Predicting Pulmonary Edema with Convolutional Neural Network Trained with Biomarker NT-pro B-type Natriuretic Peptide

Abstract:

One of the main challenges researchers face when developing convolutional neural networks (CNNs) for classifying medical images is label acquisition. Manual labels created by physicians can be time-consuming and difficult to obtain. Our paper of interest, *Deep Learning Radiographic assessment of pulmonary edema: Optimizing Clinical Performance, training with serum biomarkers* [1] tackled the problem using patient blood serum biomarkers B-type natriuretic peptide (BNP) and NT-pro B-type natriuretic peptide (BNPP) as continuous labels to train ResNet models.

We reproduced most of the results from Huynh et al., with only BNPP measurements and lower image resolution ranges. We used a similar set of metrics (MAE, F1 score, AUC, and Pearson r) to compare our model performance with the paper of interest. We replicated Fig. 4, Fig. 5, Fig. 6, and Fig. 11 from the original paper. We also experimented with VGG 16 and VGG 19 to compare the performance with ResNet 152, trained on image resolution 64 x 64 and 224 x 224. Among our models, VGG 19 with image resolution 64 x 64 performed the best.

Introduction:

Deep learning algorithms, especially convolutional neural networks (CNNs), have been widely used in medical image interpretation. In recent years, multiple research groups have shown the applicability of CNNs in pulmonary disease detection via chest radiograph interpretation and classification [2]-[4]. Although these researches have shown promising results, their models are all trained on manually annotated chest radiographs, which could make the data collection time-consuming, labor-intensive, or even not perfectly reliable for certain diseases like pulmonary edema [5]. In order to overcome this obstacle, our paper of interest,

Deep Learning Radiographic assessment of pulmonary edema: Optimizing Clinical Performance, training with serum biomarkers [1] employed NT-pro B-type natriuretic peptide (BNPP) and B-type natriuretic peptide (BNP) values collected from laboratory blood serum tests as the label of each patient's chest radiograph. Our goal is to largely replicate the result that Huynh et al. obtained, but due to the lack of BNP values, our model is only trained on BNPP values rather than both. Additionally, due to hardware limitations, simpler models like VGG 16 and VGG 19 are used in addition to selected complex models like ResNet 50 and ResNet 152 used by Huynh et al. Our models are also trained only on the lower resolution images (64x64, 128x128, 256x256, 512x512).

Using blood serum biomarkers has been a standard procedure in clinical diagnosis for over 70 years [6]. BNPP is a common serum cardiac biomarker, which is indicative of acute heart failure. Blood serum biomarkers could be collected concurrently or retrospectively, and thus are more likely to provide sufficient datasets for model training compared to manually labeled radiographs. The BNPP samples used to train the original model are collected alongside the radiographs with a mean absolute difference of 2 hours, and 22 minutes.

Pulmonary edema is a condition caused by an increase in the extravascular water content of the lungs when the rate of fluid filtration exceeds the rate of lymphatic removal. A common type of pulmonary edema is cardiogenic pulmonary edema, which is the result of elevated pulmonary capillary pressure from left-sided heart failure [7]. As BNPP is indicative of acute heart failure, it is then also indicative of cardiogenic pulmonary edema [5][8]. Therefore, using BNPP as the label of chest radiographs is a reasonable and more efficient substitute for manually labeled radiographs for training a CNN model that could help with clinical pulmonary edema detection.

Being the major diagnostic technique for pulmonary diseases, the chest radiograph is the most common input data for CNN model training in the domain. However, in most published literature, training data has been down-sampled from original images [9]-[11]. In real-world

disease diagnosis situations, physicians would focus on specific features like the existence of alveolar flooding for severe cases, or interstitial Kerley B lines and peribronchial cuffing for mild cases. These features could be very subtle on the radiographs and require high resolution for identification. Therefore we evaluated the performances of models trained on radiographs with resolutions varying from 64x64 to 512x512 to test the effect of resolution on model performance.

The loss function used in the training process is a customized MAE function. Moreover, in order to be more clinically relevant, our models are also evaluated on Pearson r values, sensitivity, specificity, and AUC scores with different cut-off thresholds. For the purposes of replicating the result obtained by Huynh et al., we will use the same metrics and aim to obtain similar results as Fig. 4, Fig. 5, Fig 6, and part of Fig. 11 and Fig. 12. However, due to the various limitations aforementioned, we expect our models to have lower performance than our paper of interest. In addition to the ResNet models that Huynh et al. employed, we are interested in testing how VGG models perform compared to ResNet based on the same set of metrics.

Methods:

VGG Models

VGG stands for Visual Geometry Group, and it is a popular deep Convolutional Neural Network architecture first introduced in 2014 for image recognition and classification. We used two versions of VGG: VGG16 and VGG19. Their structures are relatively similar: VGG16 has 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers, and VGG19 has 16 convolution layers, 3 fully connected layers, 5 max-pooling layers, and 1 Softmax layer. We trained both models with image dataset with 64 x 64 and 224 x 224 resolutions, with Adam optimizer, 20 epochs, batch size being 8, and learning rate being 0.00005. All the layers were unfrozen to tune for best parameters. The models were trained on 15,164 images for training set, 1,913 images for the validation set, and 1,823 images for the test set.

ResNet Models:

ResNet stands for Residual Neural Network, the first ultra-deep neural network with hundreds of layers. Similar to the authors of the publication that we are trying to replicate, we used ResNet 152 to train on our image data with 64 x 64 and 224 x 224 resolutions, with Adam optimizer, 20 epochs, batch size being 8, and learning rate 0.00005 for the lower resolution and 0.0005 for the higher resolution images. In contrast to VGG models, only the last 3 layers and 3 blocks of ResNet 152 were unfrozen, which was determined empirically. The dataset was exactly the same as VGG models.

Results:

As shown in Figure 1, among the 6 combinations of our models, the best one is VGG 19 on image with 64 x 64 resolution. It has the lowest Test MAE of 0.478, highest Pearson r of 0.695, highest AUC of 0.858, and highest F1 score of 0.707. It is still slightly worse than Justin's models from his publications, but the statistics are very similar.

	VGG16, 64	VGG16, 224	**VGG19, 64**	VGG19, 224	ResNet152, 64	ResNet 152, 224	ResNet 152, 256 (Public-ation)
Test MAE	0.480	0.502	**0.478**	0.493	0.822	0.512	/
Pearson r	0.687	0.677	**0.695**	0.686	0.085	0.668	0.699
AUC	0.854	0.844	**0.858**	0.852	0.792	0.843	0.872
F1 score	0.706	0.692	**0.707**	0.701	0.633	0.681	0.721

Figure 1: Overall Performance of Different Models with Different Resolutions

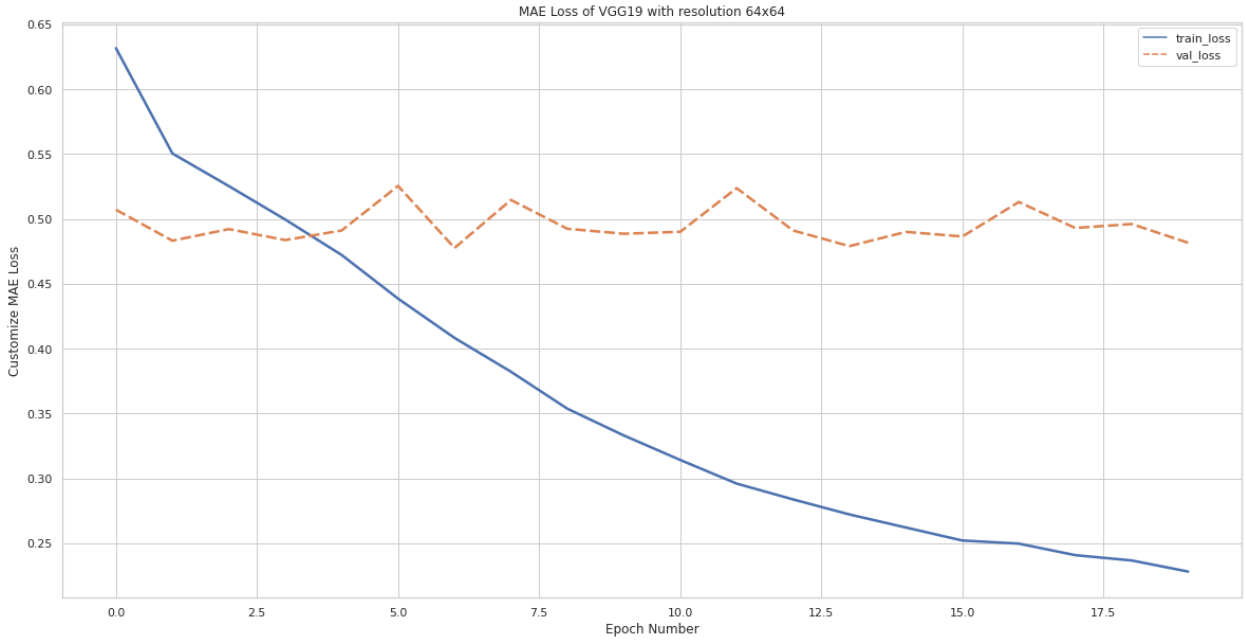

Figure 2: Training and Validation Loss of the best model, VGG 19 with 64 x 64 resolution.

Figure 2 shows the training and validation loss from the 20 epochs training the VGG 19 model with 64x 64 resolution images. The exponential decay of the training loss shows that the model is learning, and the fluctuation of the validation loss without drastically increasing shows that the model is not overfitting.

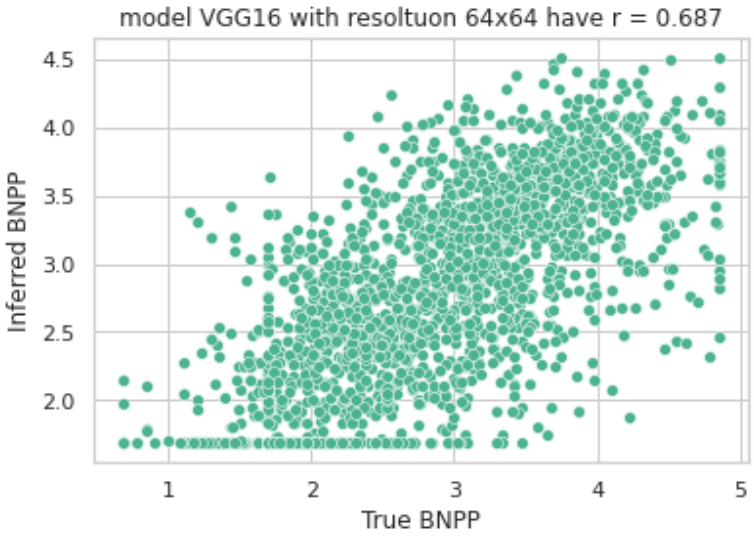

Figure 3: Pearson correlation graph for the best model, VGG 19, trained on 64 x 64 resolution

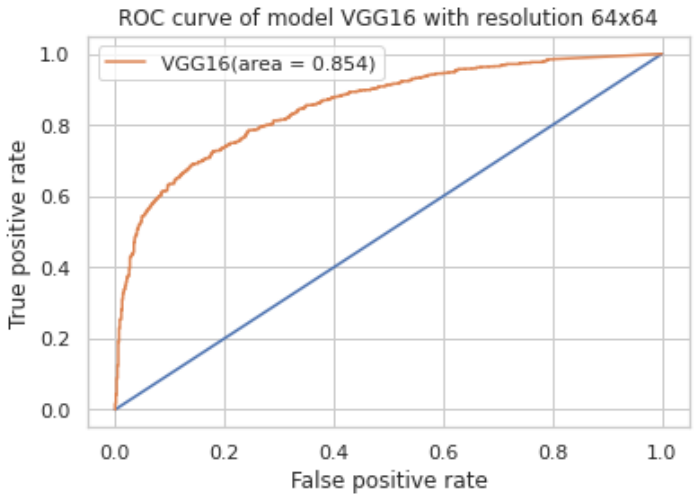

Figure 4: ROC curve for the best model, VGG 19, trained on 64 x 64 resolution

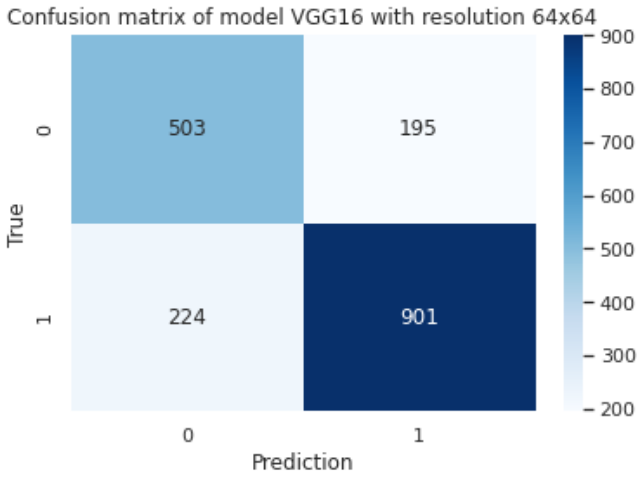

Figure 5: Confusion matrix for the best model, VGG 19, trained on 64 x 64 resolution

Figure 3-5 shows the plots we reproduced from the best model, VGG 19 trained on 64 x 64 resolution images. They are the scatter plot with Pearson correlation between predicted and true BNPP, the ROC curve, and the confusion matrix respectively, based on the best model.

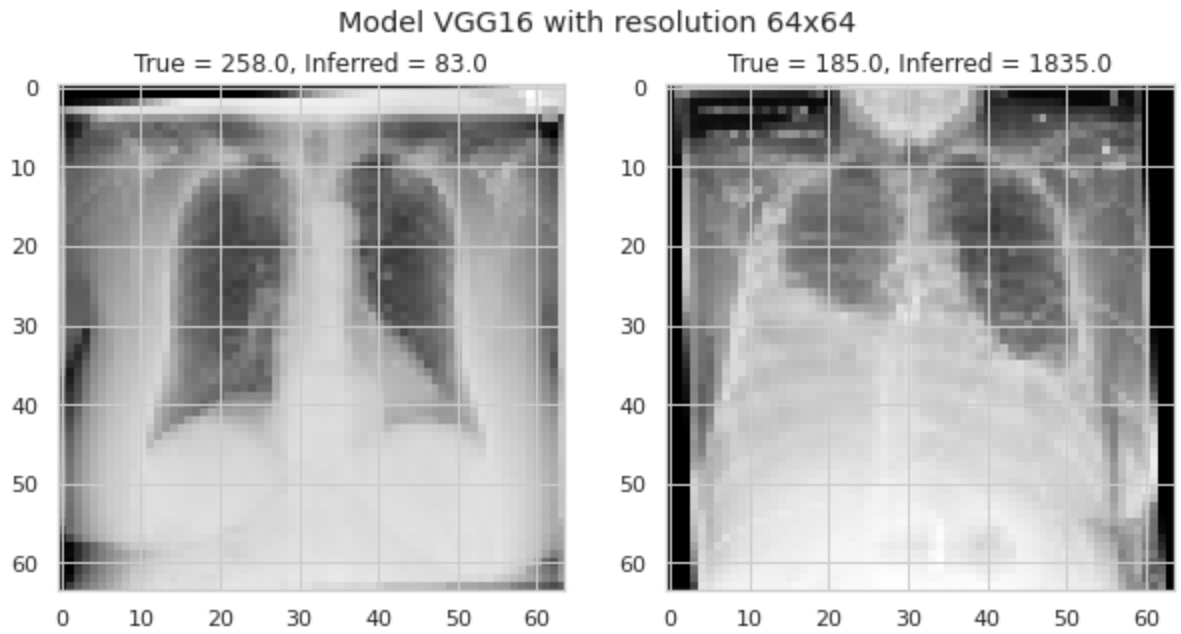

Figure 6: Sample input image data with true and inferred BNPP values

Figure 6 above shows two examples of input image files with their true and inferred BNPP values labeled. The true BNPP for the left image is 258.0, and the inferred BNPP is 83.0. For the right image, the true BNPP is 185.0, and the inferred one is 1835.0.

Discussion:

Due to memory constraints, we were only able to train models with image resolutions 64 x 64 and 224 x 224, compared to the previous result from the publication with resolutions up to 1024 x 1024. Since Justin, the author of the paper we were trying to replicate, did not test the VGG models, we could not compare our result with his. However, with the same model, ResNet 152, with similar resolutions, we were able to achieve similar performance results as the previous publication. All five models, exclusing the ResNet 152 model trained on 64 x 64 resolution, had fluctuations in their performance, and their differences in performance metrics was insignificant when taken into consideration their variations during training and validation. Therefore their performances were relatively similar. The ResNet 152 model trained on the lower resolution performed significantly worse, possibly due to overfitting a too complex model onto images with resolutions too low.

In the future, it would be ideal to use class activation visualizations like Grad-CAM to see which parts of the input images are influencing the model performance the most. In this way, we would be able to visualize and differentiate which models are truly learning on the useful parts of the radiographs and which models are focusing on extraneous parts.

Citations:

[1] Huynh, J., Masoudi, S., Noorbakhsh, A., Mahmoodi, A., Kligerman, S., Yen, A., Jacobs, K., Hahn, L., Hasenstab, K., Pazzani, M., & Hsiao, A. (2022). Deep Learning Radiographic assessment of pulmonary edema: Optimizing Clinical Performance, training with serum biomarkers. *IEEE Access*, *10*, 48577–48588. https://doi.org/10.1109/access.2022.3172706

[2] P. Lakhani and B. Sundaram, ‘‘Deep learning at chest radiography: Auto- mated classification of pulmonary tuberculosis by using convolutional neural networks,’’ *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.

[3] L. Wang, Z. Q. Lin, and A. Wong, ‘‘COVID-Net: A tailored deep con- volutional neural network design for detection of COVID-19 cases from chest X-ray images,’’ *Sci. Rep.*, vol. 10, no. 1, pp.1–12, Dec. 2020, doi: 10.1038/s41598-020-76550-z.

[4] E. J. Hwang, J. G. Nam, W. H. Lim, S. J. Park, Y. S. Jeong, J. H. Kang, E. K. Hong, T. M. Kim, J. M. Goo, S. Park, K. H. Kim, and C. M. Park, ‘‘Deep learning for chest radiograph diagnosis in the emer- gency department,’’ *Radiology*, vol. 293, no. 3, pp. 573–580, Dec. 2019, doi: 10.1148/radiol.2019191225.

[5] G.E.Duggan,J.J.Reicher,Y.Liu,D.Tse,andS.Shetty,‘‘Improvingref- erence standards for validation of AI-based radiography,’’ *Brit. J. Radiol.*, vol. 94, no. 1123, Jul. 2021, Art. no. 20210435, doi: 10.1259/bjr.20210435.

[6] H. A. Krebs, ‘‘Chemical composition of blood plasma and serum,’’ *Annu. Rev. Biochem.*, vol. 19, no. 1, pp. 409–430, Jun. 1950, doi: 10.1146/annurev.bi.19.070150.002205.

[7] J.F.Murray,‘‘Pulmonary edema: Pathophysiology and diagnosis,’’*Int.J. Tuberculosis Lung Disease*, vol. 15, no. 2, pp. 155–160, 2011.

[8] P. Ray, M. Arthaud, S. Birolleau, R. Isnard, Y. Lefort, J. Boddaert, and B. Riou, ‘‘Comparison of brain natriuretic peptide and probrain natriuretic peptide in the diagnosis of cardiogenic pulmonary edema in patients aged 65 and older,’’ *J. Amer. Geriatrics Soc.*, vol. 53, no. 4, pp. 643–648, Apr. 2005, doi: 10.1111/j.1532-5415.2005.53213.x.

[9] I. Pan, A. Cadrin-Chênevert, and P. M. Cheng, ‘‘Tackling the radi- ological society of North America pneumonia detection challenge,’’ *Amer. J. Roentgenol.*, vol. 213, no. 3, pp. 568–574, Sep. 2019, doi: 10.2214/AJR.19.21512.

[10] S.Jaeger,S.Candemir,S.Antani,Y.-X.J.Wáng,P.-X.Lu,andG.Thoma, ‘‘Two public chest X-ray datasets for computer-aided screening of pul- monary diseases,’’ *Quantum Imag. Med. Surg.*, vol. 4, no. 6, pp. 475–477, Dec. 2014, doi: 10.3978/j.issn.2223-4292.2014.11.20.

[11] J. C. Y. Seah, J. S. N. Tang, A. Kitchen, F. Gaillard, and A. F. Dixon, ‘‘Chest radiographs in congestive heart failure: Visualizing neural net- work learning,’’ *Radiology*, vol. 290, no. 2, pp. 514–522, Feb. 2019, doi: 10.1148/radiol.2018180887.

Responsibility in Group Project:

For the coding part, I have been implementing and testing the VGG 16, VGG 19, and ResNet 152 models with Angela Wang. We collaborated to preprocess the image data and optimized the models together.

For the code part, I am responsible for debugging the run.py file.

For the written part, we first discussed the main goals of our project. After coming to an agreement, Angela drafted out the report, which was later edited by me.