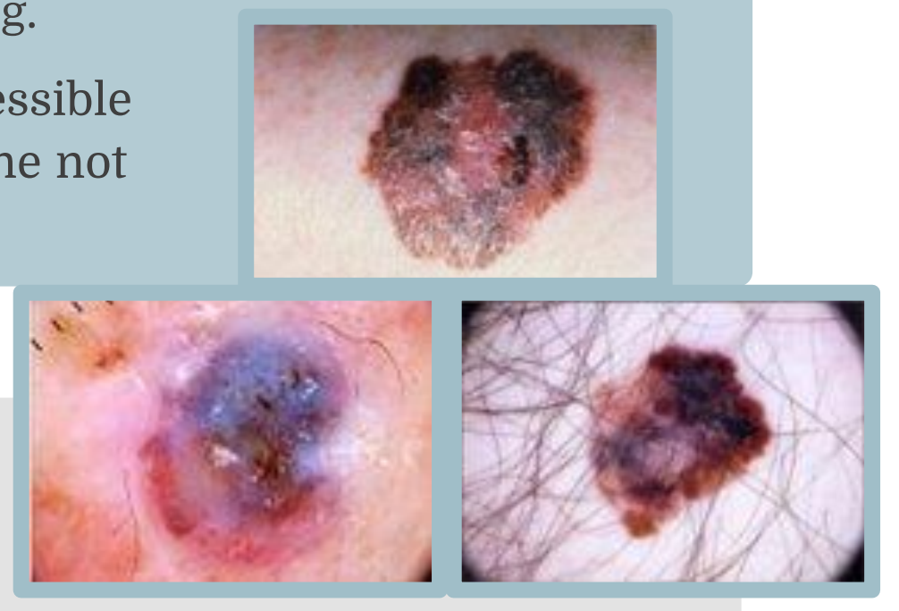
Skin Cancer Detection and Interpretation

Anastasia Anichenko, Manuel Bradicic, Felix Olesen, Michal Sitarz
COM3025 Group 8

Introduction

- Skin cancer is an extremely common disease, affecting people all around the world.
- Detecting skin cancer early is key to recovery, an early diagnosis can raise the survival rate up to 95% [14].
- The aim of this research is to combine the medical knowledge with SOTA Machine Learning techniques, to create highly accurate models and gain deeper understanding of their decision-making.
- This will aid in detection of cancer, making it more accessible (only requiring images), and hopefully outperforming the not ideal predictions from the dermatologists



Literature Review

Data Preprocessing

There are several techniques used in medical literature to handle noise, camera resolution, anatomic site and illumination:

- De-noising the images [1][2][3]
  - Removing artifacts: hair (Dull Razor), ink marking, air bubbles, black frames. (K-Means clustering)
  - Fixing lighting and contrast: contrast enhancement [4] and illumination variations. (Monte-Carlo non-parametric model)
- Data pre-processing [2][5][3]
  - Real-time data augmentations are used for increased variety in the dataset such that less overfitting occurs. (geometric transformations, scaling and normalization)
  - Data Balancing is implemented to improve the balance between classes, as some are largely underrepresented. [6] [7]
  - Duplicate Removal refers to the process of identifying and eliminating augmented images in the database in similar approach to [32].
- Segmentation [2][8]
  - Partitioning the lesion from the skin area surrounding it by applying masks.
  - Segmentations help to decrease complexity and reduce unnecessary noise.

The ABCD rule is used in dermatology for melanoma detection. We explored the possibility of applying **features** related to shape and color to help with the training and understanding of lesions in the following ways:

- The features are used with Machine Learning models, such as an ANN or an SVM, to classify malignant and benign images [9]
- Similarly, color features have been extracted from the images and combined with metadata, to an ensemble model with a CNN to increase accuracy based on the manual feature extraction and patient data [2]

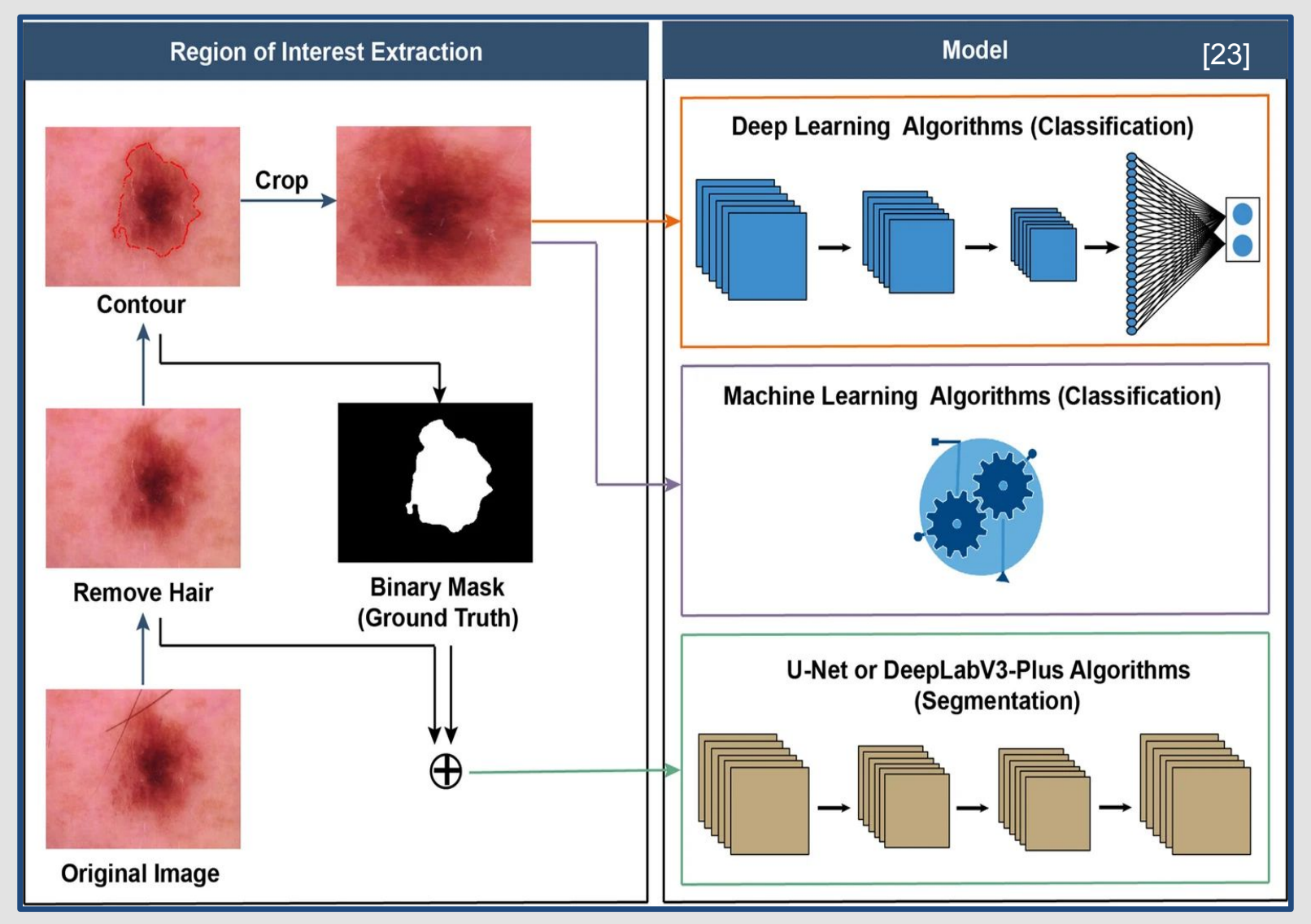
Models

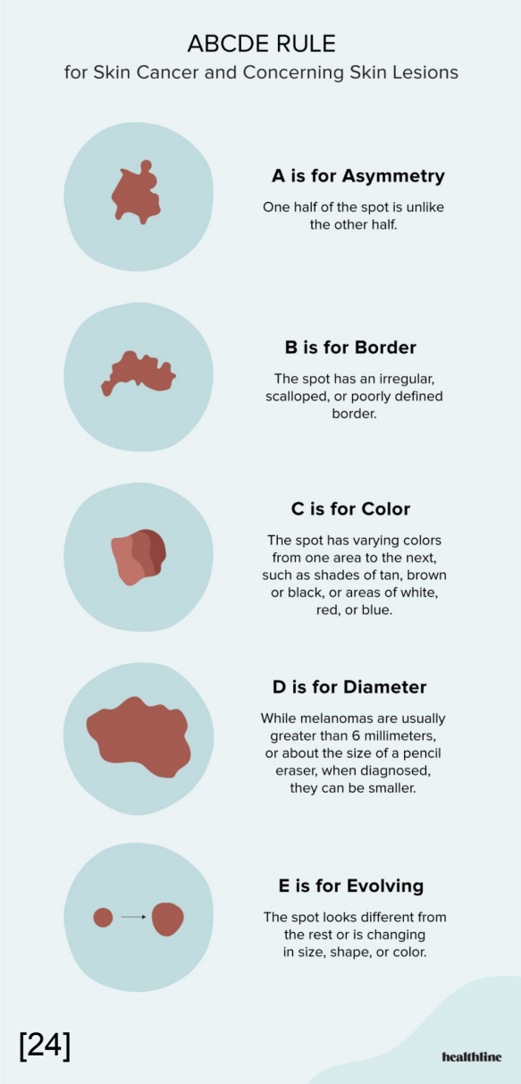
CNNs are a popular approach to skin cancer classification with many **outperforming dermatologists** [10]. Amongst them, ImageNet pre-trained EfficientNet is a commonly used backbone [5, 11, 2]. For instance, experiments evaluating B0 to B7 EfficientNets on HAM10000, where B4 performed the best with F1-score of 87% and accuracy of 87.91% [5].

Ensemble of networks are used to improve on the models and prevent overfitting. One strategy searched for an optimal subset of multi-resolution EfficientNets (with metadata) to create an ensemble [13]. Finally, a **medical vision transformer** [6] achieves 96.14% on HAM10000, whilst FixCaps [12], currently holds SOTA performance (96.49%).

Supervised methods are limited by the cost of acquiring and annotating high quality data. Popular self-supervised methods such as contrastive learning tackle this problem by taking sample pairs of unlabelled data and learning an embedding space so that similar samples are close together and dissimilar ones are far apart [15]. A simplified contrastive network - SIMCLR [16] achieved SOTA results on ImageNet. Chen et al. [16] note that the composition of data augmentation plays a crucial role in the success of their proposed network. Application of contrastive learning to medical imaging proves to be more challenging as the various classes tend to contain a smaller amount of distinguishing features, to tackle this multimodal supervised approaches have been developed pairing images with text data [17, 18].

In addition the researchers approached the challenge of limited and imbalanced data using **synthetic data generation**. Both GANs [19, 20] and Diffusion models [21] showing impressive photo realistic outcomes. A study performed by Akrout et al. [22] suggests that diffusion models are capable of generating high quality skin cancer lesions that aid the augmentation and balancing of training data. Alternatively, many of the mentioned approaches use tools like weighted cross-entropy, to increase the weight of the underrepresented classes [2, 13].


Region of Interest Extraction
Model
[23]
Crop
Contour
Remove Hair
Binary Mask (Ground Truth)
Original Image
Deep Learning Algorithms (Classification)
Machine Learning Algorithms (Classification)
U-Net or DeepLabV3-Plus Algorithms (Segmentation)


ABCDE RULE
for Skin Cancer and Concerning Skin Lesions
A is for Asymmetry
One half of the spot is unlike the other half.
B is for Border
The spot has an irregular, scalloped, or poorly defined border.
C is for Color
The spot has varying colors from one area to the next, such as shades of tan, brown or black, or areas of white, red, or blue.
D is for Diameter
While melanomas are usually greater than 6 millimeters, or about the size of a pencil eraser, when diagnosed, they can be smaller.
E is for Evolving
The spot looks different from the rest or is changing in size, shape, or color.
[24]
healthline

Problem Analysis

Dataset

Two popular datasets for skin cancer classification are HAM10000 and ISIC 2019 with 10,015 and 25,331 training samples respectively. ISIC 2019 is an expansion on HAM10000 by concatenating HAM1000, BCN_20000 and MSK into a single dataset [25]. ISIC 2020 is the largest ISIC dataset released to date with 33,126 training images but unlike the previous two datasets, it focuses on skin cancer classification as a binary problem of benign vs malignant [26].

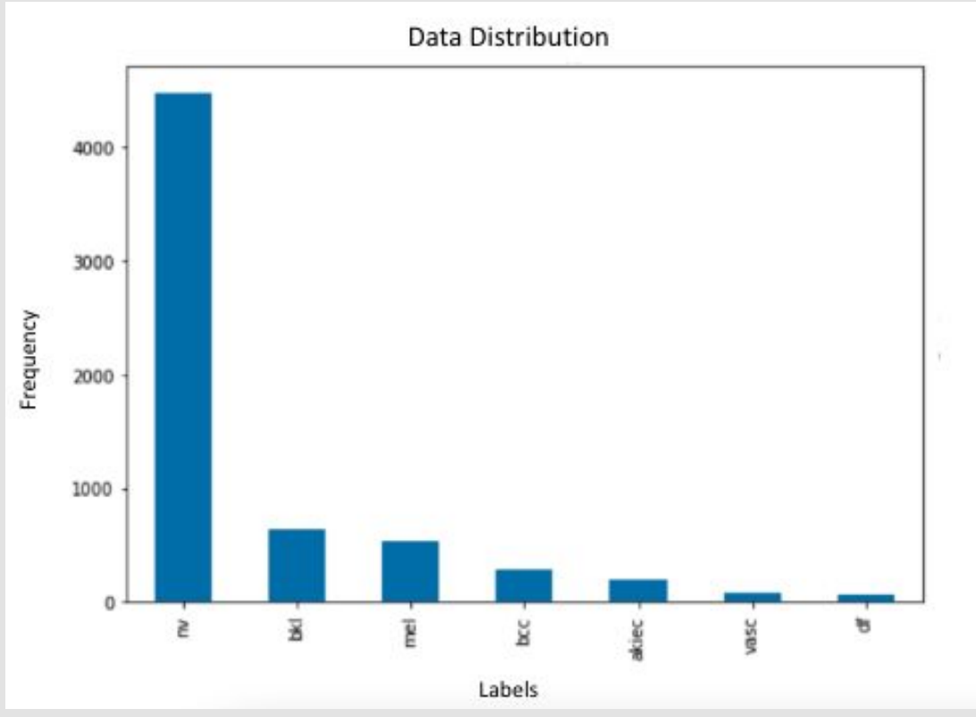
We decided to use HAM10000 as we wanted to tackle a multi-class problem and explore the effects SOTA methods from literature such as Diffusion Models [21] and segmentation [2][8]. Additionally, HAM10000 was chosen over ISIC 2019 so that we can focus on the quality of the data and models rather than simply using a larger dataset to boost performance.

General Challenges with ML and Skin Lesion Classification

When it comes to specifically skin lesion classification it can be very difficult to classify some types of cancer. Even dermatologists struggle with classifying certain cancers purely based on appearance alone. Some more prominent ones, but non-malignant, such as seborrhoeic keratoses, can share features with less common, but malignant ones [14]. This is a difficult challenge to address but we attempt to tackle it by implementing feature extraction following the ABCD rule as well as integrating metadata into the pipeline.

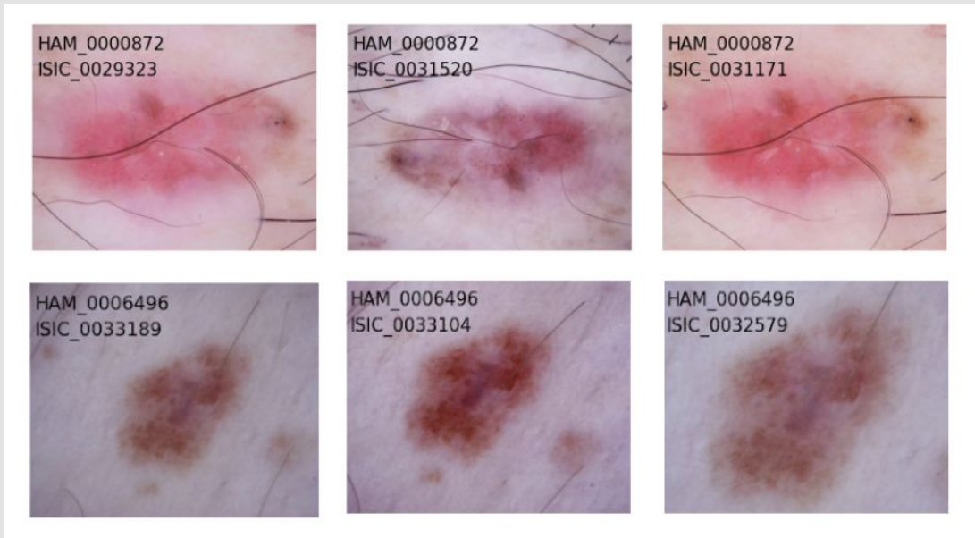
Dataset Imbalance

One of the biggest problems with the HAM10000 dataset is its major class imbalance. Nevus makes up almost 67% of the entire dataset whilst DF makes up less than 12%. We tackle this challenge with data augmentation an generating synthetic data for underrepresented classes using GANs and Diffusion Models. In addition to imbalance, there is a lot of noise in the data which is addressed with various pre-processing techniques such as hair removal and normalization.



Duplicates

Additionally the 10,015 training samples only contain 7,860 unique lesions. Which can introduce further bias and lead to overfitting of the model. Duplicates share the same HAM_ID which could have made the task of duplicate removal very easy. Unfortunately this could lead to data loss as some duplicates are cropped to different areas and also appear to sometimes show the lesion at a different point in time. We address this with data pre-processing and the training split.



Overall Approach & Evaluation Metrics

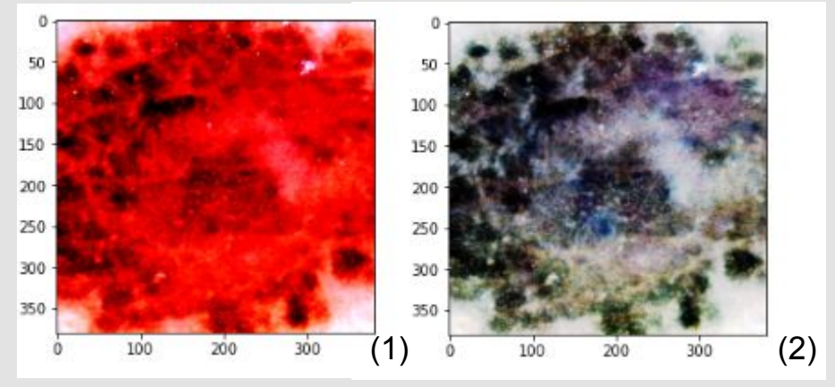
Our overall approach focuses on different methods for skin cancer classification. including CNNs, Vision Transformers and contrastive learning. We aim towards building an effective data-preprocessing pipeline by implementing literature-based techniques to evaluate their efficacy on the aforementioned ML architectures. Our main evaluation metric will be recall rate which is a direct reflection of the number of overall positive samples. In medical imaging, a high recall rate is one of the most desirable [28] properties as we want to include as many True Positive samples as possible even if it comes at the cost of classifying more False positives and therefore producing a lower precision.

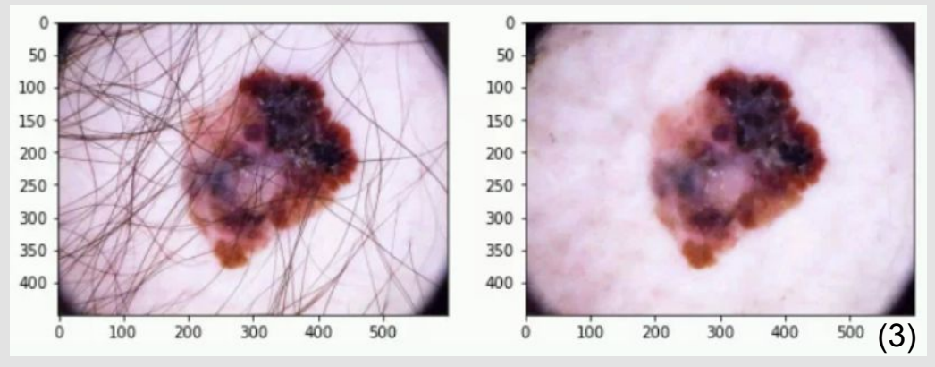
Basic Data Pre-processing

- Correctly preparing the dataset is a crucial first step to training a model which can classify the images on the test set that is provided. Below is a summary of the main techniques that were used in this project to improve the classification performance.

Preparing Images:

- Normalization (1) - normalizing the image with mean and standard deviation, in order to set a range and reduce data skewness
- Histogram Equalization (2) - using image histograms to improve the contrast in the image





De-Noising:

- Hair and Artifact Removal (3) - removing unwanted elements using Dull-Razor algorithm, which applies a series of morphological operations to the image to create a mask
- Filtering images - using a median blur for removing noise in the images

Augmentations:

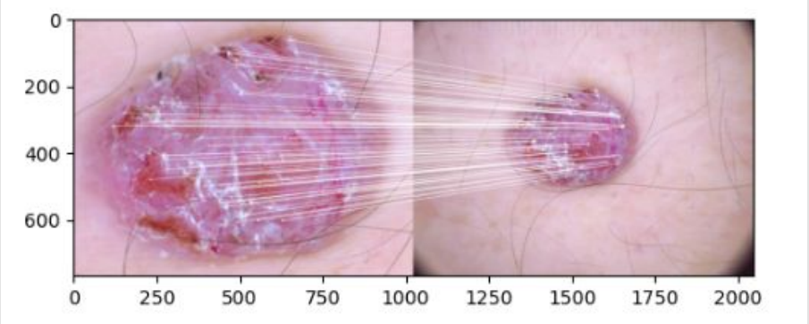
- Mainly, real-time geometric augmentations were used to tackle the small dataset problem and train the models to predict better on previously unseen data.
- In contrast, we also implemented class balancing by creating random transforms of the provided images to match the class distribution for all of the classes

Data Split:

- When splitting the dataset we ensure that all images that share the same lesion id were put either into training or validation to ensure there was no data leak.

Duplicate Removal

The dataset appears to have hundreds of invariants of images in the dataset. This characteristic, data leakage, usually results in model overfitting. We used Scale Invariant Feature Transform (SIFT) to identify and eliminate duplicates.



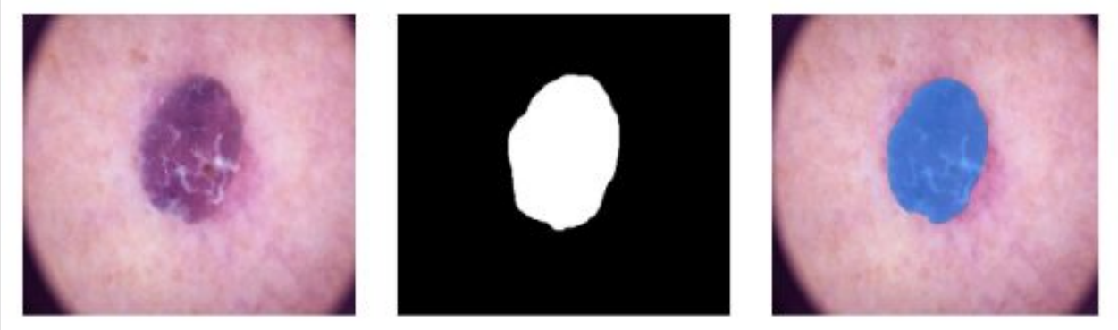
Advanced Data Pre-processing

Diffusion models

To tackle the problem of an unbalanced dataset, GANs and **Diffusion models** were introduced. Amongst the classes, dermatofibroma (df) was particularly challenging to classify due to insufficient data, resulting in poor performance. By using such models, researchers can overcome the challenge of data scarcity for certain classes. 'Improved Denoising Diffusion Probabilistic' [30] study, proposed by OpenAI, was implemented on the aforementioned class in the dataset. Resulting in 256 newly generated images of dermatofibroma

Image Segmentation

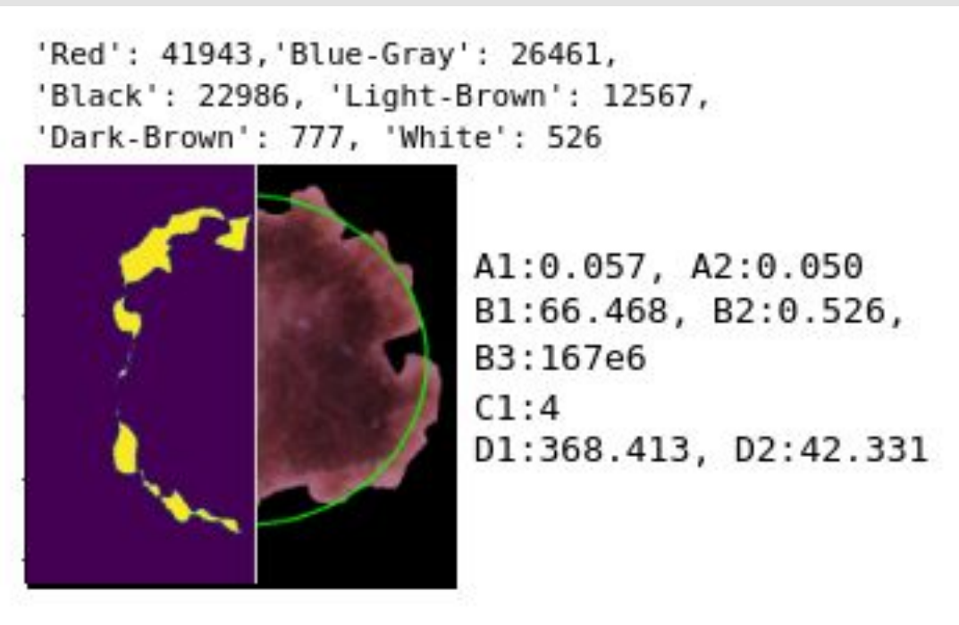
Another challenge this team faced was generating **segmentations** of test images, which are essential for localising lesions and determining their exact boundaries. Firstly, ResNextUNet was implemented to produce bounding boxes of lesions for the test set. These bounding boxes were then utilized to create segmentations of the lesions using recent research published by Meta; 'Segment Anything' [29]. This technique allows models such as feature maps to focus on the lesion shape, the 'ABCD' rule, as well as removal of additional 'noise' for classification - skin and hair around the lesion.



Feature extraction

An additional exploration was by looking into manual **feature extraction** following the ABCD rule.

- The extracted features used: asymmetry in the area, border sizes, number of malignant- associated colors and diameter irregularities.
- An ensemble of EfficientNetB1 and an ANN was trained to classify the images alongside the extracted features and metadata.
- An SVM was trained on the extracted features and the provided medical metadata to classify the lesions.



```
'Red': 41943,'Blue-Gray': 26461,
'Black': 22986, 'Light-Brown': 12567,
'Dark-Brown': 777, 'White': 526
```

```
A1:0.057, A2:0.050
B1:66.468, B2:0.526,
B3:167e6
C1:4
D1:368.413, D2:42.331
```

Models

There was an extensive series of models that were trained and tuned on different permutations of the data pre-processing pipeline. This included supervised and unsupervised techniques for tasks such as classification, which is discussed below.

CNN Classification:

- Initially, **EfficientNetB1** was used as a baseline. A pre-trained model with the weights trained on ImageNet was used as a feature extractor to optimize the final classification layer for the skin cancer problem. This iteration of the model is optimized for 244x244 image sizes.
- However, we switched to using **EfficientNetB4** which is a better perfmoning model, and it uses higher resolution images (380x380) which can yield more details during training. Similarly, we switched to fine-tuning the models, as medical data and ImageNet are not in the same domain.
- The batch size had to be decreased from 16 to 8 to fit the images on the GPU. Both iterations of the EfficientNets have been experimented on with different data pre-processing techniques, hyperparameters and the advanced features.

Vision Transformer Classification:

- As an alternative to CNNs we also used a pre-trained **Vision Transformer** (ViT). This used pre-trained weights provided by Google, and image size of 244x244 and patch size of 16x16. A ViT with an image size of 384x384 was also tested.
- The ViT was fine-tuned without freezing using a batch size of 16 and a learning rate of 1e-4.

Unsupervised Classification:

- One of the approaches that was taken towards classifying the HAM10000 dataset was the use of contrastive learning through SimCLR V1. This involved training a base encoder network (ResNet18) for feature encodings.
- These encodings would then be fed into a projection head (2 layer MLP) for mapping encodings into a latent space where contrastive loss would be applied. This model was deemed suitable for this problem as its contrastive nature provided algorithmic insights into how the data pre-processing pipeline alters noticeable features in each image.
- This approach heavily relies on good data pre-processing and augmentations to work effectively, as this is what determines that the landscape of the latent space through which SimCLR learns. InfoNCE loss was used as the similarity metric when comparing images. Once SimCLR was trained, the projection head was then removed, leaving the feature encoder which then became a feature extractor.
- These features were then used to train two classifiers: Logistic Regression and SVM. These simple classifiers were used as deeper ones tended to heavily overfit on the train data.

Hyperparameters

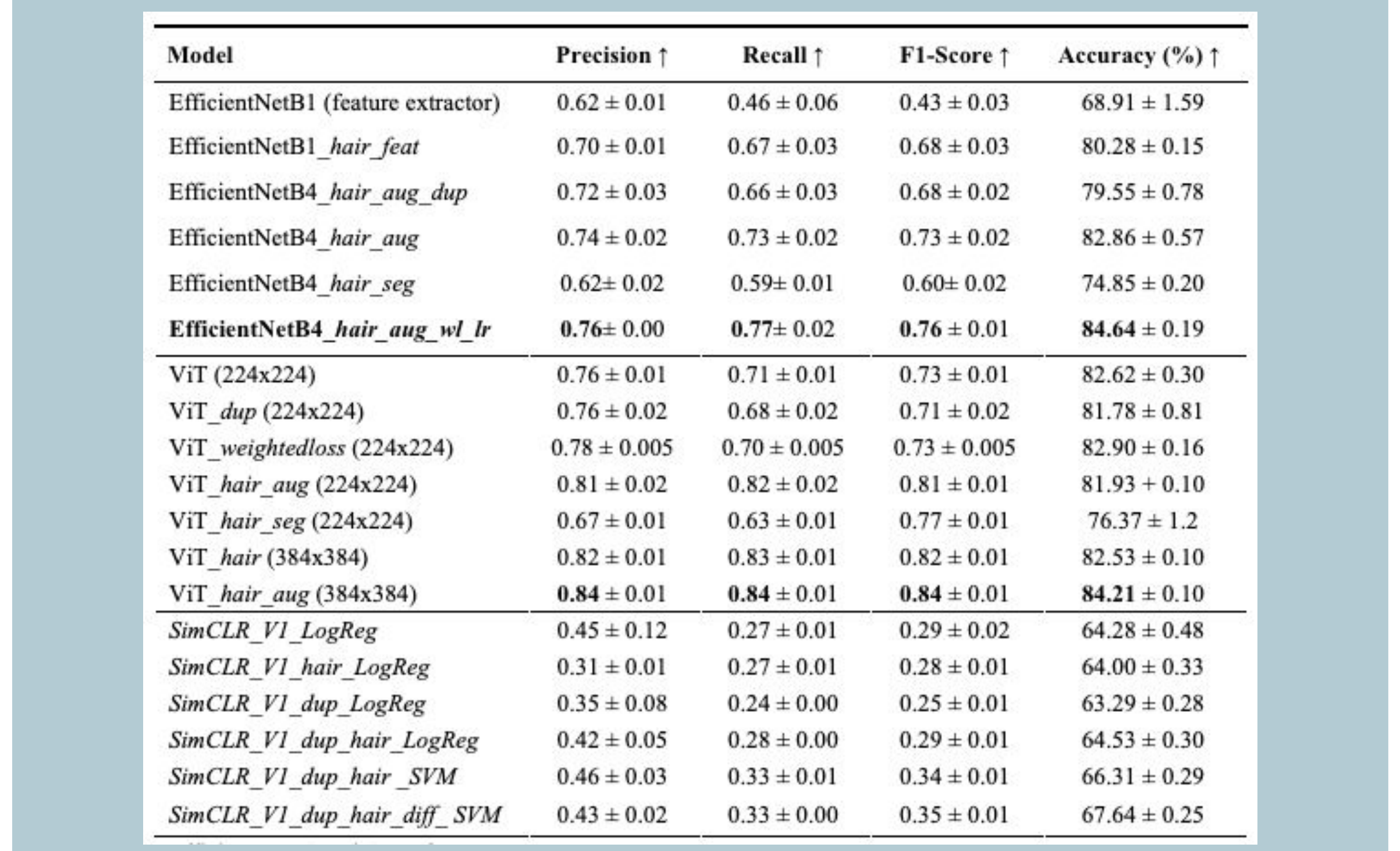
The cost function used to evaluate the model performance was the cross entropy loss, where all of the predicted class probabilities are compared to the actual class desired. Furthermore, weights were added based on the probability distribution of the dataset for balance.

$$L_{CE} = -\sum_{i=1}^{n} \omega_i p(x_i) log_e(q(x_i))$$

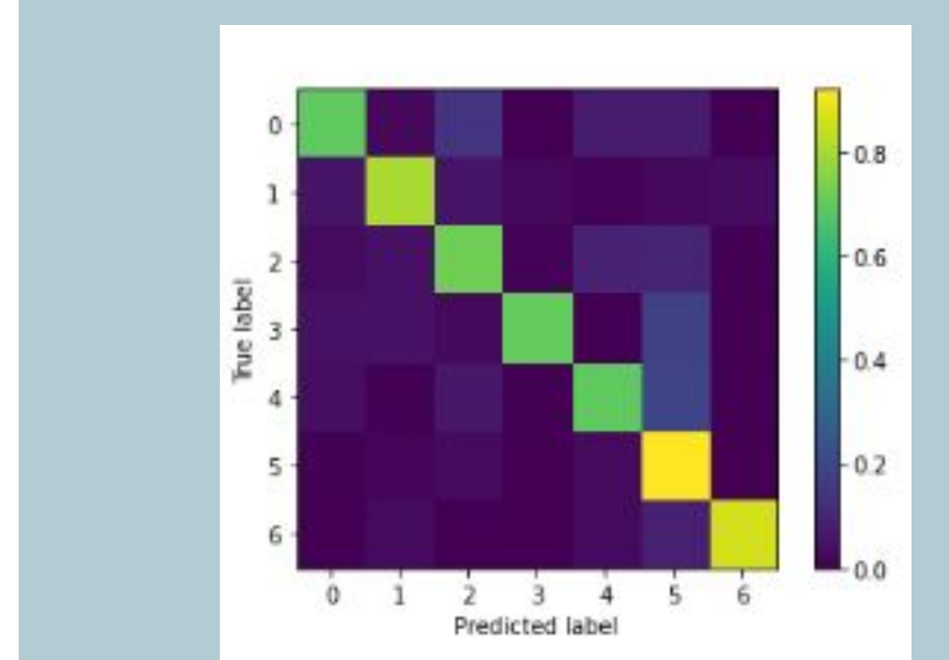
[34]

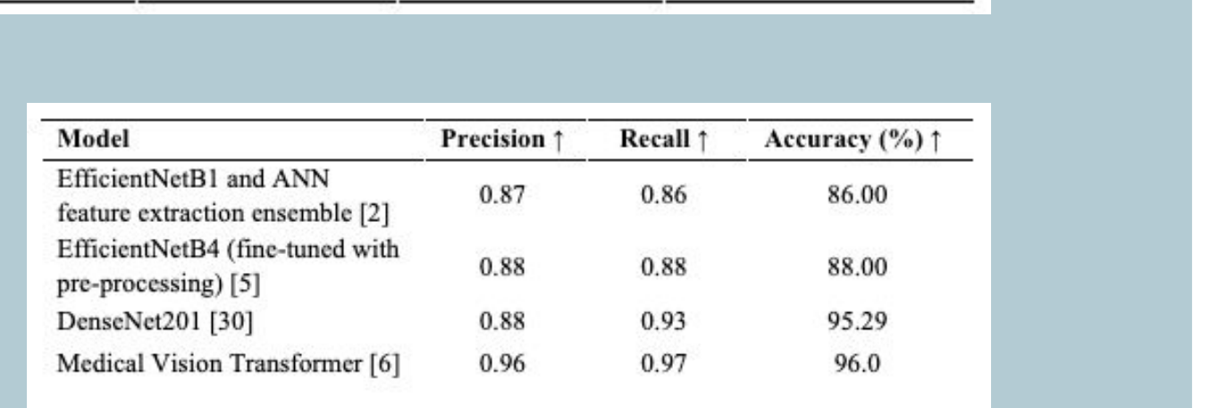
Most of the research uses SGD optimizer with momentum, as even though Adam usually converges faster, whilst SGD oftentimes can converge to more optimal solutions (better generalization). We set the learning rate to 1e-4, but also experimented with higher learning rates for fine-tuning the models, and then decreasing it with schedulers.

Results

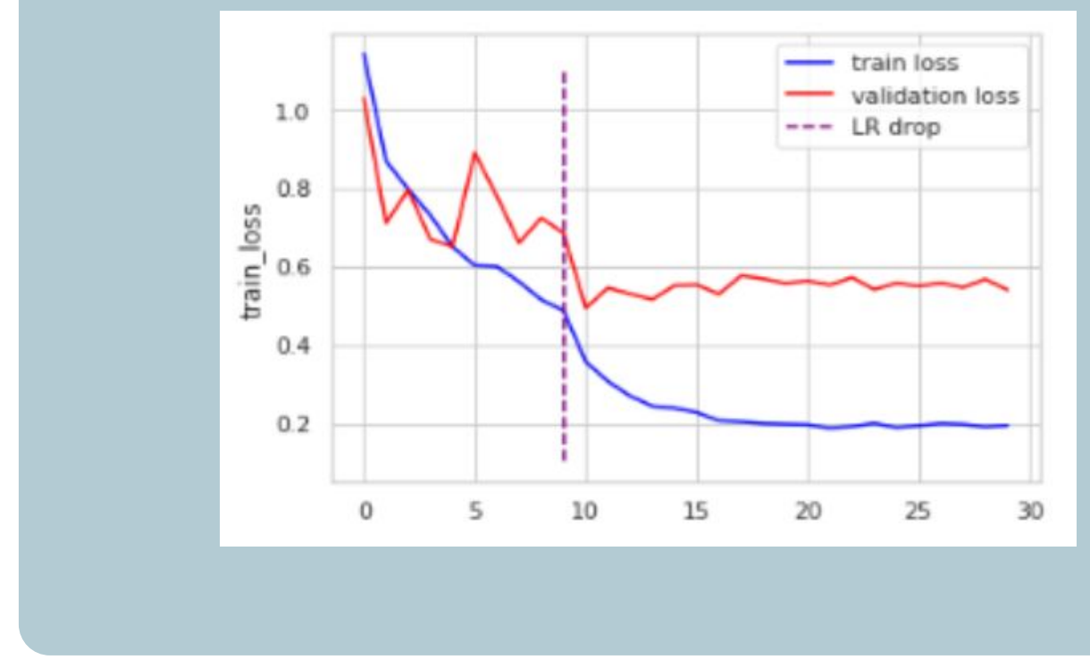


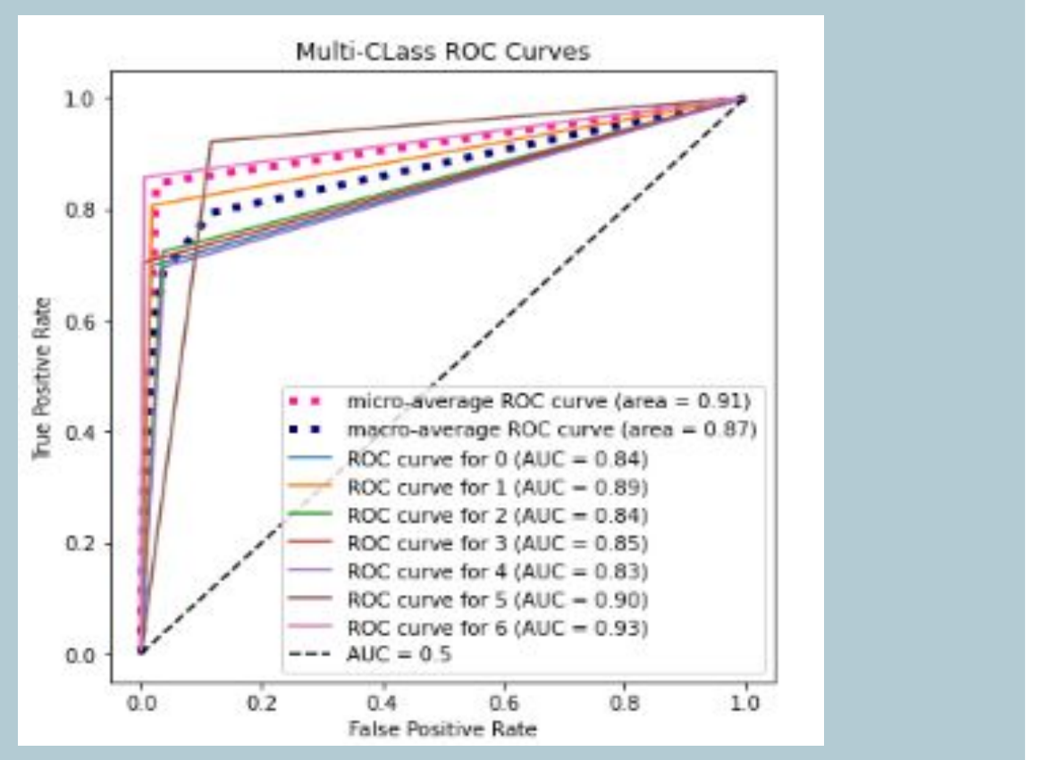
Model	Precision ↑	Recall ↑	F1-Score ↑	Accuracy (%) ↑
EfficientNetB1 (feature extractor)	0.62 ± 0.01	0.46 ± 0.06	0.43 ± 0.03	68.91 ± 1.59
EfficientNetB1_*hair_feat*	0.70 ± 0.01	0.67 ± 0.03	0.68 ± 0.03	80.28 ± 0.15
EfficientNetB4_*hair_aug_dup*	0.72 ± 0.03	0.66 ± 0.03	0.68 ± 0.02	79.55 ± 0.78
EfficientNetB4_*hair_aug*	0.74 ± 0.02	0.73 ± 0.02	0.73 ± 0.02	82.86 ± 0.57
EfficientNetB4_*hair_seg*	0.62± 0.02	0.59± 0.01	0.60± 0.02	74.85 ± 0.20
EfficientNetB4_*hair_aug_wl_lr*	**0.76**± 0.00	**0.77**± 0.02	**0.76** ± 0.01	**84.64** ± 0.19
ViT (224x224)	0.76 ± 0.01	0.71 ± 0.01	0.73 ± 0.01	82.62 ± 0.30
ViT_*dup* (224x224)	0.76 ± 0.02	0.68 ± 0.02	0.71 ± 0.02	81.78 ± 0.81
ViT_*weightedloss* (224x224)	0.78 ± 0.005	0.70 ± 0.005	0.73 ± 0.005	82.90 ± 0.16
ViT_*hair_aug* (224x224)	0.81 ± 0.02	0.82 ± 0.02	0.81 ± 0.01	81.93 + 0.10
ViT_*hair_seg* (224x224)	0.67 ± 0.01	0.63 ± 0.01	0.77 ± 0.01	76.37 ± 1.2
ViT_*hair* (384x384)	0.82 ± 0.01	0.83 ± 0.01	0.82 ± 0.01	82.53 ± 0.10
ViT_*hair_aug* (384x384)	**0.84** ± 0.01	**0.84** ± 0.01	**0.84** ± 0.01	**84.21** ± 0.10
SimCLR_V1_LogReg	0.45 ± 0.12	0.27 ± 0.01	0.29 ± 0.02	64.28 ± 0.48
SimCLR_V1_hair_LogReg	0.31 ± 0.01	0.27 ± 0.01	0.28 ± 0.01	64.00 ± 0.33
SimCLR_V1_dup_LogReg	0.35 ± 0.08	0.24 ± 0.00	0.25 ± 0.01	63.29 ± 0.28
SimCLR_V1_dup_hair_LogReg	0.42 ± 0.05	0.28 ± 0.00	0.29 ± 0.01	64.53 ± 0.30
SimCLR_V1_dup_hair _SVM	0.46 ± 0.03	0.33 ± 0.01	0.34 ± 0.01	66.31 ± 0.29
SimCLR_V1_dup_hair_diff_ SVM	0.43 ± 0.02	0.33 ± 0.00	0.35 ± 0.01	67.64 ± 0.25





Model	Precision ↑	Recall ↑	Accuracy (%) ↑
EfficientNetB1 and ANN feature extraction ensemble [2]	0.87	0.86	86.00
EfficientNetB4 (fine-tuned with pre-processing) [5]	0.88	0.88	88.00
DenseNet201 [30]	0.88	0.93	95.29
Medical Vision Transformer [6]	0.96	0.97	96.0



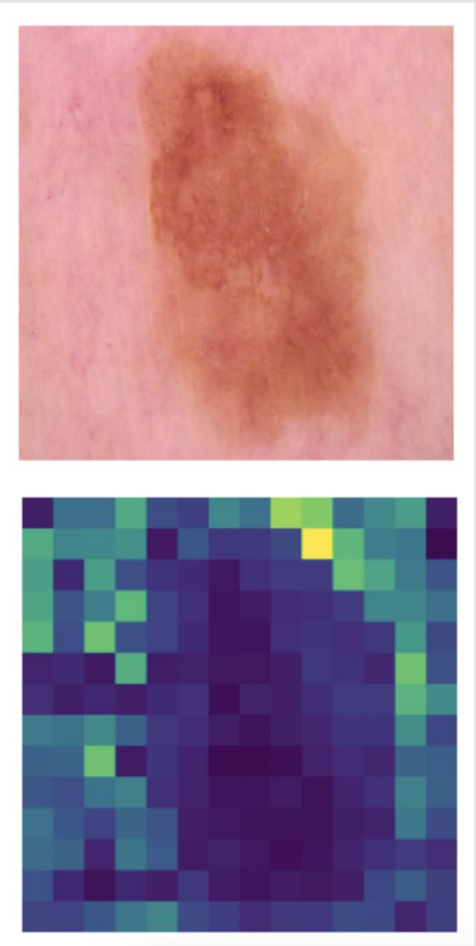


Discussion and Findings

The best performing model overall was a larger EfficientNet (B4), with higher resolution images (380x380). Specific data processing helped improve predictions, including data augmentations, normalization and hair removal. Furthermore, using a learning rate scheduler yielded the best result, reaching a local minima due to the drop in learning rate during training.

Transformers performed adequately, however most of the pre-processing was not as impactful compared to the pre-processing on CNNs. Since techniques used in this research were adapted from research on CNNs, it is surmised that the researched techniques are not directly transferable to ViTs. There are also other types of ViTs to be explored (medical or masked), which could perform better on this type of data.

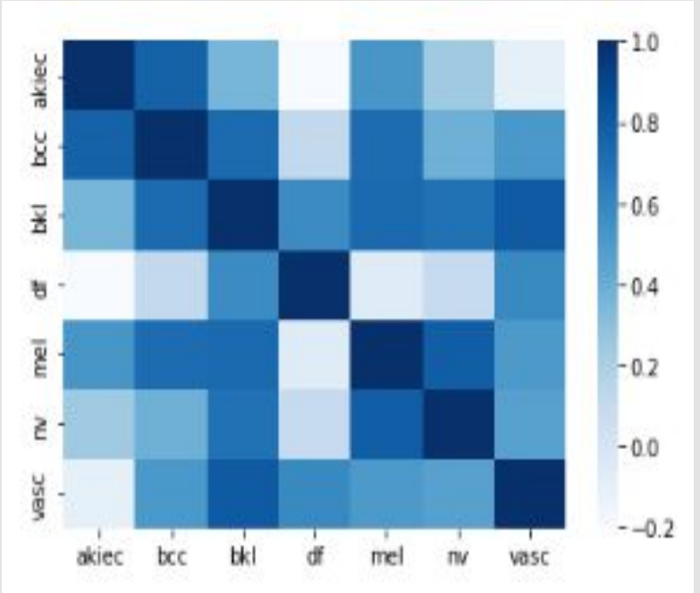
The AUC curve demonstrates that underrepresented classes have a high TPR and a low FPR indicating a low misclassification rate (further supported by the confusion matrix). The high AUC value for Nevus indicates less FPs, which is a common issue with overrepresented classes. This was achieved with the help of weighted loss and the pre-processing techniques which aided our overall goal of high recall rates.



In contrast to most SOTA research, our best performing models do not use segmentation, or some of the pre-processing techniques. Judging by the validation loss increase of 5-6% from the rest of the models, the ground-truth masks have high positive impact on the dataset, but the test dataset masks predicted by our segmentation model do not boost the performance. However, based on the attention maps extracted from our ViT, many of the models focus on unexpected parts of the image.

The use of contrastive learning with SimCLR aimed to provide an alternate approach to classification and helped demonstrate significant improvements with uncategorised diffusion data, featuring a 1.33% increase from adding unlabeled synthetic data. Further improvements could be made with more robust SOTA unsupervised techniques and more exploration into synthetic data generation.

In general, feature extraction has not provided much value to the models. Pearson coefficient was used to find similarities between the extracted features in the correctly predicted classes, to show similarities between the features. For the most classes, it does not see, to have big impact, but it seems to have some for melanoma (most deadly one) and nevis (non- malignant). They seem to share some high level features that the network picks up on (likewise, it is the most misclassified by practitioners [33]), as they get misclassified as one another the most. Those two classes should be addressed in the future research to get better distinctions between them, which would improve the network predictions significantly, and raise the recall of the predictions.



References

[2] 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, 2023.
[3] H. C. Reis, V. Turk, K. Khoshelham, and S. Kaya, 'InSiNet: a deep convolutional approach to skin cancer detection and segmentation', Medical & Biological Engineering & Computing, pp. 1–20, 2022.
[5] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, 'Multiclass skin cancer classification using EfficientNets--a first step towards preventing skin cancer', Neuroscience Informatics, vol. 2, no. 4, p. 100034, 2022.
[6] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, 'An effective skin cancer classification mechanism via medical vision transformer', Sensors, vol. 22, no. 11, p. 4008, 2022.
[33] R. C. Maron et al., 'Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks', European Journal of Cancer, vol. 119, pp. 57–65, Sep. 2019.