AARHUS UNIVERSITY

Multi-Method Analysis for Optimal Skin Lesion Detection: Images, Meta-Data, and Fusion Models

Group 6: Lucie Van Roy, Ginevra Bozza, Grazia Cossu, Saijal Singhal

December 11, 2024

Introduction

Traditional diagnostic method of skin cancer is based on visual inspection. To improve this process, we have implemented various models to classify skin lesion images and metadata. The target labels for classification are benign and malignant. The objective of this project is to discuss the results obtained through three different approaches:

1. **Images-Only Approach**: Uses skin lesion images only, without any additional clinical information
2. **Metadata-Only Approach**: Uses the clinical metadata of the lesion, such as patient demographics and lesion-specific measurements
3. **Fusion-Based Approach**: Combines both image and metadata through multimodal fusion techniques, using both early and late fusion

DataSet

To achieve this, we have chosen the SLICE-3D dataset from the ISIC Archive, created by the International Skin Imaging Collaboration (ISIC). This dataset is particularly well-suited for our project due to its comprehensive and diverse data collection over a span of 10 years from nine leading dermatological centers worldwide. It includes both high-quality lesion images and detailed metadata for each sample, such as patient age, sex, and lesion-specific attributes like color irregularity, border symmetry, and anatomical location. The dataset provides diagnostic labels, classifying lesions as benign or malignant, with strong labels derived from histopathology and weak labels based on clinical judgment.

Methodologies

The project employs different network architectures tailored to the three proposed approaches:

Image-Only Approach

Network structure

This approach utilizes a pre trained model, the *ResNet-18*, a residual network architecture with 18 layers. Convolutional layers are commonly used for image processing because they are translation-invariant meaning that they can recognize patterns in different positions.

Using transfer learning, we apply the knowledge acquired from the weights pre-trained on the ImageNet dataset to our specific task of binary classification of skin images. To achieve this, we freeze the pre-trained parameters and replace the last fully connected layer with a new one designed to fit our output size of two classes.

Training approach

To use transfer learning, we need to apply transformations to make the dataset compatible with the pre-trained model. Specifically, we resize the images and normalize them to match the format expected by the model. To improve generalization, we introduce random horizontal and vertical flips to the images, as convolutional networks are not inherently rotation-invariant.

Before training, we address the issue of class imbalance by downsampling the majority class (class 0 or benign) and assigning specific weights to the loss function. It penalizes more errors on the minority class.

Metadata-Only Approach

Network structure

For this approach, we chose a Multi-Layer perceptron (MLP), because of its efficiency, flexibility and suitability for tabular data. It has an input layer and three fully connected hidden layers, followed by batch normalization, and using ReLU activation function. The output layer is a single neuron with a Sigmoid activation function, producing a probability score.

Training approach

To improve the model performance and avoid overfitting, the data is preprocessed before the training, using the following strategies:

- Downsampling: the dataset is sampled to address the class imbalance problem and equalize the number of positive and negative samples
- Features encoding: the features are encoded using One-Hot encoding (for low cardinality categorical data) or standardized (for numerical data)
- Features selection: the features that are non-relevant are dropped and an importance analysis is performed using a correlation matrix and a Random Forest model, to identify and keep only the most relevant features

Fusion-Based Approach

Network structure

The fusion-based approach combines image and metadata features using both *early fusion* and *late fusion* techniques:

- **Early Fusion**: The model is designed to handle both image and metadata inputs as one. The image data is pre-processed using transformations such as resizing and normalization, and the metadata is encoded numerically, with categorical variables transformed via one-hot encoding. These two types of data are then concatenated and passed through fully connected layers (dense layers). To enhance learning, batch normalization, ReLU activation, and dropout are applied throughout the layers to prevent overfitting.

- **Late Fusion**: The Late Fusion architecture integrated a Multilayer Perceptron (MLP) for processing metadata and a ResNet-18 backbone for image feature extraction. The MLP developed previously was incorporated here. A pre-trained ResNet-18 with its final fully connected layer replaced by an Identity layer, extracting a 512-dimensional embedding. In the fusion stage, embeddings from the MLP (32-dim) and ResNet-18 (512-dim) were concatenated and passed through two fully connected layers (544 → 64 → 1), using ReLU activation and a Sigmoid function at the final layer for binary classification.

Training approach

Early Fusion Model:

The model uses cross-entropy as the loss function and the Adam optimizer. A learning rate scheduler is applied to adjust the learning rate over time to improve convergence during training. Additionally, early stopping is implemented to halt training when there is no improvement in performance, thus avoiding overfitting.

Late Fusion Model: Metadata was preprocessed through steps such as balancing the dataset, feature selection using Random Forest, and feature standardization. Image data underwent transformations like random cropping, flipping, and normalization to augment the training process. Metadata and image data were loaded into PyTorch DataLoader objects using customized dataset classes to ensure alignment between the two modalities through a shared identifier (isic id). The training leveraged Binary Cross-Entropy Loss (BCELoss), with class weights to handle the initial class imbalance. The optimizer used was Stochastic Gradient Descent (SGD) with momentum, helping to stabilize learning. The training followed a typical loop structure involving mini-batch updates for gradient calculation during training phases and performance evaluation on validation data at the end of each epoch.

Results and Comparison

0.1 Metadata-Only MLP Model

The MLP model suing only meta-data performed exceptionally well, achieveing an accuracy of 0.8592 and F1-score of 0.8571. The training and validation loss curves indicate smooth convergence, with minimal overfitting as evidenced by the small gap between the training and validation accuracy curves. This demonstrates that metadata provides structure and meaningful features for the classification task, enabling the model to generalize effectively. Its balance precision and recall further highlight its robustness in both identifying positive cases and avoiding false positives. This further implies that meta-data only is very powerful and provides significant information in skin cancer prediction.

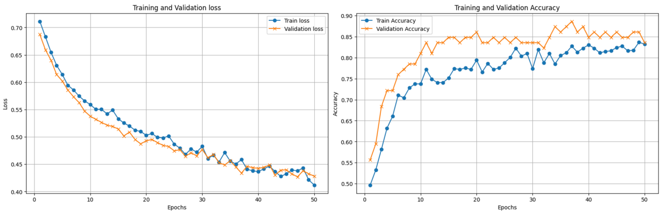


Figure 1: Training v/s Validation Accuracy and Loss for MLP Model

0.2 ResNet-18 Image-only Model

The ResNet-18 mode, trained on image data alone, showed the weakest performance amoung all models, with an accuracy of 0.6962 and an F1 of 0.6907. Training and validation loss curves exhibit considerable fluctuations and the validation accuracy lags behind the training accuracy. This suggests challenges in effective learning from only visual data. There were also considerable fluctuations inclusing sharp peaks and drops, particularly in validation loss. This behavious suggests instability in learning or sensitivity to small data variations. Model's reliance on visual features faces challenges possibly because of the high complexity of the image data and limited number of malignant images avaible in contrast to benign images. Despite this, the model managed a balance precision and recall of 0.7087 and 0.6962, respectively, highlighting its moderate ability to identify both positive and negative cases.

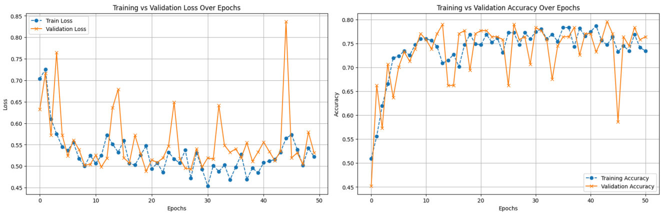


Figure 2: Training v/s Validation Accuracy and Loss for ResNet Model

0.3 Early Fusion Model

The early fusion model, combining meta-data and image features at an early stage, achieved an accuracy of 0.7468 and an F-score of 0.6774. The model showed high precision (0.9130) but struggled with recall (0.5385). This indicates that there is a bias toward correclty classifying the dominant class while missing several positive cases. The validation loss curves diverge from the training loss, suggestig potential overfitting. This reflects the difficulty of effectively fusing heterogeneous data modalities early in the network even though, the model performed early stopping at epoch 14 preventing more overfitting. The model did combine the metadata and image features but it may have underutilized the complementary information from both modalities.

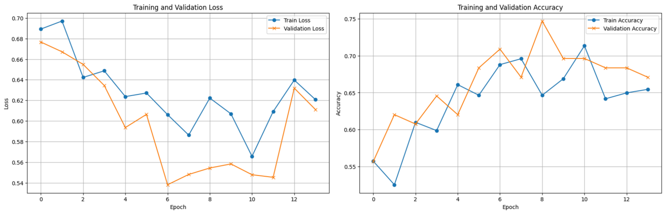


Figure 3: Training v/s Validation Accuracy and Loss for Early Fusion Model

0.4 Late Fusion Model

The late fusion model delivered the best overall performance, with an accuracy of 0.8608 and an F1-score of 0.8642. By processing metadata and images independently before combining their embeddings, the model effectively leveraged the strengths of both modalities. The training and validation curves were smooth and closely aligned, indicating strong generalization and minimal overfitting. It achieved the highest recall (0.8974), demonstrating its ability to detect true positives making it ideal for applications where missing positive cases is critical while maintaining balanced precision (0.8333). The late fusion model outperforms the early fusion model across all metrics, particularly in recall and F1-score, suggesting that separating the processing pipelines for metadata and images before combining them may result in better feature representation and decision-making.

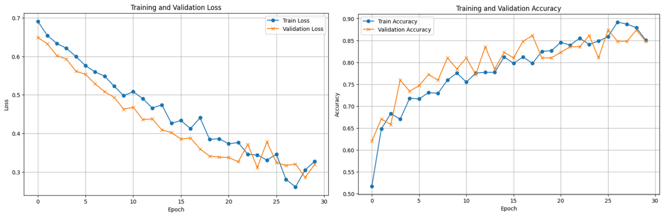
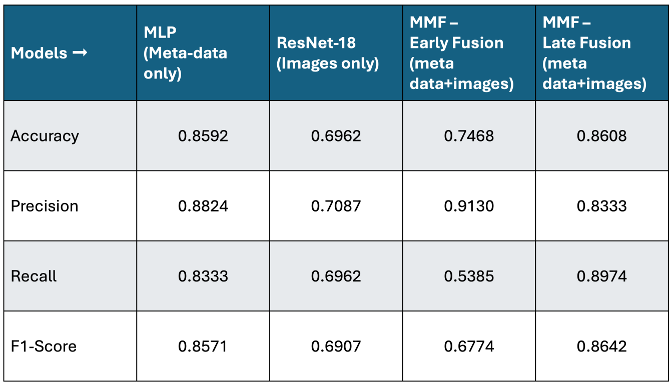


Figure 4: Training v/s Validation Accuracy and Loss for Late Fusion Model

Conclusion

When comparing all four models, it is evident that combining metadata and images yields better performance than using either modality alone. The MLP model excelled in analyzing metadata, achieving high accuracy and precision, showing that metadata provided valuable features. However, the ResNet model, relying only on images, underperformed, likely due to the inherent challenges of extracting robust features from medical images alone.



Models ➞	MLP (Meta-data only)	ResNet-18 (Images only)	MMF – Early Fusion (meta data+images)	MMF – Late Fusion (meta data+images)
Accuracy	0.8592	0.6962	0.7468	0.8608
Precision	0.8824	0.7087	0.9130	0.8333
Recall	0.8333	0.6962	0.5385	0.8974
F1-Score	0.8571	0.6907	0.6774	0.8642

Figure 5: Comparison of Performance Metrics for all models

The fusion models demonstrated the power of multimodal learning. The early fusion model prioritized precision but sacrificed recall, leading to a moderate F1-score. The early stopping suggests that while the model combined features from both modalities, it struggled to capture synergistic relationships effectively. In contrast, the late fusion model successfully combined metadata and image embeddings at a later stage, leveraging their individual strengths. This approach resulted in the highest accuracy, recall, and F1-score, making it the most balanced model.

Although the model achieves high performance, the improvement over the metadata-only MLP model is relatively modest (0.8608 vs. 0.8592 in accuracy, for instance). This suggests that the addition of image data may not always yield significant incremental benefits in cases where metadata is already highly informative. This raises questions about the added value of combining image data with metadata when metadata alone provides sufficient discriminatory power.

In summary, the late fusion model outperformed others by effectively integrating metadata and image information, highlighting the importance of a well-designed fusion strategy for multimodal tasks.