

# Improving Chest X-Ray Classification by RNN-based Patient Monitoring

1<sup>st</sup> David Biesner  
*Fraunhofer IAIS and University of Bonn*  
 Sankt Augustin and Bonn, Germany  
 david.biesner@iais.fraunhofer.de

2<sup>nd</sup> Helen Schneider  
*Fraunhofer IAIS*  
 Sankt Augustin, Germany

3<sup>rd</sup> Benjamin Wulff  
*Fraunhofer IAIS*  
 Sankt Augustin, Germany

4<sup>th</sup> Ulrike Attenberger  
*University Hospital Bonn*  
 Bonn, Germany

5<sup>th</sup> Rafet Sifa  
*Fraunhofer IAIS*  
 Sankt Augustin, Germany

**Abstract**—Chest X-Ray imaging is one of the most common radiological tools for detection of various pathologies related to the chest area and lung function. In a clinical setting, automated assessment of chest radiographs has the potential of assisting physicians in their decision making process and optimize clinical workflows, for example by prioritizing emergency patients.

Most work analyzing the potential of machine learning models to classify chest X-ray images focuses on vision methods processing and predicting pathologies for one image at a time. However, many patients undergo such a procedure multiple times during course of a treatment or during a single hospital stay. The patient history, that is previous images and especially the corresponding diagnosis contain useful information that can aid a classification system in its prediction.

In this study, we analyze how information about diagnosis can improve CNN-based image classification models by constructing a novel dataset from the well studied CheXpert dataset of chest X-rays. We show that a model trained on additional patient history information outperforms a model trained without the information by a significant margin.

We provide code to replicate the dataset creation and model training.<sup>1</sup>

## I. INTRODUCTION

Chest X-ray, the most common image examination method in the world, is used by radiologists to aid in the diagnosis of a wide variety of life-threatening conditions like Pneumonia or Pneumothorax. The automatic analysis of chests radiographs has therefore the potential to optimize the clinical workflow, for example through clinical decision support systems or improved workflow prioritization [1]. Due to the increasing shortage of skilled professionals and the ever rising number of examinations conducted in hospitals, improvements in clinical workflows are of major importance to ensure a consistent quality of patient care.

State-of-the-art methods for the detection of thoracic diseases in chest X-rays focus on the evaluation of a single, usually frontal, chest X-ray. However, due to the time-, cost- and radiation-efficiency of X-ray, multiple image data are often taken of a patient [2]. Once a new image is taken,

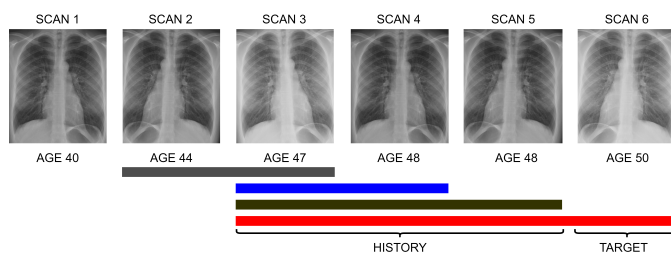


Fig. 1: Diagram of our dataset creation process for  $n_{\max\_age\_diff} = 3$  and  $n_{\min\_images} = 2$ . Here, 6 individual scans were taken of a single patient at varying ages. The colored bars under the images mark individual datapoints in our dataset. The bottom datapoint (red) contains 4 images. Scan 6 at age 50 is the target scan, scans 3 to 5 at ages 47 to 48 are the history. Note that scans can be part of the dataset both as target and as history. Additionally note that some scans from the full dataset are not present in the new patient history dataset. Scan 2 is not a target image in the dataset since it does not have any history images with age difference less than 3. Scan 1 is not part of the new dataset at all, since it does not have any history images and is not part of any history. Image from [4].

the previous images still contain valuable information that a clinical physician can use to monitor the development of certain pathologies and better classify the new information.

In addition to the sequential image data, a report text is available for each retrospective image, which contains the classification information of the image. While these unstructured reports themselves are not easily integratable in an image classification pipeline, very capable natural language processing (NLP) algorithms have been developed to automatically classify report texts and thus image data [1], [3]. It can therefore be assumed that machine-readable pathology information for older examinations are available. Within this work, we denote this combination of previous sequential image data with the corresponding findings as the patient history.

Previous approaches do not sufficiently exploit the potential of the patient history information by processing one image

<sup>1</sup>To be published in proceedings of IEEE International Conference on Machine Learning Applications IEEE ICMLA 2022.

at the time. Physicians also take a patient’s history i.e. both older image recordings and reports texts, into account when making a diagnosis. This leads to the motivation that the automatic analysis of chest X-rays can also be improved by the patient history. The sequential processing of the patient’s image data exploits the data potential of the clinical practice during modeling.

The aim of this work is therefore to combine elements of time series analysis with image processing methods and thus to enable the processing of the entire available information. This includes

- the description of the curation process for generating a novel dataset of patient histories from the CheXpert dataset of chest X-ray images,
- an overview of model architecture and argue for the inclusion of previous expert labels in the classification process,
- an evaluation of the effectiveness of the proposed method against baseline methods in which we show that, when available, patient history provides valuable information to the system and
- an outlook into open questions for further research.

Code to replicate all experiments in this study is available at [https://github.com/fraunhofer-iaais/rnn\\_patient\\_monitoring](https://github.com/fraunhofer-iaais/rnn_patient_monitoring).

## II. RELATED WORK

The evaluation of chest X-rays with regard to the automatic detection of diseases is an ongoing field of study. The Stanford Machine Learning Group developed a comprehensive public image dataset, CheXpert, for this task using automatically generated labels [1]. The dataset has enabled the development of various machine learning methods for the classification of chest diseases, e.g. implementation of deep neural networks, transfer learning or label smoothing methods [5]–[8].

However, most of the models mentioned do not consider all the image information available for a patient. Multi-view classification models can be developed for example by taking into account the lateral scan in addition to the frontal view [9], [10], improving the AUROC score of the baseline one-view model by up to 2%. In [11], sequential image data analysis techniques are implemented to improve lung disease classification. The sequential aspect of the input is artificially generated by processing the same image with different networks, and does not arise from the consideration of temporally different acquisitions.

As described in the introduction, image data of the same modality with associated report texts are often available for a patient in everyday clinical practice. To the best of our knowledge this sequential image data analysis of the patient history was not considered in the evaluation of chest X-rays yet. We would like to fill this gap through our research and investigate whether higher performance can be achieved if the image data potential of a patient is better exploited.

Our models are based on the DenseNet [12] architecture, which is a type of convolutional neural network [13] already

applied for the analysis of chest x-rays in previous studies [6], [14]–[16]. We model the patient history using a recurrent neural network called GRU (gated recurrent unit) [17], a architecture that has already been applied to a various of time series applications [18]–[20].

## III. DATA

The CheXpert dataset contains a total of 224 316 chest X-ray scans from 65 240 individual patients. There are 14 observation available relating to a variety of thoracic diseases, parsed from the corresponding reports written by clinical radiologists. For details on the data acquisition process see [1].

From the 14 observations, the CheXpert challenge proposes a subset of 5 pathologies as classification target due to their prevalence and clinical importance:

- 1) Cardiomegaly,
- 2) Edema,
- 3) Consolidation,
- 4) Atelectasis and
- 5) Pleural Effusion.

In order to keep our methods comparable to literature we restrict our model training and evaluation on the 5 pathologies as well. We further restrict the dataset to only frontal (as opposed to lateral) scans, which reduces the size of the dataset to 191 027.

Each observation in the dataset is labeled as either positive (pathology detected in the scan), negative (pathology not detected in the scan) or uncertain. The term ‘uncertain’ can refer both to the uncertainty of the radiologist in diagnosing and also the ambiguity of the report [1]. To deal with uncertain labels and convert the prediction problem into a binary classification problem, we adopt the method of [6]. The uncertain labels of the pathologies ‘Edema’, ‘Atelectasis’ are mapped to a positive, and all uncertain labels of ‘Cardiomegaly’, ‘Consolidation’ and ‘Pleural Effusion’ to a negative label. The mapping was selected to optimize performance on the hold-out hand-labeled test set provided by CheXpert, which does not contain uncertain labels [1].

For each scan the dataset provides additional metadata, like the patients sex and age. We utilize the age information present for each scan to create a novel dataset for patient monitoring.

We sort all scans by patient ID. For each scan, we filter the patient scans for scans with age at least  $n_{\text{max\_age\_diff}}$  years less than the target scan age. We remove all scans with higher age. We filter all scans from the dataset for which we find less than  $n_{\text{min\_images}}$  previous images.

For  $n_{\text{max\_age\_diff}} = 3$  and  $n_{\text{min\_images}} = 1$  (i.e. scans with at least one additional image that is not more than 3 years old) we end up with a dataset of 82 220 scans with corresponding patient history.

We split the patient monitoring dataset into a training, a validation and a test split. We split the dataset into training (80%), validation (10%) and test (10%) by patients, meaning scans from a single patient are only present in one of the splits. We receive a dataset of 73 340 target scans in the training split,

Dataset	Images	Datapoints	$R_0$	$R_1$	$R_2$	$R_3$	$R_4$
train	105 179	73 340	13.9	37.5	8.2	31.1	52.0
valid	13 270	9299	13.3	38.4	7.9	32.5	51.9
test	12 715	8717	12.6	37.6	8.2	31.2	52.9
full	191 027	191 027	12.3	33.2	6.8	31.2	40.4

TABLE I: Full statistics on our proposed datasets. Sets train, valid and test refer to our training, validation and test split for images with available patient history. These datasets contain more total images than target images (i.e. datapoints), since some images are only part of a patients history, see also Figure 1. Statistics  $R_0, \dots, R_4$  depict the ratios of positive labels to negative labels in the split for the labels Cardiomegaly, Edema, Consolidation, Atelectasis and Pleural Effusion respectively. For example, of the 82 220 target images in the training split, 11 252 images (13.7%) were positive for Cardiomegaly. We additionally provide statistics for the full dataset without any restrictions on patient history (full).

9299 target scans in the validation split and 8717 target scans in the test split. See Table I for a full overview of the dataset statistics, with additional information on the ratio of positive examples for each class.

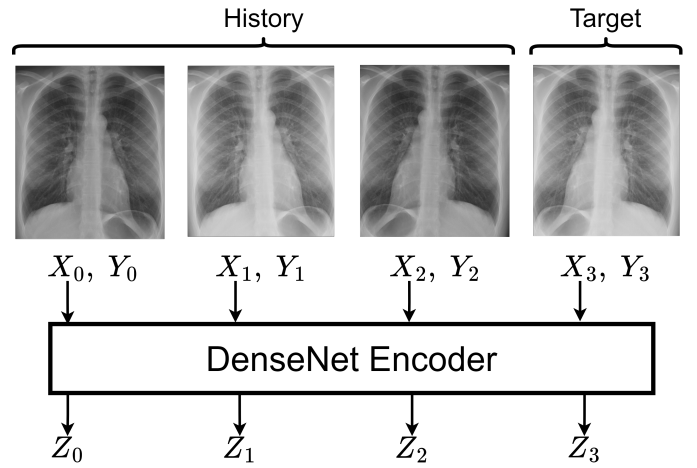
Note that a single scan can be present multiple times in the dataset: as a target image with corresponding patient history or as part of a patient history for a later target image. But an image can not be part of two dataset splits, since we split the dataset by patients. See Figure 1 for a diagram of the dataset creation.

We preprocess each image as follows. We first resize the image to 224x224 pixels, the size of the ImageNet dataset [21]. This is done in order to maximize the effect of using image classification models pretrained on the ImageNet dataset. Pretraining on ImageNet has been shown to improve classification performance in downstream tasks in various domains, including medical images [6], and that this size is sufficient for classification of the various pathologies present in the dataset [6], [7]. Further we apply a normalization scheme as in [6], which normalizes each image in the dataset to fit the mean and variance of color present in ImageNet. Before each training step we apply a random rotation, shift and zoom to each image. We do not apply random transformations to the images in the validation and test split.

#### IV. MODELS

The base architecture for all experiments in this study is DenseNet [12]. Based on previous work [5]–[8] we use a single DenseNet model, pretrained on ImageNet, with a linear classifier for all 5 labels, as a baseline.

The patient monitoring model we propose in this study employs a DenseNet architecture as well. For each image with associated patient history images we encode each image into a latent vector using the same pretrained DenseNet as the baseline model. The latent vectors are then passed through a bidirectional recurrent neural net (RNN). The output of the last



(a) Diagram of the encoding process for one datapoint with one target image  $X_3$  and three history images  $X_0, X_1, X_2$  (with corresponding labels  $Y_0, \dots, Y_3$ ). The DenseNet encoder processes each  $X_i$  individually and returns a latent representation  $Z_i$ .



(b) Diagram of the prediction architecture. We concatenate each encoded image  $Z_i$  with the corresponding labels to the previous image  $Y_{i-1}$ . The first image  $Z_0$  is concatenated with a vector of all zeros. The image-label pairs are passed through a recurrent neural net. The output of the RNN is passed through a linear layer to map it onto a logit vector of size of the target label vector and a final sigmoid layer maps the logits to a probability vector  $\hat{Y}_3$ . During training, we calculate the cross binary entropy between prediction  $\hat{Y}_3$  and  $Y_3$ .

Fig. 2: Schematic representation of both parts of the prediction architecture.

layer of the RNN is then passed to a linear classification layer. Finally the logit outputs of the classification layer are passed through a sigmoid layer, outputting probabilities for each label. As RNN architecture we use a GRU (gated recurrent unit [17]), which is similar to an LSTM (long short term memory) but contains less parameters. Our experiments have shown that the lower parameter count makes the model less prone to overfitting and the rather short sequences of images do not necessitate the specific long term memory of LSTMs.

In order to make use of information given by radiologists when examining the previous images, we enrich the RNN with the label information. For this, we append a binary vector over all labels of the corresponding image to each encoded image passed to the RNN. See Figure 2 for a more detailed description of the proposed model.

We argue that this method is valid when evaluation on a validation set and when predicting labels in a clinical setting. Each X-ray scan conducted in a hospital will eventually be examined by an expert radiologists. Methods such as ours for automatic pathology prediction only aide the clinician in their decision making process. Therefore all previous scans available at the hospital will have the relevant pathology

Model	Validation			Test		
	AUROC Macro	AUROC Micro	AUROC Weighted	AUROC Macro	AUROC Micro	AUROC Weighted
DenseNet Baseline	71.6 $\pm$ 5e-2	80.3 $\pm$ 7e-2	72.7 $\pm$ 1e-1	74.6 $\pm$ 5e-2	81.1 $\pm$ 3e-2	75.5 $\pm$ 5e-2
RNN Image	68.8 $\pm$ 1e-1	79.5 $\pm$ 7e-2	70.6 $\pm$ 7e-2	69.6 $\pm$ 7e-2	79.7 $\pm$ 1e-1	71.0 $\pm$ 1e-1
RNN Image + Label	<b>73.8</b> $\pm$ 8e-2	<b>82.1</b> $\pm$ 5e-2	<b>75.3</b> $\pm$ 1e-1	<b>74.8</b> $\pm$ 1e-1	<b>82.5</b> $\pm$ 6e-2	<b>75.7</b> $\pm$ 1e-1
RNN Label	72.4 $\pm$ 8e-2	81.6 $\pm$ 5e-2	73.7 $\pm$ 8e-2	73.6 $\pm$ 8e-2	82.1 $\pm$ 8e-2	74.5 $\pm$ 7e-2

TABLE II: Evaluation of the proposed RNN-based methods against the baseline DenseNet method. We evaluate on the validation and test split of our novel patient monitoring dataset. We compute the AUROC score and display the score over all 5 classes, as macro average, micro average or weighted average. For each split, we apply 10x bootstrapping to receive a statistically robust performance metric. We provide the mean score over all bootstrapped subsets and the standard deviation.

information present. We can therefore assume that our system will have access to this information during inference time.

While most examinations of chest X-rays only result in written reports by radiologists, not structured label information, there has been much work conducted on the automatic extraction of labels from free text reports. The dataset considered in this study, CheXpert [1], itself provides labels extracted by rule-based methods from free text reports. Further studies [3] have examined the use of sophisticated transformer-based language models [22] for label extraction and further improved the accuracy. Application of such algorithms to automatically analyze written reports and use extracted information to enhance the image classification system is feasible in a clinical setting.

In order to examine whether the decision making process of the model is only dependent on the previous labels, we train an additional RNN-based model only on the label information, we therefore ignore the actual images and pass the binary label vector directly to the RNN.

## V. TRAINING AND EVALUATION

### A. Training Details

For this evaluation we train 4 distinct models:

- *DenseNet Baseline*: The baseline DenseNet model used in various previous studies [1], [6],
- *RNN Image*: The proposed RNN-based model with only the encoded images passed through the recurrent neural net,
- *RNN Image + Label*: The proposed RNN-based model with both the encoded images and corresponding labels passed through the recurrent neural net, and
- *RNN Label*: The proposed RNN-based model with only the labels passed through the recurrent neural net.

We train each model on the training split of our patient monitoring dataset until convergence of the binary-cross-entropy loss and choose the model with the lowest validation loss for evaluation on the test set. We train the models for a maximum of 30 epochs with a learning rate of 1e-3 and the reduce-on-plateau learning rate scheduler using the Adam [23] optimizer.

### B. Evaluation

We evaluate all trained models against the validation and test split of our patient monitoring dataset. To receive a statistically

robust estimate the model performance, we apply Poisson bootstrapping [24] to resample 10 validation and test splits. We evaluate the metrics over all bootstrapped datasets and compare mean and standard deviation.

Our models output a probability vector over all labels for each image. The probability denotes the confidence of the model that a certain pathology is visible in the image. In practice, one would apply a threshold to classify pathologies as present or not present, depending on the probability output of the model. However, choosing a threshold introduces bias towards one type of classification error, and therefore thresholds must be determined with a certain use case in mind. Metrics such as Precision, Recall, Accuracy and F1-Score require a certain classification threshold to be set. We therefore consider a threshold independent metric in this study, the area under the receiver operating characteristic curve (AUROC [25]). A perfect AUROC score is achieved if every positive sample is assigned a higher probability than every negative sample, i.e. if a threshold with 100% accuracy is possible. The AUROC score then denotes the ratio of correctly sorted examples in the dataset.

We average the AUROC score over the 5 classes using three methods:

- macro averaging: unbiased average of all five scores,
- weighted averaging: average of all five scores, weighted by class support, i.e. number of positive examples in each class, and
- micro averaging: global score over all samples and classes.

We evaluate our proposed methods against the DenseNet baseline in Table II. Considering the AUROC metrics on the test set, we see that the RNN-model with both image and label information (*RNN Image + Label*) scores higher than the baseline for all averaging methods. This holds as statistically significant considering the 10-factor bootstrapping and the displayed standard deviation. The most significant difference shows in the micro-averaged AUROC score on the test set, in which the proposed model scores 1.4 percentage points higher than the baseline.

We do not see this full effect when we restrict the information the RNN model receives to only the encoded images or even the label vectors. The *RNN Image* model scores lower than both the baseline and the *RNN Image + Label* model in all evaluations. The *RNN Label* model scores lower than

the *RNN Image + Label* model in all evaluations, and only slightly higher than the baseline in some. Both RNN-based models with only partial patient history information are not able fully utilize the additional data and reliably improve on the baseline. In order to take full advantage of the patient history information available in the novel dataset, we must present both the encoded images and the previous label vectors to the model.

## VI. CONCLUSION AND OUTLOOK

In this study we proposed a novel method for classification of chest X-ray images and argued how taking into account information from the patient history can aide image classification systems in the prediction of clinical pathologies found in chest radiographs.

The proposed model consists a recurrent neural net, which processes the sequence of encoded patient history images with corresponding pathology labels. We determine that the inclusion of real labels from the dataset is justifiable in this setup, since in a clinical setting previous X-ray images will always have been examined by an expert radiologist.

To test the method we implemented a novel dataset based on the well-known CheXpert dataset of chest X-rays. The method consistently improves the classification performance if such patient history is available for a given image. We release the code for training and dataset reproduction, such that more future research can investigate the effect of inclusion of patient history in the image classification pipeline.

Open research questions include:

- Does the inclusion of patient history benefit image models trained on other chest X-ray datasets or other medical imaging datasets?
- We see an improvement of the baseline method when increasing the number of training images to the full dataset. How does the performance of the proposed model improve when trained on more data?
- Are more sophisticated sequence models (like transformer architectures) better at parsing information from the encoded sequences or does the increase in parameter count only work towards overfitting the model?
- Can the direct comparison of patient history image data further increase the performance of the model?

We plan to address these questions in future work and hope to implement our methods in practice applications soon.

## REFERENCES

- [1] Jeremy Irvin et al., “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 590–597.
- [2] Daniel S. Kermany, Michael Goldbaum, et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.
- [3] Akshay Smit et al., “Chexpert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert,” *arXiv preprint arXiv:2004.09167*, 2020.
- [4] Mikael Häggström, “Chest radiograph,” [https://commons.wikimedia.org/wiki/File:Normal\\_posteroanterior\\_%28PA%29\\_chest\\_radiograph\\_%28X-ray%29.jpg](https://commons.wikimedia.org/wiki/File:Normal_posteroanterior_%28PA%29_chest_radiograph_%28X-ray%29.jpg), 2017, last accessed 08.09.2022.
- [5] Helen Schneider, David Biesner, et al., “Improving intensive care chest x-ray classification by transfer learning and automatic label generation,” in *European Symposium on Artificial Neural Networks ESANN 2022*, 2022.
- [6] Hieu H Pham et al., “Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels,” *Neurocomputing*, vol. 437, pp. 186–194, 2021.
- [7] Zhuoning Yuan et al., “Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3040–3049.
- [8] Matthew BA McDermott et al., “Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output,” in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 913–927.
- [9] Xiongfeng Zhu and Qianjin Feng, “Mvc-net: Multi-view chest radiograph classification network with deep fusion,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 554–558.
- [10] Jonathan Rubin, Deepan Sanghavi, et al., “Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks,” 04 2018.
- [11] Md. Zabirul Islam et al., “A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images,” *Informatics in Medicine Unlocked*, vol. 20, pp. 100412, 2020.
- [12] Gao Huang et al., “Densely connected convolutional networks,” *CoRR*, vol. abs/1608.06993, 2016.
- [13] Keiron O’Shea and Ryan Nash, “An introduction to convolutional neural networks,” 2015.
- [14] Nkechinyere N. Agu, Joy T. Wu, et al., “Anaxnet: Anatomy aware multi-label finding classification in chest x-ray,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part V*, Berlin, Heidelberg, 2021, p. 804–813, Springer-Verlag.
- [15] Helen Schneider et al., “Towards symmetry-aware pneumonia detection on chest x-rays,” in *IEEE Symposium Series on Computational Intelligence SSCI 2022*, 2022.
- [16] Jianhong Cheng et al., “Classification and detection of covid-19 x-ray images based on densenet and vgg16 feature fusion,” *European Radiology*, 02 2022.
- [17] Junyoung Chung, Caglar Gulcehre, et al., “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- [18] Jianlong Xu, Kun Wang, et al., “Fm-gru: A time series prediction method for water quality based on seq2seq framework,” *Water*, vol. 13, no. 8, 2021.
- [19] Daria Lavrova et al., “Using gru neural network for cyber-attack detection in automated process control systems,” in *2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2019, pp. 1–3.
- [20] Ekaterina Kalinicheva, Dino Ienco, et al., “Unsupervised change detection analysis in satellite image time series using deep learning combined with graph-based approaches,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1450–1466, 2020.
- [21] Jia Deng, Wei Dong, et al., “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [22] Jacob Devlin, Ming-Wei Chang, et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.
- [23] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” 2014.
- [24] James A. Hanley and Brenda MacGibbon, “Creating non-parametric bootstrap samples using poisson frequencies,” *Computer Methods and Programs in Biomedicine*, vol. 83, no. 1, pp. 57–62, 2006.
- [25] Francisco Melo, *Area under the ROC Curve*, pp. 38–39, Springer New York, New York, NY, 2013.