
13. A. K. Gairola, V. Kumar, and A. K. Sahoo, "Mlff-net: a multi-model late feature fusion network for skin disease classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 3, pp. 1906–1914, Sep. 2024.
14. A. Khan, N. Chandran, and D. R. Gangodkar, "Integration of multimodal data sources for enhanced skin disease classification and cancer prediction: A study leveraging pre-trained models on ham_10000 metadata and squamous cell carcinoma (scc) images," 2024, unpublished.
15. P. Tschandl *et al.*, "Ham10000: A large dataset of multi-source dermatoscopic images of skin lesions," https://doi.org/10.7910/DVN/DBW86T, 2018, accessed: 2024.
16. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017.
17. D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992.
18. Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, Jun. 2012.
19. L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

Table 1. Summary of Multimodal Approaches for Skin Cancer Classification

Study	Model(s) Used	Dataset	Accuracy
[8]	ResNet-50 and MLP with self-attention and cross-attention	PAD-UPES-20	76.8%
[9]	EfficientNet-B1 and MLP with color space transformations	HAM10000	86%
[10]	VGGNet-19, ResNet-50, DenseNet-121, Inception-V3 and FC layer with attention mechanisms	PAD-UPES-20	80%
[11]	DenseNet-161, Inception-v4, ResNeXt-50 and MLP with modified cross-entropy loss	ISIC Archive	85.2%
[12]	Ensemble of CNNs (DarkNet-53, DenseNet-201, GoogLeNet, Inception-V3, InceptionResNet-V2, ResNet-50, ResNet-101, Xception) using weighted decision fusion	DermIS, ISIC 2020	89% / 82%
[13]	DenseNet-121 and Vision Transformers (ViTs) fused via Multi-Receptive Field Feature Fusion Block (FFB)	HAM10000	86%
[14]	CNNs and LSTMs to process images and textual metadata from multimodal sources	HAM10000 + 100 SCC images	92%

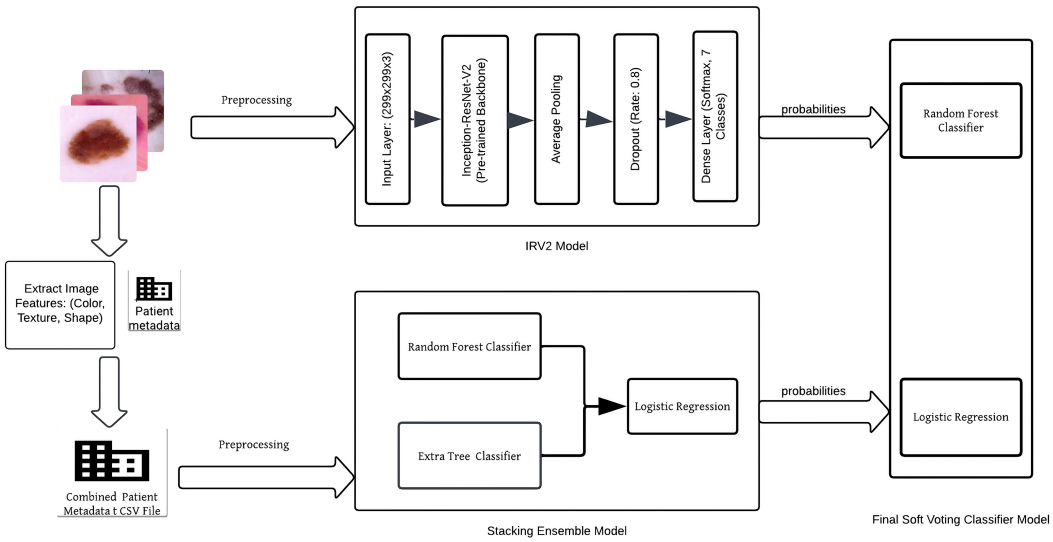

Figure 1. The proposed model architecture.

Table 2. Distribution of images across classes in training and testing after stratified split and data augmentation.

Category	Training Data	Testing Data
Vasc	7096	10
Df	5877	6
Mel	7903	34
Bkl	7931	66
Akiec	6992	23
Nv	8042	663
Bcc	7858	26
Total	51699	828

Table 3. Summary of the created metadata using image features and patient information.

Column Name	Description
lesion_id	Unique identifier for each lesion.
image_id	Unique identifier for each image.
dx	Diagnosis label (e.g., 'bcc', 'mel', 'nv').
dx_type	Method used for diagnosis (e.g., 'histo').
age	Age of the patient.
sex	Sex of the patient ('male' or 'female').
localization	Location of the lesion on the body (e.g., 'scalp', 'back').
HSV_mean	Mean value of the HSV color space for the image.
HSV_median	Median value of the HSV color space for the image.
RGB_mean	Mean value of the RGB color space for the image.
RGB_median	Median value of the RGB color space for the image.
YCbCr_mean	Mean value of the YCbCr color space for the image.
YCbCr_median	Median value of the YCbCr color space for the image.
contrast	Contrast value extracted from the Gray Level Co-occurrence Matrix (GLCM).
dissimilarity	Dissimilarity value from the GLCM.
homogeneity	Homogeneity value from the GLCM.
energy	Energy value from the GLCM.
correlation	Correlation value from the GLCM.
ASM	Angular Second Moment (ASM) value from the GLCM.
lbp_0 to lbp_9	Local Binary Pattern (LBP) histogram values for texture analysis.

Table 4. Classification report for the image model, showing precision, recall, F1-score, and support for each class.

Class	Precision	Recall	F1-Score	Support
akiec	0.54	0.65	0.59	23
bcc	0.90	0.73	0.81	26
bkl	0.72	0.52	0.60	66
df	0.40	0.33	0.36	6
mel	0.64	0.41	0.50	34
nv	0.94	0.98	0.96	663
vasc	0.83	1.00	0.91	10

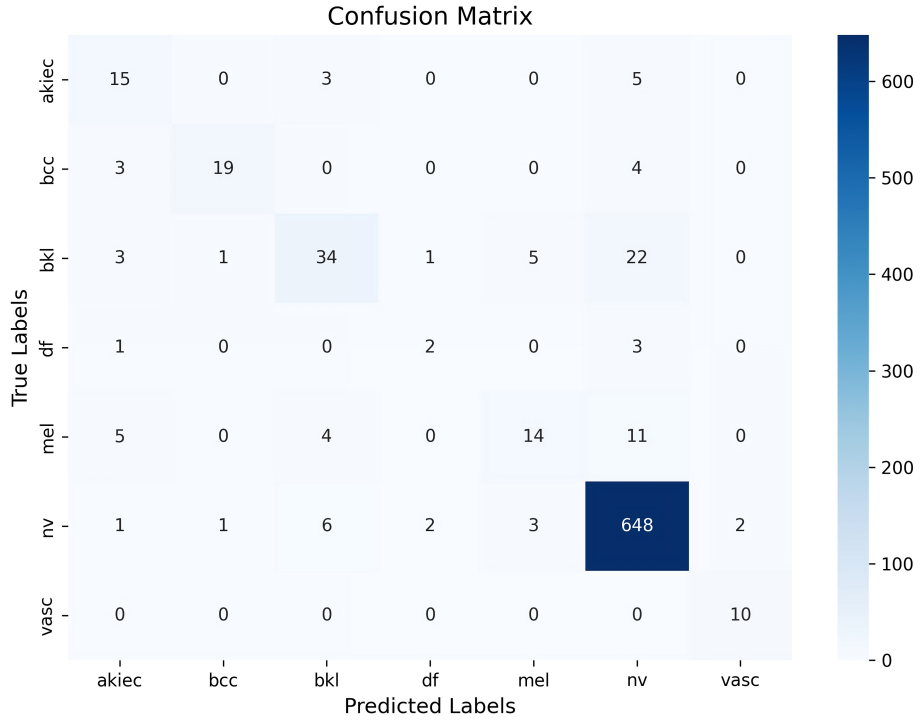

Figure 2. Confusion Matrix for the Image Model, showing the true positives, false positives, and false negatives for each class.

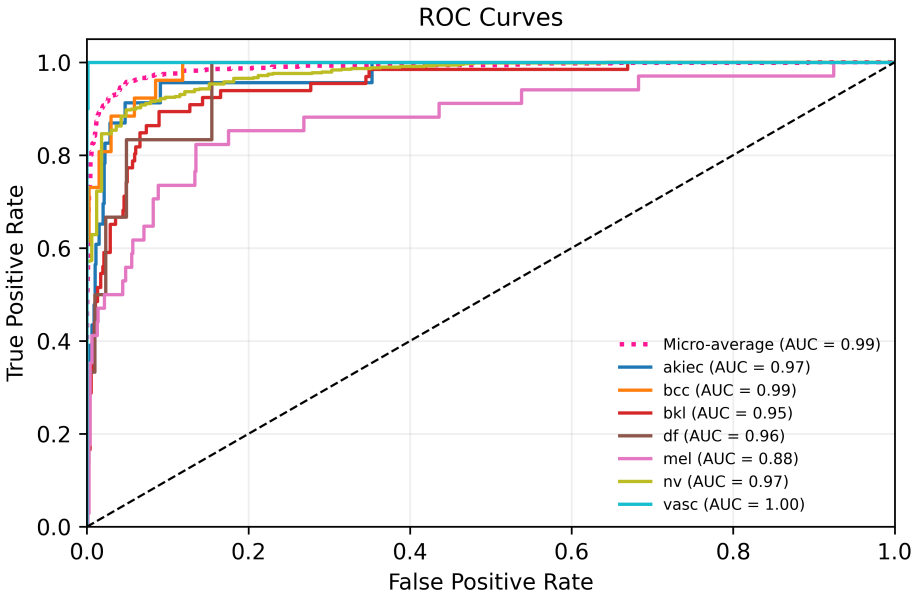

Figure 3. ROC Curve for the Image Model, illustrating the model's ability to differentiate between classes and its overall discriminative power.

Table 5. Classification report for the stacking model combining Random Forest, Extra Trees, and Logistic Regression classifiers.

Class	Precision	Recall	F1-Score	Support
akiec	0.00	0.00	0.00	23
bcc	0.33	0.19	0.24	26
bkl	0.38	0.21	0.27	66
df	0.33	0.17	0.22	6
mel	0.32	0.24	0.27	34
nv	0.86	0.97	0.91	663
vasc	0.25	0.10	0.14	10

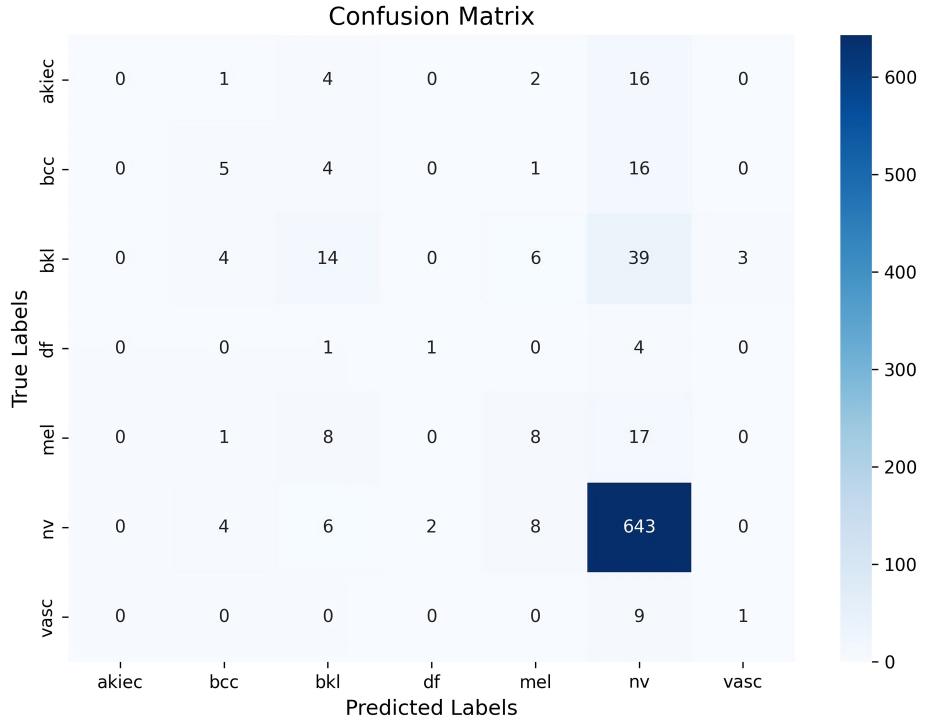

Figure 4. Confusion matrix for the stacking model combining Random Forest, Extra Trees, and Logistic Regression classifiers.

Table 6. Classification report for the fusion model (Soft Voting Classifier).

Class	Precision	Recall	F1-Score	Support
akiec	1.00	0.57	0.72	23
bcc	1.00	0.77	0.87	26
bkl	1.00	0.86	0.93	66
df	1.00	0.83	0.91	6
mel	1.00	0.88	0.94	34
nv	0.95	1.00	0.97	663
vasc	1.00	0.60	0.75	10

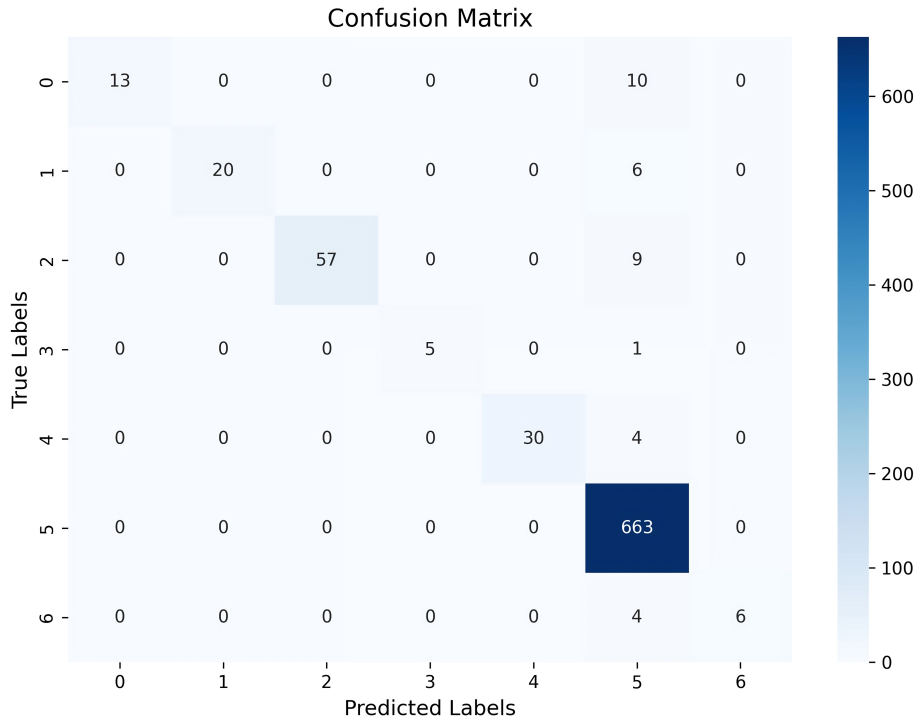

Figure 5. Confusion matrix for the Fusion Model (Soft Voting Classifier).

Table 7. Comparison of the standalone models and the Fusion Model.

Model	**Accuracy**	**Macro Avg. Precision**	**Macro Avg. Recall**	**Macro Avg. F1-Score**
Image Model	90.0%	0.77	0.68	0.72
Stacking Model	81.0%	0.35	0.27	0.30
Proposed Fusion Model	**95.9%**	**0.99**	**0.79**	**0.87**

Table 8. Comparison of the proposed fusion model with literature approaches.

Model	**Architecture**	**Accuracy**
[9]	EfficientNet-B1 (Image) + MLP (Metadata)	86.00%
[13]	DenseNet-121 + Vision Transformer (Image) + Feature Fusion	86.00%
[14]	Pre-trained CNNs (Image) + LSTMs (Metadata)	92.00%
The Proposed Model	Inception-ResNet-v2 (Image) + Stacking (Metadata) + Soft Voting Classifier	**95.9%**