DSC Capstone Quarter 1 Project

Joshua Lee
jdlee@ucsd.edu

Zoya Hasan
zohasan@ucsd.edu

Brian Huynh
bth001@ucsd.edu

Albert Hsiao
a3hsiao@health.ucsd.edu

Abstract

Pulmonary edema contributes to more than 15 million hospitalizations worldwide each year, and a prompt diagnosis is crucial to preventing respiratory failure. Accurate assessments of chest radiographs are needed, but tend to be restricted by subjective human interpretation, making it a significant difficulty to obtain reliable radiographic labels for pulmonary edema. The purpose of this study is to reproduce and evaluate the use of serum biomarkers (BNP/BNPP) as continuous labels for the assessment of convolutional neural networks (CNN) based pulmonary edema. The data source was provided by UC San Diego and is deprived of 27,748 frontal chest radiographs of 16,401 patients (2017-2020) paired with laboratory results from BNP or BNPP. To carry out the task of creating a regression model that outputs BNPP values, the dataset was split into a train, validation, and test set. All images were downscaled to 256×256 from 1024×1024 due to time and computing resources. Both VGG16 and ResNet50 architecture pretrained on ImageNet were fine tuned for regression using mean absolute error to predict BNPP values, and the model was evaluated using Pearson correlation between predicted and actual BNPP values. The main findings of the original article were that increasing the resolution of the image improves AUC and lung attention. This approach may reduce the dependence on manual radiologist labels while improving the interpretability of the model.

Code: https://github.com/brianthuynh10/dsc180-capstone-recreate

1 Introduction

1.1 Intro Paragraph

Pulmonary edema, a condition where fluid accumulates in the lungs, causes people to have difficulty breathing and can quickly become life-threatening. Although radiographic imaging is normally used for monitoring and diagnosis, accurate interpretation remains a challenge, especially when differentiating mild from severe cases where visual differences are subtle. In recent years, deep learning has been proving it's role in medical imaging by offering opportunities to improve diagnostic efficiency and consistency. Despite this, these models require reliable labeled data, which are time-consuming and costly to obtain. To address this challenge of label scarcity and improve diagnostic accuracy, Huynh et al. (2022) explored using serum biomarkers, specifically B-type natriuretic peptide (BNP) and NT-proBNP, as objective and continuous training labels for convolutional neural networks (CNNs) assessing pulmonary edema on chest radiographs.

1.2 Literature Review

Traditional methods include manual radiologist labeling, which is time-consuming and inconsistent, and alternative weak supervision such as semi/self-supervised learning or Natural Language Processing (NLP) derived labels. Early applications of CNN include its use for pneumonia and pneumothorax classification. Current limits of deep learning in radiology include coarse labels as well as the lack of severity quantification. The novelty of the contribution of Huynh et al. was their use of BNP/BNPP biomarkers as labels instead of the typical method of using text/manual annotations. Huynh et al. also explored the impact of resolution and explainability metrics (Grad-CAM, Saliency, XRAI) on deep learning in radiology. The key finding that they found was that higher resolution results in better lung attention and slightly better AUC. For clinical implications, this means improved interpretability and diagnostic objectivity for cardiogenic pulmonary edema.

1.3 Relevant Data

The data used in this study is from a UC San Diego institutional dataset of 27,748 frontal chest radiographs from 16,401 patients (2017-2020) paired with BNP or BNPP lab results. The variables in this study include the radiograph (input image), the biomarker value (BNP or BNPP, pg/mL), as well as the metadata (patient ID, date/time). The splits used in this study are 80% for the training set, 10% for the validation set, and 10% for the test set. The labeling rationale for this study was that BNP values over 100 and BNPP values over 400 indicate acute heart failure, which was used as the thresholds for binary classification. In our replication, we reduced the image resolutions to 256x256 and used log transformation for loss stabilization. For the evaluation metrics, we used Pearson r and mean absolute error (MAE).

2 Methods

2.1 Data Sources and Cohort Construction

We used a UC San Diego institutional dataset containing 27,748 frontal chest radiographs from 16,401 patients collected between 2017–2020, each paired with a BNP or NT-proBNP (BNPP) laboratory measurement obtained within a clinically relevant window. Metadata included unique study identifiers, patient age, and timestamp information. Following the original study design, we applied an 80/10/10 patient-level split to create training, validation, and test sets, ensuring no patient overlap across partitions.

All radiographs were stored in 10 HDF5 files ($\sim$100 GB). We enumerated image keys across all files and retained only images whose unique identifiers appeared in the provided train/validation/test CSVs. This resulted in final sets of 2,634 training images, 329 validation images, and 331 test images.

2.2 Image Preprocessing

To ensure computational feasibility, all radiographs were downsampled from their native 1024×1024 resolution to 256×256 using mean pooling across non-overlapping 4×4 blocks. Images were then scaled to the $[0, 1]$ range and converted from single-channel grayscale to three-channel format to match ImageNet-pretrained CNN architectures.

During training, we applied light data augmentation consisting of random rotations up to 10° to increase robustness to minor pose variation. All images were resized to 224×224 and normalized using ImageNet channel statistics, consistent with standard transfer-learning practice in radiology imaging tasks.

2.3 Biomarker Label Processing

BNPP values were log-transformed to stabilize the regression objective. Labels were then standardized using the mean and standard deviation of the training set only. The same transformation parameters were applied to the validation and test labels to avoid data leakage.

2.4 Model Architectures

We evaluated four VGG architectures—VGG11, VGG13, VGG16, and VGG19—to assess whether model depth influenced the learned radiograph–BNPP relationship. All networks were initialized with ImageNet pretrained weights.

For regression, we replaced the original classification head with a custom multilayer perceptron consisting of:

- Linear($25088 \rightarrow 1024$), ReLU, Dropout(0.3)
- Linear($1024 \rightarrow 256$), ReLU
- Linear($256 \rightarrow 1$)

Following the original study, all convolutional feature extractor layers were frozen during training, leaving only the regression head trainable.

2.5 Training Procedure

Training was performed in PyTorch with GPU acceleration when available. For each model, we used:

- Optimizer: Adam
- Learning rate: 1×10^{-5}
- Loss: composite of L1 loss (MAE) and a correlation-based loss
- Epochs: 50
- Batch size: varied (8, 16, 32, 64, 128)

The correlation component of the loss was formulated as:

$$\mathscr{L}_{\text{corr}} = 1 - r,$$

where r is the Pearson correlation between predicted and true labels within a batch. This encourages correct rank-ordering of biomarker severity, while MAE ensures calibration.

All experiments were logged using Weights & Biases (W&B), including batch losses, epoch metrics, and scatter plots of predicted vs. true BNPP values.

2.6 Experimental Design

Architectural Comparison Experiment

To evaluate model depth, each VGG architecture was trained on the same half-split of the training set. All models used identical seeds to ensure consistent sampling, augmentation, and dataloader ordering. Validation and test sets were kept constant across experiments.

Each architecture was trained across five random seeds to assess performance stability. The primary evaluation metric was Pearson's r between predicted and true BNPP values.

Batch Size Experiment

To isolate the effect of batch size, we trained the VGG16 model on the full training set with batch sizes of 8, 16, 32, 64, and 128. All other hyperparameters were held constant. For each batch size, five independent runs were conducted using different random seeds.

We evaluated:

- Predictive performance (Pearson's r)
- Mean Absolute Error (MAE)
- GPU training time per epoch

This experiment quantified computational trade-offs associated with varying batch sizes.

2.7 Evaluation Metrics

Performance was assessed using:

- **Pearson's correlation coefficient (r)** — primary metric for strength of the radiographic–biomarker relationship.
- **Mean Absolute Error (MAE)** — secondary metric for calibration accuracy.
- **Scatter plots of predicted vs. true BNPP** — qualitative evaluation of model behavior across epochs.
- **Training time** — used only in the batch size analysis.

3 Results

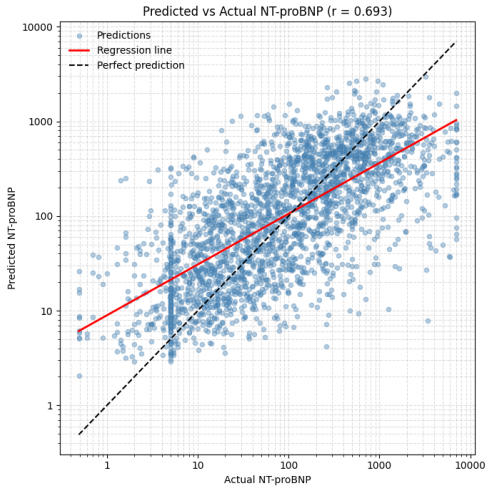
3.1 Examining Model Depth

In Figure 1, we show the relationship between the measured laboratory BNPP values and the BNPP values inferred by each VGG model for a single random seed. Each model was trained on the same half-split of the original training set, selected using an identical seed to ensure consistent sampling across models. The validation and test sets remained unchanged, and all models were trained using 256×256 input images. Among the four architectures, VGG19 produced the strongest relationship between inferred and measured BNPP values (r = 0.702), while VGG13 showed the weakest relationship (r = 0.687).

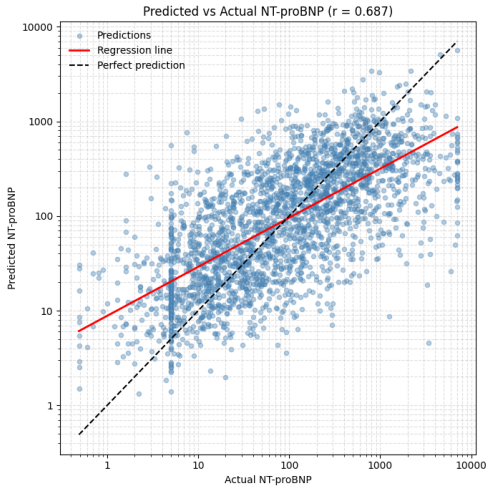
When repeating the experiment across multiple random seeds for all four VGG architectures, a clear trend emerges: deeper VGG models consistently produce stronger relationships between inferred and measured BNPP values. As shown in Figure 2, there is a noticeable increase in Pearson's rrr from VGG11 to VGG13 and then to VGG16. The improvement from VGG16 to VGG19 is smaller, suggesting that the performance gains may be approaching a plateau within the VGG family. To achieve further improvement, a deeper or more expressive architecture, such as ResNet, may be necessary to capture relationships that the VGG models are unable to represent effectively.

3.2 Evaluating Effect of Batch Size

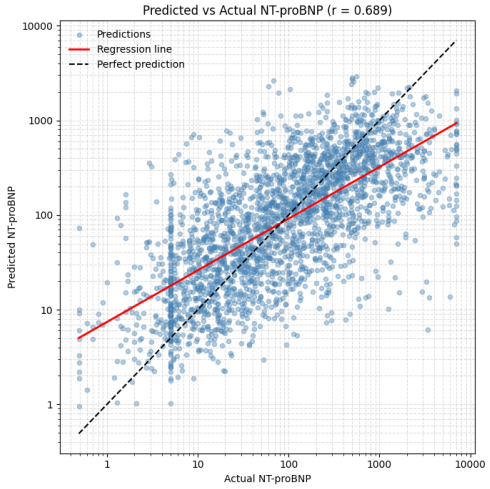
Training a VGG16 model on the full dataset revealed only a weak relationship between batch size and model performance. Extremely large (batch size 128) and very small (batch



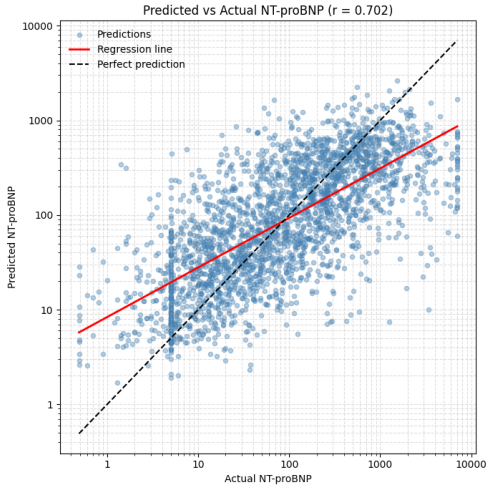
(a) VGG11



(b) VGG13



(c) VGG16



(d) VGG19

Figure 1: Test set scatter plots for VGG11, VGG13, VGG16, and VGG19 trained on half training data

size 8) batches showed slightly lower average Pearson-r values compared to moderate batch sizes (16, 32, 64). However, across five random seeds, the variability in Pearson-r was large for all tested batch sizes. This finding indicates that these differences are not entirely meaningful.

In contrast, training time exhibited a strong dependence on batch size (Figure 3): larger batch sizes led to significantly faster training, with the largest improvements occurring between the smaller batch sizes. Considering both predictive performance and computational efficiency, a batch size of 32 represents a balanced choice, offering stable Pearson-r values while significantly reducing training time relative to smaller batches.

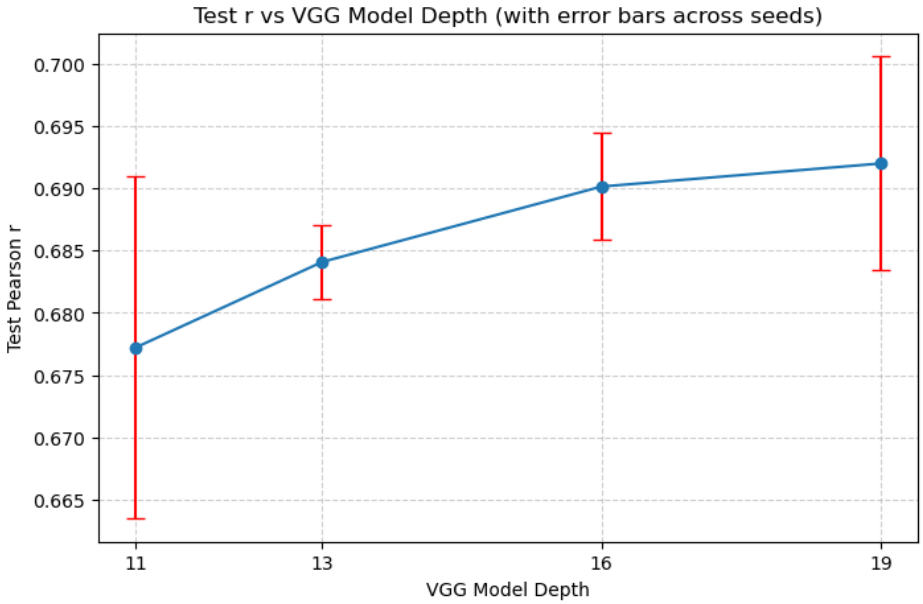
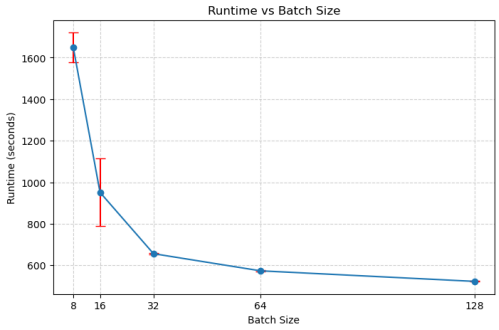


Figure 2: Line graph showing the trend and variability of Pearson's r between inferred and measured BNPP values across different VGG model depths.

(a) Minimal difference in captured Pearson-r across different batch sizes



(b) As batch size increases, runtime decreases with greater drops in training time at smaller batch sizes

Figure 3: Runtime and Pearson-r graphs when training with different batch sizes

4 Discussion

Our results support the central finding of Huynh et al. (2022): serum biomarkers such as BNPP provide a meaningful, objective signal that CNNs can learn from chest radiographs to estimate pulmonary edema severity. The stronger correlations achieved by deeper VGG architectures indicate that model capacity plays an important role in capturing the subtle visual patterns linked to BNPP, reinforcing the value of continuous biomarker labels over subjective ordinal grades.

The batch size experiments showed that predictive performance remained relatively stable across configurations, while training time decreased substantially with larger batches. This highlights a practical takeaway: once hyperparameters fall within a reasonable range, computational efficiency may guide model-building decisions more than marginal changes in accuracy.

Overall, reproducing and extending the original study allowed us to develop a clearer understanding of the full neural-network pipeline—from data preprocessing and architectural selection to training, evaluation, and interpretation. By aligning our findings with the original paper, we demonstrate both the robustness of biomarker-driven supervision and our ability to carry out a complete deep learning workflow for a clinically relevant radiology task.

5 Conclusion

In this project, we implemented and developed an entire methodological framework for reproducing and extending the findings of Huynh et al. (2022) on using serum biomarkers (BNP/BNPP) as continuous labels for pulmonary-edema assessment from chest radiographs. The goal for this project was to assess whether deep convolutional neural networks can learn clinically meaningful relationships between radiographic features and biomarker values, while also taking a look at how architectural depth and training configurations influence the performance of the model. In Quarter 1, we focused on building a reproducible data pipeline, defining our model approach, and conducting controlled experiments that showcase the usability of BNPP-based regression.

Across our methodological decisions, we used a UCSD institutional dataset of 27,748 radiographs paired with BNP/BNPP values, applied standardized preprocessing including downsampling and log-transforming labels, and evaluated several ImageNet-pretrained VGG architectures. Our experiments show a consistent trend: deeper VGG models produced stronger correlations between predicted and measured BNPP values, with VGG19 achieving the highest Pearson r on the test set. Our batch-size analysis further demonstrated that predictive performance remains relatively stable across reasonable configurations, while training time decreased substantially with larger batch sizes: an important practical consideration for scaling future experiments.

Although the results in our Quarter 1 project are preliminary, they uplift the central hypothesis that serum biomarkers provide an objective, learnable indication for pulmonary-edema severity. The patterns we observed in our assessment hint that increasing model capacity yields incremental improvements, and that computational trade-offs should be considered alongside predictive accuracy when performing experiments. Taking everything into account, these discoveries reinstate the promise of biomarker-supervised models as a complement to subjective radiologist labels.

The work we did for this project also comes with several limitations. The use of downsampled 256×256 radiographs likely constrained model expressiveness compared to high-resolution analyses in the original paper. The appearance of performance variability across

random seeds indicates the sensitivity of BNPP regression to initialization and data sampling. Also, the current pipeline freezes convolutional layers, which may inhibit our ability to fully see the nuanced features associated with pulmonary edema. These limitations encourage more extensive experimentation in the next phase in Quarter 2.

In Quarter 2, we will extend our current pipeline by including explainability mechanisms backed up by large language models (LLMs) to better interpret model predictions and explain insights to non-technical and clinical stakeholders. Building upon our existing regression framework, we intend to pair radiographic features and saliency outputs with LLM-generated narrative explanations that give context for BNPP predictions, showcase relevant imaging regions, and articulate model confidence in a clinically meaningful manner. This will include experimenting with multimodal LLMs and prompt engineering tactics to explain quantitative outputs (e.g., Pearson r, MAE, Grad-CAM maps) into fluid, stake-holder friendly summaries. By adding LLM-driven interpretability with traditional visualization tools, in Quarter 2 we will focus on producing explanations that support decision making, improve transparency, and make the model's behavior more available to researchers, physicians, and hospital administrators.

References

Huynh, Justin, Samira Masoudi, Abraham Noorbakhsh, Amin Mahmoodi, Seth Kligerman, Andrew Yen, Kathleen Jacobs, Lewis Hahn, Kyle Hasenstab, Michael Pazzani, and Albert Hsiao. 2022. "Deep Learning Radiographic Assessment of Pulmonary Edema: Optimizing Clinical Performance, Training With Serum Biomarkers." *IEEE Access* 10: 48577–48588. [Link]