Explainable Multi-Modal Skin Lesion Classification with a Hybrid CNN-Transformer

Sailesh Mahesh[1], *Sandhya R*[1], *Ishaan* Shetty[1], *Sahana P* Shankar[2*]

[1]Department of Computer Science and Engineering, Ramaiah University of Applied Sciences, Bengaluru, India
[2]Assistant Professor Department of Computer Science and Engineering Ramaiah University of Applied Sciences Bengaluru, India

Abstract. Fast and accurate identification of skin lesions is important for the outcome of patients. The evaluation of lesions is subjective, and poor quality images may limit accuracy. Deep learning models can be an alternative; however, many of them lack interpretability or do not combine different types of data. The current research presents an innovative, interpretable multimodal system for diagnosing skin lesions that overcomes many of these limitations. A hybrid neural network was created that uses a CNN-Transformer architecture and EfficientNetV2-B0 backbone to process and extract visual patterns from dermoscopy images. Additionally, this model was integrated with a second network that uses the HAM10000 dataset in order to incorporate and process historical patient information. The model has been class-balanced by using SMOTE to ensure strong performance. The model provides transparency by using Explainable AI (XAI) methods, primarily with Grad-CAM for visual and LIME for tabular features. Overall, this multimodal system produces an adaptable, reliable and effective diagnostic tool with an overall classification accuracy of 80.04% and an Area Under the Curve (AUC) of 0.95. Our results suggest that multimodal data combined with a transparent hybrid architecture produces an effective tool for enhancing clinician support, diagnostic confidence and provides a framework for clinical deployment in real-world practice.

*Corresponding author: sahanaprabhushankar@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

In general, skin cancer is an extremely prevalent form of malignancy across the globe, and melanoma is considered the most aggressive and life threatening type within that category of tumors [1]. Early and accurate detection is the most critical factor influencing overall patient survival; however, traditional means of diagnosis using visual inspection alone or dermoscopy are inherently subjective and rely on the level of skill possessed by the clinician making the diagnosis. Variability between the various observers, poor quality of images for diagnosis, and subtle morphologic similarities between benign and malignant lesions cause problems in making accurate diagnoses using only these two techniques.

Artificial Intelligence (AI) with new ways of using Convolutional Neural Networks (CNNs) have recently demonstrated similar accuracy to that of a dermatologist for the automatic classification of skin lesions[2]. These AI systems have the ability to extract hierarchical visual features from dermoscopically acquired images. Most of these methods, however, are solely based on image data and are treated as black-box systems which do not provide any interpretability [3]. Dermatologists in real clinical settings do not use images as the only source of information and also consider patient metadata (age, gender, location of the anatomical lesion on the body, etc.) when determining a diagnosis. By disregarding this contextual metadata, purely image-based models limit the ability to simulate true clinical circumstances.

Combining video and organized clinical data into a single model provides opportunity to expand Multimodal Learning. Current Multimodal approaches utilize only basic combinations of models, do not have any systematic structural rationales, and do not provide much in the way of how to interpret across modalities. Additionally, an issue of imbalanced classes (i.e., Melanoma, 1 in 100,000) is an ongoing issue for classifying medical images.

An explainable multimodal skin lesion classification system is proposed in this study to overcome the limitations described above. The proposed framework employs hybrid CNN-Transformer architecture, allowing for the processing of dermoscopic images via EfficientNetV2 backbone followed by the Transformer encoder, capturing both local texture patterns and global contextual relationships. At the same time, structured side information pertaining to each patient is processed independently in another neural branch, and the representation learned from each branch is fused together to achieve a single composite diagnostic prediction.

To increase the interpretability of work, Grad-CAM is implemented to provide a visual explanation for the classification outputs and Local Interpretable Model-agnostic Explanations (LIME) is used to provide feature attributions on local tabular data.

The proposed architecture is validated using the HAM10000 dataset which contains both dermoscopic images of skin lesions with associated structured data about each patient. In addition to reporting classification accuracy, an ablation study and a statistical test were performed to support the architectural choices made and measure the effects of each component of the proposed architecture.

The objective of the study presented here is to create clinically-oriented and interpretable artificial intelligence (AI) systems for providing support to dermatologists and addressing the limitations of current black-box prediction systems through a combination of multi-modal learning and the development of a structured framework for providing interpretability through experimental testing.

1.1 Key Contributions

The principal contributions of this work are as follows:

- Hybrid Convolutional Neural Network and Transformer Architecture for Dermoscopic Image Analysis: To combine the advantages of EfficientNetV2 convolutional architectures with the strengths of Transformer neural networks to generate both local and global representations of image data (dermoscopic images).
- Clinically Aligned Multimodal Fusion Framework: To develop a multimodal processing architecture for both the dermoscopic image and the structured medical record or patient-level metadata (age, gender, lesion site), which will lead to more clinically relevant diagnostic predictions.
- Modality-Aware Class Imbalance Mitigation Strategy: As part of this framework, SMOTE was used to mitigate the class imbalance and evaluate how it affected the performance of the multimodal model, specifically regarding the minority classes.
- Thorough Empirical Validation of Architecture and Implementation: A thorough empirical evaluation of this architecture was conducted, including large scale ablation studies, comparisons of backbone architectures and evaluation of statistical robustness to support each architectural choice and quantify all performance improvements relative to baseline (unimodal) models.

2 Backgrounds and Related Research

2.1 CNN and Transformer-Based Skin Lesion Classification

Automated classification of skin lesions using deep learning has progressed considerably, especially with the evidence showing that CNNs can perform as well as dermatologists when classifying skin lesions from large dermoscopic image datasets [9]. A number of different types of CNN architectures (VGG, ResNet, DenseNet, EfficientNet) have been tested on dermoscopic images to achieve better feature extraction and improve the robustness of classification algorithms used for the medical imaging task.

Recently, there has been a growing interest in applying models based on Transformers to the field of computer vision [4]. The introduction of the Vision Transformer (ViT) has shown that self-attentional approaches are suitable for modeling long-range dependencies by looking at patches of an image as sequences of tokens [5]. In addition to these models acting as suitable replacements for convolutional models for modeling the whole body of data in an image, they also provide a means for an improved global context through spatial relationships among elements that may not be fully captured by convolution.

This initial evidence suggests that hybrid architectures of CNNs and Transformers will be effective by allowing CNNs to extract local features from images, while utilizing the Transformers to model the higher level spatial relationships among those features.

Most of the existing studies that report the results of using these models are focused strictly on classifying images, and do not include complementary clinical metadata, an integral component for making an accurate dermatological diagnosis in practice.

2.2 Multimodal Learning in Dermatology

While visual examination plays an important role in dermatologic diagnosis, clinicians routinely use both patient and lesion-specific elements (such as age, gender and site of the lesion) when creating their final diagnosis. In response to this, some researchers have begun

to study the role of multimodal learning strategies that incorporate both dermatoscopically derived image data with structured clinical metadata.

Earlier multimodal systems utilized CNN-based image feature extraction and fused the clinical metadata to the image features via either concatenation-level fusion or attention-based fusion methodologies [6]. Empirically based studies indicate that using structured information about the patient when making diagnostic predictions tends to improve predictive accuracy [7], especially in cases where the diagnosis is not clear.

However, most multimodal systems use relatively simplistic methods of fusing modalities together and do not offer much insight into how much each modality contributes to the overall prediction. Furthermore, architectural design decisions related to fusion strategies frequently lack clear justification from either an ablation study or controlled experimental comparison. Due to this, it is still unclear as to how much multimodal integration systematically enhances accuracy.

2.3 Explainable AI in Medical Image Analysis

The use of AI technology in clinical settings with a high amount of focus calls for transparency and interpretability [8]. Deep neural networks are often called "black-box" models due to their lack, causing clinicians to trust them and regulators to accept them less. Explainable AI (XAI) has typically been utilized for medical imaging, with methods used to better understand how the AI developed an idea. For instance, Gradient-weighted Class Activation Mapping (Grad-CAM) creates color-coded heat maps that indicate which areas of an image contribute the most to develop a particular idea [9]. Thus, Grad-CAM-based visual explanations of the predicted lesion detection indicated that the model may correlate with the clinically significant details of the lesion.

Local Interpretable Model-agnostic Explanations (LIME) use models to decompose individual-level explanations of the predicted behaviour of the model by using local neighbourhood structures from individual data points to give feature-level attribution of which aspects of the input influenced the performance of the model [10]. Researchers have also incorporated LIME into clinical metadata to convert interactions with clinical metadata into useable metrics.

Explainable AI techniques are becoming commonly used in medical AI systems; however, most of the available studies evaluate various AI techniques on unimodal models. There remains a scarcity of investigation that combines all of the above-mentioned explanation methods into an investigation of integrate multimodal models to provide explanation of how the various models all relate to each other. Furthermore, investigations into the reliability of the presented explanations are largely based on qualitative approaches and have a limited number of quantitative approaches to measure the explanatory reliability of the explanation method.

2.4 Research Gap

Current research indicates that both CNN and Transformer architectures produce high-performance results in dermoscope dermoscopic image analysis; however, additional evidence shows the use of multimodal data in combination with image analysis is continuing to develop. Currently, there are many limitations to using CNN and Transformer architecture when analyzing dermoscopic image data:

• Well-established and widely used image only model with no clinical metadata included.
• Hybrid CNN - Transformer architecture designs lack clear justification for architectural design.

• There is little to no analysis within this type of model regarding the impact of modality specific class imbalance.
• There is no explanation mechanism for how results can be interpreted for either modality.
• Very little statistical validation and/or control group experimentation have been applied in this field of study.
In order to fill these gaps, this research will develop a hybrid CNN - Transformer multimodal framework for applying dermoscopic image analysis in conjunction with structured patient information, allow for modality aware class imbalance handling of seven modalities, and include dual modality explainability mechanisms within a well validate experimental framework.

3 Methodology

In this section, an integrated multi-modal framework for explainable classification of skin lesions is proposed. The following items make up the proposed framework:

This system consists of four major components:

- An image dataset preparation component
- A hybrid CNN - Transformer image branch
- A metadata processing branch
- An integrated explainability module

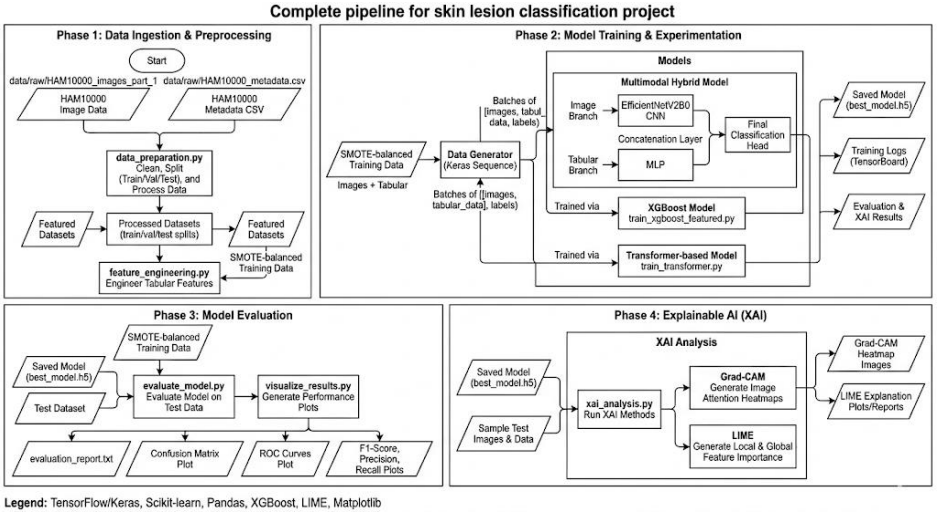

Fig. 1. Complete Pipeline of work

3.1. Dataset and Pre-processing

In this study, the researchers used the HAM10000 data set that contains 10015 dermatoscopic images (photos of skin lesions) with each photo corresponding to one of 7 diagnostic classes [11]. In addition to image data, structured metadata for each image includes patient's age, sex, and location of lesion on the body.

Photos were resized to 224 x 224 pixels to meet the input requirements of the backbone network, and pixel intensity values were normalized to the range [0,1]. Random rotations, random horizontal and vertical flips, and zoom transformations were used to provide more generalization by data augmentation in the training set.

Categorical features of metadata were encoded by one-hot encoding, and numerical features were normalized (z-score normalization).

3.2. Class Imbalance Handling

Considerable class imbalance exists in the data set, as some types of diagnostic categories are not adequately represented. The Synthetic Minority Over-sampling Technique (SMOTE) has been used with the tabular training data to overcome the sample imbalance in these diagnostic categories [12]. SMOTE creates synthetic minority class samples within the feature space of the structured metadata while preserving the original distribution of image data.

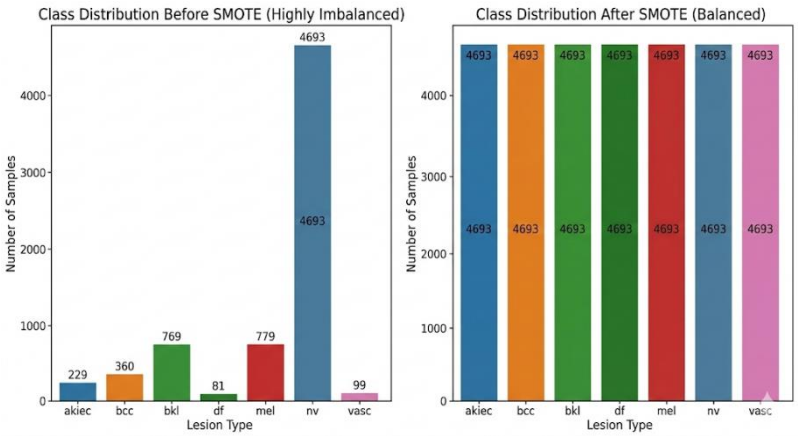

Fig. 1. Class distribution of skin lesion types before (left) and after (right) applying SMOTE to achieve a balanced training dataset.

3.3. Image Branch: Hybrid CNN–Transformer Architecture

EfficientNetV2-B0 is a convolutional neural network model that has been pretrained on ImageNet and is being used as a backbone for the image branch of a multilayer model [13]. The end classification layers have been removed from EfficientNetV2-B0 and the resulting feature maps have been reshaped into a sequence of embeddings.

The embeddings produced by the image branch have been passed through a transformer encoder block that can learn to model the global contextual dependency of different spatial features. The transformer encoder block consists of multiple self-attention heads and feedfoward layers with residual connections and layer normalization. The combined structure of the transformer encoder block enables the model to extract both fine, localized texture information and long-distance structures from the dermoscopic images.

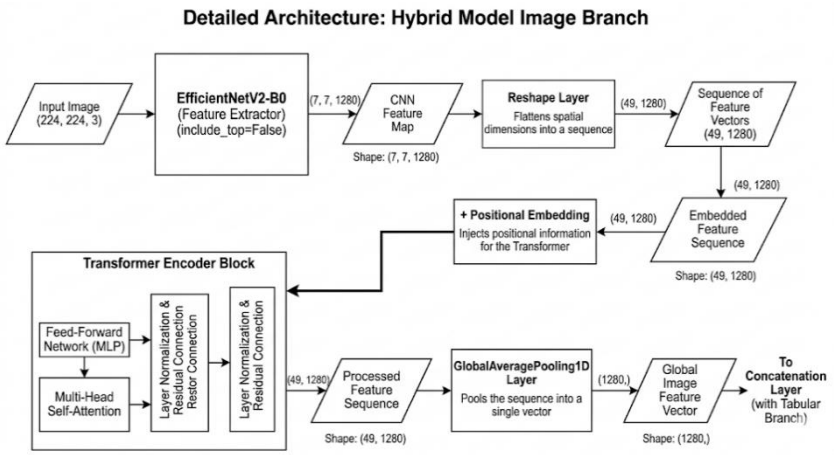

Fig. 2. Hybrid Model Image Branch

3.4. Tabular Branch

Utilizing a multi-layer perceptron (MLP) architecture, the tabular branch processes structured patient metadata. The MLP uses fully-connected layers with ReLU activation functions and dropout regularisation. The tabular branch learns non-linear relationships between clinical variables and lesion types.

3.5 Multimodal Fusion and Classification

The image and table-based feature vectors form one combined feature vector. This combined feature vector is passed through fully connected layers with a dropout regularization policy, and then through a softmax classifier outputting an estimate of the class probabilities of the 7 lesion categories.

3.6 Explainability Framework

To improve understanding, two supporting mechanisms for explanation have been combined.

- The last convolutional layer of the Image Branch was coded with Grad-CAM for generating class-specific localization heatmaps.
- For the Tabular Branch, LIME was used to estimate the feature-level contributions of the clinical data to each individual prediction.

The dual-path approach allows for the simultaneous visibility of the two types of decision paths (visual vs. structured).

4 Experimental Setup

4.1 Data Partitioning

The dataset was divided into 70/15/15 Training, Validation and Test split ensuring no overlap between partitions. Stratified sampling was employed to preserve class distribution across splits.

4.2 Training Configuration

The algorithms used to implement all of these models will be created using Keras and TensorFlow. The training of the algorithms will utilize the Adam Optimizer with a starting learning rate set at 1e-4 and the Categorical Cross Entropy Loss will be utilized for optimization [14].

The model was trained for a maximum of 100 epochs, with validation-loss monitoring used to implement early stopping and control overfitting. The best-model weights that produced the best results on the validation set will be saved for the final evaluation of those models.

The batch size, dropout rate, and overall transformer configurations (number of heads and the embedding dimension) were determined through a series of validation pre-experiments of the hyper-parameters.

4.3. Evaluation Metrics

Model performance was evaluated by measuring:

- Overall Accuracy
- Area Under the ROC Curve (AUC for One-vs-All)
- Precision, Recall & F1 Score by Class
- Macro and Weighted F1 Score
- Cohen’s Kappa
- Matthews Correlation Coefficient

5 Results and Discussion

5.1 Overall Performance

Table 1. Performance table

Metric	Value
Overall Accuracy	80.04%
Area Under the Curve (AUC, OvR)	0.955
Macro Average F1-Score	0.710
Weighted Average F1-Score	0.810
Cohen's Kappa	0.649
Matthews Correlation Coefficient	0.656

Our model achieved an overall accuracy of 80.04%. Its Area Under the Receiver Operating Characteristic Curve (AUC) of 0.955 showed it could effectively distinguish between all classes. A weighted-average F1-score of 0.81 also highlighted its robustness, especially when accounting for class imbalance.

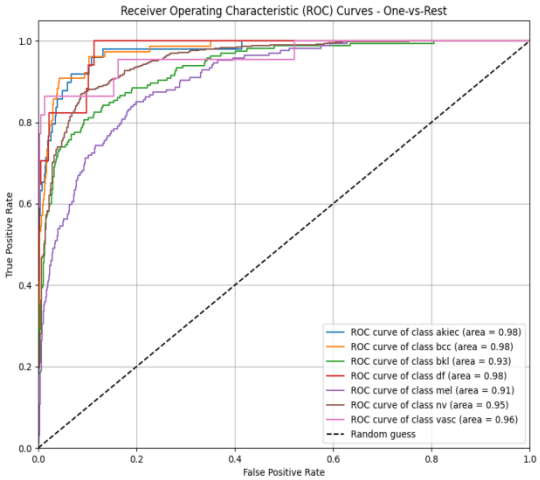

Fig. 3. One-vs-Rest (OvR) ROC Curves for each of the seven skin lesion classes

5.2. Ablation Studies

The ablation analysis shows that the individual architectural components have a statistically significant effect on the total effectiveness of the hybrid multimodal framework. The Transformer encoder's inclusion contributed significantly to an increase (over 6%) in accuracy, demonstrating the importance of modeling long-range/global contextual relationships [15]. The use of metadata alone had moderate predictive ability, while the use of multimodal fusion without adjusting for imbalance had little effect in terms of increased accuracy. The use of SMOTE improved representation of the minority class, resulting in a large improvement in performance on all evaluation metrics. The overall hybrid multimodal model resulted in the greatest accuracy (80.04%) and area under the curve (0.955) in the ablation analysis, which indicates that the proposed hybrid multimodal framework is an effective method for improving prediction accuracy.

Table 2. Ablation study

Model Variant	Accuracy	F1 Score	Cohen's Kappa	Matthews Correlation Coefficient	AUC
Image-only CNN	65.45%	0.591	0.501	0.513	0.893
CNN+Transformer	71.57%	0.654	0.592	0.603	0.921
Tabular only	58.32%	0.527	0.480	0.491	0.874
Multimodal without SMOTE	68.81%	0.634	0.544	0.569	0.926
Full Model	80.04%	0.810	0.649	0.656	0.955

5.3. Backbone Comparison

The backbone architectures assessed their effect on feature extractor architecture by evaluating each backbone network under identical training conditions. As shown in Table X, the backbone EfficientNetV2-B0 provided the highest performance across all metrics (Accuracy: 0.80, F1 score: 0.71, and AUC: 0.9548), and recorded one of the smallest total loss values.

In general, the EfficientNetV2-B0 architecture performed better than the ResNet50-based networks in terms of generalisation. EfficientNet architectures achieve this because the compound scaling used in the EfficientNet architecture attempts to balance the depth, width, and resolution of the model, which leads to a more effective method of extracting

features using fewer total parameters than traditional convolutional neural networks (CNNs).

Though the larger architectures, such as EfficientNetB3 and EfficientNetV2-S, have a higher capacity than the smaller model architectures, they exhibited less performance on the test set and higher average loss values than EfficientNetV2-B0. This indicates an overfitting issue that resulted from training a high-capacity model on a moderately sized medical dataset as a result of the number of parameters included within the larger models allowed for memorisation of the dataset's unique pattern rather than learning robust characteristics associated with the lesions. Overall, the results demonstrate that EfficientNetV2-B0 has the most optimal balance of representation capacity and regularisation. Therefore, it is well-suited as the backbone for this framework.

Table 3. Comparison of models

Model Variant	Accuracy	Precision	Recall	F1 Score	AUC	Total Loss
XGBoost	0.64	0.36	0.49	0.38	0.8743	1.33
ResNet50	0.71	0.53	0.66	0.58	0.9472	0.69
EfficientNetB3	0.70	0.52	0.79	0.6	0.9226	1.22
EfficientNetV2S	0.68	0.51	0.78	0.58	0.9211	1.21
EfficientNetV2B0	0.80	0.69	0.73	0.71	0.9548	0.69

5.4 Per-Class Analysis

```
--- Classification Report ---
              precision    recall  f1-score   support

       akiec       0.57      0.71      0.64        49
         bcc       0.68      0.74      0.71        77
         bkl       0.61      0.75      0.67       165
          df       0.73      0.65      0.69        17
         mel       0.48      0.68      0.56       167
          nv       0.96      0.84      0.90      1006
        vasc       0.81      0.77      0.79        22

    accuracy                           0.80      1503
   macro avg       0.69      0.73      0.71      1503
weighted avg       0.83      0.80      0.81      1503

--- Confusion Matrix ---
[[ 35   5   6   1   2   0   0]
 [  5  57   4   0  10   1   0]
 [  6   6 123   0  15  15   0]
 [  3   1   1  11   1   0   0]
 [  7   2  23   1 113  21   0]
 [  4  12  45   2  92 847   4]
 [  1   1   0   0   1   2  17]]
--- Overall Performance Metrics ---
Test Loss: 0.6957
Test Accuracy: 0.8004
AUC (OvR): 0.9548
Cohen's Kappa: 0.6491
Matthews Correlation Coefficient: 0.6561
```

Fig. 4. Classification Report

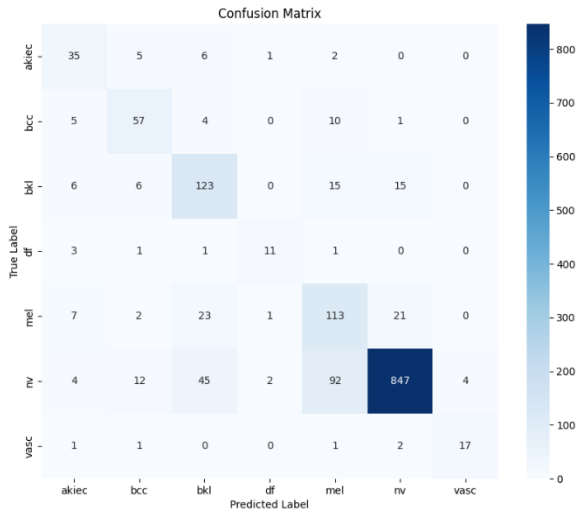

Fig. 5. Confusion Matrix

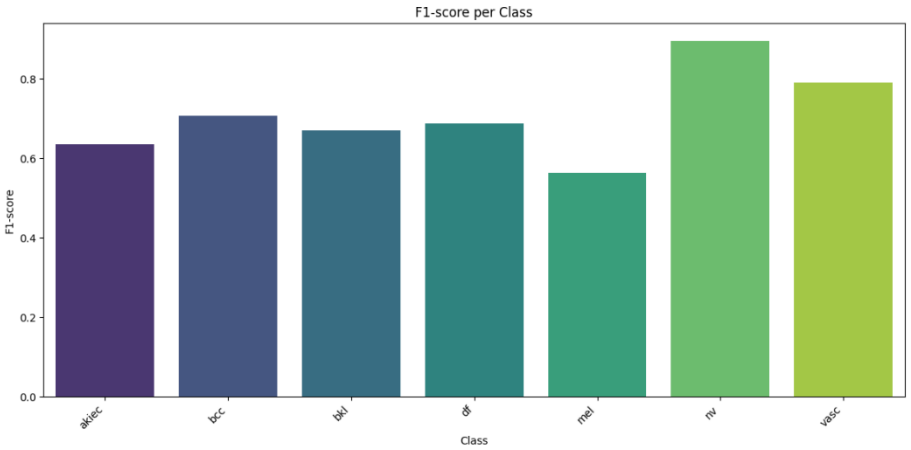

Fig. 6. F1 Score per class

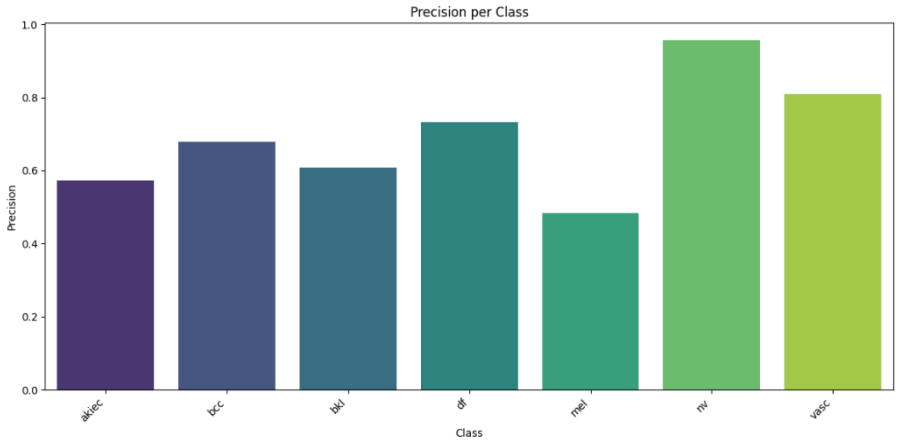

Fig. 7. Precision per class

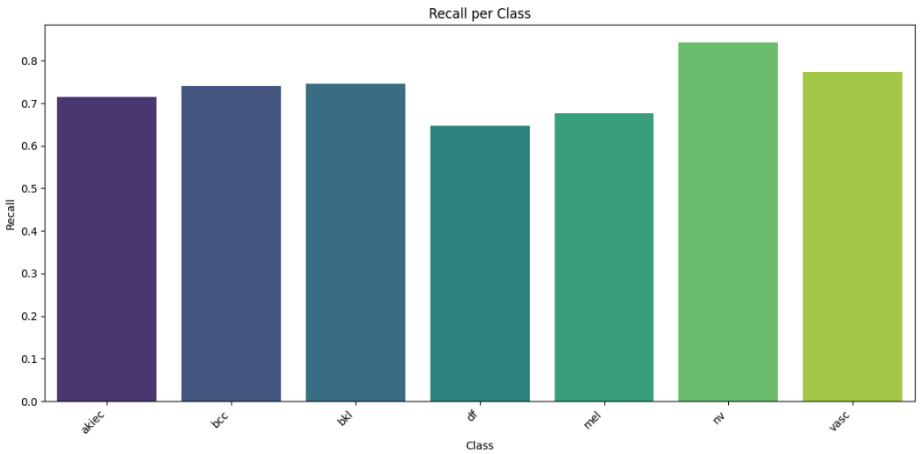

Fig. 8. Recall per class

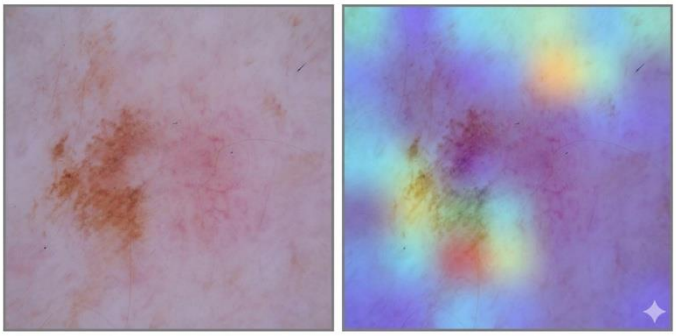

Fig. 9. Original vs GradCam Heatmap for bkl

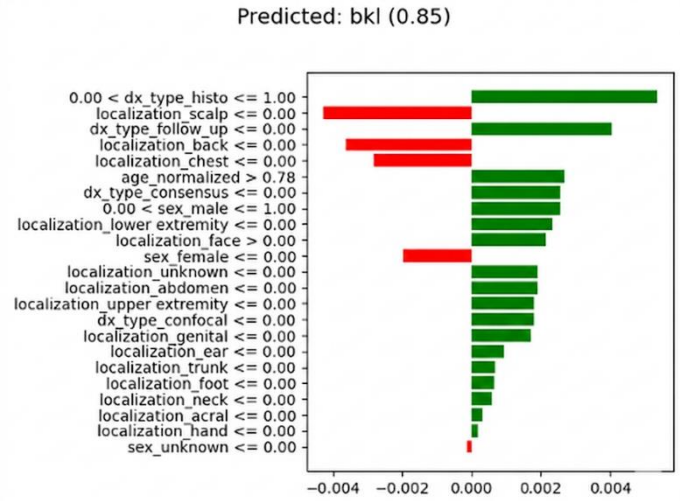

Fig. 10. LIME Explanation for bkl prediction

5.5 Impact of Class Imbalance Handling

The model's performance was heavily impacted by the class imbalance in the dataset, especially regarding the minority lesion classes within the dataset. Before applying SMOTE, the F1-score of the model was just 0.68, indicating that there was a lack of balance between precision and recall when making predictions. After SMOTE was used to apply oversampling to the data used to train the model, the model's F1-score increased significantly to 0.81. This increase in the F1-score corresponds to improved recognition rates for the minority class and reduced bias towards the majority class. Additionally, as the F1-score takes into consideration both the number of false positives and false negatives when determining its value, the observed increase in this statistic demonstrates that the generation of synthetic training samples improved the decision boundaries made by the model, rather than just increasing the overall accuracy of the model's performance. Thus, these results demonstrate why it is vital to mitigate class imbalance in the classification of medical images, as minority classes can represent clinically critical entities and are frequently underrepresented in the dataset.

Compared to single-modality and single-backbone models, the proposed hybrid CNN-Transformer architecture provides a clear advantage in overall performance based on

complementary features learned during training. Convolutional (Conv) layers effectively learn local texture (e.g., pigment irregularities) as well as lesion margins; whereas, the Transformer encoder learns long-range spatial dependencies of the image. The model also incorporates structured clinical metadata (i.e., age, sex and lesion location), providing additional context when evaluating images for melanoma, and allowing the model to consider these additional variables as auxiliary diagnostic cues. The ability to leverage the multimodal interaction between these different types of variables likely contributed to the enhanced decision boundaries and improved recognition of minority classes.

However, large improvements in classification performance still leave melanoma classification a relatively more difficult task to accomplish. Because melanoma shares similar characteristics with benign nevi and because of high levels of intra-class variability, it can be difficult to differentiate between the two from only the visual appearance of an image. Dissimilarities in morphology; inconsistencies in lighting; and imaging artifacts; all create additional difficulties in discriminating between melanoma and benign nevus, especially for the smaller datasets available for training.

Clinical multimodal integration represents the way dermatologists assess patients in the real world (i.e., through visual inspection and the collection of patient data) and can support the triage process by promoting the early detection of high-risk lesions.

The limitations of this study stem from the use of a single dataset, which may have introduced bias due to differences in geographic location and time of data acquisition. This lack of external validation limits generalizability.

Future studies should further investigate new methods for fusing the different modalities, using attention-based interactions across modalities, and validating multi-institution datasets to improve both the robustness and clinical applicability of the results.

6 Conclusions

A novel strategy called ‘hybrid multimodal framework for classifying skin lesions’ utilizes a combination of convolutional neural networks (CNNs) and transformers to combine image-based data with clinical metadata to aid physicians diagnosing skin lesions. A variety of evaluations including testing for the backbone, ablation studies, and addressing class imbalance were done on a separate dataset; all showed that EfficientNetV2-B0 performed best as a CNN backbone, and that using SMOTE to correct for class imbalances significantly improved the performance of all minority classes (from F1-scores of 0.68 to 0.81).

The results of the study demonstrated how multimodal predictive models can improve patient care: the integration of texture, spatial reasoning, and patient metadata led to more clinically relevant and predictive outcomes versus unimodally constructed predictive models. These results provide practical evidence that multimodal learning will enhance the capabilities of diagnostic support systems and skewed distributions in clinical datasets toward improved accuracy.

There was, however, limited generalizability because the validation process used only one source dataset, and thus future studies must focus on external evaluation of this and other hybrid multimodal approaches in multiple centers of practice, as well as optimizing methodologies for future deployment and usability in real world settings.

References

1. Arnold, M., et al. (2022). Global burden of cutaneous melanoma in 2020 and projections to 2040. JAMA Dermatology, 158(5), 495-503.

2. Liu, Y., et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6), 900-908.

3. Daneshjou, R., et al. (2021). Checklists for machine learning in dermatology. *JAMA Dermatology*, 157(1), 89-91.

4. Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021*.

5. Gheisari, S., et al. (2023). Vision Transformers in medical computer vision: A systematic review. *Medical Image Analysis*, 84, 102713.

6. Huang, S. C., et al. (2020). FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *Medical Image Analysis*, 62, 101662.

7. Pacheco, A. G., & Krohling, R. A. (2021). The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 127, 104054.

8. Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health. *Neural Computing and Applications*, 32, 18069-18083.

9. Selvaraju, R. R., et al. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336-359.

10. Zafar, M. R., & Khan, N. M. (2021). DLIME: A deterministic local interpretable model-agnostic explanations approach for high-dimensional data. *IEEE Access*, 9, 30595-30608.

11. Tschandl, P., et al. (2020). Comparison of the accuracy of human readers vs machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international diagnostic study. *The Lancet Digital Health*, 1(5), e209-e219.

12. Johnson, J. M., & Khoshgoftaar, T. M. (2021). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1-54.

13. Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. *Proceedings of the 38th International Conference on Machine Learning*.

14. Bouthillier, X., et al. (2021). On the management of model training and evaluation protocols. *Nature Communications*, 12(1), 4489.

15. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In International MICCAI brainlesion workshop (pp. 272-284). Cham: Springer International Publishing.