Task 1a

Dataset and Model: We used three datasets DermaMNIST; BloodMNIST and PathMNIST. The DermaMNIST set is the smallest set for initial training and longer evaluation. The BloodMNIST to see if the same CAM methods provide similar results on a different dataset with similar class numbers to predict. The PathMNIST is used to see if a larger dataset and more classes have an impact on the CAM methods. We used a pre-trained resnet50 with target layer *layer3* but also played around with a CNN architecture we implemented ourselves.

To assess the performance of the different CAM-methods: There are different options to do this.

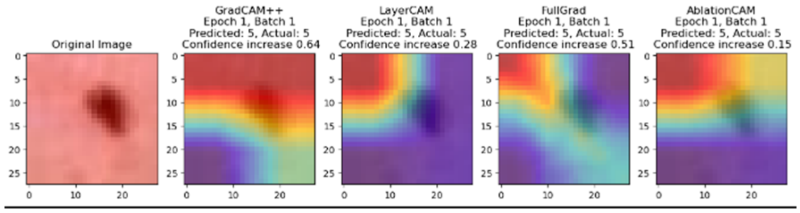
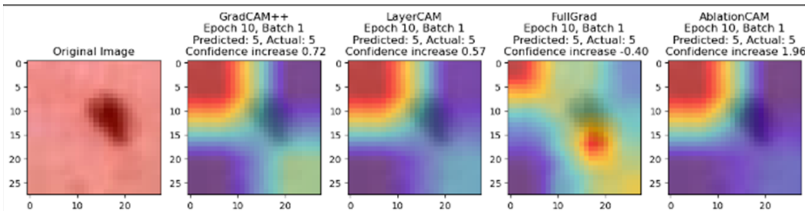
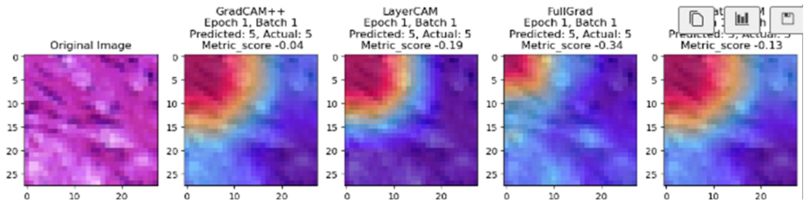
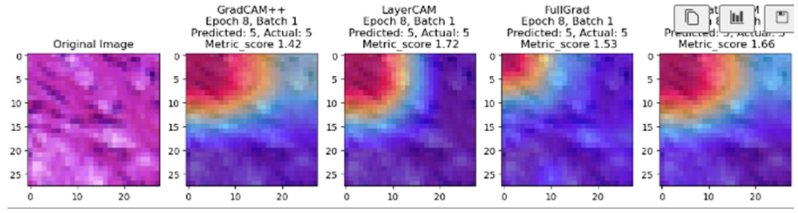
- The first is to compare **ground truth annotations** for what regions of the image are important for each class to the CAM output. However, as there are no ground truths provided, Intersection over Union (IoU) or Jaccard Score are not metrics we could use.
- Another possibility is **qualitative** assessment: By adding a function which prints the CAM-Method output overlayed on the image, then experts can rate the quality of the CAM outputs in terms of interpretability and alignment with human understanding.
- For **quantitative** assessment, we used *ROADCombined(percentiles=[20, 40, 60, 80])* and *ClassifierOutputSoftmaxTarget* metric. The first method combines evaluation from multiple perspectives, considering how well the CAM highlights pixels across different activation thresholds. A higher score would typically indicate that the CAM is effectively highlighting the pixels most relevant to the model's prediction, while a lower score might indicate that the CAM is less effective or highlighting irrelevant areas. That is why we decided to give the CAM with the highest ROADCombined score the most points. we also measured the change in the confidence, after softmax, that's why we also used *ClassifierOutputSoftmaxTarget*. And then logged the confidence increase/decrease.
- The last performance assessment was done through Computational **Efficiency**: By recording the time taken by each CAM method to process an image we evaluated how efficient each method is.

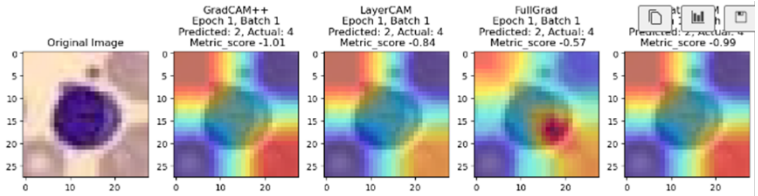
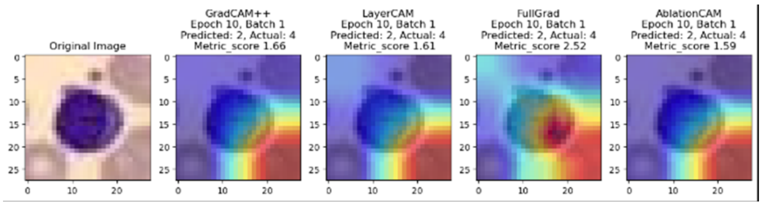
Dataset	Evaluation	GradCAM++	LayerCAM	FullGrad	AblationCAM
DermaMNIST	ROAD Combined	0.72	0.57	-0.4	**1.96**
	Avg. Time	0.05	**0.04**	0.46	0.53
	Confidence increase	-2.24	-2.38	-2.69	**-2.19**
	Model ACC:	71. 57			
PathMNIST	ROAD Combined	1.42	**1.72**	1.53	1.66
	Avg. Time	0.49	**0.46**	0.56	0.65
	Confidence increase	-0.91	-0.95	-1.00	-0.65
	Model ACC:	94.37			

BloodMNIST	ROAD Combined	0.04	-0.04	**0.211**	-0.66
	Avg. Time	**0.033**	0.04	0.44	0.58
	Confidence increase	1.66	1.6	**2.52**	1.58
	Model ACC:	89.56			

Which specific method outperforms the rest? From our results, Generally, LayerCam was the fastest method and GradCAM++ was the second in terms of time, In terms of cam metric, each dataset had a winner, AblationCAM worked best with DermaMNIST, the LayerCAM was suitable had the highest metric scores with PathMNIST and FullGrad was the best for BloodMNIST. It is worth mentioning that FullGrad method is using only all bias layers so that means that the bias layers were helpful for the BloodMNIST

Task 1b

Dataset	Epoch	Visualization of highlighted features
DermaMNIST	Epoch 1	 Original Image GradCAM++ Epoch 1, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.64 LayerCAM Epoch 1, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.28 FullGrad Epoch 1, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.51 AblationCAM Epoch 1, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.15
	Epoch 10	 Original Image GradCAM++ Epoch 10, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.72 LayerCAM Epoch 10, Batch 1 Predicted: 5, Actual: 5 Confidence increase 0.57 FullGrad Epoch 10, Batch 1 Predicted: 5, Actual: 5 Confidence increase -0.40 AblationCAM Epoch 10, Batch 1 Predicted: 5, Actual: 5 Confidence increase 1.96
PathMNIST	Epoch 1	 Original Image GradCAM++ Epoch 1, Batch 1 Predicted: 5, Actual: 5 Metric_score -0.04 LayerCAM Epoch 1, Batch 1 Predicted: 5, Actual: 5 Metric_score -0.19 FullGrad Epoch 1, Batch 1 Predicted: 5, Actual: 5 Metric_score -0.34 Metric_score -0.13
	Epoch 8	 Original Image GradCAM++ Epoch 8, Batch 1 Predicted: 5, Actual: 5 Metric_score 1.42 LayerCAM Epoch 8, Batch 1 Predicted: 5, Actual: 5 Metric_score 1.72 FullGrad Epoch 8, Batch 1 Predicted: 5, Actual: 5 Metric_score 1.53 Metric_score 1.66

BloodMNIST	Epoch 1	 Original Image GradCAM++ Epoch 1, Batch 1 Predicted: 2, Actual: 4 Metric_score -1.01 LayerCAM Epoch 1, Batch 1 Predicted: 2, Actual: 4 Metric_score -0.84 FullGrad Epoch 1, Batch 1 Predicted: 2, Actual: 4 Metric_score -0.57 Predicted: 2, Actual: 4 Metric_score -0.99 0 5 10 15 20 25 0 10 20
	Epoch 10	 Original Image GradCAM++ Epoch 10, Batch 1 Predicted: 2, Actual: 4 Metric_score 1.66 LayerCAM Epoch 10, Batch 1 Predicted: 2, Actual: 4 Metric_score 1.61 FullGrad Epoch 10, Batch 1 Predicted: 2, Actual: 4 Metric_score 2.52 AblationCAM Epoch 10, Batch 1 Predicted: 2, Actual: 4 Metric_score 1.59 0 5 10 15 20 25 0 10 20

- ☐ Substantial difference between the features that are selected for explanation?
  - ○ For PathMNIST the features selected for explanation stayed quite constant through the epochs, and the reason for this that PathMNIST is a large dataset for the number of iteration in the first epoch was larger in comparison with the other datasets.
  - ○ For DermaMNIST LayerCAM and AblationCAM used similar features selected for explanation throught the training while FullGrad and GradCAM++ changed quite a lot which features they selected for explanation .
  - ○ For BloodMNIST the FullGrad method selected as only method different aspects for explanation. All the other, especially GradCAM++ and LayerCAM agreed on the features selected for explanation, and probably this is because the target layers for the FullGrad method are all bias layers.
- ☐ At what point can onestop training because features look good enough
  - ○ For PathMNIST after epoch 7 this can be related to the way larger training set than in the other datasets
  - ○ For both BloodMNIST and DermaMNIST, between the epochs the CAM's output is different features selected for an explanation, therefore the point at which the training should stop would be when the validation score does not improve anymore.
  - ○ Also for all datasets, we believe that the ROAD metric stayed positive for multiple epochs ~ 5 epochs, and there are no huge jumps in the metric value.