Identify Pneumonia in Chest X-Rays with Machine Learning methods

PneumoniaDetector

April 2022

- Ning Wan; ningwan; Email: ningwan@seas.upenn.edu
- Zheng Han; hanzheng; Email: hanzheng@seas.upenn.edu
- Zhonghao Lian; zhlian; Email: zhlian@seas.upenn.edu

Abstract

Pneumonia is an infection of one or both of the lungs caused by bacteria, viruses, or fungi. In medical image diagnosis, pneumonia often presents itself as increased opacity in chest X-Rays. In order to improve the efficiency and accuracy of identifying pneumonia from chest X-Rays, various models have been proposed for this task. In this work, we examine multiple non-deep and deep learning models for classifying chest X-Ray images. We first feed pre-processed data to a family of traditional machine learning models, including SVM, Random Forest, Gradient Boosting, etc. Then, we compare the results of several deep learning models, which include a self-designed CNN, AlexNet, and ResNet18. While most previous work focuses on CNN-based models, we introduce Swin Transformer, which is a state-of-the-art computer vision model that leverages ideas like patch embedding. Statistical results obtained demonstrate that Swin Transformer can generate significantly better classification results for chest X-Rays than CNN-based models.

1 Introduction

With the rapid development of digital images, image classification plays an essential role in the medical area. Effectively and efficiently classifying medical images is instrumental in clinical treatment. Moreover, with the spreading of Covic-19, pneumonia has been a growing concern within the public and the medical society. Our work aims at classifying pneumonia images and identifying images with signs of pneumonia using machine learning and deep learning techniques.

In clinical diagnosis and treatment for pneumonia, the chest X-Ray is the most commonly used examination. Pneumonia usually presents itself as area of increased opacity on chest X-Rays. In order to identify pneumonia based on chest X-Ray images, professional radiologists are required. However, for this time-consuming task, the number of experienced radiologists is limited. Also, there exist other factors that can lead to increased opacity in chest X-Rays, such as bleeding, lung cancer, etc. Therefore, introducing advanced machine learning and deep learning techniques into this task can be of substantial impact.

In the field of identifying pneumonia with chest X-Rays, several machine learning and deep learning frameworks are explored in previous work. Among the models, CNN-based methods are the most frequently discussed, due to their inherent advantages in dealing with image classification tasks. In our work, we first present and compare results for several traditional machine learning models, which serve as our baseline. Then, we examine and compare results for several CNN-based deep learning models, including our self-designed CNN, ResNet18, and AlexNet. Finally, we apply Swin Transfomer to this task, which achieves significant improvements compared to the baseline.

2 Related Work

Many types of research have been conducted in the domain of pneumonia chest X-Ray image classification. Among the studies, many find CNN to be the optimal base model for this task. Stephen, Okeke, et al (2019)[5] proposed a baseline model based on CNN, which is trained on chest X-Ray images with several data augmentation techniques. Varshni, Dimpy, et al (2019)[6] proposed a technique to utilize pre-trained CNN as feature-extractors, followed by different classifiers to identify abnormal chest X-Rays.

Other researches also focus on other models in the deep learning domain. Kong, Lingzhi, and Jinyong Cheng (2021)[1] introduced a deep learning technique based on the combination of the Xception neural network and long-term short-term memory (LSTM), which also follows a two-step process that firstly extracts features from X-Ray, then performs the classification. Yadav et al (2019)[8] proposed a transfer learning-based approach to classify pneumonia images, which shows that retraining specific features on a new target dataset is essential to improve performance.

In order to make some huge improvements in detection performance, some cutting-edge deep learning models should be deployed. Swin Transformer is one of them. According to the research conducted by Ze Liu, etc.[3] The basic idea about that is to use patches from images as sequences and pass them into the transformer layers as the convolution procedure. This is a great innovation that applies NLP-like methods to Computer Vision problems.

3 Dataset

We use the CXR dataset from Pneumonia Detection Challenge published on Kaggle, which is provided by the Radiological Society of North America

(RSNA). All the images in the dataset are in DICOM format.

The main dataset consists of CXR images from 30227 patients. Among the pictures, the ones that have clear evidence of pneumonia are marked with bounding boxes, which highlight the opaque areas in the CXR images. The bounding boxes are denoted by the index of their upper-left corners. In contrast, those without definite evidence of pneumonia do not have such bounding boxes.

Since the focus of this study is on image classification, we eliminate the bounding box information, while focusing on the CXR images themselves. The CXR data are RGB images of various sizes, each has its corresponding patient id and label. The dataset is unbalanced: pneumonia positive images account for one-third of the dataset. We randomly assign 10% of the images as the test dataset, the rest are used to train the models. For data preprocessing, we applied Random Horizontal Flip on the images to increase generality, and also resize all the images to size 224x224.

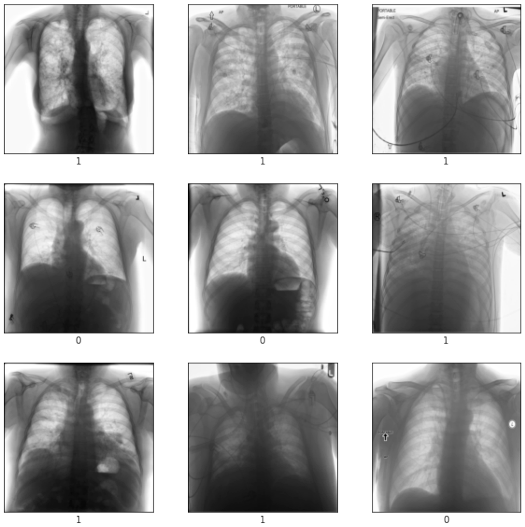

Figure 1: Sample data in the chest X-Ray dataset

4 Methodology

In this section, we deploy the non-deep learning methods, base deep learning methods, and advanced deep learning architectures.

4.1 Non-deep Learning Method

4.1.1 Data Pre-processing

For non-deep learning models, e.g., KNN, SVM, etc., the original X-ray image dataset is too large to deal with. Therefore, we extract features from the original image dataset. The features we choose are *mean, standard deviation, area, perimeter, equivalent diameter, irregularity, and HU moments.* In order to get these features, we construct a feature extraction pipeline as shown in Figure 2. First, we calculate the mean value and standard deviation value of the original X-ray grey image. Then we enhance the image: we use 1) histogram equalization to show a good contrast of the lungs, 2) high pass filter to sharpen the image, exhibiting more image details, 3) Otsu thresholding to smoother lung edges, 4) Sobel filter to detect and extract the lung edges and get a contour. We also segment the lungs by using the contour, and in order to better recognize the lung opacity position, the center of moment of the segment is needed. Moreover, during X-ray inspection, patients have different body sizes and may move the body, leading to slightly scaled and rotated images. However, we want the center of moment to be invariant to scale and rotation, so we take Hu moments as features. After calculating Hu moments (7 values in total), the 3rd value and 7th value are removed since the 3rd moment depends on the other moments and 7th moment is relevant to mirror images but we do not have mirror images in the dataset. Therefore 5 values of Hu moments are applied.

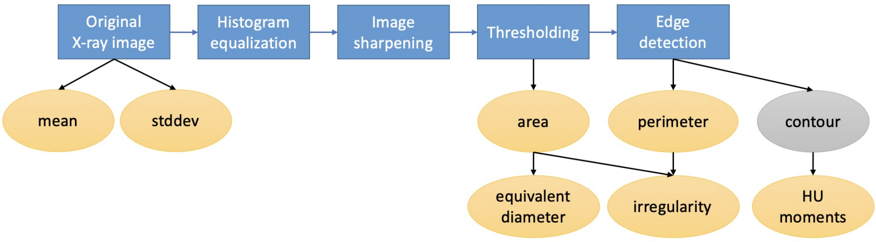

Figure 2: Feature extraction pipeline used in the non-deep method. Blue rectangles are pre-processing operations, the grey oval is an auxiliary feature, and yellow ovals are chosen features we will feed in non-deep models.

4.1.2 Non-deep Learning Models

As for non-deep learning models, we select 1) classic machine learning models: logistic regression, KNN, and SVM; and 2) ensemble machine learning models: random forest and gradient boosting. We apply the scikit-learn package [4] to deploy non-deep learning models.

We use Grid Search [2] to optimize the hyperparameters. We prepare 2 or 3 candidate values for important hyperparameters. And the followings are the best hyperparameter combinations we explored for each model.

Logistic regression. We utilize LogisticRegression function. In logistic regression model, we set 'penalty' to 'l2', regularization strength 'C' to '1', maximum number of iterations 'max_iter' to '100'. Other hyper parameters remain default.

KNN. We utilize KNeighborsClassifier function. In this model, we set 'n_neighbors' to '10', and other hyper parameters remain default.

SVM. We utilize SVC function. In this model, we set 'kernel' to 'rbf', regular-

ization strength 'C' to '10'. Other hyper parameters remain default.
Random Forest. We utilize RandomForestClassifier function. In this model, we set estimator numbers 'n_estimators' to '600', the maximum depth of the tree 'max_depth' to '10', the minimum number of samples required to split an internal node 'min_samples_split' to '2'. In order to speed up, we enable all the precessors to run in parallel. Other hyperparameters remain default.
Gradient Boosting. We utilize GradientBoostingClassifier function. In this model, we set estimator numbers 'n_estimators' to '400', 'learning_rate' to '0.01', maximum depth of individual regression estimators 'max_depth' to '5', minimum number of samples required to split an internal node 'min_samples_split' to '2'. Other hyper parameters remain default.

4.2 Deep Learning Method

4.2.1 Data Pre-processing

We split the training data into truly used training data and testing data with a ratio 9:1. We have done the data augmentation here to make the images convenient to be trained. For example, we have implemented RandomHorizontalFlip for the image and resized it to 224. We also have done visualization to show some samples of transformed data.

4.2.2 Deep Learning Models

We have implemented a self-designed CNN model without any pretrained steps. Besides, we fine-tuned pretrained deep learning models like AlexNet and ResNet18. The fanciest model we implement here is Swin Transformer. The learning rate is controlled at 0.001 for all models except Swin Transformer performs well. For Swin, we tuned that hyperparameter and find out 0.00001 is good for training.
Self-Designed CNN. This CNN is a network with 8 convolution layers and 3 fully connected layers.

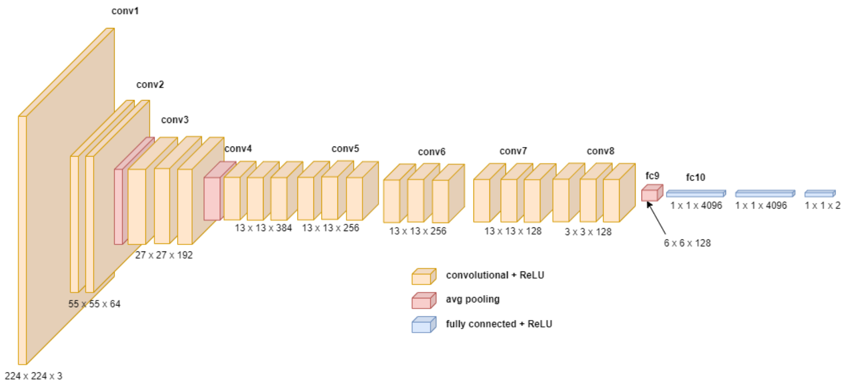

Figure 3: Self-Designed CNN.

AlexNet. We used the pretrained AlexNet and then made a transfer-learning on this new classification problem by changing classifier parameters and setting the class number to be 2. The training data are passed through this model.

ResNet18. We used the pretrained ResNet-18 and then made a transfer-learning on this new classification problem by setting the class number to be 2. The training data are passed through this model.

Swin Transformer. Developed by Ze Liu, etc.[3] as Figure 4, Swin Transformer is a state-of-art deep learning model that applies transformers to computer vision problems. It is to regard patches divided from images as sequences, which are passed through linear embedding, then put into transformers for training. Unlike first-generation vision transformer, this one uses an idea like patch embedding in order to make the trained patches grow from small size to large size with similar behavior as the traditional CNN. Also, the shifting window multi-head self-attention created here is an innovative idea to build connections between neighboring patches by making use of the advantage of the shifting window. It also reduces it from $O(n^2)$ (relationship between every two tokens) in the traditional NLP transformer to $O(2n)$ (relationship between each patch and its neighbors), which saves much time complexity. Here, we are using the Swin-L (C = 192, layer numbers =2, 2, 18, 2) for both untrained and pretrained model, where C is an arbitrary dimension projected to from the linear embedding layer.

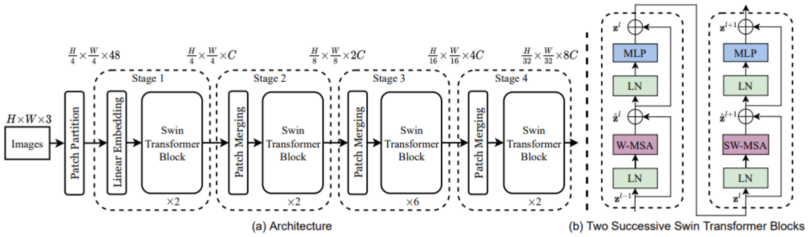

Figure 4: The basic architecture of Swin Transformer. There are four-stage pieces of training with embedding and transformers inside.

5 Results

In this section, we show the results of the above methods described in Section 4. We choose Accuracy, Precision, Recall and F1 score as performance metrics.

5.1 Non-deep Learning Method

We train 5 non-deep learning models - Logistic Regression, KNN, SVM, Random Forest, and Gradient Boosting - with hyperparameters described in section 4.1.2. Then we compare the performance of these models, and the result is shown in Table 1. From this table, we can conclude that Gradient Boosting has the best performance among all the non-deep learning models listed above on all metrics. And the confusion matrix of Gradient Boosting is shown in Table 2.

Table 1: The summary of the performance of different non-deep learning models. And Gradient Boosting has the best performance among all non-deep learning models listed above.

Non-deep learning models	*Accuracy*	*Precision*	*Recall*	*F1-score*
Logistic Regression	0.661	0.658	0.335	0.444
KNN	0.688	0.654	0.486	0.557
SVM	0.663	0.704	0.281	0.407
Random Forest	0.788	0.771	0.677	0.721
Gradient Boosting	**0.792**	**0.772**	**0.689**	**0.728**

Table 2: The confusion matrix of the Gradient Boosting. The name of each row in bold text is the "Real Class", and the name of each column in the italic text is the "Predicted Class".

	Normal	*Pneumonia*
Normal	3035	506
Pneumonia	732	1673

5.2 Deep Learning Method

We train 4 deep learning models, self-designed, AlexNet, ResNet, and Swin Transformer. The result is shown in Table 3. We use the cross Entropy to calculate the loss and do backpropagation. The self-designed one does not have great performance since it does not use any transfer learning here, which is implemented from scratch. The other three perform better than that. And Swin Transformer performs best among them, which has an accuracy of around 0.88. Actually, for that model, we have tried both the original version and pretrained version and kept the pretrained one, because it shows a better final result and has faster improvement during the training. The confusion matrix of the Swin Transformer is shown in Table 4. The training process is shown in Figure 5.

Table 3: The summary of the performance of different deep learning models. And Swin Transformer has the best performance among all deep learning models listed above.

Deep learning models	*Accuracy*	*Precision*	*Recall*	*F1-score*
Self-Designed CNN	0.774	0.730	0.478	0.577
AlexNet	0.832	0.729	0.740	0.734
ResNet	0.842	0.740	0.787	0.762
Swin Transformer	**0.879**	**0.807**	**0.798**	**0.803**

Table 4: The confusion matrix of the Swin Transformer. The name of each row in bold text is the "Real Class", and the name of each column in the italic text is the "Predicted Class". Note that the data size used for the deep learning model and the non-deep learning one is different.

	Normal	*Pneumonia*
Normal	1910	178
Pneumonia	189	746

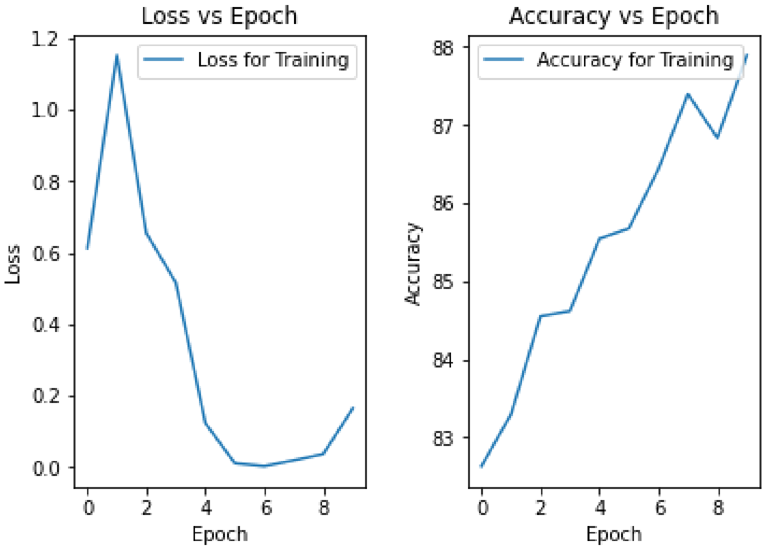

Figure 5: Loss vs Epoch and Accuracy vs Epoch for Pretrained Swin Transformer

6 Discussion

6.1 Findings

For non-deep learning models, we find that ensemble methods, especially Gradient Boosting, achieve better results. The reason is that the ensemble method can prevent the classifier from being stuck in a local optimum since it can start from much different initialization. From another perspective, any classifier has some risk of misclassification, therefore, combining multiple classifiers can reduce the risk.

For deep learning models, Swin Transformer shows its great power in analyzing images. This is because it makes use of the innovative shifting self-attention property of the transformer and regards CV as a sequential problem so that more inner structure could be detected here.

For all the models, it is obvious that the deep learning models outperform non-deep learning models. This is because deep learning models can extract deeper and more abstract features, which better simulates humans for abstract understanding.

6.2 Limitations and Ethical Considerations

A fundamental ethical concern in the use of medical image data lies in privacy protection. Since medical images are closely associated with patients' personal privacy, image collection should follow regulations and guidance, including

European General Data Protection Regulation (GDPR)[7], and also under the consent for use and reuse of patients.

Another ethical concern is about how the results of the classification model satisfy medical transparency standards. In a medical setting, doctors have to disclose meaningful details about medical treatment to patients. Thus, if the diagnosis is replaced by black-box machine learning systems, it is hard to tell why the system outputs a certain diagnosis. As we can see in the study, we can barely tell the difference between an abnormal chest X-Ray image and a normal one. Thus, it is crucial that there's at least some level of human intervention involved.

6.3 Social Impact Analysis

With the spreading of Covid-19 throughout the world, pneumonia is more of a substantial concern than ever before. Traditionally, the doctors have to manually look at the chest X-Rays to identify possible signs of pneumonia, which could be a burden. With the introduction of such image classification techniques, tens of thousands of images can be classified in seconds, which significantly expedite the process of looking through an array of chest X-Rays.

While this work mainly focuses on identifying pneumonia through chest X-Rays, similar thought processes can be easily migrated to the diagnosis of other diseases, such as heart failure, heart valve disease, etc. With more advanced and accurate classification models being developed, medical image classifications can contribute to the medical world in a stronger sense.

However, as such techniques being more widely introduced in day-to-day medical treatment process, related regulations and guidelines must be developed. We believe the best practice for applying such technique is to classify the images first with models, and then have the doctors to examine the suspicious images and make a conclusion. Otherwise, some patients that develop early stages of pneumonia might not be diagnosed in time, which can have serious consequences.

6.4 Future Research Directions

Due to the limitation of time and compute power, there is still room for improvement in the future.

First, we can optimize the hyperparameters more carefully. We may try more hyperparameters in Grid Search and may apply other optimization methods - e.g., Random Search, Bayesian Optimization, etc. - to achieve better results.

Then, we can also try to generate medical image captions. This is a good way to improve the diagnosis confidence. If the model thinks that the patient has pneumonia, it must give some evidence, explaining the reason - for example, the algorithm can generate sentences like this: ”There is opacity at the bottom of the right lung.” And this can assist radiologists to diagnose effectively and efficiently.

Further, we hope to propose more computer vision models like ViT and Swin Transformer, which are inspired by NLP models like Transformer. To some extent, the pixels or patches in images are also sequences, thus some NLP models have great potential to be utilized in the computer vision area. We want to contribute not only to medical images in the application area but also to general model backbones.

7 Acknowledgment

We would like to thank Professor Lyle Ungar, Konrad Kording, and Teaching Assistant Yifei Li for their guidance during this semester. We would especially thank TA Yifei Li for his guidance in our dataset selection and model recommendation.

8 Conclusions

The project implements not only traditional but also cutting-edge machine learning models. The traditional non-deep learning methods mostly take advantage of the mathematical property of the images. For the deep learning models, different sorts of networks are implemented and show better outcomes than traditional models. Meanwhile, the Swin Transformer shows the best performance and the most potential in learning from images. We expect more NLP-like models could be well utilized for different machine learning problems for further explorations. We hope our models, especially Swin Transform model, could help detect Pneumonia in healthcare industry.

References

[1] Li-Qin Kong and Jinyong Cheng. Based on improved deep convolutional neural network model pneumonia image classification. *PLoS ONE*, 16, 2021.

[2] Steven M. LaValle and Michael S. Branicky. On the relationship between classical grid search and probabilistic roadmaps. In *WAFR*, 2002.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, 2019, 2019.

[6] Dimpy Varshni, Kartik Thakral, Lucky Agarwal, Rahul Nijhawan, and Ankush Mittal. Pneumonia detection using cnn based feature extraction. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7, 2019.

[7] Effy Vayena, Alessandro Blasimme, and Ivan Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15, 2018.

[8] Samir S. Yadav and Shivajirao Manikrao Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6:1–18, 2019.