

Impact of Non-Visual Data on Multiclass Medical Radiology Image Classification Performance

Pratham Shah, Luke O'Donnell, MD

1. Purpose

Chest X-ray (CXR) analysis traditionally relies solely on the image for multiclass disease classification (pneumonia, pneumothorax, etc.). This study investigates if incorporating non-visual data (patient demographics, X-ray acquisition details) alongside the image improves diagnoses. This project analyzes diverse models: multilayer perceptron (MLP), convolutional neuronal network (CNN) with Batch Normalization (BN), pre-trained InceptionResnetV2, CNN with recurrent neuronal network (CRNN), and Dynamic Affine Feature Map Transform (DAFT) CNN. These models are evaluated on an unbalanced test dataset mimicking real-world scenarios, the research explores whether a more comprehensive approach, similar to how doctors utilize patient context, can enhance the accuracy of machine learning models for CXR analysis.

2. Introduction

Multiclass classification tasks are workhorses in many areas, including medical imaging analysis. (Castellino, 2005) In CXR analysis for example, a model might be trained to classify an image as depicting pneumonia, pneumothorax, cardiomegaly, or other conditions. (Kaviani et al., 2022) This classification is based solely on the image itself, a convenient approach but one that overlooks the complexities of real-world medical diagnosis.

Physicians don't make diagnoses in isolation. They consider a patient's entire medical picture, including age, health history, and even seemingly mundane details like how the X-ray was taken. (Nixon et al., 2024) This richer context is crucial for interpreting test results accurately. The Bayesian Pretest/Post-Test Probability (BPP) framework helps quantify this process. Here, a physician first estimates the initial likelihood of a specific condition in the patient (pretest probability) based on their background information. Then, a diagnostic test like a CXR is conducted. Finally, by combining the pretest probability with the test results, the physician arrives at a final diagnosis (post-test probability). (Nixon et al., 2024)

Consider a CXR suggesting pneumonia. This image might be interpreted differently depending on the patient's age and health. In a young, healthy patient, the pretest probability of pneumonia would be quite low. Even if the CXR showed some concerning features, a physician might reasonably conclude the image is normal. Conversely, the same image in a clinically sicker, older patient with symptoms suggestive of pneumonia would likely be interpreted as confirmation of the disease due to the higher pretest probability.

This study delves into the potential of incorporating such patient metadata, functioning as pretest probability information, into machine learning models for CXR analysis. The dataset used in this research includes details like patient age, gender, and even the viewpoint (PA or AP) from which the X-ray was taken. Notably, CXR viewpoint can be an indicator of a patient's overall health, with PA views preferred for mobile patients and AP views often used for sicker, hospitalized patients. (de Lacey et al., 2008)

By including this non-visual data alongside the CXR images during training and evaluation, the researchers aim to quantify the impact on the models' predictive abilities. Models like MLP, CNN, InceptionResnetV2, CRNN, and DAFT are evaluated with and without this additional data. This

investigation sheds light on how leveraging a more holistic view of the patient, similar to how physicians approach diagnosis, can potentially improve the accuracy and effectiveness of machine learning models in CXR analysis.

3. Hypothesis

Including non-visual data in medical diagnostic images will improve a model’s predictive abilities for multiclass image classification. This additional information functions as pretest probability, which influences the accuracy of the post-test probability and final classification.

4. Data

The National Institutes of Health (NIH) CXR Dataset includes 112,120 images in PNG format with 15 possible labels. One CXR can be assigned multiple labels if several diagnostic findings are present.

The preprocessing of this dataset included filtering for six labels. These labels were carefully selected to focus on three distinct anatomic locations in the chest. The label "cardiomegaly" pertains to the heart, while the labels "atelectasis," "consolidation," "fibrosis," and "mass" describe pathologies in the lungs. The label "pneumothorax" focuses on the area between the chest wall and the lung tissue.

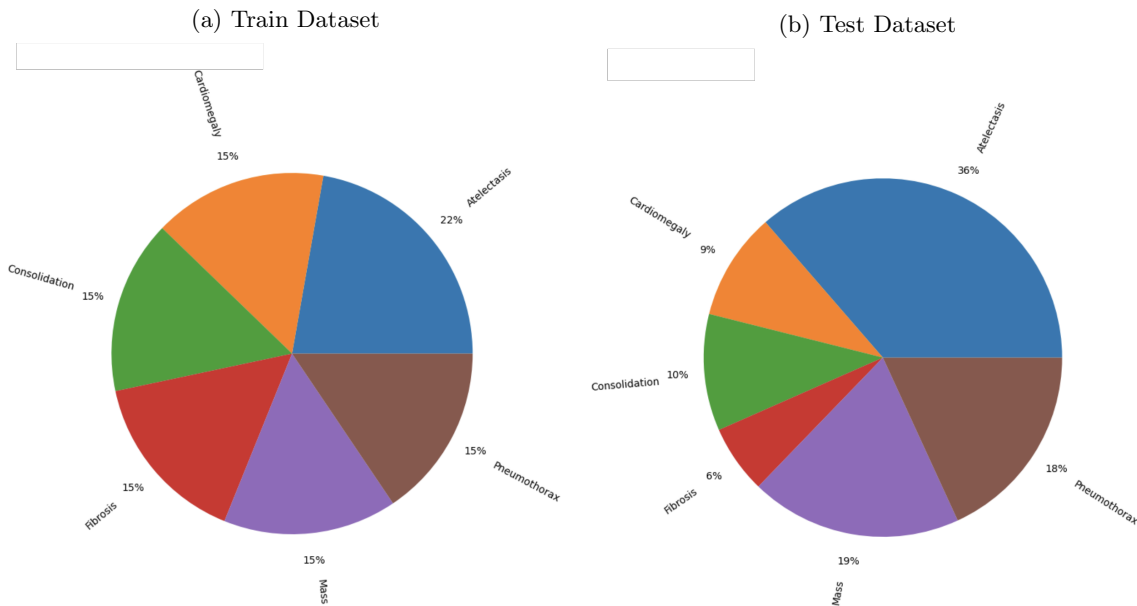


Figure 1: Labels were balanced in the (a) train dataset so the number of samples for each label was at least 70 percent of the predominant label with the use of upsampling. The (b) test dataset was left unbalanced and only used unique data for each label.

This filtered dataset contained 11,678 unique CXRs, which were subsequently divided into train, validation, and test datasets in a ratio of 70 percent, 10 percent, and 20 percent, respectively. To ensure the robustness of our model, we took steps to mitigate potential overfitting during the training phase. The train and validation datasets underwent partial upsampling, ensuring that the sample

number of images for each label was at least 70 percent of the most represented sample number. In contrast, the test dataset was intentionally left unbalanced to better mimic real-world clinical practice.

The non-visual metadata for each CXR included the patient’s age, sex categorized as male or female, and CXR viewpoint as either AP or PA. The patient’s age was normalized. The patient’s sex and the CXR viewpoint were made binary.

The data was structured in a dictionary that included the true label, the CXR as a normalized 1x1024x1024 matrix, and non-image data made into a 1x3 vector.

5. Material and Methods

Each model underwent two training sessions: one using only CXR image data and another using CXR image and non-visual data.

The hyperparameters for all models were meticulously standardized, enabling a fair cross-model comparison. A batch size of 16, shuffled before each train and validation epoch, was chosen to facilitate frequent weight and bias adjustments. MLP, CNN with BN, RCNN, and DAFT underwent 12 epochs of training and validation, with the Cross-Entropy loss function used to calculate loss after each step. Given computational needs, InceptionResnetV2 underwent 6 epochs. The Adaptive Moment Estimation (Adam) optimizer function. With the exception of InceptionResnetV2, a learning rate of 0.01, was employed for updating weights and biases. InceptionResnetV2 has a learning rate of 0.001 given significant increased depth of model.

The evaluation stage of each model was conducted using the weights and biases that yielded the highest accuracy in the validation dataset during training. This approach assessed the model’s predictive capabilities using the best-performing parameters.

The location for including non-visual data depended on each model’s architecture.

MLP used flattened images for the input layer and three subsequent hidden layers. The node count for the hidden layers was successively 512, 256, and 128, with ReLU applied after each layer. When using non-visual data, the vector containing information was concatenated to the front of the input flattened image before being processed through the hidden layers.

CNN with BN was designed with five convolution layers with successive 16, 32, 64, 128, and 254 output channels. Each convolutional layer used a kernel size of 3x3 with no padding for progressive size reduction. ReLU and BN follow each convolution. The data was ultimately flattened with an adaptive pool layer and fed into a subsequent fully connected (FC) layer for final classification. When included, non-visual data was added to the flattened output of the adaptive pool with an additional BN before the data was fed into the FC layer.

InceptionResNetV2 has 164 layers with residual connections, and the Inception architecture includes multiple convolutions of varying sizes concatenated together. (Szegedy et al., 2016) The pretrained InceptionResNetV2 model allows transfer learning and is trained on a million images of 1000 unique objects from the ImageNet database.(Szegedy et al., 2016)

The pretrained InceptionResNetV2 model was downloaded from the timm library available on Hugging Face. This downloaded model was fine-tuned for one channel of input data as opposed to three channels. Output was adjusted for the six possible labels. When incorporating non-visual data, this

vector was concatenated to the end of the flattened output of the adaptive layer. This concatenated data underwent subsequent BN before being fed into the FC layer.

For the CRNN model, the process starts with convolutional layers, conv1 and conv2, which extract features from the chest X-ray images. (Zhao et al., 2019) Subsequently, max-pooling layers reduce the spatial dimensions of the feature maps while retaining essential information. The output is then reshaped to prepare for input into the LSTM layer, which captures sequential information and dependencies across the image. Finally, a fully connected layer maps the LSTM output to the output classes, facilitating classification.(Zhao et al., 2019)

Conversely, the DAFT CNN begins with DAFT layers, daft1 and daft2, incorporating dynamic affine transformations to enhance feature extraction and spatial robustness. (Pölsterl et al., 2021) These transformations are followed by max-pooling operations to further refine the features. After flattening the feature maps, two fully connected layers process the features, with the final layer producing class probabilities through a softmax activation function.(Pölsterl et al., 2021)

6. Results

This study evaluated the performance of various machine learning models for classifying multiple lung diseases in chest X-rays (CXRs). A threshold of 0.5 was used to distinguish between positive and negative classifications when calculating accuracy and F1 scores. Due to the intentional imbalance in the test dataset, mimicking a real-world clinical setting, overall model performance was assessed using weighted F1 scores and weighted area under the curve (AUC). Additionally, label performance in each model was evaluated for each label using a one-vs-rest (OvR) approach to calculate label specific F1 scores and AUC scores.

	Weighted F1 Score	Atelectasis F1 Score	Cardiomegaly F1 Score	Consolidation F1 Score	Fibrosis F1 Score	Mass F1 Score	Pneumothorax F1 Score
MLP Image Only	0.31	0.53	0.00	0.00	0.00	0.32	0.31
MLP Image and Non-Visual Data	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CNN with BN Image Only	0.32	0.50	0.17	0.30	0.02	0.13	0.30
CNN with BN Image and Non-Visual Data	0.37	0.56	0.23	0.32	0.22	0.32	0.21
InceptionResnetV2 Image Only	0.45	0.47	0.52	0.29	0.26	0.41	0.44
InceptionResnetV2 Image and Non-Visual Data	0.48	0.61	0.51	0.27	0.24	0.40	0.50
CRNN Image Only	0.19	0.53	0.00	0.00	0.00	0.00	0.00
CRNN Image and Non-Visual Data	0.06	0.00	0.00	0.00	0.00	0.32	0.00
DAFT Image Only	0.19	0.53	0.00	0.00	0.00	0.00	0.00
DAFT Image and Non-Visual Data	0.01	0.00	0.00	0.00	0.12	0.00	0.00

Table 1: Weighted F1 scores for each model trained on images alone and on images with non-visual data. Also included are F1 scores for each label in each model.

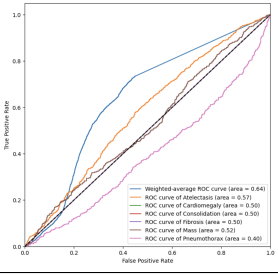
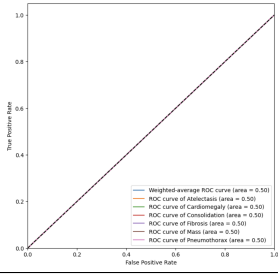
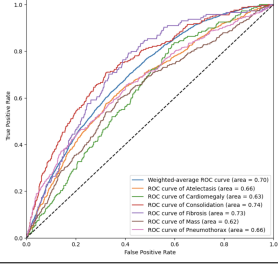
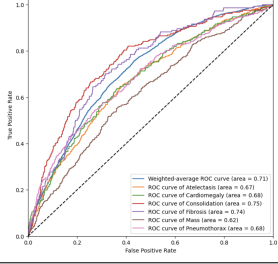
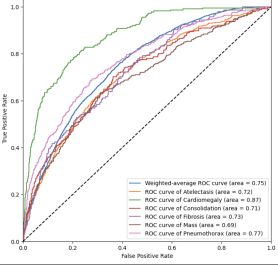
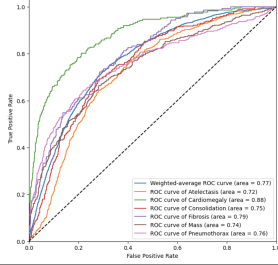
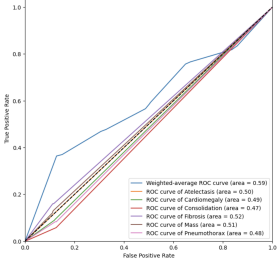
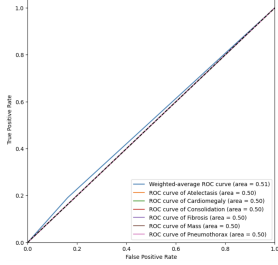
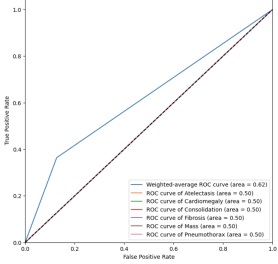
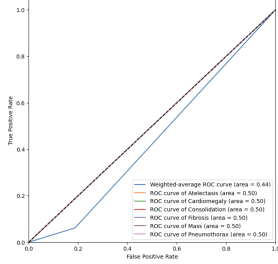
Model	Image Data	Image and Non-Visual Data
MLP		
CNN with BN		
Inception-Resnet-V2		
CRNN		
DAFT		

Table 2: Multiclass Receiver Operating Characteristic (ROC) with weighted area under curve (AUC) and one-vs-rest (OvR) label-specific AUC for each model trained on image data only versus trained on image and non-visual data.

Among the models tested, only the Convolutional Neural Network (CNN) with Batch Normalization (BN) and the InceptionResnetV2 architecture demonstrated any significant ability to differentiate between disease classes. For both of these models, incorporating non-visual patient data alongside the CXR image during training and evaluation led to slightly improved performance. The CNN with BN showed an increase in weighted F1 score from 0.32 to 0.37 and a rise in weighted AUC from 0.70 to 0.71. InceptionResnetV2 achieved an even greater improvement, with its weighted F1 score rising from 0.43 to 0.44 and its weighted AUC increasing from 0.75 to 0.77.

InceptionResnetV2 emerged as the best performing model overall, being the only one to approach an F1 score of 0.5, which indicates reasonable performance. The high weighted AUC of 0.77 achieved by InceptionResnetV2 suggests that F1 scores could potentially be further improved by optimizing the decision threshold used for binary classification. Currently, a standard threshold of 0.5 is applied across all models.

Furthermore, InceptionResnetV2 achieved the best performance in separating "cardiomegaly" from other disease classes, with an impressive OvR (one-vs-rest) AUC of 0.88 when trained on both image and non-visual data. "Pneumothorax," a condition related to air around the lungs, was the second-best differentiated disease with an OvR AUC of 0.76 using InceptionResnetV2 trained on a combination of image and non-visual data. Conversely, diseases like "atelectasis," "consolidation," "mass," and "fibrosis," all located within the lung tissue, proved to be the most challenging to distinguish between, resulting in lower performance.

The remaining models evaluated, including MLP, CRNN, and DAFT, exhibited no meaningful predictive abilities. When trained solely on image data, MLP produced a misleadingly high weighted F1 score of 0.57. However, this was due to the model simply predicting "atelectasis" for all images. It's important to note that "atelectasis" was the over represented class even in the balanced training dataset, constituting 22 percent of the samples compared to only 15 percent for other classes. The low AUC of 0.4 for "pneumothorax" further highlights the difficulty this model had in learning and generalizing from the data.

Interestingly, when MLP was trained on a combination of image and non-visual data, its performance completely collapsed. The model ended up predicting zero for all possible labels, essentially failing to classify any image. This suggests that the inclusion of irrelevant or incompatible non-visual data might have hampered the model's ability to learn.

CRNN and DAFT also displayed a lack of predictive ability. When trained only on images, both models predicted all samples to be "atelectasis," again reflecting the over representation of this class in the data. When non-visual data was incorporated during training, CRNN began predicting all labels as "mass," while DAFT consistently predicted "fibrosis" for all images. These nonsensical predictions indicate that these models were unable to leverage the additional non-visual information effectively.

7. Discussion

This study aimed to evaluate the impact of non-visual data on the performance of various machine learning models for lung disease prediction using chest images. We compared different models, including MLP, CNN with BN, InceptionResnetV2, CRNN, and DAFT. While InceptionResnetV2 and CNN with BN achieved the best overall performance, we intentionally used simplified CRNN and DAFT architectures to prioritize observing the effect of non-visual data on model behavior rather than achieving peak performance with these architectures.

Our main finding was that incorporating non-visual data generally improved model performance for InceptionResnetV2 and CNN with BN. InceptionResnetV2, for example, showed an increase in both weighted F1 score (from 0.43 to 0.44) and weighted AUC (from 0.75 to 0.77) when trained with both image and non-visual data compared to using images alone. This suggests that relevant non-visual information can contribute to more accurate disease prediction.

It is also worth noting that our study focused on multiclass prediction, aiming to differentiate between various lung diseases. This inherently presents a more complex challenge compared to a binary classification task (disease present or absent). Additionally, medical images often differ from natural images in terms of information content. (Varoquaux and Cheplygina, 2022) Extracting relevant features from medical images can be more challenging due to factors like variability in image quality and subtle disease manifestations. Also image labeling is subjected to dataset availability, bias, and human error in labeling. (Varoquaux and Cheplygina, 2022)

Furthermore, it is important to acknowledge that the inclusion of non-visual data might not always be beneficial. Some models, like CRNN and DAFT, performed poorly when trained with both data types. It is possible that the specific non-visual data introduced noise or irrelevant features, making it harder for these models to learn meaningful representations for disease classification.

Future research should focus on identifying the most relevant non-visual data points and exploring techniques to improve model robustness to potential noise in the data. In addition, while InceptionResnetV2 had the best performance, it also required high computational costs needing at least an hour for each epoch of training. Future research can further investigate lighter models allowing for highly detailed image processing and analysis.

8. Contribution Statement

Pratham Shah worked on CRNN and DAFT model architecture building. Pratham also contributed to the final paper writeup with additions to the result section and discussion section. Pratham was also provided the final text editing of the paper.

Luke O'Donnell worked on MLP, CNN, InceptionResnetV2 architecture. Luke also worked on database set-up on the HPC and writing needed code for training and evaluating models. Luke also contributed to the final paper writing and Overleaf formatting.

References

- Ronald Castellino. Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17–19, 2005.
- Gerald de Lacey, Simon Morley, and Laurence Berman. *The Chest X-Ray: A Survival Guide*. Saunders Ltd, United Kingdom, 2008.
- Parisa Kaviani, Mannudeep Kalra, Subba R Digumarthy, Reya Gupta, Giridhar Dasegowda, Ammar Jagirdar, Salil Gupta, Preetham Putha, Vidur Mahajan, Bhargava Reddy, Vasanth K Venugopal, Manoj Tadepalli, Bernardo C Bizzo, and Keith J Dreyer. Frequency of missed findings on chest radiographs (cxrs) in an international, multicenter study: Application of ai to reduce missed findings. *Diagnostics (Basel)*, 12(2382), 2022.
- Michelle Pistner Nixon, Farhani Momotaz, Claire Smith, Jeffery Smith, Mark Sendak, Christopher Polage, and Justin Silverman. From pre-test and post-test probabilities to medical decision making. *medRxiv [Preprint]*, 2024.
- Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. *Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform*, page 688–698. Springer International Publishing, 2021. ISBN 9783030872403. doi: 10.1007/978-3-030-87240-3_66. URL http://dx.doi.org/10.1007/978-3-030-87240-3_66.
- Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. URL <http://arxiv.org/abs/1602.07261>.
- Gael Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit. Med.*, 2022.
- Bin Zhao, Xuelong Li, Xiaoqiang Lu, and Zhigang Wang. A cnn-rnn architecture for multi-label weather recognition, 2019.