BIS 568 Final Report

Automated Diagnosis of Cardiovascular Diseases

Using Multi-Phase Cardiac MRI and Convolutional Neural Networks

Instructor: Dr. Wade Schulz

Group Members:

Yuchen Liu, Rong Sun, Ruoxi Teng, Abbey Yuan

Introduction and Problem Specification

Cardiovascular disease is the leading cause of death worldwide [1], and timely, accurate diagnosis is critical for effective management and improved patient outcomes [2]. MRI is widely regarded as the gold standard for cardiac imaging, offering detailed views of cardiac anatomy and function [3]. However, interpreting these images is time-consuming, requiring trained experts. This presents a significant opportunity for AI: by automating the diagnostic process, AI has the potential to reduce workload, improve diagnostic consistency, and facilitate early detection of cardiac conditions.

Therefore, our project leverages a cardiac MRI dataset to perform image classification. The goal of this project is to develop a CNN-based model to classify cardiac cine-MRI images into clinically relevant categories. From a regulatory perspective, this AI system would be considered Software as a Medical Device and would need to meet FDA guidelines and Good Machine Learning Practices.

The model aims to serve as a decision support tool for end-users such as cardiologists and radiologists, assisting in accurate, rapid, and reproducible diagnosis from routine MRI scans. Ultimately, the project aims to build a reliable, interpretable, and clinically meaningful AI system to assist in the diagnosis and classification of cardiac diseases using MRI data.

Methods

Data Preprocessing: The data was sourced from the Automated Cardiac Diagnosis Challenge (ACDC) dataset, including 150 patients evenly distributed across five cardiac conditions [4]. Inclusion criteria required high-quality short-axis cine-MRI scans with clearly labeled end-diastolic (ED) and end-systolic (ES) phases. Patients with corrupted or incomplete files were excluded automatically during preprocessing. Each patient's 3D MRI volumes were processed by extracting 2 mid-slices along the z-axis for both ED and ES frames. All pixel intensities were normalized to the [0, 1] range. The model used pre-segmented label masks rather than raw MRI images, incorporating expert-defined anatomical regions and reducing complexity by focusing on structural contours.

Model Development and Evaluation: We initially built and evaluated a 5-class multi-head CNN Model 1 distinguishing NOR (Normal), MINF (Systolic Heart Failure with Infarction), DCM (Dilated Cardiomyopathy), HCM (Hypertrophic Cardiomyopathy), and ARV (Abnormal Right Ventricle). After observing moderate performance, we reformulated the task into a binary classification problem: Normal (NOR) vs. Abnormal (ABN), resulting in Models 2–4. These binary models progressively improved via increased network depth and optimized hyperparameters. All models used convolutional layers with ReLU activation and max pooling, followed by fully connected layers and dropout for regularization. We used the Adam optimizer and binary cross-entropy loss. Model 1 featured 5 output heads, one per class. Models 2-4 featured a single output head for binary classification.

Model performance was assessed using accuracy, precision, recall, F1-score, confusion matrices, and AUC metrics (ROC-AUC and PR-AUC). For the 5-class model, we computed per-class and macro-average F1-scores. For binary models, we included calibration curves to assess confidence reliability. Visual interpretability was ensured using Grad-CAM and saliency maps.

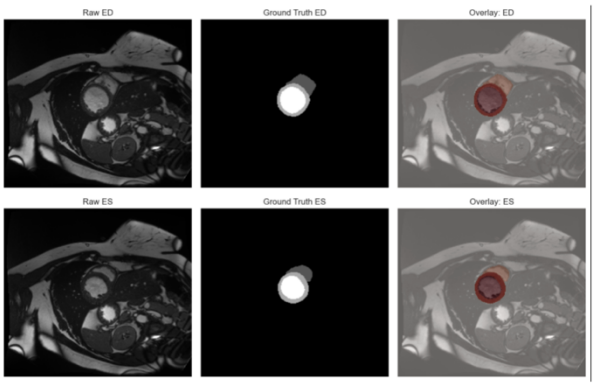
MLOps Integration: We integrated MLflow into our model development pipeline to log hyperparameters, training metrics, and model artifacts. Key parameters such as learning rate, batch size, and number of epochs were tracked alongside performance metrics like validation loss and top-1 accuracy. This approach allowed reproducibility and easy comparison between experimental runs, which enhanced transparency, streamlined collaboration, and provided a robust framework for iterative model improvement and deployment readiness.

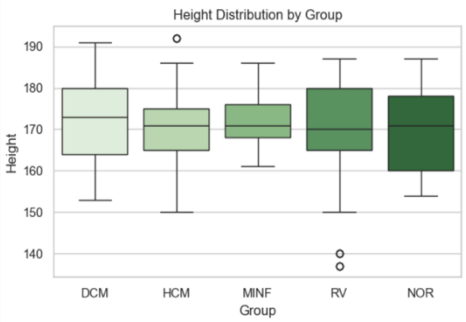
Results

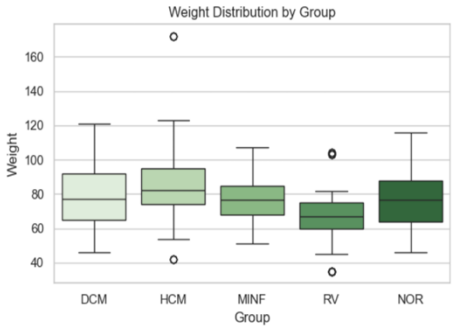
Exploratory Data Analysis: The cohort includes image data of five diagnostic categories: DCM, HCM, MINF, RV, and normal. Each category contains 30 patients, resulting in a balanced class distribution. For demographic balance across disease groups, height was consistently distributed, and while weight varied slightly, the normal group's median was only marginally higher. Histograms comparing

normal vs. abnormal groups revealed substantial overlap, suggesting no strong demographic separation. These results confirm the dataset's representativeness and suitability for unbiased classification without extensive preprocessing.

Fig. EDA: Sample Image Data and Demographics



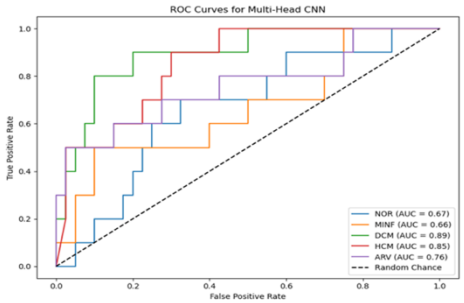


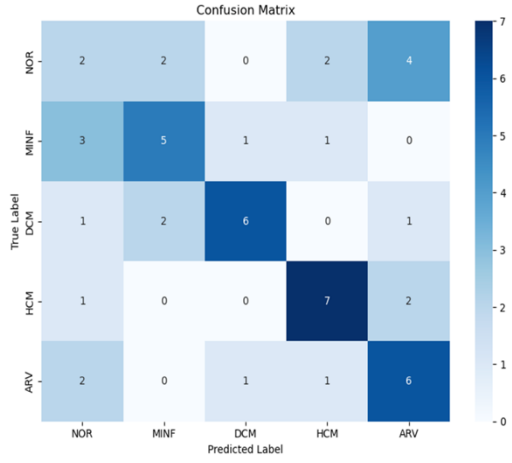


Model Performance Overview: We evaluated four CNN models: one multi-class classifier (Model 1) and three binary classifiers (Models 2-4). Model 1 achieved an overall accuracy of 52.0%, with highest F1-scores for DCM and HCM (both 0.667), but poor performance on NOR (F1 = 0.211). Binary models performed significantly better. Model 2 improved accuracy to 74.0%, followed by Model 3 at 76.0%, and the best performance was achieved by Model 4 with 86.0% accuracy and a macro F1-score of 0.773. ROC-AUC and PR-AUC metrics showed consistent improvement across models, with Model 4 reaching 0.865 and 0.949 respectively.

Tabel : Model 1 Classification Report

	precision	recall	f1-score	support
NOR	0.222	0.2	0.211	10
MINF	0.556	0.5	0.526	10
DCM	0.75	0.6	0.667	10
HCM	0.636	0.7	0.667	10
ARV	0.462	0.6	0.522	10
accuracy	0.52	0.52	0.52	0.52
macro avg	0.525	0.52	0.518	50
weighted avg	0.525	0.52	0.518	50

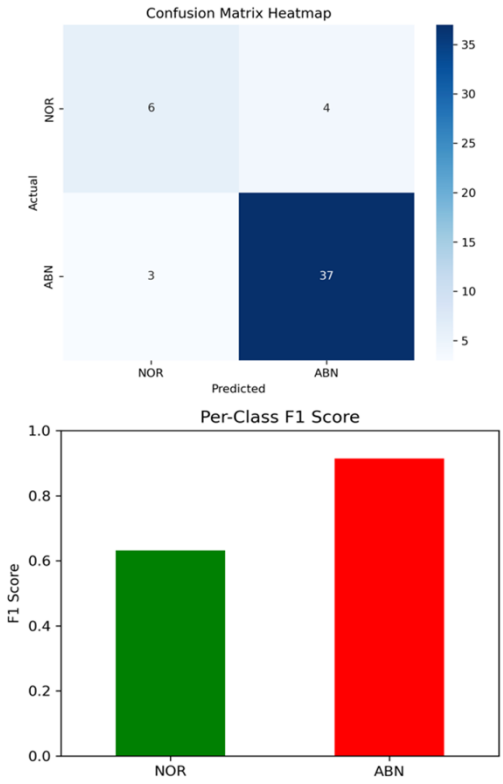




Bias Assessment: Bias was assessed using per-class F1 scores and confusion matrices. For model 4, despite strong overall performance, it demonstrated class-specific disparities: 37/40 abnormal

cases were correctly classified, while only 6/10 normal cases were identified. This suggests a tendency toward false positives for normal samples. The NOR class had an F1-score of 0.63, while ABN scored 0.92. These discrepancies may stem from class imbalance, labeling inconsistencies, or scanner variability. Mitigation strategies include data augmentation for NOR cases, multi-rater label validation, and preprocessing standardization.

Fig. Model 4 Performance



Interpretability and Explainability: To ensure transparency, we applied GradCAM and Saliency Maps to Model 1. GradCAM visualizations (Fig. GradCAM Plot) showed that the model focused on anatomically relevant myocardial regions across different disease categories. Saliency Maps (Fig. Saliency Plot) highlighted edge-focused patterns, indicating that predictions were driven by structural contours in the MRI slices. These techniques offer clinicians visual insight into what the model prioritizes, increasing trust in AI-assisted diagnostics.

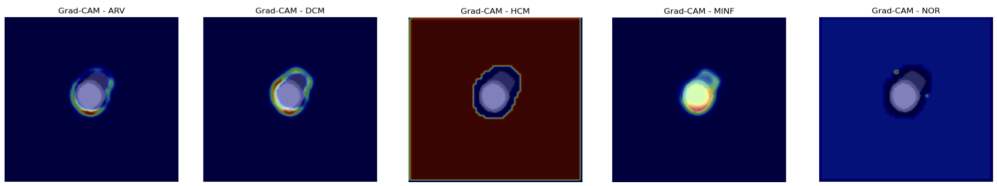


Fig. GradCAM Plot: Anatomic Regional Contribution to Prediction

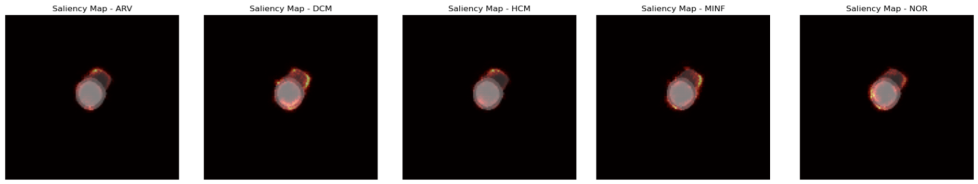


Fig. Saliency Plot: Pixel Level Abnormalities

Conclusion

While the multi-head CNN performed 2.6 times better than random guessing, it still struggled to distinguish between subtle disease classes – likely due to limited data and high inter-class variability –

suggesting that more data or advanced modeling techniques are needed for reliable multi-class diagnosis. In contrast, binary models, especially the improved deep CNN (Model 4), achieved high accuracy (86%) and balanced performance, making them well-suited for clinical decision support. Overall, binary classification represents an effective first step in automated cardiac MRI interpretation, with strong potential for detailed multi-class classification as larger datasets become available.

Discussion

Evaluation Plan: To assess post-deployment performance, we propose monitoring several key metrics. Top-1 accuracy will track overall performance stability as patient populations or imaging protocols evolve. Per-class F1-scores will detect degradation in specific categories, especially underrepresented ones like NOR. Prediction confidence will help flag increased uncertainty, indicating potential data drift or emerging issues. Calibration error will assess how well predicted probabilities reflect true outcome likelihoods, essential for clinical decision-making.

Implementation Plan: Inference will occur whenever a patient receives a cardiac cine-MRI scan, aligning with routine cardiovascular imaging workflows. The model operates in a batch inference mode and does not require real-time processing. A standardized pipeline will also handle mid-slice selection, normalization, and ED/ES phase stacking. The production architecture includes MRI acquisition, segmentation, feature extraction, classification, and interpretable output delivery via Grad-CAM, supporting modular updates and continuous refinement.

Dissemination Strategy: To ensure meaningful impact, we will engage both clinical and research stakeholders. Target audiences include cardiologists and radiologists who may use the tool in practice, and AI/ML researchers focused on medical imaging. We plan to disseminate results through peer-reviewed journals and AI/healthcare conferences, as well as through clinical implementation channels such as hospital technology committees and continuing medical education (CME) seminars. These efforts aim to promote adoption, feedback, and further development of our model in both research and healthcare settings.

Github Repository Link

All code can be found in our Github Repository under final_project, including model1.ipynb and model 2_3_4.ipynb.

https://git.yale.edu/qy88/BIS568

References

1. Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022). The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. Journal of the American College of Cardiology, 80(25), 2361–2371. https://doi.org/10.1016/j.jacc.2022.11.005

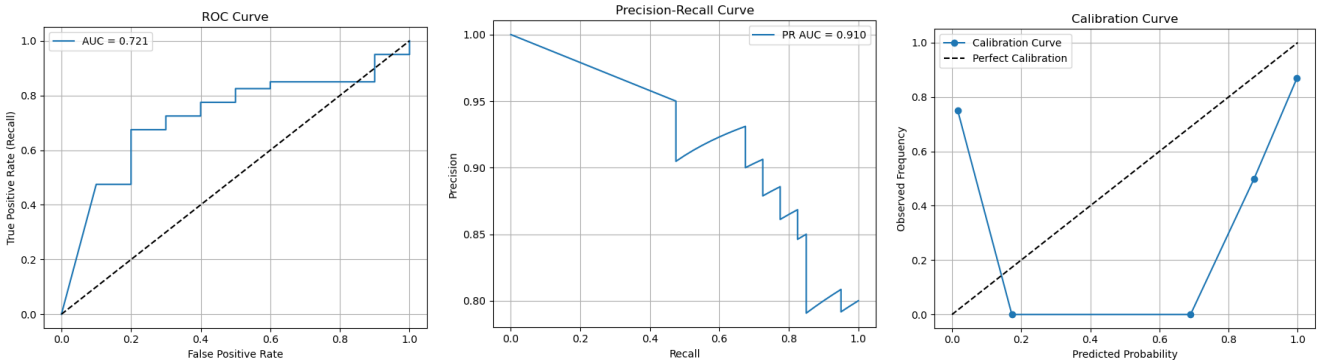
2. Taylor, C. J., Hartshorne-Evans, N., Satchithananda, D., & Hobbs, F. R. (2021). FASTer diagnosis: Time to BEAT heart failure. BJGP open, 5(3), BJGPO.2021.0006. https://doi.org/10.3399/BJGPO.2021.0006

3. Wang, Y. J., Yang, K., Wen, Y., Wang, P., Hu, Y., Lai, Y., Wang, Y., Zhao, K., Tang, S., Zhang, A., Zhan, H., Lu, M., Chen, X., Yang, S., Dong, Z., Wang, Y., Liu, H., Zhao, L., Huang, L., Li, Y., … Zhao, S. (2024). Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nature medicine, 30(5), 1471–1480. https://doi.org/10.1038/s41591-024-02971-2

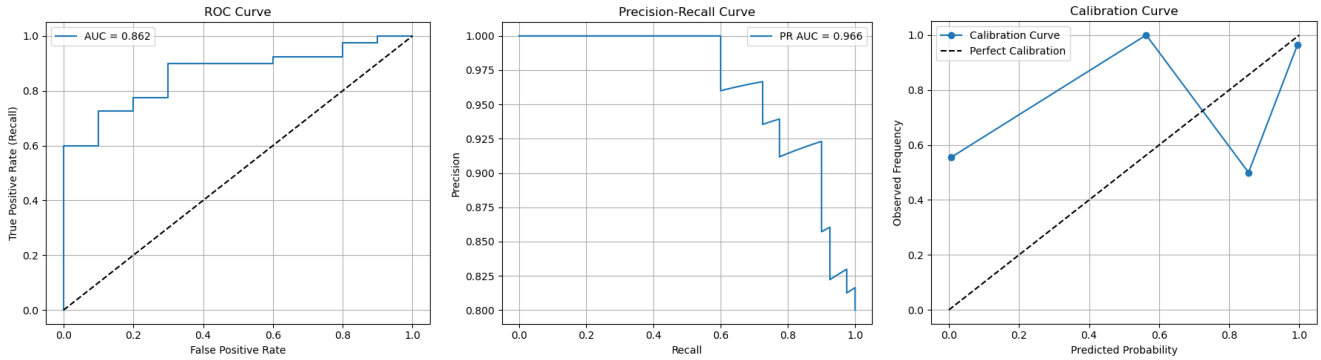
4. Dazel S. *Automated Cardiac Diagnosis Challenge (MICCAI 2017)* [Internet]. Kaggle; [cited 2025 May 3]. Available from: https://www.kaggle.com/datasets/samdazel/automated-cardiac-diagnosis-challenge-miccai17?resource=download

Appendix

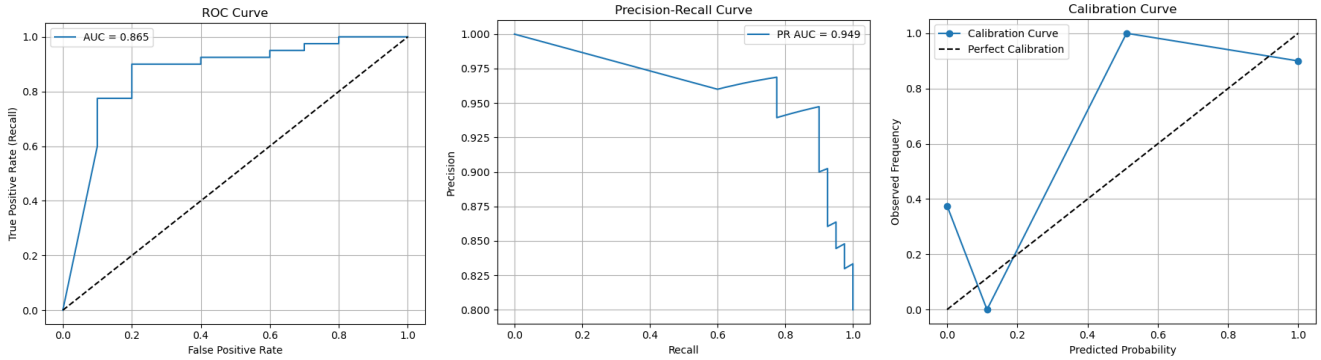
Sup.Fig. Model 2: Simple Binary CNN (Collapsed to NOR vs. ABN)



Sup.Fig. Model 3: Deep Binary CNN (Increased Depth)



Sup.Fig. Model 4: Improved Deep Binary CNN (Best Model)



Sup.Fig. Performance Matrix Comparison of All CNN Models