Fine-Grained Skin Disease Classification Enhanced by Multi-Scale Attention Mechanisms

Jaswanth Kranthi Boppana (jboppana@iu.edu)
Pranay Reddy Gundala (vgundala@iu.edu)
Ranvir Singh Virk (Rsvirk@iu.edu)
Venkata Ramakrishna Reddy Dwarampudi (vedwar@iu.edu)

1. Abstract

This project aims to develop an end-to-end skin disease recognition and classification system that integrates convolutional networks and attention mechanisms to enhance the detection and classification of skin lesions, including melanoma and seven other diseases. The system employs segmentation techniques to extract diseased skin regions, enabling classification on both original and segmented images, with a comparative analysis to evaluate their performance. Leveraging spatial, channel, and self-attention mechanisms, along with multi-scale feature extraction and hierarchical classification, the model distinguishes between benign and malignant lesions before identifying specific diseases. Using datasets such as ISIC and HAM10000, this approach aims to improve classification accuracy and interpretability, advancing early diagnosis and clinical decision support. Extensive testing demonstrated that models incorporating attention mechanisms significantly outperformed others. The Inception ResNet V2 with soft attention achieved the highest performance, excelling in both multi-class and binary classification tasks, with an accuracy of 89.60% in identifying malignant lesions, making it the most effective model for clinical application.

2. Introduction

Skin diseases are among the most prevalent health concerns globally, impacting over 900 million people at any given time, with conditions ranging from benign lesions to malignant melanomas. Melanoma alone accounts for 1% of all skin cancer diagnoses but is responsible for the majority of skin cancer-related deaths, emphasizing the need for early and accurate diagnosis [3]. Despite advancements in dermatoscopy and digital imaging, the differentiation of benign from malignant skin lesions remains a significant challenge due to overlapping features, inter-class similarities, and high intra-class variability [2]. Traditional diagnostic approaches, often reliant on clinical expertise, suffer from limitations in scalability and subjectivity, leading researchers to explore automated methods for skin disease classification [1].

Numerous attempts have been made to leverage deep learning for skin lesion analysis. Initial efforts employed basic convolutional neural networks (CNNs), achieving moderate success but struggling with class imbalance and nuanced pattern recognition [3]. More sophisticated approaches, such as ResNet50 with transformers, introduced multi-head self-attention mechanisms to model spatial dependencies, improving performance in handling underrepresented classes [4] . However, even these advanced models fell short of optimal recall for malignant cases—a critical metric for clinical utility. Researchers have also explored attention mechanisms and multi-scale feature extraction, yielded promising results but left room for further improvement.[5]

This work builds upon these advancements by proposing an end-to-end skin diseased recognition system that integrates convolutional backbones with attention mechanisms to address the shortcomings of existing models. Our system incorporates segmentation

techniques to isolate diseased skin regions, enabling a comparative analysis between original and segmented images. By employing spatial, channel, and self-attention mechanisms, coupled with hierarchical classification, the model not only distinguishes between benign and malignant lesions but also identifies specific disease types. Notably, we focus on Task 3 of the ISIC Challenge—disease classification—while also addressing lesion segmentation and attribute detection to create a holistic solution [1].

The ISIC Challenge, a leading benchmark in skin lesion analysis, provides datasets such as ISIC 2018 and HAM10000 for evaluating automated systems [1,2]. This challenge comprises three tasks: lesion segmentation, attribute detection, and disease classification, with Task 3 being the most clinically relevant. Our approach achieves significantly equal performance to the state-of -the-art on Task 3, leveraging the Inception ResNet V2 architecture with soft attention mechanisms to attain an accuracy of 89.60% in malignant lesion classification, the highest among tested models. By advancing interpretability and accuracy, our system holds significant promise for clinical decision support and early intervention in dermatological care

3. Dataset

3.1 Dataset Overview

The **HAM10000 dataset** consists of **10,015 high-resolution dermatoscopic images** of skin lesions categorized into seven distinct classes:

- **nv**: Melanocytic nevi
- **mel**: Melanoma
- **bkl**: Benign keratosis-like lesions
- **bcc**: Basal cell carcinoma
- **akiec**: Actinic keratoses and intraepithelial carcinoma
- **vasc**: Vascular lesions
- **df**: Dermatofibroma

Each image is associated with metadata that includes:

- **Lesion ID**: Unique identifier for skin lesions
- **Image ID**: Unique identifier for images
- **Diagnosis (dx)**: Class label
- **Diagnosis Type**: Histopathological confirmation
- **Patient Age** and **Gender**
- **Localization**: Body location of the lesion

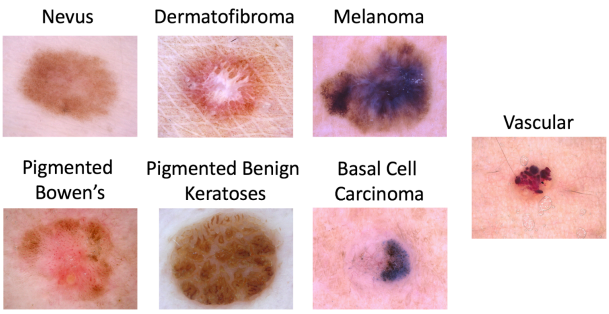

Fig 1: Dataset Image sample

3.2 Dataset Statistics

- **Class Distribution**: The dataset is imbalanced, with the majority of images belonging to the "nv" class. The distribution is visualized in **Figure 5**, where "nv" dominates, while classes like "vasc" and "df" are underrepresented.
- **Age Distribution**: The dataset includes patients aged 0 to 85, with most cases concentrated in the 40-60 age group. **Figure 2** illustrates a right-skewed distribution where middle-aged individuals form the majority of cases.
- **Gender Distribution**: The dataset contains both male and female patients, with slight male predominance, as shown in **Figure 3**. A negligible portion of data lacks gender information.
- **Localization**: Lesions are localized on various parts of the body, including the scalp, face, back, abdomen, and limbs. The "back" and "lower extremities" are the most common locations, as shown in **Figure 4**.
- **Localization vs. Diagnosis**: A heatmap in **Figure** shows the correlation between localization and diagnosis. Notably, melanocytic nevi (nv) are most common on the back and lower extremities.

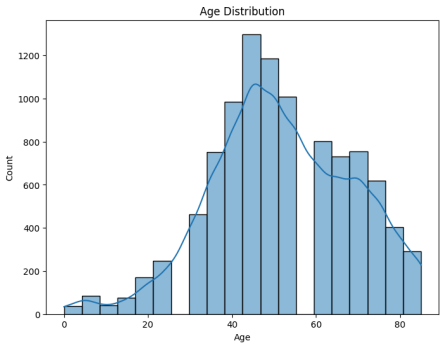

Fig 2: Age Distribution

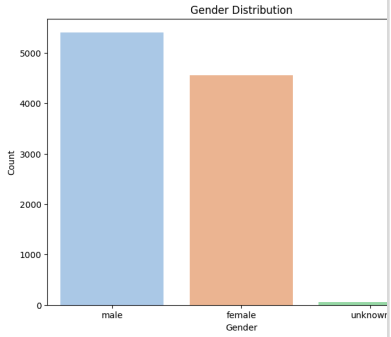

Fig 3: Gender Distribution

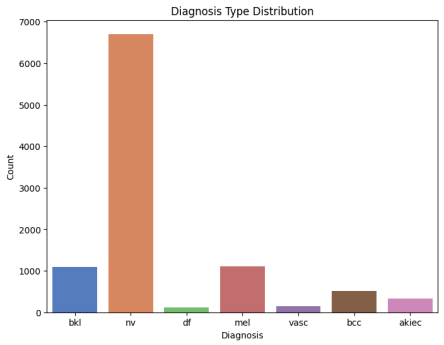

Fig 4: Diagnosis type

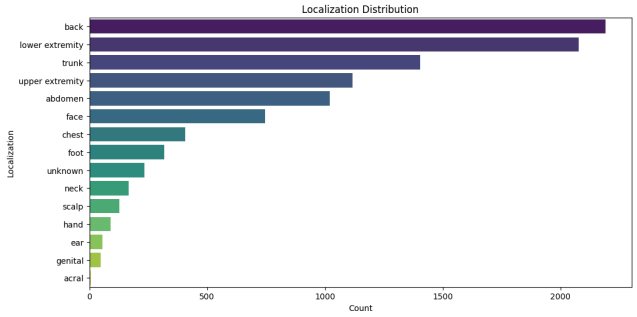

Fig 5: Localization Distribution

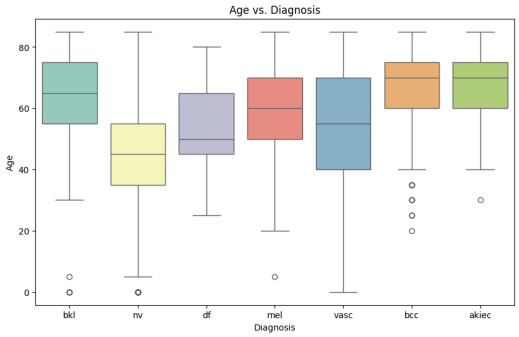

Fig 6: Age vs Diagnosis

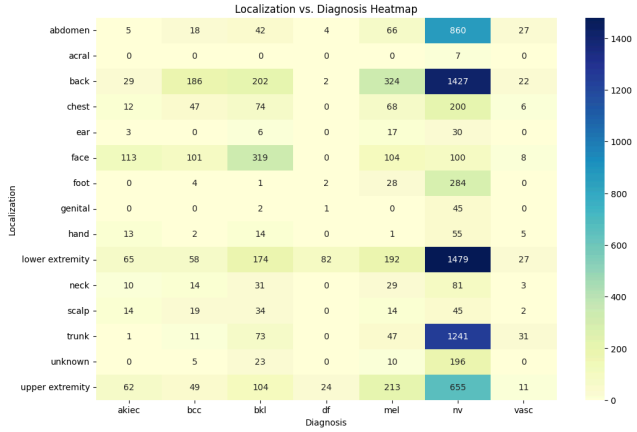

Fig 7: Localization vs Diagnosis Heatmap

4. Methodology

4.1 Problem Formulation

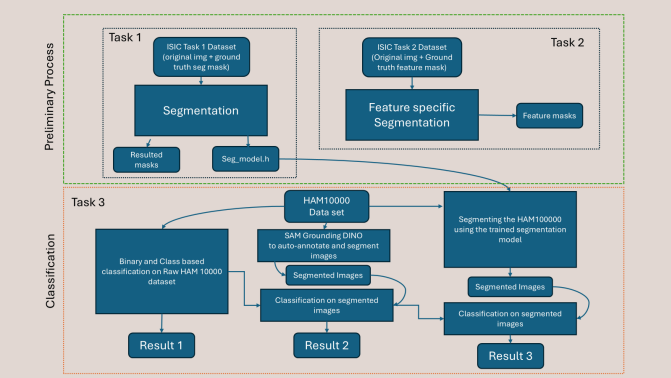

Fig 8: Process FlowChat

As we have discussed in the previous section of this paper we are trying to provide an end to end skin disease detection and classification system. We are doing this by solving the key problems and tasks that are proposed by ISIC challenge . Fig 1 shows the flow of how the problem is approached with a constructive analysis. Lets now dig into the core problem we are trying to solve as a part of each task and understand the proposed approaches. We have divided the problem into stages namely the preliminary process and the Classification. The Preliminary analysis is based on the taks 1 and 2 of the ISIC 2018 Challenge which are the Lesion Boundary Segmentation, Lesion Attribute Detection respectively. We have specific datasets to work up on for each task, for the Lesion Boundary Segmentation we have around 2500 images that are randomly selected from the HAM10000 dataset and its ground truth masks. For Lesion Attribute Detection we have 5 attributes to detect or segment so the data consists of 2594 images and 12970 corresponding ground truth response masks (5 for each image). Once the preliminary analysis is done we use the model that we have developed in task 1 to further segmentation. Here in this project our main focus lies solving the classification part of the problem. We are intended to perform both binary and multi- class classification. For Binary classification we need to define if the skin-disease in the image is melanin or benign. Where as for multi-class classification it is expected to classify the image in to one of the class among the 7 classes which are Melanoma ,Melanocytic nevus , Basal cell carcinoma , Actinic keratosis / Bowen's disease, Benign keratosis , Dermatofibroma, Vascular lesion. So, to implement this we have tried various architectures , methods and different fine tunings techniques. As shown in the Flow chart 3 different flows for classification are proposed.In the next part of the report we will be unfolding and describing various methods that are implemented and their differences compared to existing works. We would like to mention that the ISIC challenge just gave us a flow and dataset to build an end to end skin disease classification system.

4.2 Data Preparation

The original dataset images were initially distributed across two folders. The images were reorganized into **seven class-specific folders** to facilitate training with ImageDataGenerator for deep learning models.

The dataset was split into **training (70%)**, **validation (20%)**, and **testing (10%)** sets using **stratified sampling** to preserve class distributions:

- **Training set**: 7010 images
- **Validation set**: 2003 images
- **Test set**: 1002 images

4.3 Data Preprocessing

4.3.1 Data Augmentation

Data augmentation was applied to the training set to address class imbalance and improve model generalization. The following augmentations were implemented using TensorFlow's ImageDataGenerator:

- **Rescaling**: Pixel values normalized to [0, 1]
- **Rotation**: Random rotation within 20 degrees
- **Zoom**: Random zoom within 15%
- **Width/Height Shift**: ±10%
- **Shear Transformation**: Intensity set to 15%
- **Horizontal Flip**: Random flips for symmetry
- **Brightness Adjustment**: 80%-120% of original brightness

Validation and test sets were rescaled without augmentation to ensure unbiased evaluation.

4.3.2 Class Balancing

To address class imbalance, **class weights** were computed based on the frequency of each class:
Class Weights: {0: 4.37, 1: 2.78, 2: 1.30, 3: 12.36, 4: 1.28, 5: 0.21, 6: 10.11}
These weights were incorporated into the model's loss function to penalize misclassifications in underrepresented classes.

4.4 Preliminary Analysis

Before starting off with classifying the skin diseases it's important to have an overview of how the diseases look and what patterns or findings we can extract and understand. So to do that we initially performed the Lesion Boundary Segmentation. Where we detect the exact boundary of the lesion and create a binary mask for the boundary. Later this boundary can be used to segment out the lesion from the image. To implement this we have used the ISIC 2018 data set consisting of 2500 random original skin lesion images along with its respective ground truth mask. Using this as the data set we have implemented a classic Unet Architecture and a ResNet50 model which uses the pre-trained weight of imageNet to train and detect the lesion boundaries. Fig 8 and Fig 9 Shows the comparison of ground truth vs the Predicted lesion boundaries.

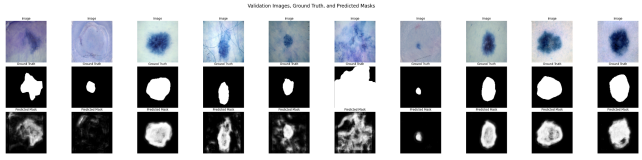

Fig 9: UNet model results

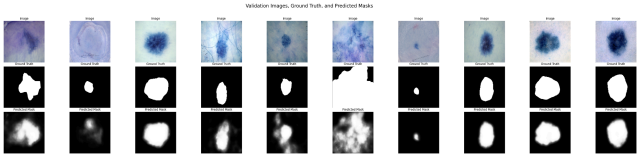

Fig 10: ResNet50 based UNet model results

We are evaluating the performance of these models using 5 important metrics such as Jaccard Index, Sensitivity, Specificity, Dice Coefficient , Validation accuracy. For our task of defining the lesion boundary to test whether it is benign or malignant we need to Focus on **Dice Coefficient** as the primary metric , Use **Jaccard Similarity** for stricter evaluations, and Monitor **Sensitivity** to ensure no parts of the lesion are missed.

Comparison table:

Evaluation Metrics	ResNet50 (ImageNet pretrained)	Unet
Dice Score	0.85	0.80
Jaccard Similarity	0.70	0.57
Sensitivity	0.95	0.70
specificity	0.92	0.98
Val_Accuracy	0.91	0.88

Table 1: ResNet50 based UNet vs UNet evaluation

Apart from identifying the lesion boundary, if we want to identify and categorize the lesions it's important to know what attributes are needed to make things understood. So we tried to perform the Lesion Attribute Detection. Where a *DeepLabV3* model is trained using the data of 2500 images which have 5 ground truth feature masks. The features that we are trying to check for are *pigment network, negative network, streaks, milia-like cysts, globules.*

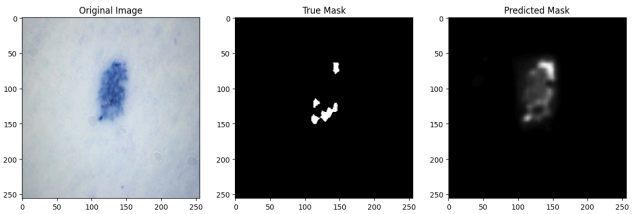

Fig 11: prediction example of globules Feature.

This implementation helps us to see what specific disease attributes exist under each given image. It plays an important role in getting a specific overview on the existence of the disease. If none of the features are highlighted in this analysis we can have a significant chance that the skin cancer is benign. Similarly it also acts as an initial understanding for multi class calcification.

So, this preliminary analysis has decent importance in getting an overview of what a random lesion is looking and how worst it could end up to be.

4.5 Classification Models

After the preliminary analysis of the images we are starting off with implementing classification methodologies. Here we are training models completely on raw images from the HAM10000 dataset. No segmentation is involved in this training process. Ideally we are performing 2 types of classifications one is the multi-class classification where we are classifying the images to 7 - different classes and the other is the binary classification which segregates the images into benign or malignant. We would like to put up various models that we implemented on this classification task.

Multi class Classification :

I. Model 1: Basic CNN

- **Architecture:** The basic CNN consisted of four convolutional layers, each followed by batch normalization and max-pooling to reduce spatial dimensions while preserving key features. The model ended with fully connected dense layers and a softmax output layer for multi-class classification. Dropout was used for regularization, and the Adam optimizer was employed for efficient training**.**
- **Motivation:** We started with this straightforward architecture to establish a baseline and observe how well fundamental convolutional operations could extract features for this dataset. This allowed us to gauge the dataset's complexity and identify areas where more advanced techniques could help.
- **Results:**
  - **Validation Accuracy**: 52.5%
  - **F1-Score**: 60.3%
  - **Recall**: 61.9% While the model captured some basic patterns, its performance on underrepresented classes was limited. This reinforced the need for more complex architectures to handle class imbalance and intricate patterns.

II. Model 2: ResNet50 with Transformer

- **Architecture**: This model leveraged ResNet50, a deep residual network pre-trained on ImageNet, as its backbone for feature extraction. The ResNet50 output was reshaped into a sequence format and processed by a transformer block, which applied multi-head self-attention to understand spatial relationships across the image. A final classification head with dense layers and a softmax activation was used for predictions.
- **Motivation**: By combining ResNet50's proven feature extraction capabilities with the transformer's ability to model long-range dependencies, we aimed to improve the model's ability to differentiate between subtle class differences. This was especially important for imbalanced datasets, where underrepresented classes require more contextual awareness.
- **Results:**
  - **Validation Accuracy:** 79.4%
  - **F1-Score:** 78.9%
  - **Recall:** 79.4%

  The addition of the transformer significantly boosted the model's ability to generalize, particularly for minority classes, as reflected by the higher F1-score and recall.

III. Model 3: Inception ResNet V2 with Soft Attention

- **Architecture:** This model combined the Inception ResNet V2 backbone with a custom soft attention mechanism. The attention layer was applied to the final convolutional block, generating attention maps to highlight important spatial regions. These attention-enhanced features were passed through global average pooling, dense layers, and a softmax output layer for classification.
- **Motivation:** Soft attention was introduced to help the model focus on the most relevant parts of the image for each class. This dynamic weighting of spatial features was particularly useful for distinguishing subtle patterns that may otherwise be overlooked, especially for imbalanced classes**.**
- **Results:**
  - **Validation Accuracy**: 82.5%
  - **F1-Score**: 83.3%
  - **Recall**: 82.5%

  This model demonstrated the benefits of attention mechanisms, outperforming the ResNet50 + Transformer model and achieving the highest recall among all models.

IV. Model 4: ResNet50 with Advanced Soft Attention

- **Architecture:** Similar to Model 2, this model used ResNet50 as its backbone but incorporated an advanced multi-head soft attention mechanism. This allowed the model to generate multiple attention maps, each capturing different aspects of the input image. The attention-modulated features were aggregated and concatenated with the original convolutional features before being passed through pooling layers and a classification head**.**
- **Motivation:** By learning multiple attention maps, the model aimed to extract diverse and complementary features, improving its ability to handle complex and subtle patterns in the dataset.

This was particularly beneficial for identifying underrepresented classes more accurately.

- **Results:**
  - **Validation Accuracy:** 80.8%
  - **F1-Score:** 80.3%
  - **Recall:** 80.8%

  This approach delivered robust performance and validated the effectiveness of attention mechanisms, achieving results comparable to the Inception ResNet V2 model.

Binary Classification:

To simplify the multiclass classification problem, we merged all lesion types into two categories: benign and malignant. This binary classification task aimed to evaluate how well the models distinguish between cancerous and non-cancerous lesions. The same three models used for multiclass classification—Basic CNN, ResNet50 with Transformer, and Inception ResNet V2 with Soft Attention—were adapted for binary classification. Below, we discuss metrics in detail.

I. Model 1: Basic CNN

- **Results**:
  - **Accuracy**: 79.80%
  - **F1-Score**: 45.99%
  - **Recall**: 44.10%
  - **Precision**: 48.04%
  - **Specificity**: 88.45%
  - **Insights**:
    - While the accuracy was decent, the F1-score indicated a trade-off between precision and recall.
    - The recall for malignant lesions was particularly low, indicating that the model struggled to identify cancerous cases effectively.
    - The high specificity highlighted its ability to correctly identify benign lesions, but this came at the cost of underperforming on malignant cases.

II. Model 2: ResNet50 with Transformer

- **Results**:
  - **Accuracy**: 88.70%
  - **F1-Score**: 67.99%
  - **Recall**: 61.54%
  - **Precision**: 75.95%
  - **Specificity**: 95.28%
  - **Insights**:
    - The accuracy improved significantly compared to the Basic CNN, along with a notable increase in F1-score.
    - Recall also increased, indicating better identification of malignant lesions. However, it was still lower than desired.
    - The transformer layers likely enhanced the model's ability to capture spatial relationships, contributing to better overall performance.

III. Model 3: Inception ResNet V2 with Soft Attention

- **Results**:
  - **Accuracy**: 89.60%
  - **F1-Score**: 76.04%
  - **Recall**: 84.62%
  - **Precision**: 69.04%
  - **Specificity**: 90.81%
  - **Insights**:
    - This model achieved the highest F1-score, reflecting a good balance between precision and recall.
    - Recall improved significantly, demonstrating the model's strength in identifying malignant lesions—a critical requirement for cancer detection.
    - The use of soft attention likely contributed to this success by emphasizing the most relevant regions of the image.

Binary classification allowed us to simplify the lesion detection problem and focus on distinguishing benign from malignant cases. Among the three models. We would also like to express why each evaluation metric is important in defining a good classification model.

General Observations

1. **Accuracy**:
   - Accuracy alone was insufficient to evaluate model performance due to the imbalanced dataset. High specificity inflated accuracy, as benign lesions were far more frequent.
2. **Recall and F1-Score**:
   - Recall was a key metric, as missing malignant cases can have severe consequences. Inception ResNet V2 excelled in recall (84.62%), making it the most effective model for identifying cancerous lesions.
   - The F1-score highlighted the trade-off between precision and recall, with Inception ResNet V2 striking the best balance.
3. **Specificity**:
   - All models maintained high specificity, particularly ResNet50 with Transformer (95.28%), indicating they were effective in identifying benign cases.
4. **Handling Class Imbalance**:

- Using class weights in the loss function helped mitigate the effects of imbalance, improving recall and F1-scores across all models.

We have now tested our developed binary and classification models where using soft attention and attention based mechanisms significantly improved the classification performance.

4.6 Auto Segmentation

Classification models are trained separately on a lesion segmented dataset. To create this segmented dataset, lesions are extracted from the images using below automated segmentation processes.

4.6.1 Segmentation with Grounding DINO + SAM

This process involves lesion bounding box detection using Grounding DINO followed by lesion segmentation using fine tuned Segment-Anything-model, guided by the DINO's bounding box as prompt for precise lesion extraction.

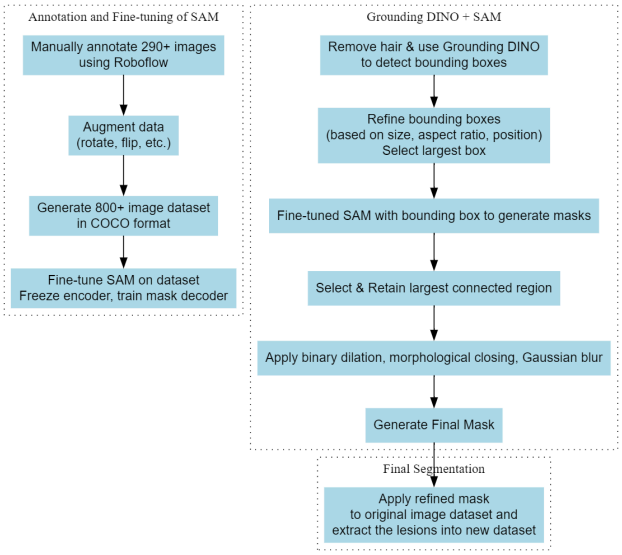

Fig 12: Segmentation process with Grounding DINO + SAM

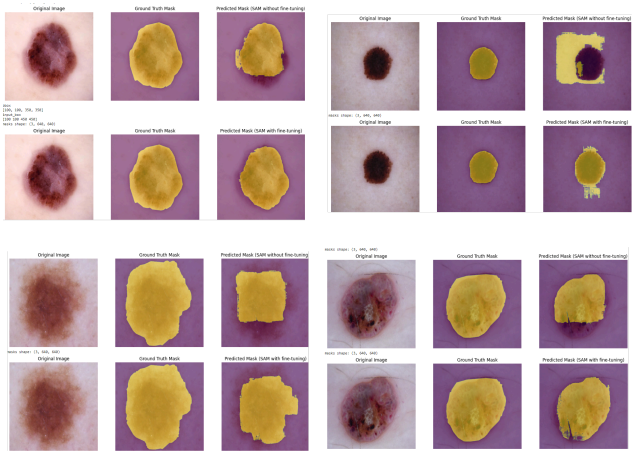

Fig 13: Segmentation results of SAM, Fined Tuned SAM

The steps are as follows:

1. **Annotated Dataset Preparation for SAM fine tuning**:
   - 300 images were manually annotated using Roboflow.
   - Augmentation and rotation techniques expanded the dataset to 800+ images in COCO format, which was used for fine-tuning a pre-trained SAM sam_vit_b_01ec64.pth (Segment Anything Model).
   - During fine-tuning, the image encoder's weights were frozen, and only the mask decoder was trained.

2. **Bounding Box Detection**:
   - Grounding DINO was employed to generate bounding box prompts for SAM.
   - A text prompt "lesion" guided bounding box detection. To minimize interference from hair follicles, images with hair removal preprocessing were used.
   - Detected bounding boxes underwent refinement based on criteria like area, aspect ratio, and position, with the largest bounding box retained as the primary lesion region.

3. **Mask Generation**:
   - The fine-tuned SAM used the bounding box as a prompt to generate segmentation masks.
   - Multiple masks were produced, and the largest one was selected for further processing.

4. **Mask Refinement**:
   - The selected mask was refined to ensure accuracy:
     - Retained only the largest connected region.

- Extended by 5 pixels in all directions using binary dilation.
- Applied morphological closing to fill any holes.
- Optionally smoothed edges using Gaussian blur.

5. **Final Segmentation**:
   - The refined mask was applied to the original image dataset to extract the lesion.
   - The segmented lesion dataset was saved for further classification tasks.

4.6.2 Segmentation with ResNet50 based U-Net

In this segmentation process, we used a pre-trained ResNet50 as the encoder, followed by a U-Net-like decoder. The ResNet50 encoder is loaded and frozen with pre-trained weights of ImageNet while the U-Net decoder is made trainable to learn and detect the lesions.

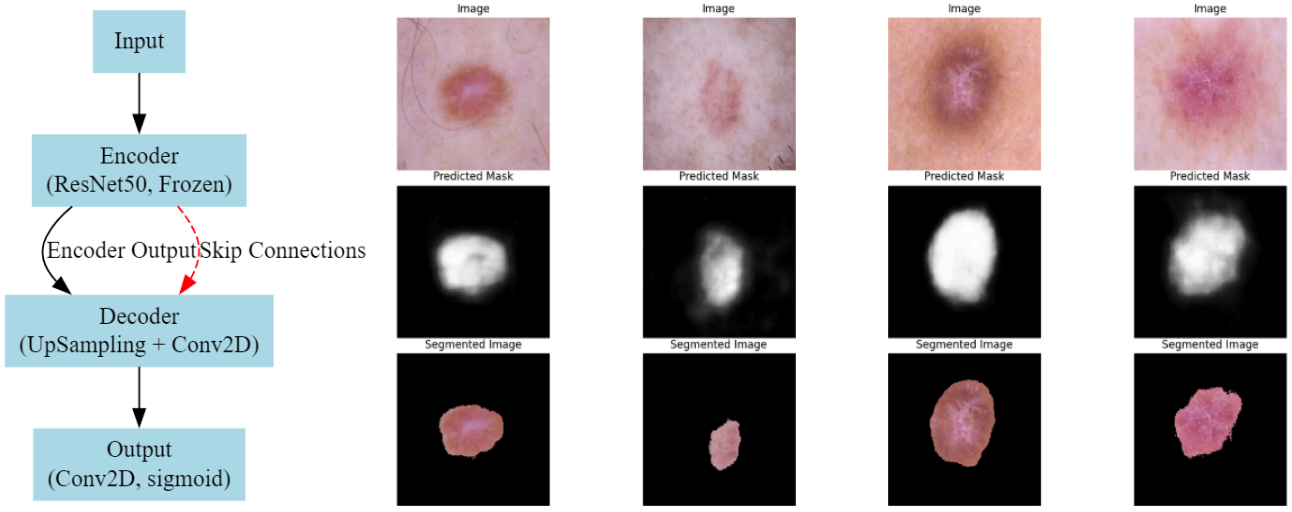

Fig 14: Segmentation process with ResNet50 based UNet and the corresponding masks, extracted lesions

5. Results

5.1 Performance comparison on raw dataset

The models proposed in section 4.5 show steady improvements as more advanced techniques were introduced. The metrics for validation accuracy, F1-score, and recall for multi classification are summarized below: MC : Multi Class

Model	Accuracy	F1-Score	Recall
Basic CNN (MC)	52.5%	60.3%	61.9%
(Binary)	79.80%	45.99%	44.10%
ResNet50 + Transformer (MC)	79.4%	78.9%	79.4%
(Binary)	88.70%	67.99%	61.54%
Inception ResNet V2 + Soft Attention (MC)	82.5%	83.3%	82.5%
(Binary)	89.60%	76.04%	84.62%
ResNet50+Advanced Soft Attention	80.8%	80.3%	80.8%

Table 2: Model performance comparison metrics

The introduction of pre-trained backbones, attention mechanisms, and transformers significantly enhanced the model's performance, especially for underrepresented classes. Models incorporating attention mechanisms (Models 3 and 4) showed improved recall and F1-scores, demonstrating their ability to prioritize critical features effectively.

- The Basic CNN provided a strong baseline but struggled with recall.
- ResNet50 with Transformer improved recall and precision, demonstrating the benefits of enhanced spatial relationships.
- Inception ResNet V2 with Soft Attention emerged as the best-performing model, with the highest recall and F1-score, making it the most effective for identifying malignant lesions.

So, on a whole it is obtained that the model **Inception ResNet V2 with Soft Attention** has turned out be the best performing model both on Multi class and binary classification tasks trained on raw HAM10000 dataset. We are also confident to say that the accuracy obtained is very close to the original paper that proposed the soft Attention mechanism []. And on a whole the

performance of this model is also out-passing many research works that are implemented in ISIC 2018 and 2019 []. Even though this classification performance is decent, we have done the training and testing on the raw image from the HAM10000 dataset where the whole image does not only contain the diseased skin . So we intend to perform segmentation on the very particular lesion , extract the lesion alone and train the proposed models on this lesion extracted dataset. As we don't have ground truth labels of HAM10000 we are proposing an Auto segmentation based approach.

Classification models mentioned in section 4.5 have been trained separately on these segmented datasets generated by both Grounding DINO+SAM, ResNet50 based U-Net. Results are presented in the following section.

5.2 Performance comparison on segmented dataset

Model	Segmentation Dataset used	Accuracy	F1-Score	Recall
Basic CNN	Grounding DINO+SAM	56.94%	59.93%	56.94%
	ResNet50-based U-Net	67.03%	53.80%	67.03%
ResNet50 + Transformer	Grounding DINO+SAM	67.53%	69.90%	67.53%
	ResNet50-based U-Net	69.63%	72.42%	69.63%
Inception ResNet V2 + Soft Attention	Grounding DINO+SAM	78.82%	79.78%	78.82%
	ResNet50-based U-Net	75.52%	77.1%	75.52%
ResNet50 + Advanced Soft Attention	Grounding DINO+SAM	73.63%	75.50%	73.63%
	ResNet50-based U-Net	70.83%	73.25%	70.83%

Table 2: Model performance comparison metrics using segmented dataset

6. Overall Comparison and Conclusion

To develop a model that works best for lesion classification, we trained and tested multiple architectures using generic and segmented datasets. After exploring various possibilities, we identified the best-performing models for each implementation path, as detailed in different sections of this report. Below, we compare the performance of these top models to determine which process flow is most suitable for classifying lesions. The following table summarizes the performance metrics

Upon reviewing the table 3, it is evident that models trained on the raw images of the HAM10000 dataset outperform those trained on segmented datasets across all metrics. However, the classification using the segmented dataset still showed promising results. We believe there are several factors contributing to the relatively lower performance of models trained on segmentation-based data. These include:

1. Differences in the size of the training datasets.
2. The use of auto-labeling models, which inherently have less expertise in annotating lesions.
3. Improper segmentation of certain images, which can introduce noise into the data and adversely affect the overall accuracy of the classification task.

Addressing these challenges, particularly by improving the quality of segmentation and ensuring proper annotation of lesions, could significantly enhance the performance of segmentation-based classification models.

Metrics	Inception ResNet V2 + Soft Attention (ResNet50-based U-Net)	Inception ResNet V2 + Soft Attention (Grounding DINO+SAM)	Inception ResNet V2 + Soft Attention (Multi Class - Original Data)	Inception ResNet V2 + Soft Attention (Binary Class-Original Data)
F1 Score	77.19%	79.78%	83.3%	76.04%
Recall	75.52%	78.82%	82.5%	84.62%
Accuracy	75.52%	78.82%	82.5%	89.60%

Table 3: Overall model performance comparison

In conclusion, the **Inception ResNet V2 + Soft Attention model trained on raw images of the HAM10000 dataset** emerges as the best-performing model for lesion classification.

7. Future Scope

In future work, we aim to incorporate demographic data such as age, gender, and lesion localization into the classification pipeline, enabling the model to leverage contextual patterns and improve its predictive performance. Enhancing segmentation techniques by addressing issues such as improper segmentation and refining lesion extraction processes will be critical for minimizing noise and boosting overall accuracy. Furthermore, integrating explainability tools, such as Grad-CAM or SHAP values, will provide interpretable visual and textual insights, increasing the clinical reliability of the model. Finally, validating the system on diverse, real-world datasets from varying populations and clinical settings will assess its robustness and facilitate its seamless integration into clinical workflows, including telemedicine platforms and dermatoscopy devices

8. References

[1] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint*, arXiv:1902.03368, 2018. [Online]. Available: https://arxiv.org/abs/1902.03368.

[2] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.161.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28117445.

[4] F. Wang and C. Wang, "Attention mechanisms for skin lesion classification," *arXiv preprint*, arXiv:1909.04525, 2019. [Online]. Available: https://arxiv.org/abs/1909.04525.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. [Online]. Available: https://arxiv.org/abs/1709.01507.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.161.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, X. Weissenborn, J. Unterthiner, A. Dehghani, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929.

[8] M. Hu, Y. Li, and X. Yang, "SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model," *arXiv preprint*, arXiv:2304.13973, 2023. [Online]. Available: https://arxiv.org/abs/2304.13973.

[9] Roboflow Blog, "Grounding DINO: Zero-shot Object Detection," Roboflow, 2023. [Online]. Available: https://blog.roboflow.com/grounding-dino-zero-shot-object-detection/.