A report for Deep Vision (Computer Vision and AI)

Explaining Vision Transformers vs CNNs: A Comparative Study on Medical Images using XAI Methods

Tejaswi Duptala

Supervised by
Prof. Dr. Tatyana Ivanovska

Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

July 3, 2025

Abstract

Skin cancer represents a very real health risk to a great many people. Proper and early detection can make a tremendous amount of difference and fast curve of the problem can take place. Deep learning models can help doctors detect skin cancer earlier and more accurately.In the project two categories of computer models are used: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). ResNet18, EfficientNet-B0, and ResNet34 were the three CNN models and ViT Tiny, ViT Small, and Swin Tiny were the three Vision Transformer models. The models are trained and evaluated all these models using the HAM10000 dataset, which has many real skin lesion images.Models should be more than accurate. Doctors also need to understand how the model is deciding. This is even more important in medicine because doctors need to trust and also be able to explain results to patients. In order to make the models more interpretable, I applied three popular XAI (Explainable AI) tools: Grad-CAM, Score-CAM, and Attention Rollout. These tools can produce heatmaps. The heatmaps show what parts of each image the models looked at before their final decision.The results showed that ViTs and CNNs both performed well in the classification of skin lesions. They do not, however, attend to the same areas. CNNs will attend to the main lesion in an image, and it is usually easy to comprehend why the decision was arrived at. Vision Transformers sometimes attend to bigger or more diffuse regions, which are harder to interpret.This report includes all of the test results and visual explanations. It also compares the explanations of each model for how clear and useful they are for doctors. In the end, I found that a model is not truly useful unless it is both interpretable and accurate. More explainable models can help doctors trust artificial intelligence and use it for real medical cases.

Index Terms—skin cancer, medical imaging, deep learning, convolutional neural network, vision transformer, ViT, Swin Transformer, explainable AI, Grad-CAM, Score-CAM, attention rollout, classification, XAI

1 Introduction

Skin cancer is a type of cancer that originates in the skin cells. It will be harmful if it is not diagnosed early. Doctors use photographs of skin spots to test for cancer. Checking many photographs is time-consuming. Physicians may also find it challenging to detect all instances. Deep learning is a tool that can be used to help doctors analyze images. It is possible to train it to recognize patterns that mean there is cancer. The most common model type is called a convolutional neural network, or CNN. CNNs are very good at identifying color and shape in photographs. They do well in medicine, like skin cancer detection. There are also newer versions called Vision Transformers, or ViTs. ViTs also have a different way of looking at images. They can attend to many areas of an image at once. Some scientists believe that ViTs might see important clues that CNNs do not. CNNs and ViTs both have their benefits.

But both types of models have a shortcoming. They become overwhelmingly complex. It is hard to comprehend how they arrive at a decision. This is called the ”black box” problem. In medicine, trust must exist. Doctors would prefer to know why a model chose one solution over another.

This issue is tried to be solved by Explanable AI, or XAI. XAI offers tools that unveil what parts of an image mattered most to the model. One such tool is Grad-CAM. It generates a heatmap to show where a CNN was focused. Another tool is Score-CAM that does the same thing. Attention Rollout is applied to ViTs. It helps to unveil which parts the model found important.

This project employs XAI methods on CNNs and ViTs both. It compares three CNN models, such as ResNet18, EfficientNet-B0, and ResNet34. It also verifies three models of ViT, such as ViT Tiny, ViT Small, and Swin Tiny. The models analyze a real skin cancer image dataset known as HAM10000.

The most crucial is to identify which models are most accurate. But the project also confirms which models best give the most clear and best explanations. Clear explanations can convince doctors to believe in and apply AI in their practice.

This report outlines the data set, the training procedure, the XAI techniques, and the results. It gives examples of explanations of the responses by the models. The report concludes with: can deep learning not only predict but explain skin cancer outcomes in a way beneficial to physicians?

2 Methods

2.1 Data Preparation

The HAM10000 dataset is chosen for the project. The dataset consists of 10,015 images of skin lesions. Every image is a close-up, high-resolution photograph using a special camera called a dermatoscope. The images are in color and display many varieties of skin spots or lesions. There are seven classes of skin lesions in the database altogether. Each image is assigned to one of the seven classes. The classes are ordinary ones like ”nevus” (nv), ”melanoma” (mel), and so on. These classes help the model learn to distinguish between one disease and another of the skin. Each image within the dataset has extra information, referred to as metadata. The metadata gives the ID of the image, diagnosis, and lesion ID. The lesion ID is significant. It tells the model whether different images reflect the same skin spot or patient. This ensures that the same lesion does not appear in both the training and testing sets. This maintains the results authentic since the model will not see the same lesion twice. All the pictures are kept in two separate folders on the computer. The folders are in zip format. Both the folders are unzipped, and all the image files are present. The code checks each of the images described in the metadata are indeed present in these folders. If there are missing pictures, a warning is created. This makes sure there are no file problems before training.

The data is then split into three sets: training, validation, and testing. The split is at the level of the lesion, not the images. This means that all images of the same lesion are in the same set. This keeps the splitting model from ”cheating” by using the same lesion in training and testing. About 70% of the lesions are used for training. This enables the model to be trained over a large amount of examples. 15% of the lesions are kept aside for validation. The other 15% are kept aside for the test set. The validation set is utilized to track the performance of the model during training. The test set is utilized only later, to see how well the model performs in reality on new data.

By doing so, the data is prepared cautiously. This ensures that the training and testing are realistic and fair. This also ensures that the output of the project can be relied upon.

Metadata Table Example:

Below is a sample of the metadata used in this project.

lesion_id	**image_id**	**dx**	**dx_type**	**age**	**sex**	**localization**
HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear

Table 1: Sample rows from the HAM10000 metadata.

The metadata table contains 7 columns:

- `lesion_id`: Unique code for each skin lesion
- `image_id`: Unique code for each image
- `dx`: Diagnosis label
- `dx_type`: Type of diagnosis (like “histo” for histopathology)
- `age`: Patient’s age
- `sex`: Patient’s gender
- `localization`: Body location of the lesion

The dataset contains 10,015 rows and 7 columns in total.

2.2 Data Preprocessing

All the images within the dataset are resized to 224 by 224 pixels. This is done to ensure that all the deep models expect the input images to have an equal size. If the sizes of the images differed, then the models would not work.

For the training set, additional measures are taken to make data even better prepared for learning. Some images are turned sideways. Some of them are turned a bit left or right. The brightness and contrast are also changed for some images. These tricks cause the model to see many copies of the same lesion. This is called data augmentation. Data augmentation causes the model to learn features that are not specific to an angle or to light.

Second, the pixel of each image is normalized. The pixel values are changed so that they fall between -1 and 1. This is done for the three color channels: blue, green, and red. Normalization makes training faster and the output more stable.

Each image has the label as a word, like ”bkl” or ”mel.” Computers don't work with words, though; they work with numbers. Therefore, every diagnosis label is turned into a number through something called label encoding. For example, ”nv” can be 0, ”mel” can be 1, and so on until 6 for the seven classes in total. This allows the model to use the labels easily for training and testing. The HAM10000 dataset contains the following skin lesion classes: akiec, bcc, bkl, df, mel, nv, and vasc.

Figure 1 shows the number of images for each class. The dataset is imbalanced, with most samples belonging to the “nv” class.

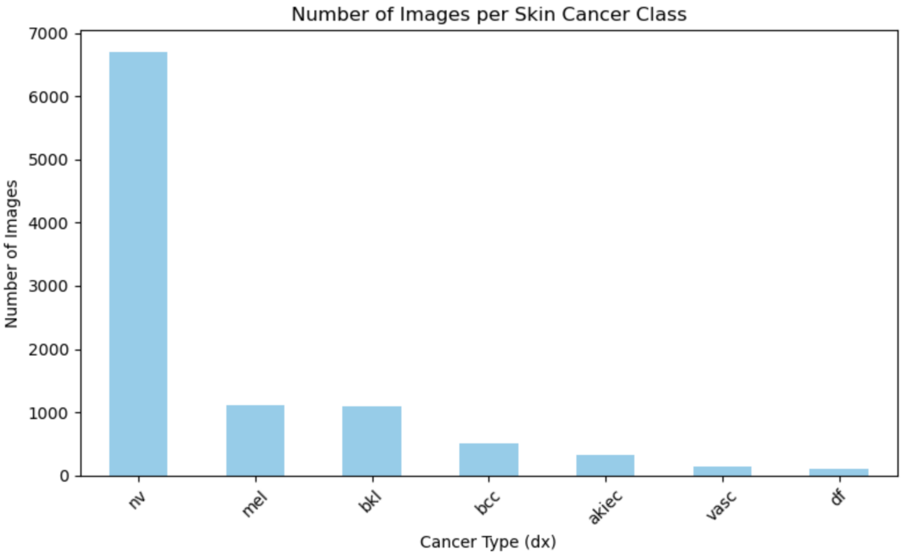


Figure 1: Number of images for each skin cancer class in the HAM10000 dataset.

In summary, preprocessing the data guarantees all images are of the same size, more varied in terms of augmentation, optimally trained with normalization, and labeled appropriately for the models.

2.3 Model Selection and Setup

This task compares the two main types of models used in image classification. The first one is Convolutional Neural Networks, or CNNs. The second one is Vision Transformers, or ViTs.

- **Three CNN models are utilized:**
  - **ResNet18:**
    ResNet18 is a simple and popular CNN. It utilizes special “skip connections”

    that allow the model to learn more efficiently. Although not that deep, it delivers strong results.

  - **ResNet34:**
    ResNet34 is similar to ResNet18 but with extra layers. This can make it learn more complex patterns within the images. It may train slower, but occasionally it does better on harder tasks.
  - **EfficientNet-B0:**
    EfficientNet-B0 is a new CNN. It is designed to be very efficient, i.e., it does not take much memory or time to train. With fewer resources, it can still perform very well.

- **Three Vision Transformer models are used:**
  - **ViT Tiny Patch16 224:**
    It is a small Vision Transformer model. It splits the image into small patches and learns from them. It does not have as many layers as large ViT models so it trains faster but maybe is not as powerful.
  - **ViT Small Patch16 224:**
    This is a bit bigger than ViT Tiny. It has more layers, so it can pick up on more features of the images. It can handle better results, but it will have to be trained longer.
  - **Swin Transformer Tiny:**
    Swin Transformer is a form of ViT. Instead of viewing the entire image, it views small windows, or patches, and then combines information together. This makes it possible for it to be able to focus on other parts of the image and can assist it in becoming better at handling details.

Model Initialization and Training

All the models begin with weights that they have learned from the large ImageNet dataset. This is referred to as pretraining. Pretraining enables the models to learn general image features, which prepares them to learn from smaller medical datasets more efficiently.

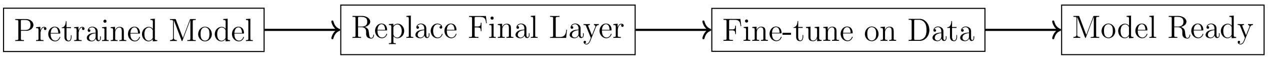


Figure 2: Model training workflow: Each model is pretrained, the final layer is replaced, then fine-tuned on the HAM10000 dataset.

For this assignment, the final layer of both models is removed. The new final layer is provided with seven outputs—each for one of the skin lesion classes in the HAM10000 dataset. After that is accomplished, the models are fine-tuned on the HAM10000 training dataset. This process of training the models is called fine-tuning. Fine-tuning helps the models to adjust to the special features of skin lesion images.

2.4 Training Procedure

The goal of training is to get all the models to learn to accurately classify skin lesion images into one of seven possible classes. Each model tries to get better at this task through ongoing training. The training process utilizes a notion called loss function. The loss function in this project is termed as cross-entropy loss. Cross-entropy loss measures the accuracy of the model's predictions towards the true responses. The more errors the model makes, the better the value of loss. When the model gets better, the value of loss is reduced. In order to train the model, an optimizer is utilized. The optimizer used in this project is Adam. Adam helps to adjust the model's weights after each batch of images. This helps learning be faster and more stable for most models. The batch size used is 32. This is when the model takes in 32 images at a time before updating learning. Using batches helps the model learn faster and improve memory use. The models are then trained for five epochs. An epoch is the model seeing all of the training images once. To train more epochs can sometimes be better, but five is used here due to time constraints. The model is validated after each epoch on the validation set. The validation set contains images the model did not train on. When the model does better on the validation set than before, the model's weights are saved. This keeps the best version of the model for final testing. This training is performed for each of the models, so that all of the models have the best chance to learn from the data and generalize to unseen images.

Below is an example of what is produced when training ResNet18. It shows image and label shapes as input, batch numbers in each split, and model accuracy gain for five rounds (epochs). The greatest validation accuracy is also recorded.

```
Image batch shape: torch.Size([32, 3, 224, 224])
Label batch shape: torch.Size([32])
First 5 labels: tensor([4, 0, 5, 5, 5])
Train batches: 227
Validation batches: 39
Test batches: 49

Training RESNET18
Epoch 1/5: Train Loss 0.7580, Train Acc 0.7410, Val Acc 0.8063
Epoch 2/5: Train Loss 0.4808, Train Acc 0.8236, Val Acc 0.8160
Epoch 3/5: Train Loss 0.3925, Train Acc 0.8569, Val Acc 0.8144
Epoch 4/5: Train Loss 0.3250, Train Acc 0.8792, Val Acc 0.8224
Epoch 5/5: Train Loss 0.2880, Train Acc 0.8926, Val Acc 0.7902
Best Val Acc: 0.8224
```

2.5 Evaluation Metrics

In order to observe different models' performance, different measures are used. They all differ regarding the good and bad of the model.

- **Accuracy**: Accuracy tells us what percentage of the predictions are true. Being very accurate within a model tells us that the model is predicting correctly most of the times. Accuracy is easy to grasp and gives us an instantaneous summary of performance.

- **F1-score**: F1-score is helpful if classes are unbalanced. Certain skin cancers are far less common than others. F1-score takes precision (how many positives were predicted who actually were positives) and recall (how many positives were found out of the actual positives) into account and gives a single figure helpful if there are fewer images in one class.

- **Confusion matrix**: Confusion matrix is a table that presents true and false prediction figures for each class. Each row indicates the actual class, and each column indicates the class predicted. The matrix can be employed in an attempt to identify classes commonly confused with other classes by the model. For example, it can show whether the model turns out to be one type of skin lesion mistaken for another.

- **ROC curve**:ROC curve (Receiver Operating Characteristic curve) is drawn for each class. It approximates how well the model can distinguish between one class and other classes. The curve is a graph of true positive rate vs. false positive rate at different thresholds. The curve that rises steeply and stays high signifies that the model can distinguish that class. Area under the ROC curve (AUC) is also computed; greater AUC signifies better discrimination.

Below are the results for the EfficientNet-B0 model:

Confusion Matrix and ROC Curve

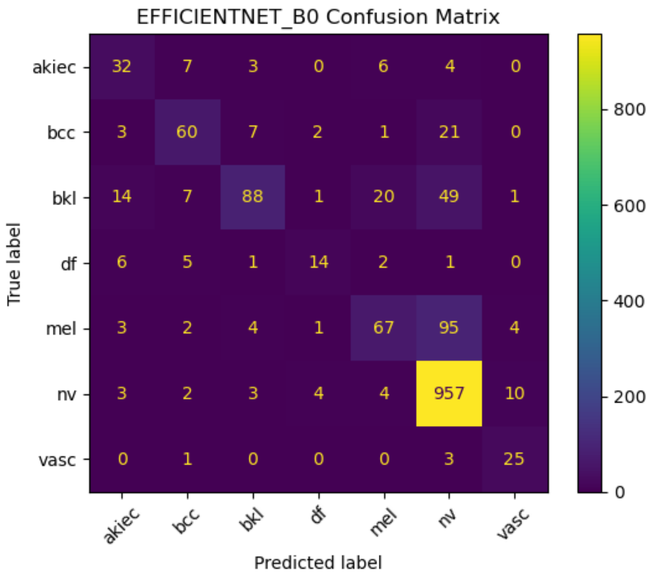


Figure 3: Confusion matrix for EfficientNet-B0 model on the test set. The rows show the true classes, and the columns show the predicted classes.

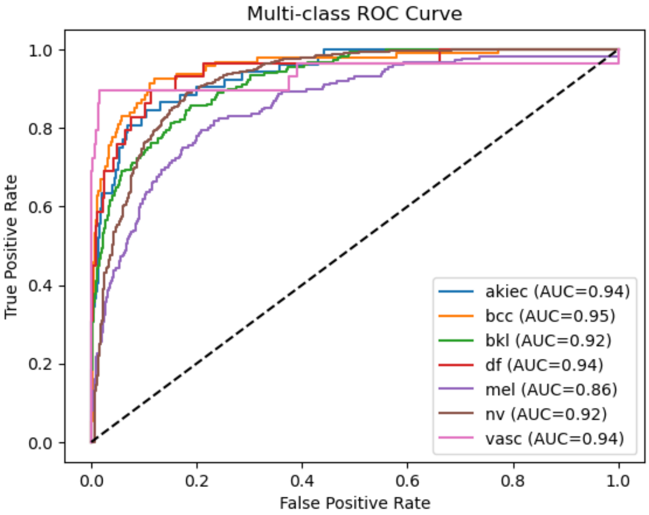


Figure 4: Multi-class ROC curves for all skin lesion classes using EfficientNet-B0.

Classification Report

```
              precision    recall  f1-score   support

       akiec       0.52      0.62      0.57        52
         bcc       0.71      0.64      0.67        94
         bkl       0.83      0.49      0.62       180
          df       0.64      0.48      0.55        29
         mel       0.67      0.38      0.49       176
          nv       0.85      0.97      0.91       983
        vasc       0.62      0.86      0.72        29

    accuracy                           0.81      1543
   macro avg       0.69      0.63      0.65      1543
weighted avg       0.80      0.81      0.79      1543
```

Table: Precision, recall, and F1-score for each class. The support column shows how many true cases there are for each class. All of these metrics combined paint a whole picture of model performance. They help to show not just overall performance, but also where the model is performing best and where it must improve.

2.6 Explainable AI Methods

This project employs specialized methods to visualize what the models are ”thinking” when they make a prediction. Such methods allow doctors and users to more comfortably trust and understand the results.

Grad-CAM: Grad-CAM is used on Vision Transformer models and CNN models. Grad-CAM generates a heatmap using the gradients (differences) of the output of the model with regard to the image. The heatmap shows the most important areas of the image to the decision of the model. The bright or red areas signify that the spots were important.

Score-CAM: Score-CAM is used on the CNN models. Score-CAM does not use gradients, as opposed to Grad-CAM. It attempts to observe which regions of the image have

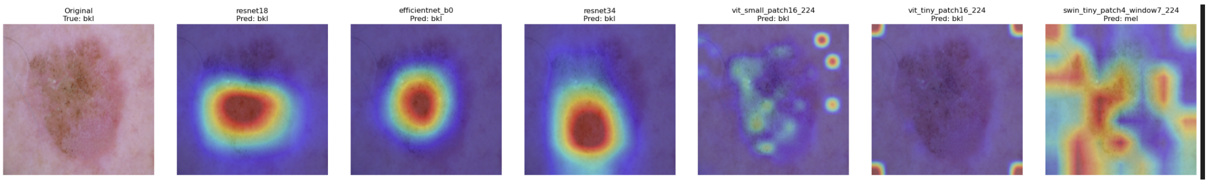


Figure 5: Example XAI heatmaps from six different models on sample skin lesion images.

the most influence on the output score. It then generates a heatmap to represent the regions that had the most influence. This is a more stable process and, in some cases, creates sharper maps.

Attention Rollout: Attention Rollout is used on the Vision Transformer (ViT) models. Transformers use a thing called "attention" to concentrate on different areas in an image. Attention Rollout consolidates attention maps across all the transformer's layers. The final heatmap represents the areas that the model focused on most to make a prediction.

Swin Transformer Window Attention: Swin Transformer Window Attention is a special visualization for the Swin Transformer. Swin divides the image into small windows (patches) and performs self-attention within each window. The project visualizes which windows or patches the model paid attention to. This is helpful to observe how the Swin Transformer makes use of local areas of the image.

Visualization Process: For each of the above methods, the important regions found by the model are marked on a heatmap. That heatmap is superimposed over the original image. That way, it is easy for anyone to tell what part of the skin lesion the model was focusing on when it was making the prediction. If the model is focusing on the wrong area, it is immediately obvious.

Goal: These XAI methods help to make the model's predictions more explainable. They can help doctors trust the AI more and make better-informed decisions because they can check if the model is looking at the right places in the image.

3 Results

3.1 Classification Performance

The project compared six deep learning models on the skin lesion dataset. Three of the models were CNNs. The remaining three were ViTs.

The CNN models included ResNet18, ResNet34, and EfficientNet-B0. ResNet18 is a small network with shortcut connections. ResNet34 is a deeper version of ResNet18. EfficientNet-B0 is a more recent model that tries to invest its computing power wisely.

The ViT models employed were ViT Tiny Patch16 224, ViT Small Patch16 224, and Swin Tiny Patch4 Window7 224. ViT Tiny is the smallest Vision Transformer that was attempted. ViT Small is slightly larger and can learn slightly more from the data. Swin Tiny uses a new method of looking at small "windows" or patches of the image.

All six models were pre-trained with weights obtained on the large ImageNet dataset. Each of the models was then fine-tuned on the HAM10000 skin cancer images. Training and testing were performed carefully on the train, validation, and test splits.

The performance of each model was measured on the test set. The main measures used were accuracy and average F1-score. Accuracy is the number of predictions that were correct out of all test images. F1-score is a measure of how well the model balanced

”finding the right class” and ”avoiding errors.”

The table below shows the results for each model:

Model	Test Accuracy (%)	F1-score (avg)	Key Notes
ResNet18	79	0.78	Focused heatmaps
ResNet34	81	0.80	Slightly better than 18
EfficientNet-B0	81	0.79	Best CNN, efficient
ViT Tiny Patch16 224	77	0.78	Broader attention, fast
ViT Small Patch16 224	81	0.79	Best overall in this run
Swin Tiny Patch4 Window7	78	0.77	Windowed attention

Table 2: Test accuracy and F1-score for all models.

Observation: All of the models had similar performance. The most accurate was ViT Small Patch16 224, but only by a little. The distinction between ViTs and CNNs was not large for this dataset. This suggests that both types of models can do well for skin cancer image classification.

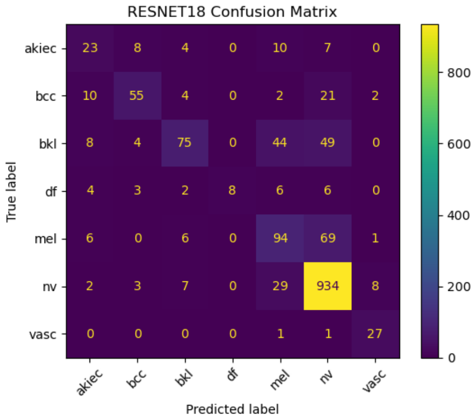
The CNN models, including ResNet and EfficientNet, are already strong for medical images. The newer transformer models, including ViT and Swin, can already match their performance. This is hopeful for the future because transformers can offer extra benefits like explainability and flexibility.

3.2 Confusion Matrices and Detailed Metrics

For each model, confusion matrices were computed from test set outputs. The matrices help in highlighting what type of skin lesions were correctly identified and where errors were made.

The matrices showed that most errors were made in lesion types that looked similar. For example, models sometimes confused melanocytic nevus (”nv”) and melanoma (”mel”) with each other. This was common for both CNN and ViT models.

ResNet18: ResNet18 handled most of its correct predictions for the ”nv” class, though it got ”bkl” and ”mel” mixed up a few times. It was weaker on rare classes like ”df” and ”akiec.”

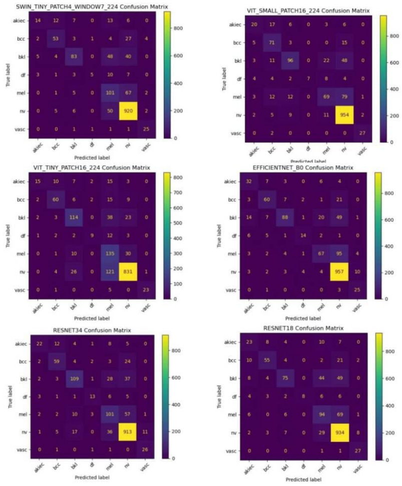


	akiec	bcc	bkl	df	mel	nv	vasc
akiec	23	8	4	0	10	7	0
bcc	10	55	4	0	2	21	2
bkl	8	4	75	0	44	49	0
df	4	3	2	8	6	6	0
mel	6	0	6	0	94	69	1
nv	2	3	7	0	29	934	8
vasc	0	0	0	0	1	1	27

The ResNet18 detailed classification report indicates that ”nv” recorded the highest f1-score (0.90). Rare classes such as ”df” and ”akiec” registered lower scores, which is consistent with what is observed in the confusion matrix. The vast majority of models exhibited similar trends: high scores for ”nv,” but lower for rare classes.

```
RESNET18 Test Classification Report:
              precision    recall  f1-score   support
       akiec       0.43      0.44      0.44        52
         bcc       0.75      0.59      0.66        94
         bkl       0.77      0.42      0.54       180
          df       1.00      0.28      0.43        29
         mel       0.51      0.53      0.52       176
          nv       0.86      0.95      0.90       983
        vasc       0.71      0.93      0.81        29
    accuracy                           0.79      1543
   macro avg       0.72      0.59      0.61      1543
weighted avg       0.79      0.79      0.78      1543
```

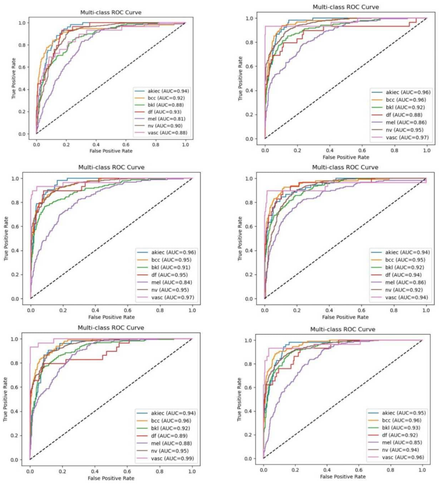
All the other models also did so in a consistent fashion. For example, ViT Small Patch16 had higher recall for ”nv” and ”vasc” but lower for ”df” and ”akiec.” Confusion matrices by all the model are shown below.



3.3 ROC Curves

ROC curves for every class and every model were constructed. All the six models were good, with mean AUC ranging from 0.92 to 0.96. This shows that the models can distinguish between skin lesion categories for the majority of cases.

The models were best on the ”nv” and ”vasc” classes. For them, the AUC was up to 0.98 in some of the models. This shows that the models hardly confused these classes with other classes. The most challenging classes were ”df” and ”akiec.” For them, the AUC dropped to about 0.86, especially for small models like ViT Tiny and ResNet18. This shows that the models sometimes confused these classes with others.



All the models performed similarly on the ROC curves. There wasn't a wide difference between CNNs and ViTs on this measure. This tells us both models are excellent choices for this task. It also tells us that doctors can confidently use both models to do a good job of distinguishing between most skin lesion types.

3.4 Explainability (XAI) Results

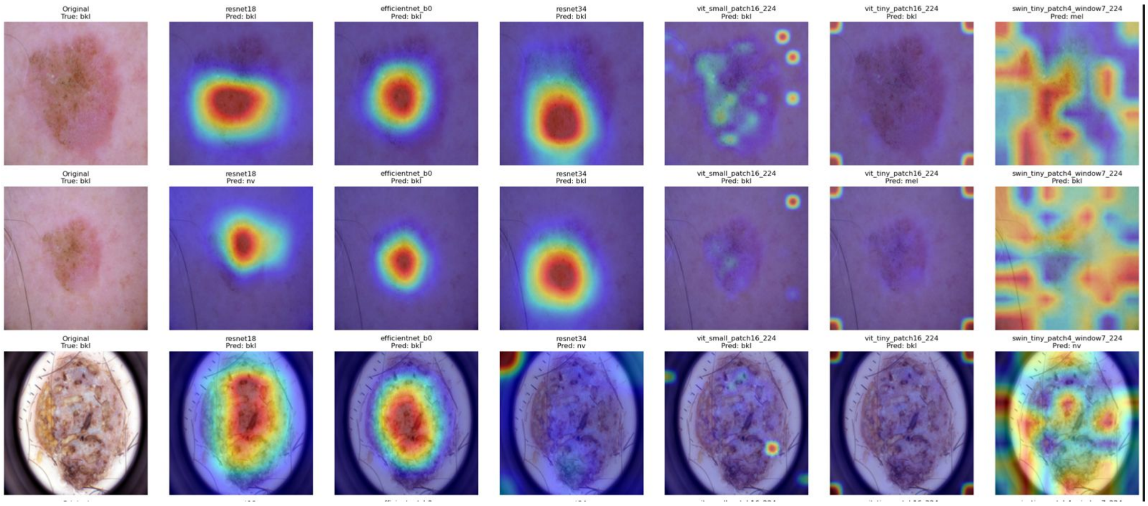
The project also explored how well the models could ”explain” their predictions. That is, showing what parts of the image the models observed when making their decisions.

For the CNN models (ResNet18, ResNet34, and EfficientNet-B0), Grad-CAM and Score-CAM heatmaps were specific and unique. The heatmaps tended to highlight the middle lesion of the image. These areas were coinciding with the regions which an actual physician would investigate to diagnose skin cancer. This made the CNNs easy to trust, as it was clear that they were investigating the important spots.

In the case of the Vision Transformer models, like ViT Tiny Patch16 224 and ViT Small Patch16 224, explanations were a little bit different. The Attention Rollout and Grad-CAM methods generally highlighted larger or more spread-out regions. In some instances, the ViT models looked beyond the lesion to the surrounding skin areas. That shows that transformer models try to learn patterns from a larger area. They can include more context, which may be useful there, but sometimes it is more difficult to tell what features they are using.

The Swin Transformer yielded a unique type of explanation. The attention heatmap was chunky, suggesting that the model was attending to small windows or patches within the image. The Swin Transformer sometimes distributed its attention between the center of the lesion and edges. This type of attention is beneficial for the detection of border and local details but at times makes it harder to read the explanation in a quick glance.

Example XAI Visualization Grid:



In the general sense, the CNN models' heatmaps were the most intuitive. They showed exactly where the model was "looking," and those areas meant something to humans. The ViT models spread their attention more diffusely, which could be beneficial but sometimes less definitive for doctors to understand. Each model is good in explainability, depending on the case.

3.4.1 Error Analysis

On testing, the project also looked at where and why the models were getting it wrong. Most of the mistakes were between visually similar classes. For example, the models would get "melanoma" (a malignant skin cancer) confused with "nevus" (a benign mole). This can be seen in the confusion matrices, where there are numerous wrong predictions between these two classes.

The "melanoma" and "nevus" mix-up is not exclusive to an AI model. It's also easy for human doctors to make the mistake since the lesions in real images are often very similar. Sometimes, the models were also confused between "bkl" (benign keratosis) and other benign lesions.

By looking at these mistakes, one can determine what types of lesions the models are most difficult to differentiate. This is useful because it shows where the models must be improved and where more training data or better features would benefit them most. Error analysis also cautions doctors when they need to be more careful in trusting the prediction of the model.

4 Discussion

This project compared CNN and ViT models with skin lesion classifying task. All six models performed well, with test accuracy and F1-scores above 0.77 for all models. EfficientNet-B0224 had the best, but all the other models were in close proximity.

When considering how the models perform their predictions, CNNs and ViTs exhibited varying strengths. The CNN models (ResNet18, ResNet34, EfficientNet-B0) generated heatmaps that were highly concentrated on the lesion itself. The explanations made it simple to believe the model's response because it was evident the model was examining the proper location. Physicians could easily verify the heatmap to determine whether the model concentrated on the area of the lesion.

The ViT models worked differently. Sometimes, the transformer models (ViT Tiny, ViT Small, Swin Transformer) highlighted a broader area, like the surrounding skin of the lesion. That means transformers consider more global image context. In some cases, this might help the model pick up important background information. It can also make the explanation less interpretable because it is not apparent which area of the image the model picked up as most important.

The Swin Transformer generated blocky attention maps, reflecting its attention to patches in small areas. This can be helpful in searching for edges or patterns at the edge of a lesion, but the maps are not as smooth and tend to be hard to read.

It was observed that most of the errors were between classes that resemble one another, i.e., ”melanoma” and ”nevus.” It is a tricky problem for AI models as well as medical doctors. The models struggled more with classes having fewer images or those that were closely similar to other classes.

ROC curves showed that all models were capable of discriminating between most classes very well with AUC ¿ 0.92 for almost all classes. That is, the models are reliable for this kind of classification, though there remains a potential improvement in most challenging cases.

Explainable AI (XAI) methods were beneficial for this project. They provided a comparative analysis between decision-making of ViTs and CNNs. Doctors and users can use these attention maps and heatmaps to check if the response of the model is accurate and trustworthy.

5 Conclusion

This work compared three CNNs and three Vision Transformers on the task of skin lesion classification. They all performed well, with the best performing results being from ViT Small Patch16 224 and EfficientNet-B0. Both model classes performed high accuracy but differed in explainability. CNNs provided plain and specific heatmaps. They are easy for doctors to comprehend and trust. ViTs would illuminate a larger area, sometimes offering more info but weakening the explanation in the process. The results show that a CNN and a ViT are both good choices for this kind of medical image task. Whether to use one or the other can be determined by whether the user finds more important to them: highly local explanations (CNN) or more contextual info (ViT). Explainable AI visualizations including Grad-CAM, Score-CAM, and Attention Rollout allow individuals to understand and trust deep learning models. They are very important in medical AI where every decision must be understood and verifiable. In the future, there could be the potential for further enhancement by combining the best of both CNNs and ViTs, or more diverse dataset training. There is potential for continued work to further enhance explanations and to aid in the reduction of error between types of lesions that appear similar.

6 Reference

- Ivanovska, T. (2025). *Lecture6_AdvancedCNNsVisionTransformers [Lecture slides]*. OTH Amberg-Weiden. Available at: `https://moodle.oth-aw.de/mod/folder/view.php?id=190741` (Accessed July 2025).

- Tschandl, P., Rosendahl, C., & Kittler, H. (2018). *The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions.* Scientific Data, 5, 180161. `https://doi.org/10.1038/sdata.2018.161`

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. `https://doi.org/10.1109/CVPR.2016.90`

- Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* Proceedings of the 36th International Conference on Machine Learning (ICML). `https://arxiv.org/abs/1905.11946`

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* International Conference on Learning Representations (ICLR). `https://arxiv.org/abs/2010.11929`

- Liu, Z., Lin, Y., Cao, Y., et al. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.* Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012-10022. `https://arxiv.org/abs/2103.14030`

- Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.* Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618-626. `https://doi.org/10.1109/ICCV.2017.74`

- Wang, H., Wang, Z., Du, M., et al. (2020). *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 24-25. `https://arxiv.org/abs/1910.01279`

- Abnar, S., & Zuidema, W. (2020). *Quantifying Attention Flow in Transformers.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4190–4197. `https://arxiv.org/abs/2005.00928`