**Heatmap evaluations about what a Covid-19 image classifier uses for predictions**

In this project I will be looking at heatmaps created by GradCam methods to evaluate what features a neural network image classifier pays attention to in chest X-ray images. The dataset for this project has been taken from Kaggle: https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset?resource=download

The dataset contains 378 images in total. The dataset is divided into train and test files, and these files both contain three files named Covid, Normal and Viral Pneumonia. The training data contains 312 images and the test data 66 images. The covid-19 and control images are from a publicly released Github account by the University of Montreal professors, and the Viral Pneumonia data is from the Radiological Society of North America -website. It is probable that all the images then are not from the same hospital, but no specific information on that is shared on the Kaggle website.

The code has been created with the help of ChatGPT, the coding language is Python, using PyTorch. The images are resized to 224x224 pixels to have a fixed input, where the image size is not a shortcut for the model to predict labels.

I used GradCam to visualize with a heatmap the areas that my deep learning model “learns” from the training images, and to see if the areas actually contain medically relevant information, or if the patterns learned by the model are something else, like for example, image features that are caused by the type of x-ray camera used in the source hospital.

Some of the following images show clear examples of information used by the model that is not medically relevant, whereas in others the input used is located in the lung areas.

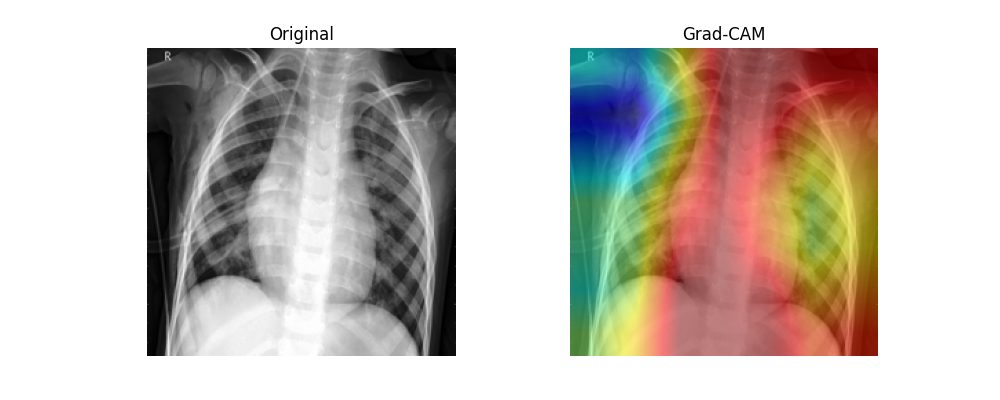


Figure 1: Image 0108.jpeg from the Covid test set. The model predicted the label of the image correctly.

In Figure 1. we can see that most of the information gathered is focused on the left side of the image, mostly to the armpit of the patient. This image was correctly classified as Viral Pneumonia. In this picture we can see the patient is not entirely in the center, leaning to the other side of the image, leaving a bigger empty area to the other side under the patient’s armpit and the side of the body. DeGrave et al. (2021) found a consistent relationship between patient positioning and Covid-19-positive and Covid-19-Negative radiographs in each of their datasets (DeGrave et al. 2021, 5). Some of these sorts of issues could be dealt with by cropping the images to not contain areas outside of the lungs, but then again, differences in disease severity could according to Ahmed et al. (2023) could have an effect on patient positioning, which a learning model could use as a shortcut (Ahmed et al. 2023, 7).

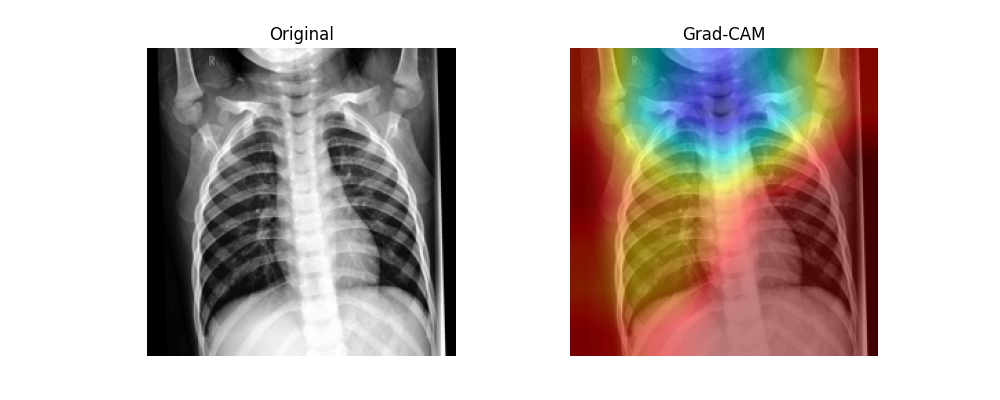


Figure 2: Image 0115.jpeg from the Normal test set. The model predicted the label correctly.

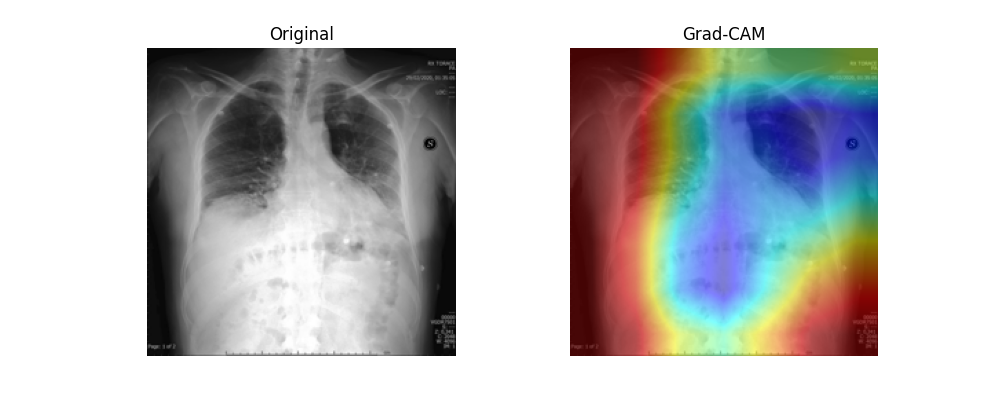
In figure 2 we can see a heatmap of a child’s X-ray showing the throat area to be the biggest hotspot for information, which seems to be an incorrect area to be looking at patterns for medical signs of Covid-19 in chest X-ray images, but of course if there is some recognizable patterns in the throat of a patient that sign to a Covid-19 infection, then this would be correct.   


Figure 3a: Image 098.jpeg from the Covid test set. Prediction label unknown.

In Figure 3a we can see that the model has taken some information from the chest and stomach area, but a large part of the image information used is from the patient’s armpit again, and from the S-mark located in the armpit. It might be a lateralization mark and DeGrave et al. (2021) say that those originate during the radiograph acquisition process, which they found to be a key visual feature that their models used to categorize their test images (DeGrave et al. (2021, 5).

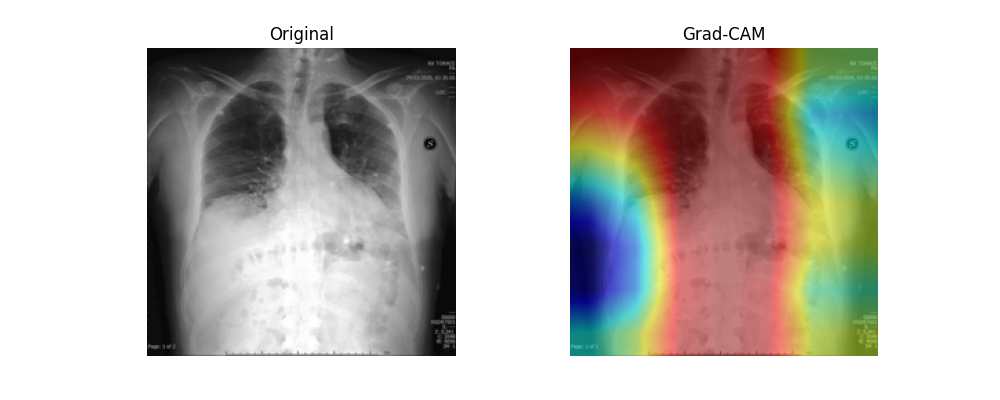


Figure 3b: Image 098.jpeg from the Covid test set. The model has been retrained in between; this time the model predicted the label to be Viral Pneumonia.

When the model is retrained, it might not use all the same features of the image to classify as it has done earlier. In picture 3b, we can see compared to the earlier heatmap of 3a of the same picture, that the model now picks details from the side of the patient, while also some features like the armpit area and the symbol S are still used. Here we can see, that the details the model uses for categorization are somewhat unstable and makes it a highly unreliable model for this type of labeling.

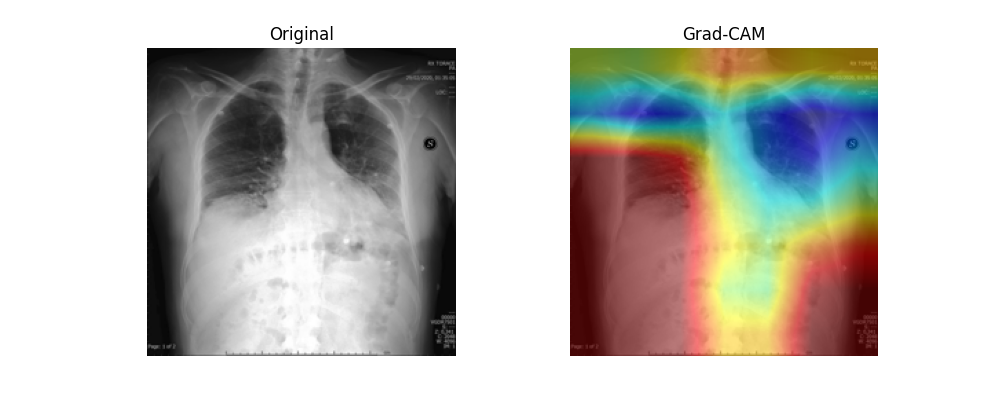


Figure 3c: Image 098.jpeg from the Covid test set. Model is trained for the third time; this time it predicted the label correctly.

In figure 3c we can see that after training the model the third time, the most consistent feature for prediction looks to be the armpit of the patient where the symbol S is also. This time the model got a 98,48% prediction accuracy on the test set. To control for this, the location where the symbol S is could be covered in all the pictures of the training and test sets, to make the data uniform in that sense.

An important detail to note also, is that we really don’t know how clear visual medical information is in every Covid positive image to know if the medically relevant visual details for diagnosing Covid-19 are present in these images. Same could be said for the Viral Pneumonia data.

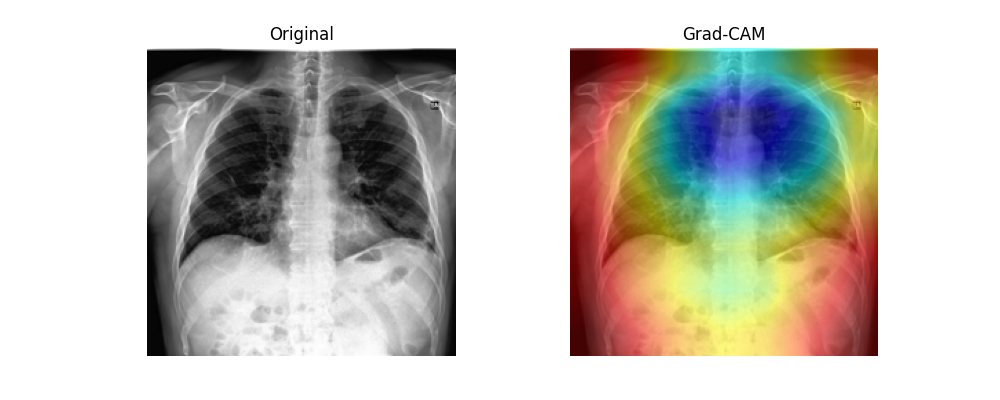
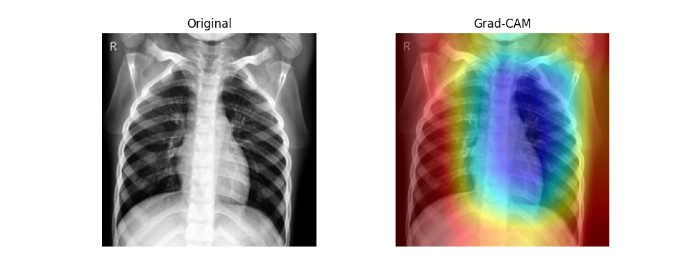
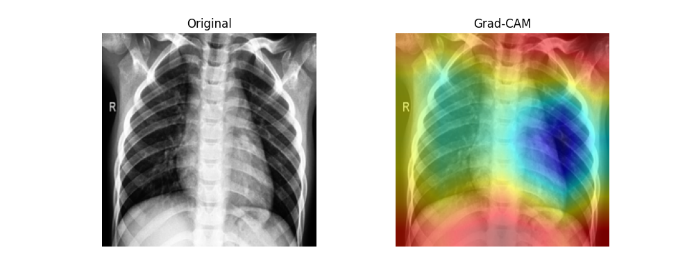
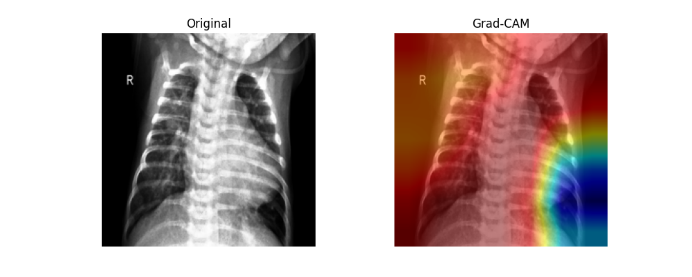


Figure 4: Image 0100.jpeg from the Covid test set, the mode categorized it correctly.

In figure 4 we can see that the model used primarily areas of the upper middle parts of the lungs, and it categorized the image correctly as Covid-19, but to analyze further if the areas contain Covid-19 pathology I would need to be a medical professional, which I am not. Then again there could be a pattern used similar to figure 2, which is the structure of the spine.

We do not have the information of sex of the patients of these images, which DeGrave (2021) found to be an easily teachable shortcut, and male sex correlates more with being Covid-19-positive than female sex (DeGrave et al. 2021, 6). It could be that this detail is affecting how my model categorizes patients also, although I did not find any obvious examples of that with a brief review. One separating factor between an adult male, adult female, and a child X-ray could be the size of the lungs.

We also do not have the details of the equipment used, like are there different types of X-ray machines used, which could lead the model not being able to generalize, like what Fernandez-Miranda (2024) found to happen with their model, when they tried to predict labels of X-ray images of machines with different types of response functions (Fernandez-Miranda 2024, 6).

Kuva, joka sisältää kohteen röntgenfilmi, Lääketieteellinen kuvantaminen, radiologia, radiografia

Tekoälyn generoima sisältö voi olla virheellistä.

Figure 5: Other examples of the X-ray heatmaps.

**Conclusion**

This model is not yet working reliably, there are many shortcuts that it can and does take advantage of, like the darker areas beside the patient. A big feature that the model seems to be paying attention to is the darker areas beside the patient and in the lungs, and if the images could be cropped so that no areas beside the lungs are usable for the model, the shortcut of the darker area beside the patient could be controlled. A clear example of a lateralization marker being used for prediction is the series of images of figure 3, where we see the model using the symbol S for all three predictions after each three rounds of retraining. This could be controlled by covering the area in all the data. In the data there are variables like patient positioning, lateralization, patient sex, hospital equipment and protocol, disease severity, and markings like lateralization markings, that need all to be paid attention to make sure the factors the model uses to make predictions are focused on the pathology of the virus.