Medical image classification with artificial intelligence

Guanfei Wang, Ishaan Pathak, Janko Vrcek, Oscar Rosman, and Sonja Kanerot
(Dated: October 11, 2024)

Medical imaging is one of the areas where artificial intelligence (AI) can be of great use, when effectively and safely implemented. In order for this to happen, research is needed to ensure the reliability and find the potential risks of doing so. One concern is that many of today's AI models lack transparency, which is what this project aims to provide for three different models: a convolutional neural network (CNN), a Vision Transformer (ViT), and a Wasserstein autoencoder (WAE). To train the models, MedMNIST which is a labeled dataset of medical images, were used. More specifically, the subsets PneumoniaMNIST containing images of healthy and inflamed lungs, OrganAMNIST with pictures of eleven different organs and DermaMNIST with skin-conditions, were chosen. The CNN and WAE were trained from scratch while transfer learning were utilized for the ViT by using a pre-trained model. Results were presented in the form of attention-maps, showing what areas of the pictures were deemed most important by the models, combined with the accuracies the models reached for each dataset.

I. INTRODUCTION:

Medical imaging plays a crucial role in modern healthcare by aiding medical professionals in diagnosing patients and monitoring treatment progress. Despite its invaluable contributions, the surge in data volume poses significant challenges, as the interpretation of medical images often relies on manual analysis by professionals. This manual process is time-consuming, prone to human error, and contingent upon the availability of specialized expertise.

To tackle these challenges, there is a growing interest in harnessing artificial intelligence, specifically machine learning, to automate and improve medical image analysis.

In this project, our aim is to explore various machine learning models to assess their ability to accurately classify medical images. Moreover, we seek transparency in these models, aiming to emphasize the features that underpin their decisions. By doing so, we aim to not only enhance the efficiency of medical image analysis but also provide insights into the decision-making process of these models.

II. BACKGROUND

Radiology is a resource-intensive aspect of modern healthcare, providing professionals with essential insights for disease diagnosis and treatment guidance [1],[2]. This field encompasses several imaging modalities, including X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) [3].

An extensive retrospective study conducted in Denmark demonstrated the potential benefits of AI assistance in radiology. In this study, radiologists using AI for mammography screening of middle-aged women experienced a 33.5% reduction in workload and a 12% increase in cancer detection rates (70% without AI, 82% with AI) [2]. The AI system used in this study was a deep convolutional neural network developed in 2019, which, when compared with radiologists, showed better performance than 61% of the 101 radiologists involved. It was concluded that the AI had comparable detection accuracy to the average specialist in mammography [4]. Similarly, another study compared AI to dermatologists in detecting various skin conditions, focusing on different types of cancer. The AI network, trained on over 2,000 diseases using supervised learning, outperformed dermatologists in specificity and matched or exceeded their sensitivity. This study emphasized the importance of early skin cancer detection and the role of diagnostic accessibility in survival rates [5].

Exploring unsupervised learning, another study highlighted the limitations of human-supervised AI training in disease diagnosis, citing bias in the data as a limiting factor. The study showed that using unsupervised learning, the AI could identify biomarkers in OCT images (retinal scans) that correlated with vision loss and other eye conditions. This approach suggested that unsupervised learning could uncover previously unknown disease biomarkers, enabling earlier diagnosis and treatment [6]. In cancer treatment, integrated AI has been shown to assist not only in diagnosis but also in improving treatment strategies. A study demonstrated that AI models could determine which areas of the body required higher doses of chemotherapy to target aggressive tumors while suggesting lower doses for other areas to enhance recovery, a method that has been proposed since the 1980s but difficult to implement [7]. Improving transparency in AI models is crucial for building trust among physicians. A project aimed at this goal developed a model called MONET (Medical Concept Retriever), which linked images with medical literature to annotate them with appropriate concepts, thereby justifying diagnosis [8].

The MedMNIST dataset, published under a Creative Commons (CC) license for educational use, contains medical images from various modalities labeled for supervised learning. It includes preprocessed images available in resolutions ranging from 28x28 to 224x224. The dataset is divided into subsets, each focusing on different

parts of the human body using modalities such as CT scans, X-rays, and MRI, which are common in today's healthcare. Benchmarks for all subsets are computed using various convolutional networks, as detailed in an accompanying paper. The subsets vary in size, with the smallest containing 780 images and the largest 236,000 images. Some subsets involve binary classification tasks, while others are multi-class problems with up to twelve classes [9].

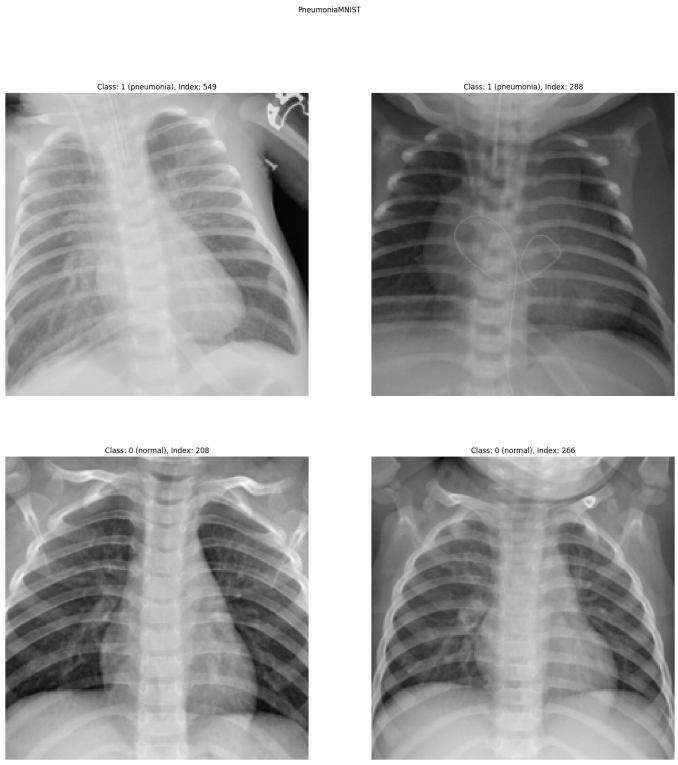


FIG. 1: Images taken from the Pneumonia MNIST dataset with their corresponding label [9].

III. LIMITATIONS

The project was constrained to exploring only three out of the twelve available subsets of the MedMNIST dataset due to limitations in computational resources and time constraints. To ensure a representative investigation of the dataset, the subsets were chosen to preserve the dataset's variability, including factors such as subset size, modality, and the number of classes. The selected subsets for analysis were Derma, OrganA, and Pneumonia MNIST. While these subsets offer valuable insights into the performance of the models, the exclusion of other subsets may limit the comprehensiveness of the findings.

Subset attributes			
Subset	**Data Modality**	**Classes**	**Samples**
Derma	Dermatoscope	7	10,015
OrganA	Abdominal CT	11	58,830
Pneumonia	Chest X-Ray	2	5,856

TABLE I: Containing the most relevant information of the subsets investigated in the project.

IV. METHOD

In this section the methodologies used to implement the machine learning models for this project are described in their respective subsection. Each approach details its unique implementation and how it is leveraged to give insight into its decision making process.

A. Auto-encoder

Auto-encoder is a type of neural network consisting of an encoder and a decoder. Encoder is a network that maps the input onto a much smaller latent space. That way the network is forced to learn the most important features of the data. The decoder is then used to decode the shrunken information back into the original data. By changing the number of latent neurons, we can see the minimal number of features needed to reconstruct the data. In this project, we specifically use the Wasserstein autoencoder (WAE), which works similarly to variational auto-encoder. Both minimize the reconstruction cost and regularizer penalizing the differences between probability distribution of the latent space and the distribution induced by the encoder. The difference is that WAE forces the encoded distribution of training samples to match the prior while ensuring the latent codes contain enough information to reconstruct the encoded data [10]. For training WAE, we used batches of 64 images. For the actual auto-encoder we used the encoder and decoder with 4 convolutional layers each. The encoder gets progressively smaller, down to the number of latent neurons, while the decoder has the mirrored structure. We train for 40 epochs with the learning rate 0.001 and the Adam optimizer. After reconstructing an image, we calculate the reconstruction loss using mean squared error and add it to the maximum mean discrepancy between the latent space distribution and the prior distribution. This way, we ensure that the encoded data in the latent space follows the normal distribution. After each training we increase the number of latent neurons by 1 starting from 1. That way we can obtain an elbow plot to see how loss varies with the number of latent neurons, and how many features are required to reconstruct images. We visualise the distribution of image classes in the latent space using PCA for the number of neurons we determine is sufficient for reliably reconstructing the images. Finally, we attempt to reconstruct images of the

two classes of the pneumonia dataset by iterating over classes and generating a random vector from the latent space for each class.

B. Convolutional Network

A CNN, as the name suggests, is a Neural Network that utilizes Convolutional layers as a core part of its architecture. These Convolutional layers act as feature extractors for the model and their deep dreams give us a rough idea of what said features represent on the original image. The convolution operation involves sliding a filter (also called a kernel) over the input data (such as an image) to produce a feature map. Each position of the filter over the input results in a single value obtained by element-wise multiplication followed by a sum. The Convolution layer is typically followed by a 2D Batch Normalization layer and an activation layer like ReLU (Rectified Linear Unit). The stack of convolution layers is followed by a fully connected layer. The output layer consists of as many neurons as there are classes in the data. This layer gives a probability of the image belonging to each class.

The CNN was trained on batches of 64 images. We trained for 50 epochs, with the Adam Optimizer at a learning rate of 0.001, and momentum at 0.9, and using the Cross Entropy loss function to backpropagate through the network and re-adjust the weights. The model training was a relatively simple task. Next, we added hooks after every convolution layer to see what their output is in order to visualize the CNN's deep dreams. Finally, we passed some of the images from the test dataset through the network and assessed the output at the hooks to get the deep dreams of those images.

C. Vision Transformer

Transformers are a recent development in the area of machine learning and started becoming dominant in the area since 2017 [11]. Since 2020 they have been adapted for computer vision. Although vision transformers can be difficult to train from scratch they have proven that they can outperform state of the art convolutional networks like ResNet by using transfer learning [12]. A vision transformer processes an image as a sequence of patches each represented by a high-dimensional vector including its location in the image. The ViT then uses a self attention mechanism to attend to what it sees as the most relevant patches. In the attention mechanism all patches are compared to the others allowing the model to learn complex dependencies across the entire image and capture long-range relationships unlike a convolutional network which works with local spatial information. By dynamically comparing all patches the model is able to determine the importance of each patch in relation to the others and can focus on the most informative areas of the image.

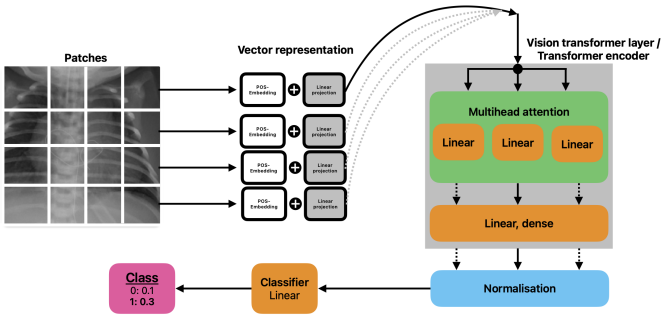


FIG. 2: Simplified illustration on how an image is divided into patches and go through the ViT model. The section with multihead attention is where patches are compared with each other and previous knowledge setting the ViT apart from other types of models.

The initial approach of this project involved designing and training a vision transformer tailored specifically for the MedMNIST datasets, inspired by a functional ViT implementation on Google Colab [13]. This attempt was unsuccessful however as the model failed to learn from the datasets and constantly assigned all samples to the same class. Various methods were tried to resolve this, including modifications to the optimisation algorithm, data augmentation techniques and troubleshooting the model architecture.

After the first approach failed the project resorted to using transfer learning instead as mentioned in literature [12]. A pre-trained model from Hugging Face [14] under an Apache 2.0 license was leveraged instead. This model uses the same model architecture as in Figure 2 but features twelve transformer layers instead of only one shown in the figure. Additionally the classifier in the end was modified for each dataset used in the project to accommodate the number of classes in the dataset. The model was also constrained to process images of the resolution 224x224 which were available in the MedMNIST datasets.

Prior to subjecting images to the ViT they underwent preprocessing by a custom composer. This included a conversion to RGB if the images was in grayscale, resizing to 224x224 resolution if the image was smaller, converting the image to tensors and finally normalising the image data.

One vision transformer was fine-tuned for each dataset. To achieve this a batch size of 16 was used for ten epochs which proved enough to get comparable performance metrics to the benchmarks to the MedMNIST dataset. For weight optimisation the Adam optimiser was used combined with Cross entropy loss. Cross entropy is a well suited loss function for multi-class classification problems and the Adam optimiser is a computationally efficient version of stochastic gradient descent. Model selection

during training was based on general accuracy on the validation set. All MedMNIST dataset are partitioned into training, validation and test set which was adhered to during training without modifications.

To further evaluate the ViTs performance on the test set additional performance metrics where calculated such as precision, recall, F1, class accuracy and confusion matrices. Finally the attention weights connecting each patch to the output token were visualised by displaying them as a heat map overlaid on the original image. This final step highlighted the ViT's focus on the original image showing which features where deemed as most important in order to make the classification.

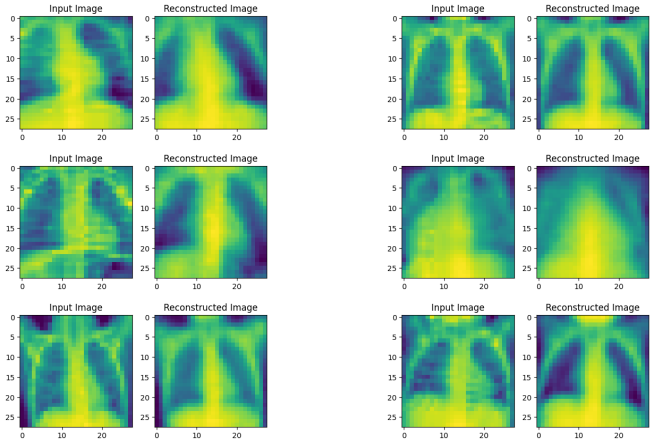


FIG. 3: The input image and the image reconstructed using the WAE. We can see that the primary features are conserved, with dark lungs remaining dark and infected lungs being bright. The WAE fails to reconstruct the ribs in all cases, but that might not be as important for classification. We could improve performance by longer training time. Peak signal to noise ratio for the reconstructed images is 23.0811. We used 8 latent neurons and 28x28 images.

V. RESULTS

In this section, we present the results of our investigation into the performance of various machine learning models for the classification of medical images. Our primary objective was to assess the ability of these models to accurately classify medical images while prioritizing transparency in their decision-making process. By highlighting the features that influence the models' decisions, we aimed to enhance the efficiency of medical image analysis and provide valuable insights into their decision-making mechanisms.

To evaluate the performance of the models, we compared their classification accuracy with the best benchmark reported in the MedMNIST publication, as shown in Table II. This comparison allows us to assess the effectiveness of our models relative to the state-of-the-art performance on the datasets.

Model accuracy			
Dataset	**CNN**	**ViT**	**Benchmark**
Derma	0.730	**0.801**	0.768
OrganA	0.871	**0.938**	**0.951**
Pneumonia	**0.937**	0.877	**0.946**

TABLE II: The obtained accuracies for the CNN and ViT models on the investigated datasets. On the right is the best benchmark on the datasets taken from the MedMNIST publication.[9]

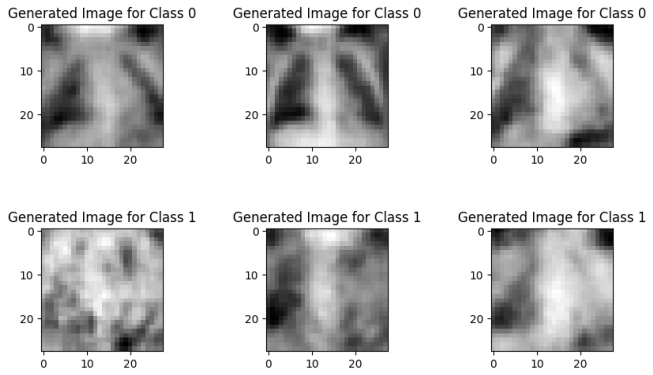
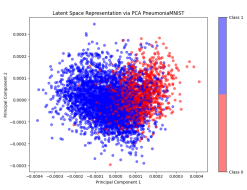


FIG. 4: Images generated using WAE. Class 0 images corresponding to healthy lungs look mostly acceptable. The shape of the lung is correctly generated and the lungs are correctly dark. Heart and diaphragm are correctly visualised. Class 1 images corresponding to infected lungs are not very well generated. The lungs are correctly coloured white, but the shape of the lungs is essentially non-existent. This is likely due to the lack of contrast between infected lungs and the rest of the chest. This might be remedied by longer training. We used 8 latent neurons and 28x28 images.

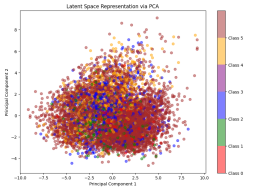
A. Auto-encoder

The reconstructed figure of pneumonia dataset is shown in Figure 3. The figures generated by sampling from the latent space are shown in Figure 4.

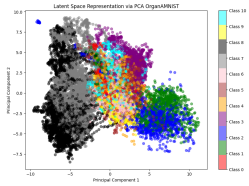
Figure 5a shows the PCA for pneumonia, Figure 5b for derma and Figure 5c for organA. Finally the elbow plot for pneumonia is shown in Figure 6a, for derma in Figure 6b and for organA in Figure 6c.



(a) PCA for the pneumonia dataset. The two classes are well segregated due to simplicity of images and only two classes. 8 latent neurons were used.

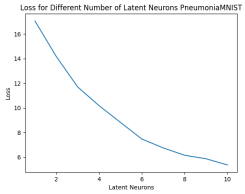


(b) PCA for the derma dataset. The six classes are not well segregated due to similarities between images and a large number of classes. 8 latent neurons were used.

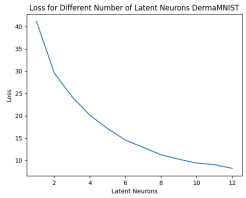


(c) PCA for the organA dataset. The eleven classes are well segregated due to the differences between images. 8 latent neurons were used.

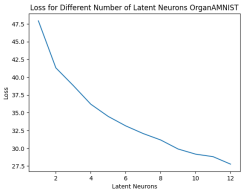
FIG. 5: PCA for the 3 datasets.



(a) Elbow plot for pneumonia dataset. 8 latent neurons seem to be sufficient to correctly reconstruct the image based on the rate at which the loss decreases.



(b) Elbow plot for derma dataset. 10 latent neurons seem to be sufficient to correctly reconstruct the image based on the rate at which the loss decreases.



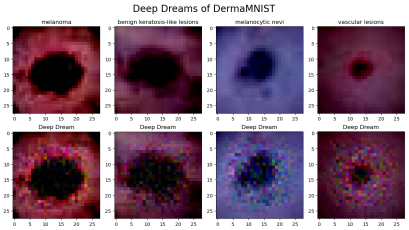
(c) Elbow plot for organA dataset. 10 latent neurons seem to be sufficient to correctly reconstruct the image based on the rate at which the loss decreases.

FIG. 6: PCA for the 3 datasets.

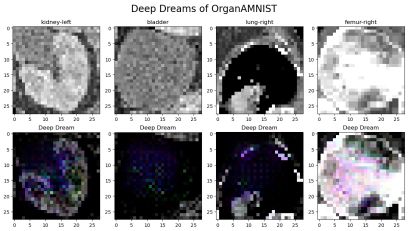
B. Convolutional Neural Network

The images shown in Fig 6, show the deep dreams of the CNN on the OrganAMNIST dataset. It is clear from looking at Fig 6a and 6b that the model focuses on the shape of the organ among other factors to make its classification.

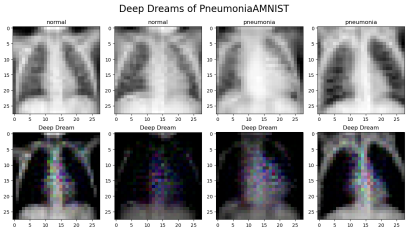
We can see a similar pattern emerge from the deep dreams of the CNN on the PneumoniaMNIST dataset as well (Fig 7a, and 7b).



(a) Deep Dream of DermaMNIST.



(b) Deep Dream of OrganAMNIST.



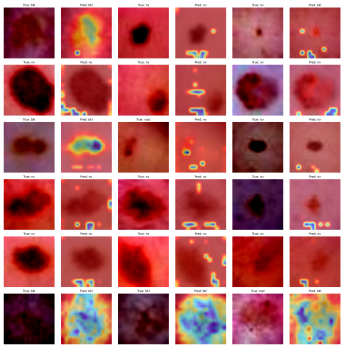
(c) Deep Dream of PneumoniaMNIST.

FIG. 7: Deep Dreams generated by the CNN, for the three datasets.

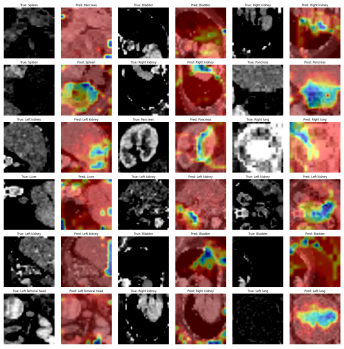
C. Vision Transformer

The ViT performance accuracy for the different datasets are presented in table II, showing that the model effectively learned to distinguish between the various categories in the datasets. In further investigation of the biological knowledge gained by the model, the attention maps provide insight into which regions of the images are most influential in the decision-making process. Figure 8 displays some biological features that the model prioritizes for the three datasets. The highlighted areas in figure 8a show intense attention at small areas in the pictures for the DermaMNIST dataset. For OrganAMNIST in figure 8b on the other hand, larger areas of attention

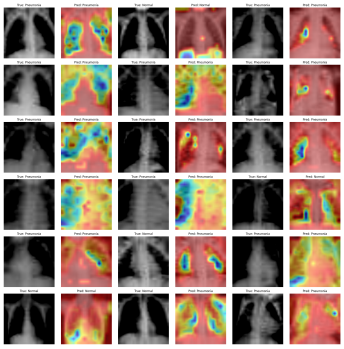
were used to distinguish between the different classes. In figure 8c showing attention maps for PneumoniaMNIST, focus is positioned in the lung area.



(a) Attention map of the ViT on DermaMNIST. Each image is paired with its respective original image highlighting the focus of the model in the classification. The attention maps did not effectively highlight the lesions, as the dataset is severely biased.



(b) Attention map of the ViT on OrganAMNIST. Each image is paired with its respective original image highlighting the focus of the model in the classification.



(c) Attention map of the ViT on PneumoniaMNIST. Each image is paired with its respective original image highlighting the focus of the model in the classification.

FIG. 8: In all attention maps one can see that the ViT has found small areas of the images it deem relevant for classification. An intuitive example are the attention maps on the pneumonia images where one can clearly see on the bottom images that the lungs are highlighted.

VI. DISCUSSION

The findings of this research project were partly aligned with the initial expectations. Regions that were deemed important for classification were similar between the models for many tasks, with some exceptions, and the accuracies reached were reasonable for a project of this extent.

PneumoniaMNIST was the dataset where the models' attention maps resembled each other the most. When comparing figure 7c with figure 8c, similar parts are highlighted in the case where the models detects pneumonia. The ViT formed a more detailed shape of the lungs which could be explained by the higher resolution used in training, compared to the CNN. When Aron Bryngelsson, a licensed physician, was asked what he would look for when classifying pneumonia, he answered: *"I would look for brighter areas in the lungs that are clearly demarcated"*. This contradicts the attention map that focuses on larger regions rather than small. An explanation to this could be that the images that were sampled to generate the attention map were not representative for pneumonia. Bryngelsson claims that he himself has difficulties detecting the infections in these particular images, and that it is normally easier to see.

When detecting healthy lungs, very much of the ViT's attention is located at the brighter areas near the bottom, which is the diaphragm. According to Bryngelsson, the appearance of the diaphragm does not give any useful information about the health of the lungs. One explaination to this might be that the model learned to focus on bright, white parts of the lungs while looking for infections, and investigates the diaphragm as it is also bright. It can also be argued that the question itself is difficult to answer, as the task is to find in what areas we see that an infection, that can be almost anywhere, is absent.

When it comes to the WAE, the reconstructed image of the healthy lungs, to the left in figure 4, were of a darker tone, aligning with Bryngelsson's description of the condition. In contrast, the reconstructed image of infected lungs, on the right side of Figure 4, is brighter and slightly blurry.

"One concern is the lack of distinction of the lungs. As a typical pair of pneumonia-infected lungs would have a small, much brighter area surrounded by otherwise normal lung tissue, it is not a very representative image of pneumonia."

A reason why there is no distinct brighter patch could be that the images were generated by a combination of many pictures of infected lungs, which all probably had infections on different regions. This could also explain why the lungs were not clearly demarcated, combined with the fact that low resolution images were used for training the WAE. Another reason could be the lack of contrast between infected lungs and the rest of the chest, resulting in a blurry image. Further training of the model could probably also increase the contrast and accuracy of the picture.

Some larger differences between the models' attention maps appeared in the OrganAMNIST dataset. While

the CNN generally more accurately pointed out the organs, the ViT often included other parts surrounding the organ, which according to Bryngelsson, can give just as much information if not more, in some cases. On the other hand, the attention maps of the ViT were often divided into small dense clusters with seemed almost randomly located. One alarming example is the attention map for the right lung that can be seen in one of the images. Most of the attention is in the upper right corner which is actually outside of the patient's body. Even if this seems counter intuitive at first, the model is rather indifferent to the meaning of the different spots on the picture. As the task for this dataset is to recognise what organ is displayed, it only looks for patterns and features that signifies the pictures of those organs. Hence, the outside of the organ or even the body, can with equal probability be of importance as the organ itself. Especially if all pictures of that organ contain that specific region. This underscores the differences in the ways humans and computers perceive the same data, which is important to consider before implementing AI in the healthcare system.

Lastly, out of the three datasets, DermaMNIST had the worst accuracy. Even if it only had seven classes, compared to OrganAMNIST which had eleven classes, the accuracy for the dataset with skin conditions were only 80.1% at maximum. When Bryngelsson was asked why he thought this was the case, he answered: *"It is quite difficult to classify a birthmark as dangerous or harmless, just from looking at it. Some of them are completely normal and thereby easy to distinguish, but many of the harmless ones can look similar to diseases which results in a large gray area, where a tissue sample is required. I would say it is easier for an experienced x-ray doctor to recognize inflamed lungs than it is for an experienced dermatologist to classify an odd birthmark"*. The previous benchmarks provided by MedMNIST strengthens this theory since the highest accuracy reached for this dataset were 76.8%. This is much lower than both PneumoniaMNIST and OrganAMNIST which both had previous accuracies above 90%.

The PCA of the pneumonia dataset shows very good segregation between the two classes. The PCA for the derma dataset shows that class 5 is grouped together quite well, but the rest of the plot is not well segregated. This is to be expected, as derma was the dataset with one of the lowest classification accuracies, indicating that the images are hard to distinguish. PCA for organA on the other hand, shows very good segregation between almost all of 11 classes. Some classes are still not well segregated which might be remedied by longer training. The three elbow plots reach the conclusion that 8 latent neurons are sufficient to reconstruct the pneumonia dataset and 10 are sufficient for derma and organA. The reconstruction accuracy of WAE for pneumonia is 23.0811, meaning that the reconstruction is good, but far from perfect.

VII. CONCLUSION

In conclusion, our study provides comprehensive insights into the performance of convolutional networks, Vision Transformers (ViTs), and auto-encoders in medical image classification tasks. Both convolutional networks and ViTs demonstrated strong performance, comparable to benchmarks established in the MedMNIST publication. Particularly noteworthy was the ViT's ability to surpass the benchmark accuracy on the Derma MNIST dataset, showcasing its potential for enhancing medical image analysis.

The visualization of ViT's attention mechanism highlighted important regions within original images, indicating its efficacy in feature extraction. While both ViT and convolutional networks successfully identified key features in decision-making, the ViT provided clearer delineation of significant image areas, enhancing interpretability.

Furthermore, the auto-encoder demonstrated notable success in reconstructing recognizable images across various datasets. Visualizations of the Pneumonia and Organ A datasets showcased clear separation of clusters, indicating robust feature representation by the auto-encoder.

Moving forward, future research should explore the integration of additional metrics and domain-specific expertise to enhance model evaluation. Furthermore, investigating the robustness of models across diverse datasets and clinical scenarios will be essential for ensuring their generalizability in real-world settings. Additionally, continued advancements in model interpretability techniques will facilitate deeper insights into model decision-making processes, fostering trust and adoption in clinical practice.

In a future work, all models could be trained on the images of highest resolution. We could also attempt to incorporate different symptoms into diagnosis, like recordings of breath and coughing alongside chest xrays for pneumonia.

Overall, our study underscores the potential of deep learning models in revolutionizing medical image analysis while highlighting the importance of comprehensive evaluation and interpretation in advancing the field.

VIII. CONTRIBUTIONS

The work was divided evenly between the group members, with all the members participating in coding and analysis of the results, as well as in writing the report. Special thanks to the project supervisor Henrik Moberg for his constructive criticism of our work, as well as to the licensed physician Aron Bryngelsson for his comments on the validity of our results.

[1] E. A. Krupinski, Current perspectives in medical image perception., Attention, perception psychophysics **72,5** (2010).

[2] A. D. Lauritzen, M. Lillholm, E. Lynge, M. Nielsen, N. Karssemeijer, I. Vejborg, and L. Moy, Early indicators of the impact of using ai in mammography screening for breast cancer, Radiology **311**, e232479 (2024), pMID: 38832880, https://doi.org/10.1148/radiol.232479.

[3] L. Pinto-Coelho, How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications, Bioengineering (Basel) (2023).

[4] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. G. Wallis, I. Andersson, S. Zackrisson, R. M. Mann, and I. Sechopoulos, Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists, JNCI: Journal of the National Cancer Institute **111**, 916 (2019), https://academic.oup.com/jnci/article-pdf/111/9/916/31963401/djy222.pdf.

[5] K. B. N. R. e. a. Esteva, A., Dermatologist-level classification of skin cancer with deep neural networks., Nature (2017).

[6] S. P. D. R. e. a. Waldstein, S.M., Unbiased identification of novel subclinical imaging biomarkers using unsupervised deep learning., Springer Nature **10**.

[7] Y. e. a. Pang, Yaru et al.Pang, Medical imaging biomarker discovery and integration towards ai-based personalized radiotherapy., Frontiers in oncology **11** (2022).

[8] G. S. D. A. e. a. Kim, C., Transparent medical image ai via an image–text foundation model grounded in medical literature, Nature medicine **30** (2024).

[9] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, Scientific Data **10**, 41 (2023).

[10] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, Wasserstein auto-encoders (2019), arXiv:1711.01558 [stat.ML].

[11] N. P. J. U. L. J. A. N. G. K. I. P. Ashish Vaswani, Noam Shazeer, Attention is all you need, Neural Information Processing Systems (2017).

[12] A. D. et al, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR (2021).

[13] DeepFindr, Vision transformers from scratch, author is an anonymous youtube channel from Germany with a linked GitHub that specialise in machine learning and reference his work to appropriate literature. Link to colab notebook: https://colab.research.google.com/drive/1P9TPRWsDdqJC6IvOxjG2 $sharingscrollTo = wdEfT7i40Eka.B.Wu, C.Xu, X.Dai, A.Wan,$ $Token-basedimagerepresentationandprocessingforcomputervision$ $https://huggingface.co/google/vit-base-patch16-224-in21k,\ arXiv: 2006.03677[cs.CV]$.

Appendix A: Code