**Data Analysis and Presentation Project: Enhancing X-Ray Imaging Analysis with Active Learning**

Nimrod Solomon ID: 206574733 E-mail: [s.nimrod@campus.technion.ac.il](mailto:s.nimrod@campus.technion.ac.il)

Matan Shiloni ID: 208634469 E-mail: [matanshiloni@campus.technion.ac.il](mailto:matanshiloni@campus.technion.ac.il)

Roni Fridman ID: 205517097 E-mail: [soldier109@campus.technion.ac.il](mailto:soldier109@campus.technion.ac.il)

**Abstract**

X-Rays, one of the most used imaging modalities, play a crucial role in diagnosing various medical conditions. However, the effectiveness of machine learning models in this area is often limited by the availability of labeled datasets, which are expensive and time-consuming to produce, as they require expert annotation. Our project focuses on leveraging Active Learning (AL) for the sake of developing a system that prioritizes the annotation of the most informative and challenging X-Ray images, thereby reducing the overall labeling effort required while improving the model's performance.

**Introduction**

**Pneumonia** is an infection of the respiratory system that primarily impacts the lungs. It arises when harmful microorganisms infiltrate the lung tissue, leading to the destruction of the pulmonary alveoli, which are responsible for oxygen absorption. As a result, these areas become filled with inflammatory fluid, impairing their functionality. Pneumonia can be triggered by a variety of pathogens, including bacteria, viruses, and parasites. While the common symptoms are concentrated around coughing, fever, chills, difficulty breathing, and chest discomfort, in more severe instances, pneumonia may result in complications such as respiratory failure, sepsis, and potentially death. Annually, pneumonia impacts approximately 450 million individuals worldwide and leads to around 4 million fatalities. Although survival rates have been significantly enhanced during the 20th century, pneumonia continues to be a primary cause of mortality in developing nations, as well as among the elderly, infants, and those with chronic health conditions.

One of the most used imaging techniques in radiology, especially for diagnosing conditions affecting the lungs and thoracic region is **chest X-rays**. Interpreting chest X-rays can be challenging due to the complex anatomy and various structures that must be evaluated systematically. In the case of Pneumonia, chest X-rays often reveal opacities in the lungs, which may present as either patchy or confluent areas depending on the extent of infection. Pneumonia typically appears as a consolidation of lung tissue, with visible air bronchograms indicating infection within the airspaces. In radiological practice, detecting pneumonia is crucial as it can lead to complications if untreated. Radiologists evaluate signs such as the silhouette sign, which helps localize lesions, and air bronchograms, which suggest alveolar disease.

Although machine learning may be valuable for radiology-related tasks, such as disease classification from X-ray images, it is often costly to build a dataset for model training. A promising approach for addressing the challenges of data scarcity and annotation costs, including but not limited to the medical domain, is **active learning**. Active learning is a subfield of machine learning and artificial intelligence where the learning algorithm selects the data from which it learns, effectively allowing it to be "curious." By choosing its training data, the algorithm can perform better with less labeled data. Active learning addresses this issue by asking an oracle, often a human annotator, to label only the most informative unlabeled instances. Particularly in the domain of X-Ray imaging, by strategically selecting the most informative images for labeling, active learning algorithms can significantly improve the performance of machine learning models while minimizing the annotation effort required.

In recent years, several active learning techniques have been applied to X-ray image classification tasks with promising results. Those include the "GOAL" strategy, proposed by Nguyen et al. (2021), focusing on selecting the samples for tagging not only based on uncertainty but also on informativeness, derived by clustering. Hao et al. (2021) proposed another data-efficient methodology that can achieve the same level of accuracy as a non-active-learning model using significantly fewer images and labels, based on CNN unsupervised representation learning, followed by a Gaussian Process (GP) classifier that conducts the active learning pipeline over the learned representations.

While these methods have demonstrated effectiveness in various X-ray image classification tasks, there is still ongoing research to develop even more efficient and robust active learning strategies. Recent advancements in deep learning and generative models have opened new possibilities for active learning, such as using generative models to synthesize new, informative images for labeling. In this project, we explore the application of active learning to the domain of pediatric radiology, specifically focusing on the classification of chest X-rays into "sick" (indicating pneumonia) or "not sick." Our focus is on developing an active learning pipeline that can effectively identify the most informative images for annotation, thereby improving the performance of a pneumonia classification model while minimizing the annotation effort required.

**Methodology**

The dataset we used in this project is Kaggle's [Chest X-Ray Images (Pneumonia)](https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia). The dataset consists of 5,863 chest X-ray images of children with and without Pneumonia, performed as part of patients’ routine clinical care in Guangzhou, China in 2018. The images are organized into three sets: train, validation, and test, and are categorized into two labels: "Normal" and "Pneumonia". The training set is imbalanced, with 1,341 normal and 3,875 pneumonia images, while the validation and test sets are more balanced.

To decide on the most appropriate approach for us to the Active Learning Pipeline, we took inspiration from common approaches from the literature. For example, Hao et al. (2021) suggested to use a Convolutional Neural Network (CNN) as a classifier, fine-tuned on a small set of images prior to the active learning phase. They have also suggested to compare the sampling strategy with a baseline model trained with random sampling. According to them, the goal should be achieving high accuracy using a small percent of the labeled dataset. Hemmer, Kühl, and Schöffer (2022) presented the idea of manipulating the last layer of a ready CNN architecture to achieve better results. Mahapatra et al. (2018) suggested the use of pre-trained model on a different dataset, and augmentation of the train set.

The architecture of our experimental framework we implemented is consisted of grid search with different machine learning models and sampling methods. At each experiment, the code loads the dataset, initializes the chosen model, and runs an Active Learning loop. The loop iteratively queries unlabeled data, trains the model on the newly labeled data, and evaluates the model's performance. Results are stored and visualized for each model and sampling method.

In the data loading part, the code resizes images to 312312 pixels, converts them to PyTorch tensors, and normalizes pixel values for better training. In addition, it splits the training dataset further into labeled and unlabeled sets, resulting in four data categories: train-labeled (the initial training set for the model), train-unlabeled (from which the model selects the informative samples to annotate), validation and test.

The core of the Active Learning loop receives the model and its configuration, labeled and unlabeled training sets, validation set, and test set, and sampling strategy as input. It iterates for a specified number of iterations:

1. Starts with a copy of the model for each iteration (avoids accumulating previous training effects).
2. Trains the model on the current labeled data
3. Selects informative samples from the unlabeled pool based on the chosen strategy.
4. Updates the training data by adding the newly labeled samples and removes them from the unlabeled pool.
5. Evaluates the model on the test data and stores the metrics.

Finally, it returns a list containing the evaluation metrics for each iteration.

Todo finish to describe model file

**References**

<https://en.wikipedia.org/wiki/Pneumonia>

<https://www.glowm.com/atlas-page/atlasid/chestXray.html>

Hao, H., Didari, S., Woo, J. O., Moon, H., & Bangert, P. (2021). Highly efficient representation and active learning framework for imbalanced data and its application to covid-19 x-ray classification.‏

Hemmer, P., Kühl, N., & Schöffer, J. (2022). Deal: Deep evidential active learning for image classification. *Deep Learning Applications, Volume 3*, 171-192.‏

Mahapatra, D., Bozorgtabar, B., Thiran, J. P., & Reyes, M. (2018, September). Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 580-588). Cham: Springer International Publishing.‏

Nguyen, C., Huynh, M. T., Tran, M. Q., Nguyen, N. H., Jain, M., Vo, T. D., ... & Truong, S. Q. H. (2021, August). GOAL: gist-set online active learning for efficient chest X-ray image annotation. In Medical Imaging with Deep Learning (pp. 545-553). PMLR.‏

Settles, B. (2009). Active learning literature survey.