Classification of Pulmonary Edema

Combining Physician Assessment and Machine Learning Inference Integrating Neural Networks with Large Language Models

AiDA lab

Dallas Plunkett dmplunkett@ucsd.edu
Jeru Paul Balares j1balares@ucsd.edu
Kendall Underwood keunderwood@ucsd.edu
Mentor: **Albert Hsiao** a3hsiao@health.ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

- Cardiogenic pulmonary edema is fluid in the lungs caused by high heart pressure.
- Many machine learning studies rely on expert labeling and validation, but expert review does not scale.
- Our primary question is can researchers use language from radiology reports to emulate expert interpretation?
- We assess those labels using NT-proBNP (BNPP), a blood measure of cardiac stress that informs edema evaluation.
- Our secondary question is can BNPP itself be estimated directly from chest X-rays?

Methods

- We fine-tuned MedGemma LLMs with LoRA to label edema from 2,390 radiology reports. About 23% of the training set were labeled as edema present.
- We trained ResNet CNNs to estimate BNPP from 21,374 chest X-ray images. Resolution was lowered to 256×256 pixels.

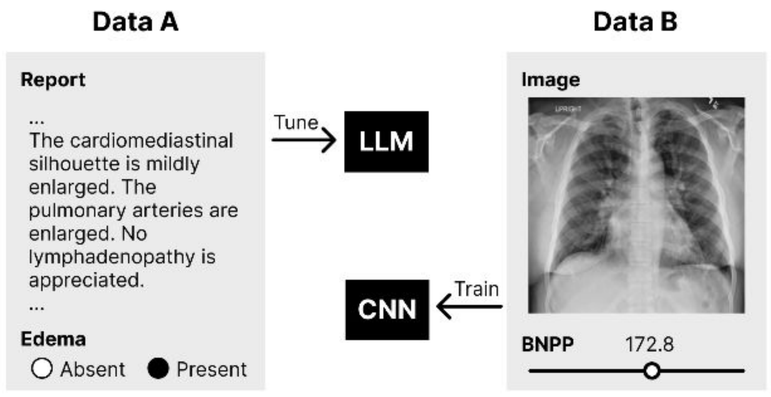

Figure 1. Data overview and model assignment.

- The CNN test set includes 2,602 cases with reports but no edema labels. We used the LLM to assign those labels to get our final dataset with predicted edema and BNPP.

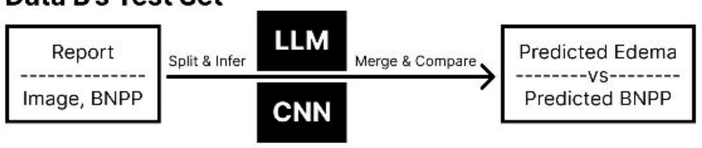

Figure 2. Study design.

Results

- A balanced 50% present and 50% absent tuning split performed best (AUC = 0.79) compared to the LLM without tuning, as well as other split strategies.

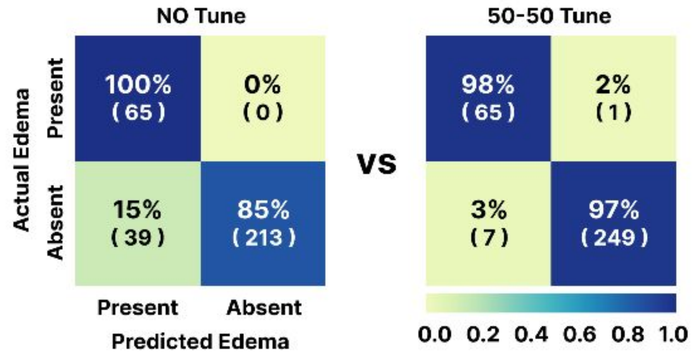

Figure 3. Confusion matrices comparing LLM results without tuning (left) and with balanced tuning (right). Percent values show how each actual edema label was distributed across predictions. Counts appear below in parenthesis.

- ResNet34 performed best among the architectures tested. Predicted and actual BNPP correlated at Pearson r = 0.705.
- The predicted distribution is slightly bimodal and broadly follows the shape of the actual BNPP distribution.

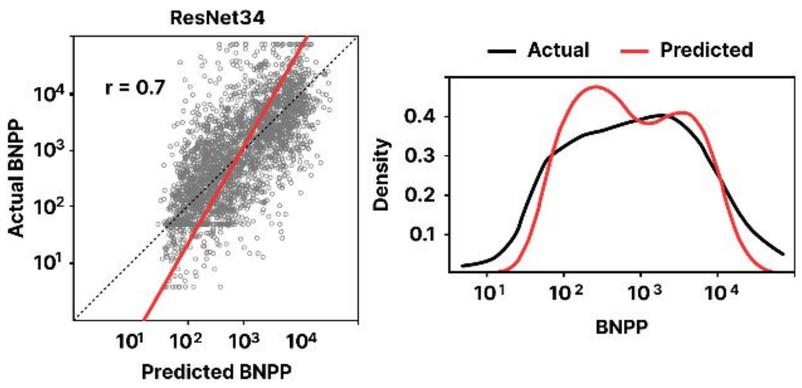

Figure 4. Predicted versus actual BNPP values (left) and their distributions (right).

- Reports labeled as present show higher BNPP values than those labeled absent. This pattern appears in both actual and predicted BNPP. Youden's J optimal threshold is 1027, suggesting the 400 standard is a conservative threshold.

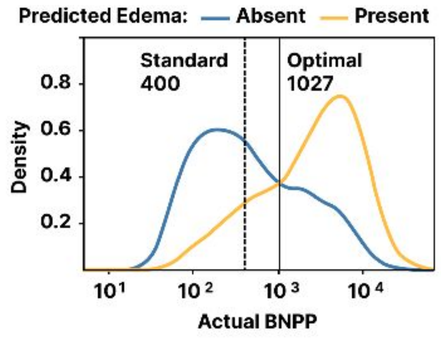

Figure 5. Actual BNPP distributions by predicted edema label.

- BNPP increases with age, so we examined separation within different age groups. Across each group, labels of present consistently showed higher BNPP compared to absent labels.

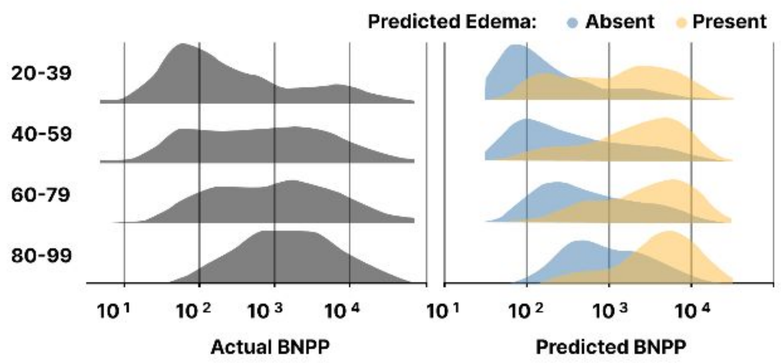

Figure 6. Age-stratified actual BNPP distributions compared to predicted BNPP and predicted edema.

Conclusion

- This study shows that a LLM can generate accurate edema labels from radiology reports align with heart stress measures. It also shows that heart stress can be estimated from X-ray images.
- By comparing edema labels with BNPP estimates, we demonstrate a practical way to evaluate edema label quality without requiring expert judgment.
- This approach offers a method for studying edema at scale while remaining grounded in physiology.