Thorax Disease Prediction via Machine Learning-based X-ray Classification

Miaolei Bao, Lingfei (Ellen) Jiang

University of Washington, Seattle, December 05, 2024

Introduction

Chest X-rays are among the most common and accessible imaging examinations for thorax disease screening and diagnosis. The growing accumulation of X-ray data and the associated interpretive challenges motivate the development of computer-aided diagnosis methods. Advanced machine learning techniques have shown promise in automating the procedure. In this study, we employed two supervised machine learning models—Support Vector Machine (SVM) and Convolutional Neural Network (CNN)—to predict the presence of thoracic diseases. We assessed and compared the prediction accuracy of these models to evaluate their performance and suitability. Additionally, we investigated the potential benefits of incorporating demographic information into the models.

Data

The data set utilized in this study originates from a publicly available random sample of the National Institute's collection of chest X-ray images in the United States. The original collection comprised 112,120 X-ray images with disease labels from 30,805 unique patients. Our sample consists of 5,606 images (Figure 1), each with dimensions of 1024 x 1024 pixels, accompanied by a dataset containing image information, disease type labels, and patient demographics (ID, age, gender, number of follow-ups). The sample was curated using Natural Language Processing techniques to extract disease classifications from associated radiological reports. The labels are expected to have an accuracy rate exceeding 90%, making them suitable for weakly-supervised learning.

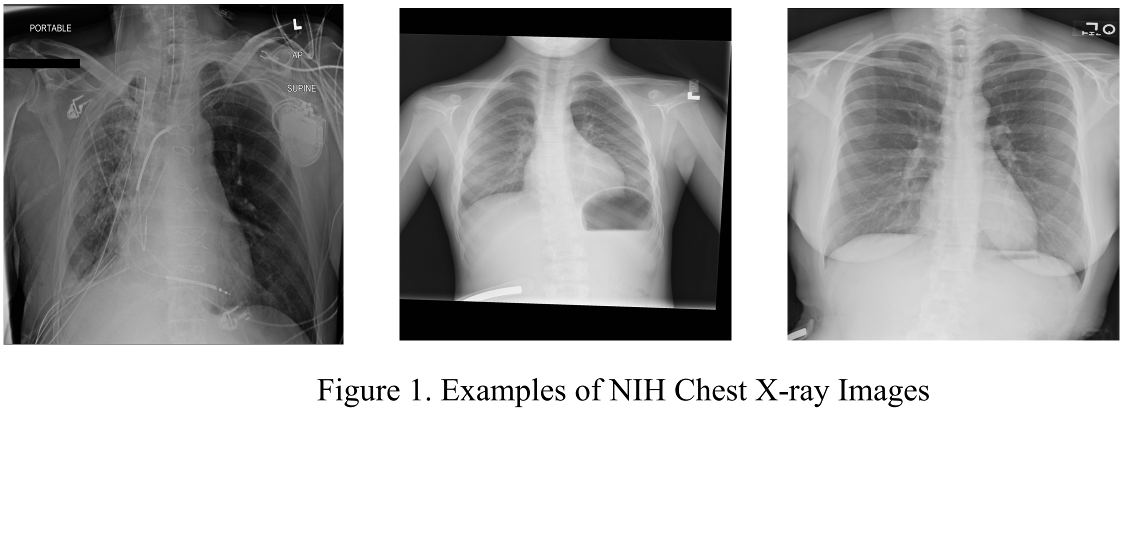


Figure 1. Examples of NIH Chest X-ray Images

Preprocessing

The data set was cleaned before being used in the analysis. First, 15 binary variables were created to indicate the presence or absence of diseases or no findings for each patient. A binary variable indicating the absence of repeated measures was created. The last letter “Y” in the characteristic age variable was removed and the age was turned into numeric. An outlier observation with an age of 411 was likely mislabeled and was treated as missing and removed. Collinearity was addressed by randomly selecting one record for patients with repeated measures. After cleaning, the final dataset comprised 4,229 observations with no missing values.

In the exploratory data analysis (EDA), we visualized the numeric variables “follow_up_num” and “age” (Figure 2,3), and disease labels with histograms (Figure 4), and created a descriptive Table 1 for all variables (Table 1). We summarized that about 43.6% of the patients had a label of “No Finding” and the diagnosis of all other thorax diseases was very low, with the lowest label “Hernia”, which only comprises 0.2% of the sample population (Figure 4).

Furthermore, the images were preprocessed to prepare for machine learning methods utilizing the R package “magick”. We read chest X-ray images into R, resized them into a fixed dimension of 64 x 64, converted them into grayscale, generated their pixel values, normalized pixel intensity to range 0 - 1, and combined the image data with the sample data set for each individual observation. We verified that no images were missing from this dataset.

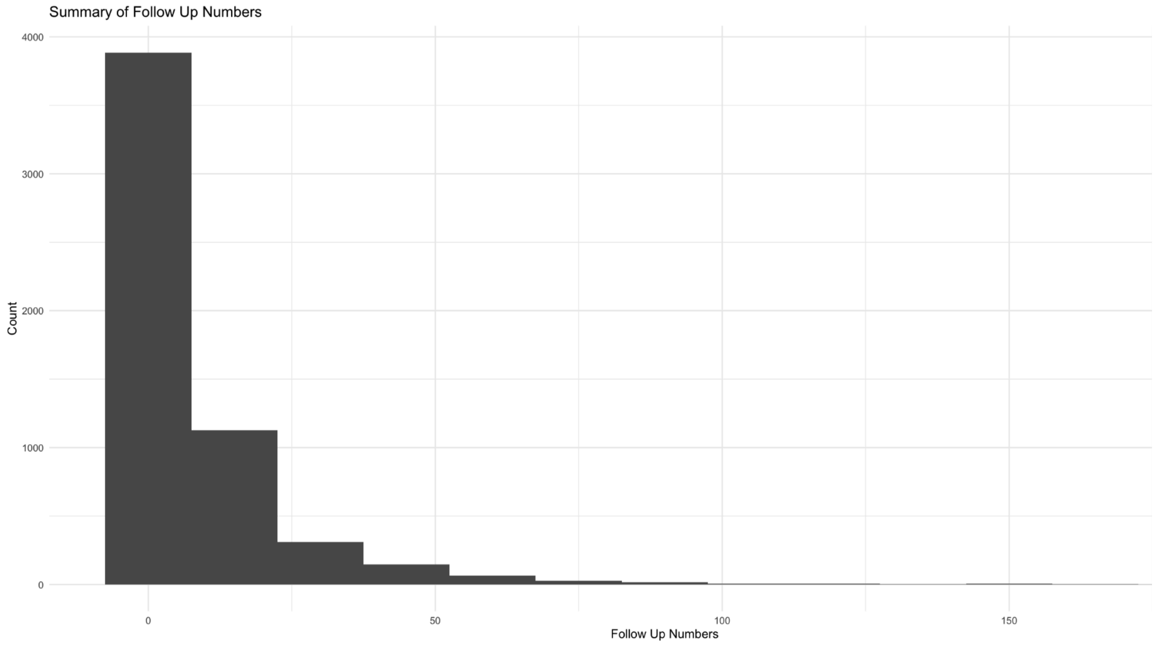
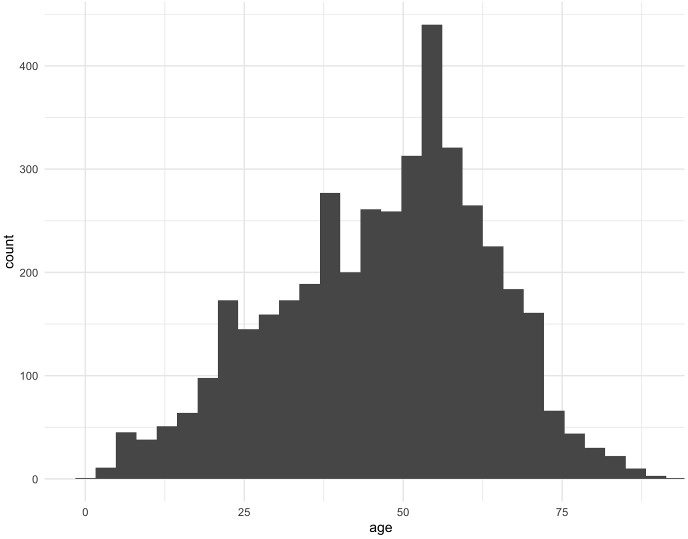


Figure 2. Summary of Follow-Up Numbers



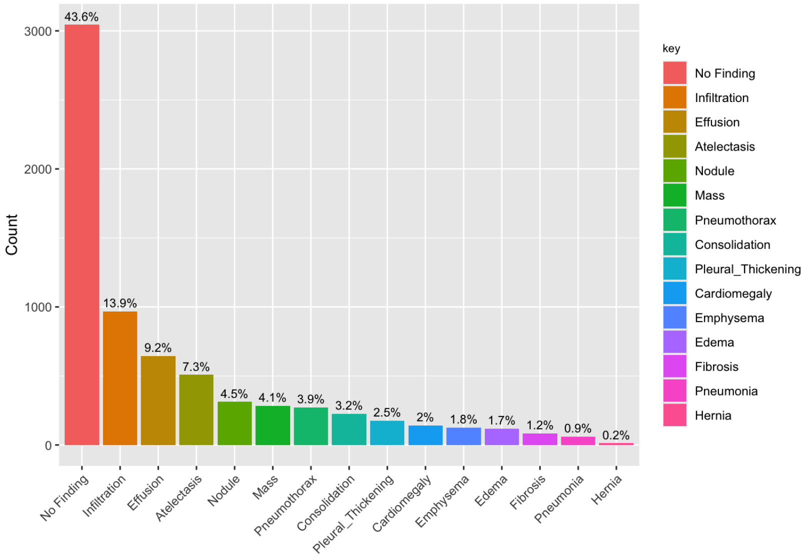
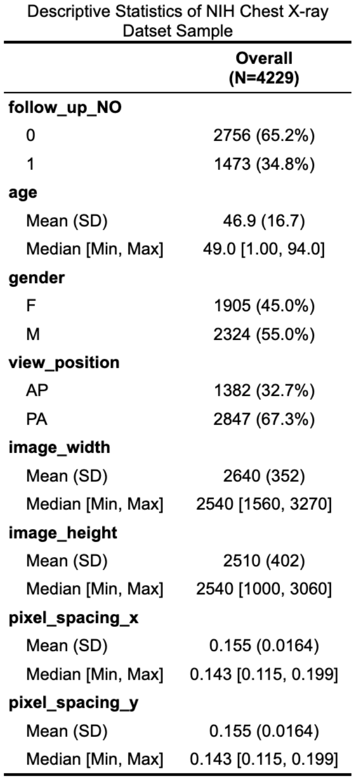


Figure 3. Summary of Age

Figure 4. Summary of Disease Findings

Table 1: Descriptive Statistics



Descriptive Statistics of NIH Chest X-ray Datset Sample

	Overall (N=4229)
follow_up_NO	
0	2756 (65.2%)
1	1473 (34.8%)
age	
Mean (SD)	46.9 (16.7)
Median [Min, Max]	49.0 [1.00, 94.0]
gender	
F	1905 (45.0%)
M	2324 (55.0%)
view_position	
AP	1382 (32.7%)
PA	2847 (67.3%)
image_width	
Mean (SD)	2640 (352)
Median [Min, Max]	2540 [1560, 3270]
image_height	
Mean (SD)	2510 (402)
Median [Min, Max]	2540 [1000, 3060]
pixel_spacing_x	
Mean (SD)	0.155 (0.0164)
Median [Min, Max]	0.143 [0.115, 0.199]
pixel_spacing_y	
Mean (SD)	0.155 (0.0164)
Median [Min, Max]	0.143 [0.115, 0.199]

Partition

To evaluate the model validity and avoid over-optimism, we randomly split the dataset into 80% training and 20% testing using a seed value of 1. We converted the training and testing data set into matrices for subsequent model implementation.

Methods

Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm that performs data classification by separating data classes with the optimal hyperplane. It has been shown to have a good performance in disease detection on chest X-ray images (Erdaw & Tachbele, 2021). We used the R package "e1071" and applied the SVM algorithm to each of the 15 disease labels. In each model, we first assumed a linear kernel and cost of 1 for training, made predictions on the testing data, and then calculated prediction accuracy from the confusion matrix for evaluation.

Convolutional Neural Network

Convolutional Neural Network (CNN) is another machine learning algorithm that can be implemented in medical image classification. Characterized by their convolutional layers, CNNs are adept at capturing spatial and hierarchical patterns in data (Alzubaidi L et al., 2021). We used the R package "keras3" to implement the CNN model. The sequential model has a architecture as described in Table 2. We adopted a ReLU activation to introduce non-linearity and a dropout layer to prevent overfitting. The final dense layer employed a sigmoid activation function to output probabilities for each disease. We trained the model on the 80% training data with a batch size of 32 over 10 epochs, while model performance was validated on the separate testing data.

Table 2. The Architecture of the CNN model

Sequence	Layer
0	Input Image
1	Convolutional Layer
2	Max-pooling Layer
3	Convolutional Layer
4	Max-pooling Layer
5	Flatten Layer
6	Dense Layer
7	Dropout Layer
8	Dense Layer (Output)

Results

We compared the prediction results for 14 diseases and “no finding” with the true labels and calculated the corresponding accuracy. The results were summarized in Figure 5 for the SVM model and in Figure 6 for the CNN model. Overall both models yielded similar outcomes. The accuracy was high for rare diseases. For instance, Hermia and Pneumonia, which account for 0.2% and 0.9% of cases in the dataset, had accuracy rates close to 1. In contrast, only about half of the “no finding” cases (43.6% in the dataset) were correctly classified. After a closer examination of the prediction results for the CNN, we discovered that it classified all cases into “no finding” in order to minimize the loss.

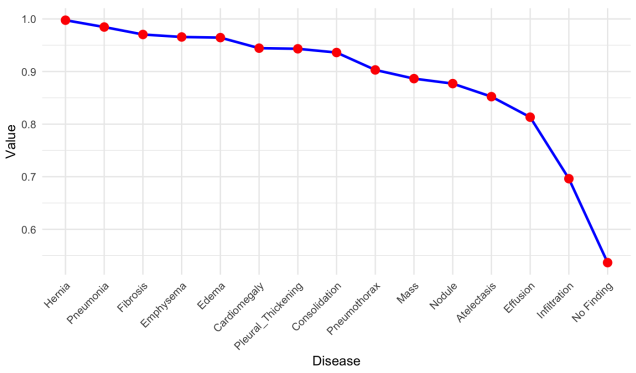


Figure 5. Disease Prediction Accuracy for SVM

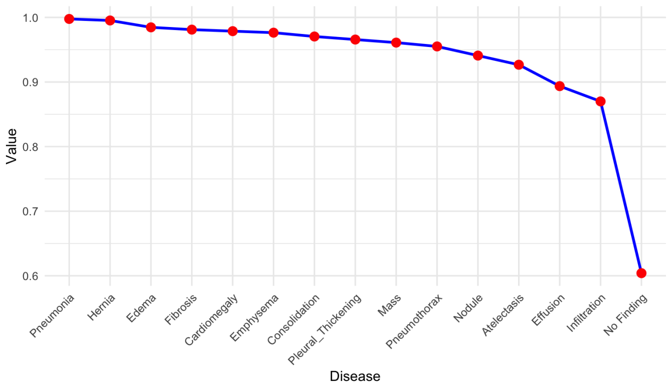
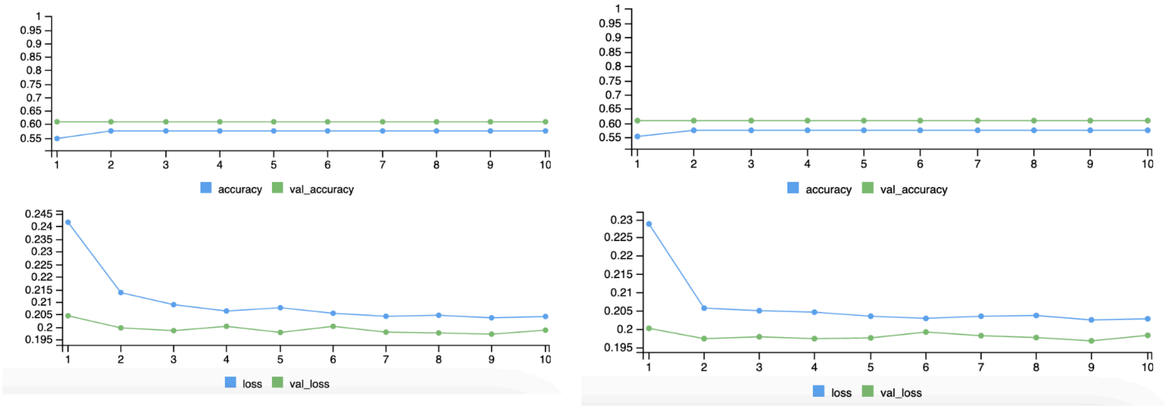


Figure 6. Disease Prediction Accuracy for CNN

To investigate the potential benefits of incorporating demographic information into the models, we added a matrix containing individual age and sex alongside the image information matrix to the model, then retrained the CNN model. The loss only changed slightly and the prediction results remained the same (Figure 7).



Modeling results using image data only

Model results using image data and demographics

Figure 7. Modeling Results Investigating Demographic Information Effects

Table 3 showed the performance of the two models.The validation dataset contains 846 images, each with 15 disease labels, resulting in a total of 12690 records. Of these, both the SVM and CNN correctly predicted the diseases in 11,012 cases. CNN was correct while SVM was wrong in 833 cases, and SVM outperformed CNN in 231 cases. In 614 cases, both models were incorrect. SVM had a correct prediction rate of 88.6% and CNN had a correct prediction rate of 93.3%. Therefore, CNN demonstrated better performance in this study.

Table 3. Model Comparison Results

	SVM correct	SVM wrong
CNN correct	11012	833
CNN wrong	231	614

Discussion

The two algorithms we used are all popular machine learning methods for disease prediction but their performance is all limited by imbalanced data and poor feature selection. For each disease finding, especially for those with rare cases, both algorithms heavily relied on the healthy groups and tended to label everyone as having no findings. In the CNN model, we used "accuracy" as the evaluation metric during the training process, which is not the optimal choice. Theoretically, F1 scores would perform better in handling imbalanced data. In this study, we also retrained the model using F1 scores as an exploratory analysis. However, CNN's prediction results remained the same—it classified all cases as "no findings." Therefore, to address the imbalance, more sophisticated methods should be employed in the future, such as oversampling the minority classes, to create a more balanced dataset.

In this study, we also explored adding demographics into the model to improve model performance. We hypothesized that age might be strongly associated with thorax diseases. However adding demographics to the dataset did not change the results. On one hand, this could suggest that age and sex did not provide additional information beyond what the model had already learned from the imaging data. On the other hand, this outcome could also be related to the imbalanced nature of the data and limitations in the model setup.

We also did an exploratory analysis to tune the parameters in SVM. We wanted to use 5-fold cross-validation on different combinations of SVM tuning parameters cost and gamma in model training and use the best model with the lowest error rate for model testing as suggested in the codes. However, the procedure was too computationally expensive and therefore we were unable to implement it. With more computational power, we can further improve the model.

References

Erdaw, Y., & Tachbele, E. (2021). Machine Learning Model Applied on Chest X-Ray Images Enables Automatic Detection of COVID-19 Cases with High Accuracy. International journal of general medicine, 14, 4923–4931. https://doi.org/10.2147/IJGM.S325609

Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Big Data 8, 53 . https://doi.org/10.1186/s40537-021-00444-8