1. Title

Cross-Institutional Validation of Deep Learning for Multi-Disease Chest X-ray Classification Using MIMIC-CXR and CheXpert

2. Project Category

Application + Algorithmic — Applies advanced deep learning techniques to automated chest radiograph interpretation while introducing algorithmic innovations in cross-institutional domain adaptation, uncertainty-aware multi-label classification, and probability calibration for clinical deployment.

3. Team Details

Principal Investigator: Peter Chika Ozo-Ogueji

Program: M.S. Data Science and Business Analytics & AI

Institution: American University, Washington, D.C.

4. Introduction/Motivation

Problem Statement: Chest radiography represents the most common medical imaging examination globally (>2 billion annually), yet radiologist shortages—particularly in resource-limited settings—create diagnostic delays and missed pathologies. While deep learning models achieve radiologist-level performance on internal validation sets, their real-world deployment remains severely limited by poor generalization across institutions. Distribution shifts from heterogeneous imaging equipment, patient populations, and institutional protocols typically cause 10-25% AUC degradation on external data, rendering models clinically unreliable despite strong development-set performance.

Deep Learning Significance: This problem exemplifies critical deep learning challenges: (1)
Domain Adaptation — models must generalize to unseen institutions without target-domain labeled data during training, requiring robust feature learning resistant to dataset shift, (2)
Multi-Label Learning — chest X-rays exhibit multiple concurrent pathologies with inter-disease dependencies demanding architectures that model joint probability distributions, (3)
Uncertainty Quantification — clinical reports contain uncertain findings requiring principled handling of label noise and ambiguity, and (4)
Calibration — predicted probabilities must align with empirical frequencies for clinical decision support and physician trust.

5. Methodology

Architecture Overview: We develop a convolutional neural network with transfer learning for uncertainty-aware multi-label classification, optimized for cross-institutional generalization:

$$\hat{y} = \sigma(f_\theta(x_s(t), x_c))$$

where x_s(t) = chest X-ray image, x_c = patient metadata (age, sex), θ = learnable parameters, σ = sigmoid for multi-label output.

Technical Components: (1) **Transfer Learning Encoder** — ResNet-50 or DenseNet-121 pre-trained on ImageNet provides generalizable visual features, fine-

tuned on MIMIC-CXR to capture radiological patterns while maintaining robustness to domain shift through aggressive augmentation, (2) **Uncertainty-Aware Multi-Label Loss** — masked binary cross-entropy excludes uncertain labels (-1) from gradient computation, only optimizing on definitive positives (1) and negatives (0):

$$\mathcal{L}(\theta) = \Sigma_i \Sigma_j m_{ij} \cdot BCE(\hat{y}_{ij}, y_{ij}) + \lambda||\theta||_2^2$$

where m_ij = mask (1 if label certain, 0 if uncertain), BCE = binary cross-entropy, λ = L2 regularization, (3) **Domain Generalization via Augmentation** — random horizontal flips, rotations (±10°), brightness/contrast jittering, and Gaussian noise simulate natural cross-institutional variations in imaging protocols and equipment settings, and (4) **Temperature Scaling Calibration** — post-hoc probability calibration learns optimal temperature parameter T on validation set to align predicted probabilities with empirical class frequencies, critical for clinical trust:

$$\hat{p}_{calibrated} = \sigma(z_i/T)$$

6. Intended Experiments

Dataset: MIMIC-CXR (v2.0) with 377,110 chest radiographs from 227,835 imaging studies covering 64,588 patients, collected 2011-2016 at Beth Israel Deaconess Medical Center. IRB-approved and HIPAA-compliant via PhysioNet credentialing. **Structured data:** 14 pathology labels (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Pneumonia, Pneumothorax, No Finding, Enlarged Cardiomediastinum, Lung Opacity, Lung Lesion, Fracture, Support Devices, Pleural Other) extracted via automated CheXpert labeler with uncertainty encoding. **External validation:** CheXpert dataset with 224,316 chest X-rays from 65,240 patients at Stanford Hospital (2002-2017), providing geographically distinct test cohort with identical label schema for evaluating cross-institutional generalization.

Experiment	Methodology	Metrics
Baseline Comparison	ResNet-50 vs. DenseNet-121 vs. EfficientNet	AUC, AUPRC
Ablation Studies	Transfer learning vs. random init, augmentation impact	Feature importance
Hyperparameter Tuning	Bayesian optimization (Optuna)	Cross-validation
Interpretability Analysis	GradCAM attention maps	Clinical relevance
Calibration Assessment	Temperature scaling	Reliability diagrams

Performance Evaluation: Temporal validation splits MIMIC-CXR chronologically (2011-2014 train, 2015 validation, 2016 test) simulating prospective deployment. Cross-institutional testing evaluates zero-shot generalization to CheXpert without fine-tuning on Stanford data. AUC quantifies discrimination across decision thresholds, AUPRC handles severe class imbalance (pneumothorax <5% prevalence), ECE measures probability calibration, and GradCAM visualizations provide clinical interpretability. Stratified subgroup analysis by pathology type, imaging view, and patient demographics reveals potential biases requiring mitigation.

7. Prior Research

Irvin et al. (2019) introduced CheXpert with 224K labeled chest X-rays demonstrating radiologist-level deep learning performance, establishing the uncertainty label paradigm we adopt—directly motivating our uncertainty-aware loss function. Rajpurkar et al. (2017) developed CheXNet achieving superior pneumonia detection through 121-layer DenseNet, highlighting transfer learning's power for medical imaging despite limited training data—aligning with our architectural choices. Johnson et al. (2023) released MIMIC-CXR enabling large-scale radiograph research with automated label extraction via NLP, providing our training foundation. Zech et al. (2018) revealed catastrophic generalization failure across hospitals (93% internal → <50% external AUC) due to confounding dataset biases, demonstrating the critical importance of our cross-institutional validation approach for clinical deployment readiness. Guo et al. (2017) showed modern neural networks produce poorly calibrated probabilities requiring post-hoc correction via temperature scaling—we adopt this for clinical reliability.

8. References

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks.
International Conference on Machine Learning, 1321-1330. PMLR.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison.
Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 590-597. https://doi.org/10.1609/aaai.v33i01.3301590

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Peng, Y., ... & Mark, R. G. (2023). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs.
Scientific Data, 10(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.
PLoS Medicine, 15(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

Dataset Reference: MIMIC-CXR Database: Johnson, A. et al. MIMIC-CXR-JPG (version 2.0.0). PhysioNet. https://doi.org/10.13026/8360-t248

Technical Stack: PyTorch 2.0 (deep learning), torchvision 0.15 (transformations), timm 0.9 (model zoo), Optuna 3.0 (hyperparameter optimization), GradCAM (interpretability), MLflow + DVC (experiment tracking).