Vision Model for Temporal Disease Progression of Chest X-ray dataset

Vaibhav Mavi VM2241@NYU.EDU
Wonkwon Lee WL2733@NYU.EDU
So Hyun Chi SC9120@NYU.EDU
Deborah Jacob DSJ9195@NYU.EDU
New York University, New York, NY, USA

Abstract

Early detection is crucial for successful treatment and prevention of severe complications. Understanding how a disease progresses over time and how it is impacted by severity and treatment can provide important insights into the efficacy of new medicines and how they can best benefit patients. A model-informed approach is an effective solution for integrating diverse sources of information and building an expert understanding of disease and treatment impacts. Disease progression modeling (DPM) uses mathematical functions and scientific principles to describe the quantitative progression of a disease over time, providing valuable insights for the development and use of medicines. Disease Progression models have the potential to improve patient outcomes, reduce healthcare costs, and accelerate the development of new treatments. One such task is to predict the three states of disease progression (improving, stable, or worsening) given the current and past multi-image frontal chest X-ray images. This work focuses on fine-tuning and evaluating the pre-trained Torch X-ray Vision model Cohen et al. (2021) for the temporal image classification task. [1]

1. Introduction

Chest X-rays are widely used in medical imaging and have been the focus of many deep learning studies. However, comparing previous work and generalizing across different datasets can be challenging due to differences in dataset organization, processing, and training procedures. Even when data and code are available, small differences can significantly impact results, making it difficult to establish meaningful baselines for researchers.

The Vision Transformer and DenseNet models are both effective tools for analyzing chest X-ray datasets. The open-source Torch X-Ray Vision library provides a convenient and standardized interface for working with these datasets. The Vision Transformer model has the potential to generalize to other related problems with minimal supervision, but its large feature size can pose challenges for downstream tasks. The DenseNet model is known for its efficient use of parameters and ability to reduce vanishing gradients during

1. The code for this study can be found at: https://github.com/vaibhavg152/TimeSeriesMedicalImageClassification

training. In this thesis, three pre-trained DenseNet models are incorporated to improve the performance of the model in detecting various medical conditions. DenseNet has achieved state-of-the-art performance in various computer vision tasks and has been applied to medical image analysis as a backbone in other deep learning models.

Time series analysis includes tasks which involve some prediction on a set of data (images and/or reports) that have a defined temporal correlation. For instance, the MS-CXR-T dataset (Bannur et al. (2023)) consists of pair of images of Chest X-rays of the same patient at two different time instances and the task is to predict whether the condition of the disease is improving, stable or worsening. It is important to use multiple external datasets to rigorously evaluate the robustness of models. In this study, we aim to fine-tune and evaluate the pre-trained Torch X-ray Vision model Cohen et al. (2021) for the temporal image classification task.

2. Resources

Data The MS-CXR-T dataset (Bannur et al. (2023)) is a temporal benchmark dataset designed to evaluate the performance of biomedical vision-language processing (VLP) models in the field of radiology. The dataset focuses on two distinct temporal tasks: temporal image classification and temporal sentence similarity, both of which are crucial for the accurate and efficient analysis of medical imaging data. By providing a rich and diverse set of chest X-ray images and associated textual reports, the MS-CXR-T dataset enables researchers to develop and assess VLP models that can accurately classify medical images and measure semantic similarity between radiology reports.

1. Temporal image classification dataset contains 1326 ground-truth labels of multi-image frontal chest X-ray images along with their radiology reports, with three states of the disease progressing labels (Improving, Stable, Worsening) for of the five distinct diseases (Consolidation, Edema, Pleural effusion, Pneumonia, Pneumothorax). The benchmark is built from the publicly available Chest ImagGenome (Wu et al. (2021)) gold and silver standard datasets.

2. Temporal sentence similarity dataset contains 361 pairs of paraphrase or contradiction sentences of disease progression of five findings. The benchmark is built from a set of sentences from the MIMIC-CXR dataset, using the Stanza constituency parser (Zhang et al. (2020)) to extract each sentence.

Models

1. **Vision Transformer** (Face (2021))
   The Vision Transformer model, developed as part of the Hugging Face Torch X-Ray Vision library, presents an effective tool for analyzing chest X-ray datasets. This open-source library provides a convenient and standardized interface and preprocessing pipeline for chest X-ray datasets. Given the model's prior exposure to chest X-ray images for pneumonia disease, it has the potential to generalize to other related problems with minimal supervision. Nonetheless, the model's features are often generated in a large size, which can pose challenges for efficient utilization in downstream tasks.

2. **DenseNet** (Huang et al. (2017))
   In this thesis, we utilize the DenseNet model architecture, which is widely known for its efficient use of parameters and its ability to reduce vanishing gradients during training. To further enhance the performance of our model, we incorporate three pre-trained DenseNet models that were trained on different datasets. These models include DenseNet121-res224-rsna, which was trained on the RSNA Pneumonia Challenge dataset, DenseNet121-res224-mimic-ch, which was trained on the MIMIC-CXR dataset, and DenseNet121-res224-chex, which was trained on the CheXpert dataset. To access these pre-trained models, we load them from the torch-xray-vision repository. By leveraging the knowledge learned from these large datasets, we can reduce the time and computational resources required to train our own model from scratch and potentially improve its performance. Incorporating these pre-trained models is expected to enhance the accuracy of our model in detecting various medical conditions.

3. Related Works

1. Densely Connected Convolutional Networks (Huang et al. (2017)).
   Proposed the DenseNet architecture, which uses dense connectivity patterns between layers to promote feature reuse and reduce the number of parameters. DenseNet has achieved state-of-the-art performance on various computer vision tasks, including image classification, object detection, and semantic segmentation. The architecture has also been applied to medical image analysis and used as a backbone in other deep-learning models.

2. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning (Rajpurkar et al. (2017))
   This paper proposes a deep-learning model that can detect pneumonia from chest X-ray images with a performance that is comparable to that of expert radiologists.

3. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Radiologists (Wang et al. (2018))
   This paper evaluates the performance of the CheXNet algorithm on a large-scale dataset of chest X-ray images and compares it to the performance of radiologists.

4. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation (Rajpurkar et al. (2020))
   This paper evaluates the performance of several deep learning models on the task of classifying chest X-ray images into abnormal or normal categories, using the MS-CXR-T dataset. The authors show that their models achieve high accuracy and outperform human radiologists in certain scenarios.

5. Temporal Convolutional Neural Network for the Classification of Pneumonia from Chest X-Rays: A Feasibility Study (Islam et al. (2020))
   This paper proposes a temporal convolutional neural network (TCN) for the classification of pneumonia from chest X-ray images, using the MS-CXR-T dataset. The

authors show that their TCN model outperforms other deep learning models and achieves state-of-the-art performance on the dataset.

4. Methodology

We build a model to perform the time series disease progression task by fine-tuning large pre-trained models. Our approach is thus divided into two steps, feature extraction and fine-tuning. We describe these two steps in this section.

4.1. Feature extraction

Since the dataset size is very small, it forbids any meaningful training of a model from scratch. Therefore, we rely on large pre-trained models' abilities to extract meaningful features from a large variety of images, which can then be used to perform several different downstream tasks. For this study, we extract features from the pre-trained DenseNet121 models. As described in section 3, DenseNet121 (Huang et al. (2017)) is a deep convolutional neural network with 121 layers, where all the convolutional layers are densely connected to each other. This enables the model to extract meaningful features from an image across different scales which is usually well suited to the requirements in the healthcare domain.

We test with three different DenseNet121 models trained on different datasets:

1. DenseNet121-RSNA: Trained on the RSNA Pneumonia Challenge.
2. DenseNet121-CheX: Trained on the CheXpert data.
3. DenseNet121-MIMIC-CH: Trained on the MIMIC-CXR dataset.

The DenseNet121 model takes an image as an input and generates a 1D feature vector of size 1024. Since this size is easy to work with, we extract these features for all the images in the dataset as part of a pre-processing step.

We also try using the Vision Transformer (ViT-base) model that is pre-trained by Google on the ImageNet-21k dataset and fine-tuned on the Chest-Xray-pneumonia dataset. However, the size of the features extracted is too large which makes the training process much more complex and time consuming.

4.2. Fine-tuning

For fine-tuning, we build a small neural network with a few dense layers to perform the time series classification using the extracted features. The data has a very small number of samples and every sample has labels for progression of one or among the 5 diseases in the dataset. However, most samples have labels corresponding to only one or two diseases. We highlight this challenge in the figure 2 This leads to a sparse data for training a single model. For dealing with this, we test three different architectures for the classification model:

- **Model1:** Train a single classifier by replacing the missing labels with the *'stable'* label.
- **Model2:** Train five separate classifiers for five diseases and only look at one disease at a time for a particular sample.

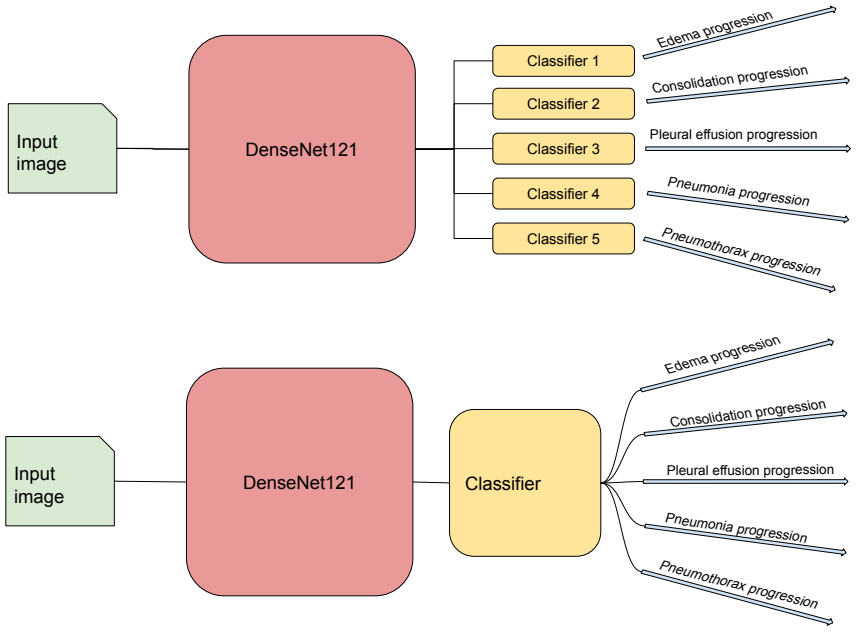

Figure 1: The two architectures proposed in this work: five separate classifiers, one for each disease (top), a single classifier predicting for every disease (bottom)

- **Model3:** Train a single model with five classification branches and use masking of the loss function to avoid back propogating gradients for those diseases in the sample which do not have a label associated with it.

5. Experiments

We perform three experiments for this study:

1. Train and evaluate the three different models explained in the previous section.
2. Train and evaluate a logistic regression baseline using the extracted features for having a fair comparison with the trained classifier.
3. Train and evaluate a logistic regression baseline on the input images directly for highlighting the usefulness of the pre-trained feature extractor used.
4. Flip the order of the two images for each sample in the test set and verify if the model is able to flip its prediction as well (from *'worsening'* to *'improving'* and vice versa). This will help us verify that the model is able to understand the temporal relation between the images and not relying on other factors (such as class imbalance) for getting the correct results.

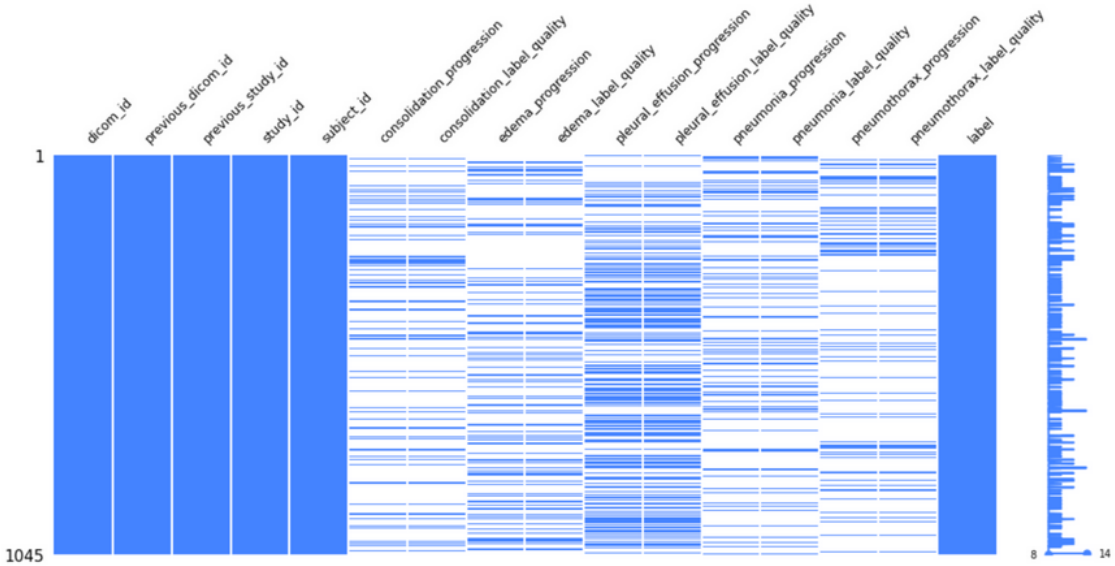

Figure 2: The figure highlights the sparse labels in the data. The white horizontal lines correspond to missing values in the data while the blue ones correspond to the existing values.

6. Results

1. **Model 1** When we try to train a simple classifier for model1, the output of the model is *'stable'* for every input. This is probably because replacing every missing value with *'stable'* makes it significantly outnumber the other labels leading to a large class imbalance. In this case, a simple model learns to predict the majority class for every input.

2. **Model 2** We show the results for model2 in Table 1.

3. **Model 3** On training the combined model with masking of the loss function, we get an accuracy of **42.91**%.

4. **Logistic regression** We show the results for model2 in Table 1. Since the DenseNet-CheX feature extractor performs the best, we show the results for logistic regression with that only.

Performance on flipped images Table 2, 3, 4 show the performances of the three models on the flipped images. Ideally, the values colored in purple should be as high as possible. Therefore, we observe that it is likely that the model3 has a better understanding of the temporal semantics of these samples than the other models.

Feature extractor	Edema	Consolidation	Pleural Effusion	Pneumothorax	Pneumonia	Average	Weighted average
DenseNet-RSNA	45.28%	45.0%	60.98%	50.0%	61.70%	53.70%	52.4%
DenseNet-MIMIC_CH	45.28%	42.5%	57.31%	**54.76%**	**63.83%**	53.40%	52.7%
DenseNet-CheX	**47.17%**	50.0%	**62.20%**	**54.76%**	**63.83%**	**56.40%**	**55.59%**
Logistic Regression	44.08%	**51.00%**	39.02%	37.21%	56.13%	45.49%	44.25%

Table 1: Results for training five separate classifier models using three different feature extractors. DenseNet121 trained on Chest X-Ray (Pneumonia) data consistently performs the best.

		Original Prediction		
		Stable	**Improving**	**Worsening**
Prediction on flipped input	**Stable**	54	45	0
	Improving	21	92	**7**
	Worsening	4	32	3

Table 2: Prediction of the Logistic regression model before and after switching the input images

		Original Prediction		
		Stable	**Improving**	**Worsening**
Prediction on flipped input	**Stable**	96	33	1
	Improving	39	58	**22**
	Worsening	1	13	1

Table 3: Prediction of Model2 before and after switching the input images

		Original Prediction		
		Stable	**Improving**	**Worsening**
Prediction on flipped input	**Stable**	48	18	**10**
	Improving	12	**41**	29
	Worsening	9	21	76

Table 4: Prediction of Model3 before and after switching the input images

7. Discussion and Limitations

One of the challenges of this project involved with the dataset. Acquiring the dataset from PhysioNet for machine learning research in the medical domain involves a series of verification processes to ensure ethical and appropriate data usage. The size of the MIMIC-IV database is enormous, originally at 4 terabytes and compressed to 500 gigabytes, posing challenges in terms of computational resources and processing time. The dataset features a complex structure, comprising five distinct types of diseases with labels distributed across three progression classes. Additionally, the dataset contains a significant number of missing values, and most samples have labels for only one or two of the five diseases. The scarcity of labels for each disease, with roughly 200 instances per disease, may impede the effectiveness of machine learning models in learning and generalizing from the data.

When employing a Vision Transformer model for feature extraction, researchers must allocate substantial computational resources and manage memory usage carefully, as the process requires considerable time and memory. The output generated by the model is extensive and cannot be easily dumped or stored due to memory constraints. Furthermore, the current model design only considers the anterior-posterior and posteroanterior views of the images, disregarding the lateral views that may contain valuable features pertinent to the findings. This exclusion of potentially informative features may result in suboptimal performance and hinder the model's ability to accurately diagnose and predict disease outcomes.

In light of these challenges and limitations, future research should focus on addressing the complexities and imbalances within the dataset, as well as incorporating lateral views into the model to enhance its predictive capabilities and improve overall performance. Researchers must also be mindful of the computational resources and memory management required when working with such a large and complex dataset.

8. Future Works

We suggest several key ideas of improvement for future research directions, particularly within the context of time-series medical image classification. First, implementing better imputation methods or exploring semi-supervised learning techniques can help overcome the limitations posed by the limited availability of medical datasets. The MS-CXR-T dataset contains 1326 samples, with a significant number of missing values. However, it is essential to exercise caution when employing automatic labeling, as it may introduce biases resulting from different schools of thought, potential disagreements between radiologists, and inherent subjectivity.

Second, the performance of the Vision Transformer model can be enhanced by conducting proper hyperparameter tuning. Although this process necessitates a significant investment in computational resources and time, optimizing hyperparameters can lead to more accurate and robust models, which, in turn, can improve the diagnosis and prediction of disease outcomes.

Lastly, the current model's limitation of considering only the anterior-posterior and posteroanterior views, while ignoring lateral views, presents an opportunity for future work. Incorporating lateral views can substantially improve the model's performance, as these views contain valuable features that are currently overlooked in the frontal chest X-ray

image dataset. By integrating all relevant multi-image views, researchers can develop a more comprehensive understanding of the chest X-ray images, ultimately enhancing the model's ability to accurately diagnose and predict disease progression.

By addressing these areas of improvement, future research can continue to advance the field of machine learning in the medical domain, ultimately leading to more effective and efficient diagnostic and treatment methods for patients.

9. Individual Contribution

Dataset Acquisition and Pre-processing: *Wonkwon, Vaibhav, So Hyun*
Fine-tuning: *Wonkwon, Deborah, Vaibhav*
Training and Evaluation: *Vaibhav, So Hyun, Wonkwon*
Results and Figures: *Deborah, So Hyun, Vaibhav*
Presentation: *All*
Project Report: *All*

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models, 2021.

Hugging Face. Vision transformer for analyzing chest x-ray datasets. https://huggingface.co/torchxrayvision/vit_chestxray14, 2021. Accessed on May 14, 2023.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017.

Md Mohaimenul Islam, Md Tariqur Rahman, Suvro Kanti Saha, and Md Rakibul Karim. Temporal convolutional neural network for the classification of pneumonia from chest x-rays: A feasibility study. *Healthcare*, 8(3):215, 2020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hetal Mehta, Tony Duan, Dong Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Dong Ding, Tony Duan, Hetal Mehta, Brandon Yang, Kaylie Zhu, Dylan Laird, Robyn L Ball, et al. Chest radiograph interpretation

with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, page 202172, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning, 2021.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. Biomedical and clinical english model packages in the stanza python nlp library, 2020.