Chest X-ray Disease Detection Using CheXpert Dataset

Barış Karaer, Mustafa Alper Sayan

***Abstract*—Large datasets have been very beneficial for the deep learning methods to achieve high levels of performance on various of medical tasks. CheXpert is a large dataset that contains 224,316 chest radiographs of 65,240 patients. A labeler has been designed to automatically detect the presence of 14 various diseases. This dataset has been carefully collected, designed and shared publicly. We have implemented various Artificial Neural Networks to evaluate and compare the performances of different kinds of models such as Single Layer Perceptron's, Multi-Layer Perceptron's and Convolutional Neural Networks. Accurately detecting the presence of multiple diseases from chest x-rays is a challenging task even with the usage of state-of-the art models. The most promising models are ensemble CNN`s.**

Keywords: Chest X-ray, CheXpert, Multi-label Classification, Single layer perceptron, Multi-layer perceptron, Convolutional Neural Network

I. Introduction

Chest Radiography is a common imaging examination globally and it is used to diagnose patients with various diseases. Predicting chest radiographs could be very helpful for radiologists and medical practitioners. With the help of neural networks, these images would be interpreted, and the results of the computations could lead great insights to the field of medicine and computer science. There has been a recent effort to use the machine learning methods to diagnose common thoracic diseases from CXR images. This is a challenging task even for a trained practitioner [1]. Therefore, there is a need for a support tool for practitioners via the usage of deep learning to help the patients diagnosed more precisely. This support tool might save lives and eliminate wrong diagnoses due to human error. This support tool especially important for regions where there exists a lacking number of trained practitioners. It will not just only fasten the diagnoses procedure but it will also give a chance to practitioners focus on more challenging tasks since their valuable time is saved using this support tool. To achieve these goals, we used the following methods single layer perceptron, multi-layer perceptron and convolutional neural networks. We trained our models on the CheXpert dataset which is one of biggest datasets publicly available. All of our models compiled using Pytorch library. [2] PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab (FAIR). It is free and open-source software released under the Modified BSD license.

II. Dataset

The original dataset consists of 224,316 images of chest radiographs, and each of these images are labelled with specific diseases. For saving some time during model training, we have used 20000 of these images. Each of them is labelled with 14 different diseases. These diseases are no finding, enlarged cardio mediastinal silhouette, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: [3] The CheXpert dataset consists of 14 labeled observations. We report the number of studies which contain these observations in the training set

Data Collection and Label Selection

[3] Stanford University collected chest radiographic studies from Stanford Hospital, performed between October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports. From these, we sampled a set of 1000 reports for manual review by a board-certified radiologist to determine feasibility for extraction of observations. We decided on 14 observations based on the prevalence in the reports and clinical relevance, conforming to the Fleischner Society's recommended glossary (Hansell et al. 2008) whenever applicable. "Pneumonia", despite being a clinical diagnosis, was included as a label in order to represent

the images that suggested primary infection as the diagnosis. The “No Finding” observation was intended to capture the absence of all pathologies.

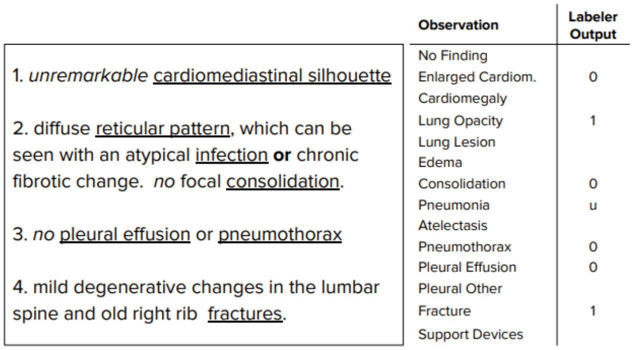

Figure 2: [3] Output of the labeler when run on a report sampled from our dataset. In this case, the labeler correctly extracts all of the mentions in the report (underline) and classifies the uncertainties (bolded) and negations (italicized).

Furthermore, Single Layer Perceptron’s, Multi-Layer Perceptron’s, Convolutional Neural Networks are used to predict the relevant disease of the patient. To further visualize the data, here is a sample image that was given with the dataset.

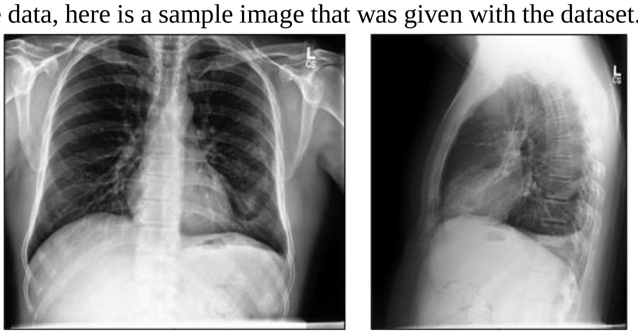

Figure 2: [2] An example chest X-ray image from multiple perspectives

We trained our models using %36 of the original dataset. This was not an arbitrary choice using %5, %10, %20 of the datasets resulted in much lower accuracy rates. With using %36 of the original dataset we maintained approximately the same accuracy. The difference was close to %1. It should also be noted that we are using downscaled version of the dataset. The original dataset is approximately 439 GB. The down-samples version is approximately 11 GB. It takes approximately 1 hours to do 10 epochs on Google Colab with GPU. This played an important role with parameter tuning.

III. Single Layer Perceptron

The perceptron is a type of supervised learning. It decides whether an input belongs to the class by using weight coefficients. It is a parametric method.[6] A Single Layer Perceptron is a feed-forward network based on a threshold transfer function. Single Layer Perceptron’s are the simplest type of ANN’s and can only classify linearly separable cases with a binary target (True, False or 1, 0). Initial weights of these SLP’s are assigned randomly, without any prior knowledge. SLP sums all the weighted inputs and if the sum is above a certain threshold (predetermined value), then SLP is set to be activated. Output should be 1 in this case. Otherwise, the output would be 0.

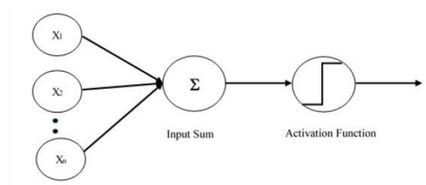

Figure 3: [8] Single Layer Perceptron Example

For $i = 1, \dots, K$, For $j = 0, \dots, d$, $w_{ij} \leftarrow$ rand$(-0.01, 0.01)$
Repeat
  For $i = 1, \dots, K$, For $j = 0, \dots, d$, $\Delta w_{ij} \leftarrow 0$
  For $t = 1, \dots, N$
    For $i = 1, \dots, K$
      $o_i \leftarrow 0$
      For $j = 0, \dots, d$
        $o_i \leftarrow o_i + w_{ij} x_j^t$
    For $i = 1, \dots, K$
      $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$
    For $i = 1, \dots, K$
      For $j = 0, \dots, d$
        $\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i) x_j^t$
  For $i = 1, \dots, K$
    For $j = 0, \dots, d$
      $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$
Until convergence

Figure 4: [4] Single layer perceptron pseudocode.

The SLP we have implemented has different accuracies for different labels. The detailed results for different diseases are explained below.

Disease	AUROC
AUROC mean	0.55
no finding	0.64
cardio mediastinum	0.45
Cardiomegaly	0.57
Lung opacity	0.56
Lung lesion	0.59
Edema	0.65
Consolidation	0.59
Pneumonia	0.50
Atelectasis	0.59
Pneumothorax	0.46
Pleural Effusion	0.59
Pleural Other	0.48
Fracture	0.53
Support Devices	0.56

Table 2: Shows the AUROC results of the single layer perceptron

We used this model as a baseline for our more complicated models. As can be seen from the table above we are reaching mean AUC of %55 which is slightly above random prediction. It is clear from this result that we need more complicated models to solve chest x-ray disease recognition task. Our Single layer perceptron model works as follows it first flattens the inputted image, the flattened input is inputted to a fully connected layer then the model applies sigmoid activation function to output the results.

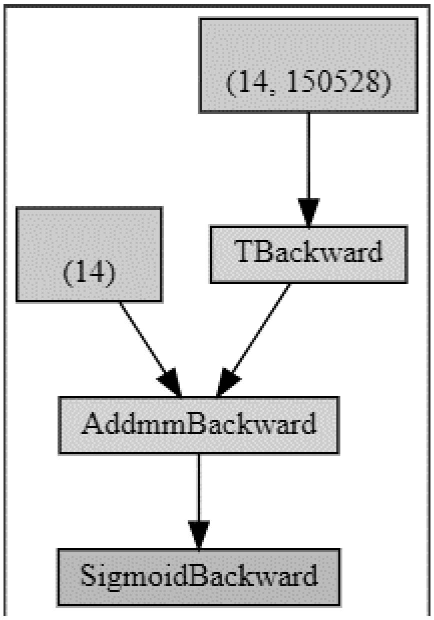

Figure 5: Shows the single layer perceptron architecture

IV. Multi-Layer Perceptron

Multi-Layer Perceptron's are very similar to SLP's, but these kinds of Perceptron's consist of more than one linear layer. If given example for three-layer networks, first layer will be the input layer, last layer will be the output layer and the middle layer would be called the hidden layer. [5] Layers after the input layer are called hidden layers because that are not directly exposed to the input. The simplest network structure is to have a single neuron in the hidden layer that directly outputs the value. The hidden layer could be increased according to the complexity of the problem. The data is fed through the input layer and the output is taken from the output layer. MLP's use a technique called backpropagation for training. Also, MLP's distinguish themselves from SLP's because of their multiple linearity.

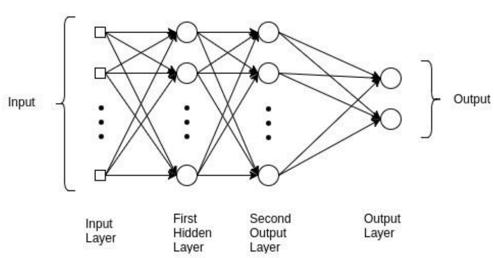

Figure 6: [9] Shows the generalized architecture of a multi-layer perceptron

The advantage of multi-layer perceptron's comes from their ability to map the features into higher dimensional spaces and they try to separate problems in those higher dimensions. We have tried 1, 2, 3 different hidden layers with 3 different hidden state neuron size for each hidden layer size. The results were rather similar to our single layer perceptron model.

Disease	AUROC
AUROC mean	0.60
no finding	0.67
cardio mediastinum	0.51
Cardiomegaly	0.57
Lung opacity	0.61
Lung lesion	0.53
Edema	0.67
Consolidation	0.53
Pneumonia	0.60
Atelectasis	0.59
Pneumothorax	056
Pleural Effusion	065
Pleural other	0.73
Fracture	0.61
Support Devices	0.57

Table 3: Shows the AUROC results of the multi-layer perceptron

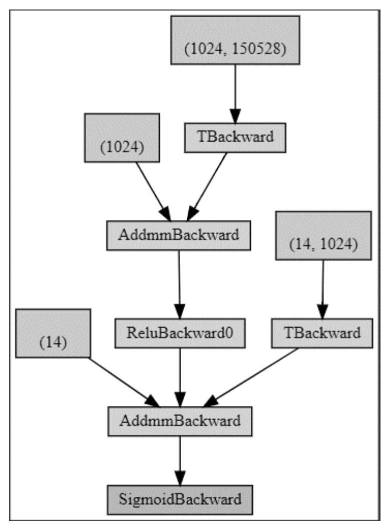

Figure 7: Best performing multi-layer perceptron architecture

V. Convolutional Neural Networks

A Convolutional Neural Network is a class of deep neural networks and it is mostly used to analyzing visual imagery. [7] It takes an input image, assign the importance (weights and biases) to various objects in the image and it can differentiate different images from one another. The preprocessing aspect of the Convolutional Network is much lower compared to other classification algorithms.

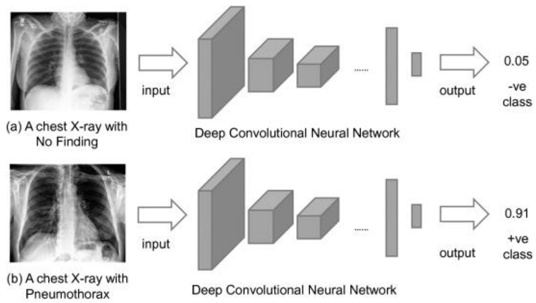

Figure 8: [10] An example illustration of the classification process

We have used 4 different CNN approaches to classify the diseases. We prepared 3 models ourselves. After seeing the complexity of the problem we have decided that we need more advanced architectures. We have opted to go with densenet121. Which is a well-known model in the field of computer vision.

Disease	AUROC
AUROC mean	0.63
no finding	0.77
cardio mediastinum	0.53
Cardiomegaly	0.59
Lung opacity	0.66
Lung lesion	0.53
Edema	0.76
Consolidation	0.56
Pneumonia	0.59
Atelectasis	0.63
Pneumothorax	0.46
Pleural Effusion	0.68
Pleural other	0.76
Fracture	0.50
Support Devices	0.75

Figure 9: Shows the AUROC results of the densenet121 model

VI. Results

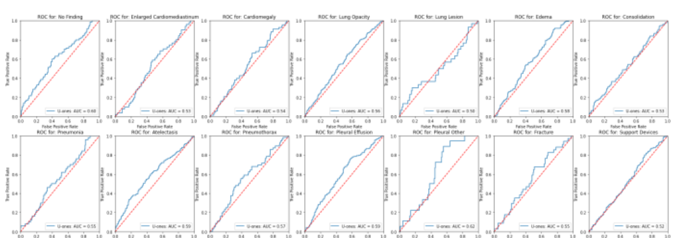

Figure 10: AUROC scores of each individual diseases in single layer perceptron

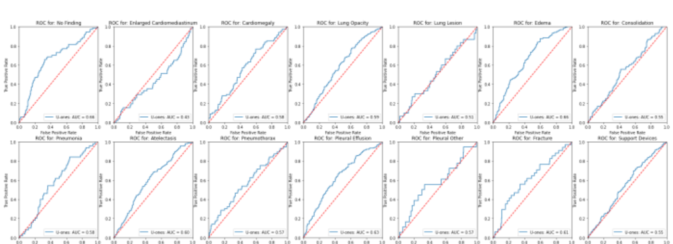

Figure 11: AUROC scores of each individual diseases in the best performing multi-layer perceptron

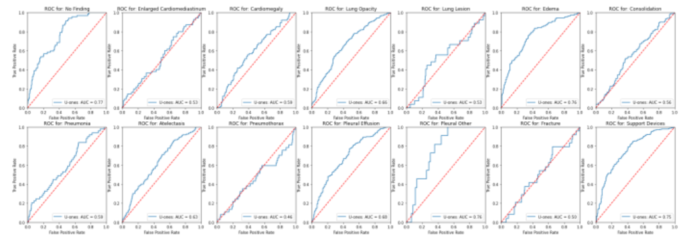

Figure 12: AUROC scores of each individual diseases in densenet121 architecture

Model	# of hidden layers	# of hidden neurons	# of convolution layers	AUROC mean
Single layer perceptron	0	0	0	0.559
Multi layer perceptron	1	100	0	0.576
Multi layer perceptron	1	150	0	0.581
Multi layer perceptron	1	1024	0	0.598
Multi layer perceptron	2	(300, 200)	0	0.569
Multi layer perceptron	3	(600, 400, 200)	0	0.531
Multi layer perceptron	3	(900, 450, 300)	0	0.521
Multi layer Perceptron	3	(1200, 900, 600)	0	0.535
Simple CNN	3	(in_size, 122, 52)	2	0.472
CNN	3	(in_size, 1000, 600)	2	0.543
CNN v2	3	(in_size, 1000, 600)	3	0.546
DenseNet121	-	-	116	0.625

Figure 13: Shows the mean AUROC scores of each tested algorithm

VII. Conclusion

We present different Artificial Neural Network implementations on the CheXpert dataset. Single Layer and Multi-Layer Perceptron's, Convolutional Neural Networks are used as models to predict the 14 different diseases that were labeled in the dataset. As it can be seen from the accuracy values, the Convolutional Neural Networks gave better results and accuracies compared to other Neural Networks. We hope that our comparison between these models show the significance and difference of each Neural Network. From the results obtained using numerous ANN`s approaches we can determine that classifying chest x-rays is a challenging task even with the usage of known architectures such as densenet121. The proposed methods on the dataset website offers us with viable solutions with the usage of ensemble convolutional neural networks. If we had more time designing our models we would consider this option to see the results. The lower scores we have received could also be attributed to down sampled version of the dataset. But using original data was not feasible without the usage of more advanced computers. Therefore, we have undesirable results. This project gave us an intuition about challenging datasets such as CheXpert. This dataset was much more different than other datasets we have used in class such as MNIST.

REFERENCES

[1] L. Delrue, R. Gosselin, B. Ilsen, A. Van Landeghem, J. de Mey, P. Duyck, Difficulties in the interpretation of chest radiography, in: Comparative Interpretation of CT and Standard Radiography of the Chest, Springer, 2011, pp. 27–49. doi: https://doi.org/10.1007/978-3-540-79942-9_2.

[2] "PyTorch," Wikipedia, 21-Dec-2020. [Online]. Available: https://en.wikipedia.org/wiki/PyTorch. [Accessed:20-Jan-2021].

[3] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," arXiv.org, 21-Jan-2019. [Online]. Available: https://arxiv.org/abs/1901.07031 . [Accessed: 20-Jan-2021].

[4] ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING. Place of publication not identified: MIT Press, 2020.

[5] J. Brownlee, "Crash Course On Multi-Layer Perceptron Neural Networks," Machine Learning Mastery, 14-Aug-2020. [Online].Available: https://machinelearningmastery.com/neural-networks-crash-course/ . [Accessed: 20-Jan-2021].

[6] "Perceptron," Wikipedia, 21-Dec-2020. [Online]. Available: https://en.wikipedia.org/wiki/Perceptron. [Accessed:20-Jan-2021].

[7] "A Comprehensive Guide to Convolutional Neural Networks- the ELI5 Way," [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [Accessed:19-Jan-2021].

[8] "Tensorflow-Single Layer Perceptron," TutorialsPoint [Online]. Available:https://www.tutorialspoint.com/tensorflow/tensorflow_single_layer_perceptron.htm [Accessed:19-Jan-2021].

[9] "A hypothetical example of Multilayer Perceptron Network," ResearchGate [Online]. Available:https://www.researchgate.net/figure/A-hypothetical-example-of-Multilayer-Perceptron-Network_fig4_303875065 [Accessed:19-Jan-2021].

[10] "tCheXNet Detecting Pneumothorax on Chest X-Ray Images Using Deep Transfer Learning" [Online]. Available:https://link.springer.com/chapter/10.1007/978-3-030-27272-2_28 [Accessed:19-Jan-2021].