# COMP9444 Project Summary

## Multi-label Chest Disease Classification on Chest X-ray Images

## MedicalMasters

Ali Ghadiri Modarres (z5485577), Hyun Wook Kim (z5483096), Jannis Becktepe (z5519475), Marla Huxhold (z5523235), Muhammad Ali (z5129534)

### I. Introduction

Modern medicine uses images such as X-rays, magnetic resonance imaging scans or computed tomography for the diagnosis of diseases. The images are evaluated by medical personnel, which need to identify parts showing signs of illnesses as well as give a prediction of the disease. These evaluations are time-consuming and difficult as diseases can occur in a lot of different shapes [1, p.1]. To address this, tools like convolutional neural network models (CNNs) have been developed, proving to be successful in image classification [1, p.2]. They can detect pathological structures in the images and assist in the diagnosing process. However, lately, vision transformers are gaining prominence as they outranked CNNs in representative visual benchmarks [2, p.12]. Both models are often too complex to understand how their process of decision-making is working [3, p.248]. For that reason, explainability methods should be applied to make the models more trustworthy and to increase their helpfulness for a user.
In our project, we trained a CNN and a vision transformer on a dataset of chest X-ray images. We compared their results to each other as well as two models pre-trained on the same dataset. Furthermore, we applied different explainability methods to test their usability. This report summarises our results and discusses the usefulness of the models for a user.

### II. Literature Review

The article “Convolutional Neural Networks in Medical Image Understanding: A Survey” (2020) gives an overview of award-winning CNNs and explains medical image tasks, and where they can be applied. It includes an explanation of tasks like classification explained for multiple diseases and variances of models.. They especially point out the importance of choosing the right architecture as well as the necessity of a big dataset of medical images [1, p.17].
The paper “Benchmarking and Boosting Transformers of Medical Image Classification” (2022) evaluates if the performance of vision transformers can be elevated by ImageNet pre-training to compete with CNNs in medical image classification by using up-to-date benchmarks. One main conclusion is that based on a good pre-trained model a transformer can compete with CNNs in medical imaging [2, p.15]. The initialization is especially important for the performance of a transformer as randomly initialised it cannot keep up with CNNs [2, p.19]. Furthermore, they conclude that domain-adaptive pre-learning leads to a better performance than ImageNet transfer learning [2, p.19].
The article “Explainable Artificial Intelligence (XAI) in deep learning-based medical image analysis” (2022) [4] provided an informative groundwork as it gives an overview of explainable artificial intelligence in medical imaging tasks. The paper discusses different methods in the categories of ease of use, validity, robustness, computational cost, necessity

to fine-tune, and open-source availability [4, p.11]. They conclude that Class Activation Mapping (CAM) gives accurate results for CNNs and use Gradient-weighted class activation mapping (Grad-CAM) as a generalization for all CNN models. Moreover, they highlight the usefulness of medical imaging of Layer-wise relevance propagation (LARP) [5, p.7].

## III. Methods

We used four models for the classification of chest X-ray images, each with a different architecture and pre-trained parameters. The first two models we chose were pre-trained on a variety of chest X-ray datasets including CheXpert, the dataset used in our project. They are both commonly known and used for image classification, which is the reason why we chose them.

DenseNet (Densely Connected Convolutional Networks) [5] is an architecture of neural networks in which each layer is connected to every other layer in a feed-forward fashion. It is known for its efficiency and compactness, especially beneficial for limited datasets. Originally designed for 224x224 images, we defined a custom transformation. Additionally, we expanded the number of channels from one to three to match the number of input channels, which we needed later for the downstream explanation methods.

The second already pre-trained model we selected is ResNet (Residual Networks) [6], an architecture containing residual blocks in the input of a block that is added to its output. It allows to train deeper networks by mitigating the problem of vanishing gradients. It introduces skip connections that bypass one or more layers, facilitating easier training. The initialization was analogous to DenseNet except for syntactical differences according to the model definition and implementation.

Furthermore, we wanted to compare the performance of the VisionTransformer (ViT) [7] to CNN models. Unlike CNNs, it is able to detect global features across images. As the ViT is pre-trained on ImageNet, we change the number of classes from 1000 (ImageNet) to 13 (CheXpert). Additionally, we unfreeze all parameters to enable transfer learning for the whole model. In the timm library, we implicitly replaced the classification head with a new linear layer by setting the num_classes parameters to the number of pathologies in our dataset.

The last model we trained is an example of ConvNeXt, a modern family of convolutional neural networks based on prior experiences with transformer models like ViT, introduced in the paper “A ConvNet for the 2020s” (2022) [8] It adopts a simpler design similar to traditional CNNs combined with several key modifications inspired by transformer models, such as LayyerScale and depthwise separable convolutions. These models enable efficiency and effectiveness for large-scale image recognition tasks, which makes them suitable as a modern competitor to the ViT in our model. In Particular, we selected the smallest version of the ConvNeXt family.

## IV. Experimental Setup

### Dataset

We use the CheXpert dataset [9] introduced by the Stanford ML Group in January 2019, which is a labelled collection of chest radiographs. The original dataset which can be found here contains 224,316 chest radiographs from 65,240 unique patients with labels for 14 common observations. Because the original dataset has a size of 471GB, we opted for a smaller version available on the Kaggle platform with a size of 12GB. The size difference is largely attributed to the resized radiograph images and consist of 191,010 images.

**Labels**
From the 14 labels of the CheXpert dataset, we only used 13 by omitting the “No Findings” label, because we concentrated on identifying the presence of pathologies and the application of explainability methods. The presence of -1 and NaN values added a layer of complexity to calculating the loss, making training a less straightforward process. -1 represents the uncertainty of the class, NaN represents unobserved class, 0 represents the absence of class, and 1 represents the presence of a class. For our task, we convert -1 to NaNs, and mask NaN values when calculating the loss considering only none-Nan classes.

**Parameters**
The number of parameters for our models are 6.9M, 23.5M, 85M, and 27M for DenseNet, ResNet, Vision Transformer, and ConvNeXt respectively. All parameters of the Vision Transformer and ConvNeXt were trained on the training dataset while only a subset of the DenseNet and ResNet were trained.

**Training, Validation, and Evaluation**
We trained our models on a subset compromising 20% of the dataset, reserving another 20% for validation. This results in 30561 instances for training and 7641 instances for validation. Model evaluation is performed on a separate test set of 202 images, which was annotated by three board-certified radiologists. We opted for a subset because retraining the models after simple errors or changes to the implementation was too costly. All the models are trained on ten epochs, employing the BCEWithLogitsLoss as the loss function to enable multi-label multi-class classification.

**Evaluation Metrics**
For evaluating the performance of our models, we use precision, recall, F1, AUC, and accuracy scores. Measured for every class and with the calculation of the total score, they provide a comprehensive assessment of the performance by capturing different aspects.

## V. Results

**General Performance:**

- **DenseNet**: This model shows the highest consistency across all three metrics, suggesting it is well-tuned for the task.
- **ResNet**: Close to DenseNet in AUC and F1 Score but slightly less in Accuracy, which might indicate it is slightly less consistent in correctly classifying all classes.
- **ViT**: It has lower scores in all metrics compared to DenseNet and ResNet, hinting at potential difficulties in extracting features from the chest X-ray images
- **ConvNeXt**: Shows a significant drop in AUC compared to the others, with moderate Accuracy and F1 Score, indicating potential issues with model generalisation.

**Class-wise Performance:**

Without specific class-wise performance data, it's hard to analyse class-wise strengths and weaknesses. If pathologies are consistently misclassified, it would suggest a need for further model training or data augmentation in those areas.

**Reasons for Lack of Performance:**

- **Data Issues**: Imbalanced datasets, presence of noise, or inadequate representation of certain pathologies can affect performance.
- **Model Complexity**: Overly complex models might overfit, while too simple models might not capture all necessary features.
- **Uncertainty and Noise**: Inherent uncertainty in medical imaging and subjective interpretations can lead to noisy labels, affecting training.
- **Hyperparameters**: Suboptimal hyperparameters can lead to poor model convergence.

**Insights About System Performance:**

- **Strengths**: DenseNet's architecture seems well-suited for this data, possibly due to efficient feature reuse.
- **Weaknesses**: ViT and ConvNeXt's lower performance might indicate that these architectures require further tuning or more data to capture the nuances of medical images.
- **Limitations**: The models may not generalise well to unseen data, especially if trained on a biased dataset.
- **Future Work**: Incorporating class weights, further hyperparameter tuning, and advanced techniques like ensemble learning could improve performance. Also, exploring uncertainty modelling and explainable AI could provide better interpretability and trust in the model's predictions.

**Visual Explanation**

Current literature reveals a wide variety of methods to generate visual explanations for deep learning models, broadly categorised into gradient-based, perturbation-based and other characteristic-based approaches. We selected GradCAM and GradCAM++ from gradient-based, ScoreCAM from perturbation-based, and EigenCAM from the third category. Our goal is to compare the explanations of these classes of methods, provide them to clinicians, and let them determine which method is superior based on their background knowledge. In this way, we can compare different classes of methods and choose the most appropriate one for this task. Our experiments suggest that GradCAM's explanations make more sense and agree with the disease characteristics based on the definitions we found.

## VI. Conclusion

The first notable contribution was the implementation of transfer learning with a vision transformer and a convolutional neural network, both pre-trained on ImageNet, for classifying chest X-ray images. Secondly, we integrated explainability methods with our trained models. The key insights gained include the learning how models need to be adapted in order to effectively apply explainability methods, identifying the method generating the most reliable explanation and determining the optimal method for each model. Our primary challenge was our limitation in computing resources, restricting us from training all models only for ten epochs, leading to subpar performance. This hindered meaningful comparisons between the pre-trained and the self-trained models. Given more time, unfreezing additional layers for the training, potentially enhances the performance.

## VII. References


1. Sarvamangala, D. R., and Raghavendra V. Kulkarni. "Convolutional neural networks in medical image understanding: a survey." Evolutionary intelligence 15.1 (2022): 1-22.

2. Ma, DongAo, et al. "Benchmarking and boosting transformers for medical image classification." MICCAI Workshop on Domain Adaptation and Representation Transfer. Cham: Springer Nature Switzerland, 2022.

3. Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317. (https://www.jair.org/index.php/jair/article/view/12228/26647)

4. Van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." Medical Image Analysis 79 (2022): 102470.

5. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

6. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

7. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

8. Liu, Zhuang, et al. "A convnet for the 2020s." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

9. Irvi, J 2019, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison