Jacob Krol BIOS 7747 2023-10-04

The Department of Cardiology is interested and exploring the possibility of identifying pathologic cardiac hypertrophy from physiological cardiac hypertrophy. They first designed a pilot study that enrolled 50 patients, for which they collected basic patient and cardiac function information to evaluate their hypothesis and the feasibility of the project. After such small pilot, they collected additional data from 200 subjects with hypertrophic cardiomyopathy.

1. Use logistic regression to evaluate the feasibility of identifying pathologic cardiac hypertrophy in the pilot dataset with 50 subjects. Use cross-validation to evaluate performance using different test group sizes (1, 10, 20, 30, 40) in terms of accuracy and area under the ROC curve. Visualize the ROC curves for different test group sizes and discuss the observed differences.

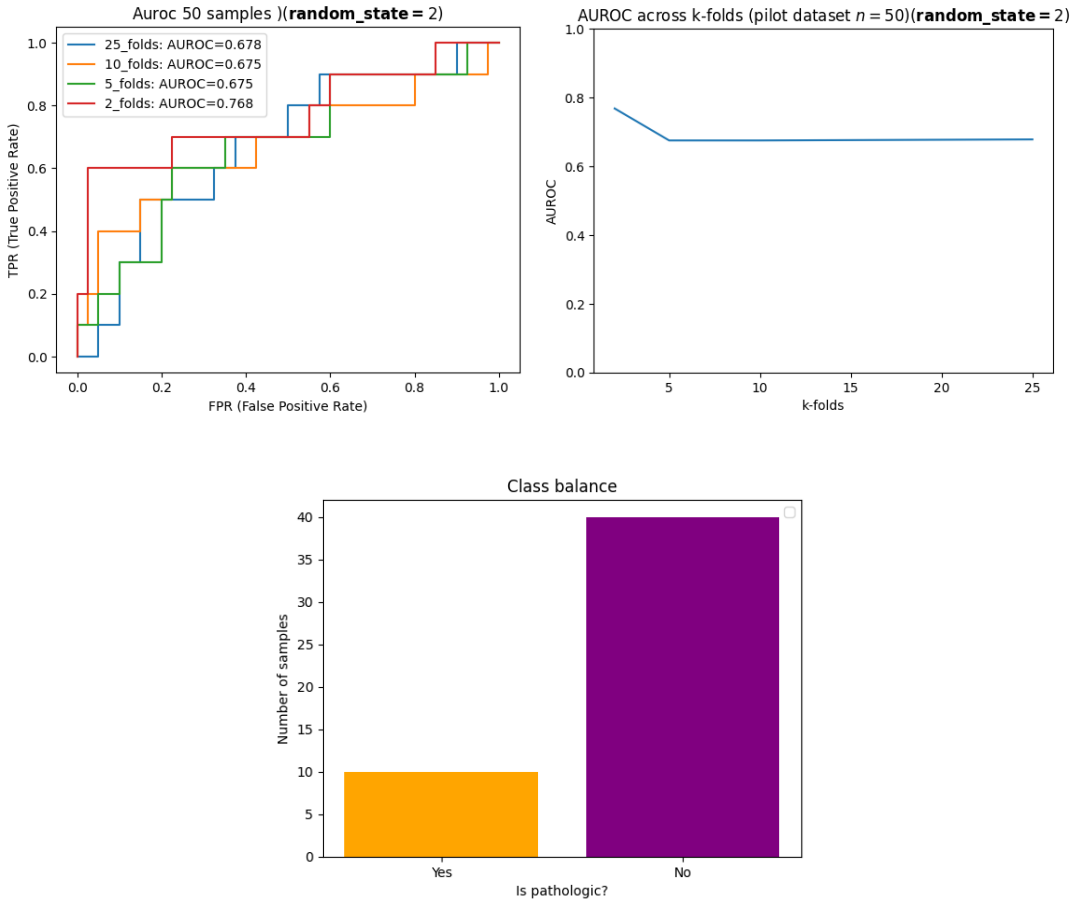
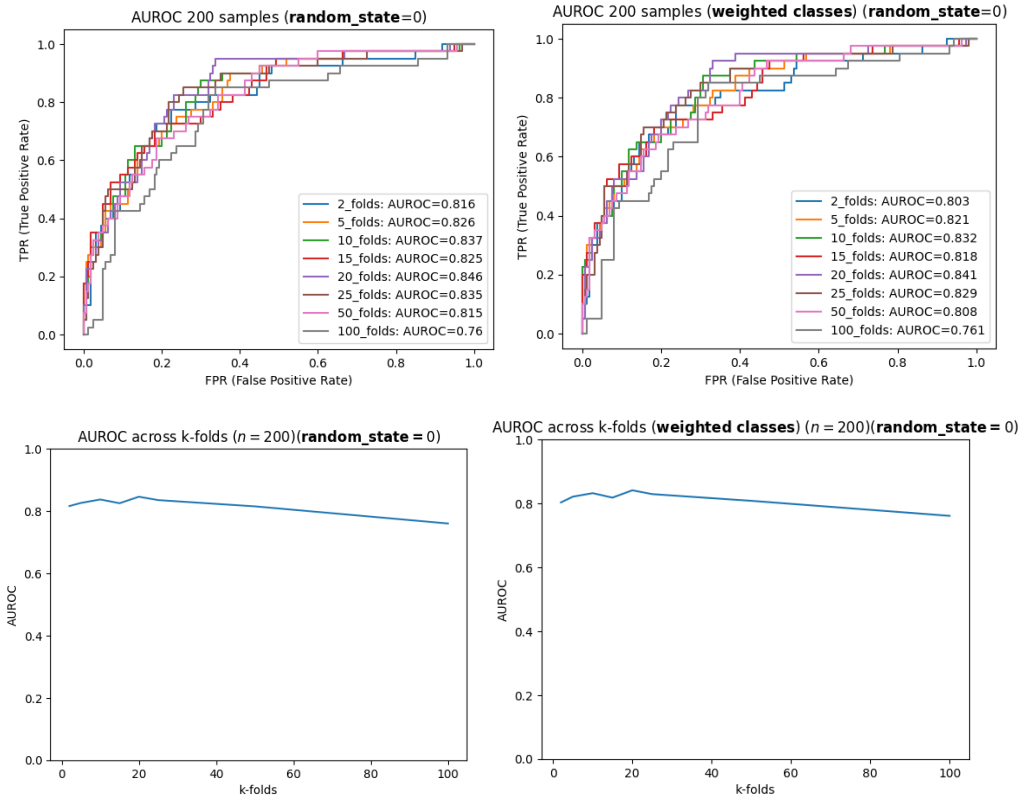


Fig 1: **Logistic Regression performance for the pilot (n=50) dataset. A (top left)**: AUROC plots for various k folds in the pilot (n=50) dataset each line represents the false positive rate (FPR) and true positive rate (TPR) across various decision thresholds aggregated to measure performance (**AUROC**) across all folds in a given k-fold split (e.g., “2_folds” [green] plots the ROC from the testing in both (k=2) fold iterations, “10_folds” [red] plots the ROC for all k=10 fold iterations, etc.).

For the pilot dataset, a logistic regression classifier was trained using stratified k-fold cross validations with various k, {25, 10, 5, 2} to predict pathological cardiac hypertrophy. The AUROC was measured using the positive labels predicted probabilities $\widehat{p_{pos}}$ and the ground truth, y, labels of all models trained in a given k-fold. (**Fig 1A**). k=2 performed best with the highest AUROC (0.768) while the cross-validations with greater folds (k=5, AUROC=0.675; k=10, AUROC=0.675; k=25, AUROC=0.678) performed worse (**Fig 1AB**). For further validation, various random states were used to shuffle the label data before making splits, and the majority of the iterations also showed the AUROC decreasing as the k-folds increased (plots available in supplementary section).

2. Repeat the study with the larger dataset collected after the pilot. Evaluate and discuss the differences compared to the initial study.



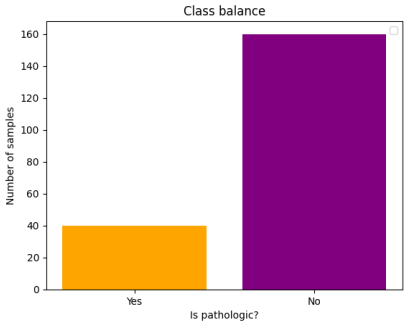
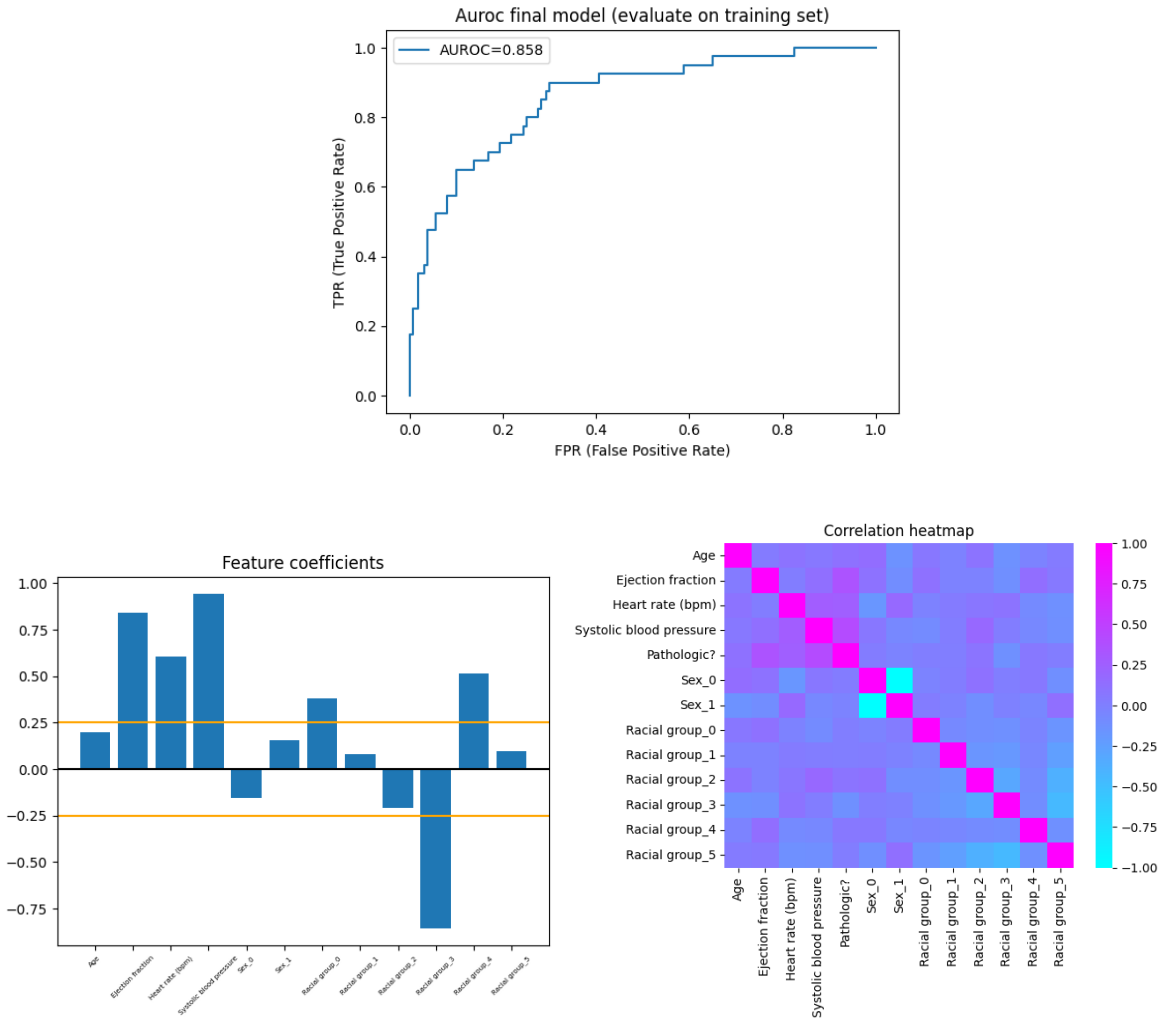


Fig 2 classification results for full dataset (n=200): **A** (top-row): AUROC plots for various k folds in the dataset with 200 samples using a classifier trained on both unweighted (top-left) and **weighted** classes (top-right). **B** (middle) Summary of AUROCs measured at different k-folds for unweighted and weighted classifiers. **C** (bottom): class balance.

Logistic regression was used to classify pathological cardiac hypertrophy in a follow up data set with more samples (n=200). To measure model performance, stratified k-fold cross validation performance in the dataset with 200 samples across various k-folds {2, 5, 10, 15, 20, 25, 50, 100} and the AUROC range across all folds was 0.76-0.837 (**Fig 2a**). A summary plot of the AUROC across k-folds shown in **Fig 2b** indicates the fold size generally had little effect on the performance. Since the classes are disproportionate with more non-pathological samples than pathological, a weighted classifier was trained which amplifies the error of the underrepresented class proportional to the ratio of the class imbalance. In this case, 160NP:40P means that errors for the pathological class (P) are scaled by a factor of 4. **Fig 2AB** show that the AUROC across k-folds is essentially the same whether or not the classes were weighted during training. **Overall, the model performed better on the n=200 dataset with a maximum AUROC of 0.846 (with relatively similar AUROCs across most folds [Fig 2AB) compared to the n=50 dataset which had a maximum AUROC of 0.768 (with higher variance of AUROC across folds [Fig 1AB]).**

3. Train a final model ready for deployment. What is the performance in the training dataset? What is the relationship between every variable and the predicted outcome?



:

Fig 3 Training and testing model on full dataset. A (top): AUROC performance when trained and evaluated on the entire dataset. **B** (bottom-left) Model coefficients. **C** (bottom-right): correlation heatmap of all variables in the dataset.

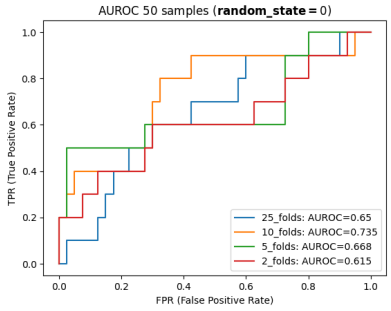
A final model was trained on the entire dataset using class weight error balancing (the same "balanced" method described in answer 2) and fit against the same data with an AUROC of 0.858 (**Fig 3A**). The feature coefficients show that age, ejection fraction, heart rate (BPM), systolic blood pressure, sex_1, racial_group_0, racial_group_1, racial_group_4, and racial_group_5 were given positive weights toward pathological cardiac hypertrophy while the remaining features were assigned negative weights toward the outcome (**Fig 3B**). L2, regularization was used to minimize the magnitude of coefficients and prevent possible overfitting when the model is applied to new data. A correlation heatmap in **Fig 3C** mirrors the regression model coefficient sign where positive correlations with the pathological variable were

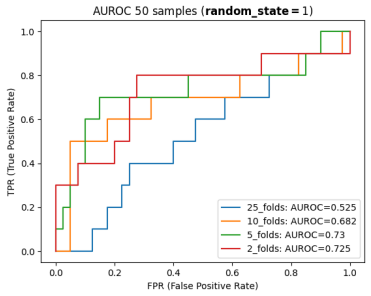
assigned positive weights in the fitted model while negative correlations with the pathological variable were assigned negative weights.

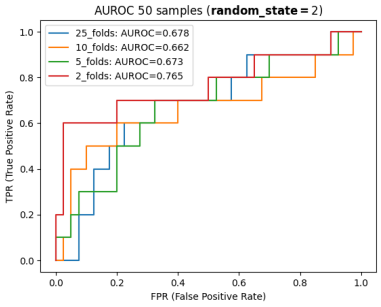
Supplementary info:

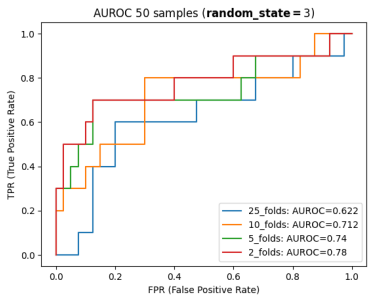
Standardization for continuous values was applied post-train/test split, for each fold iteration, to prevent adding bias in the test set. For example, each fold split would partition the predictor data into two sets (train and test) and standardization is applied within each partition separately.

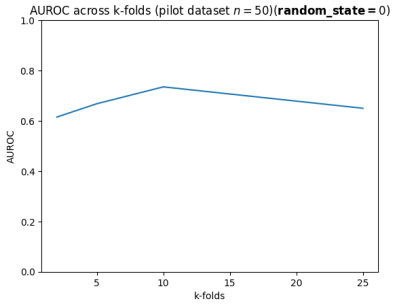
50 samples performance across multiple random shuffles

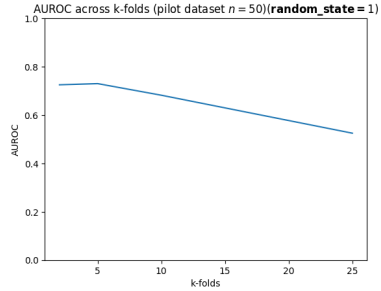


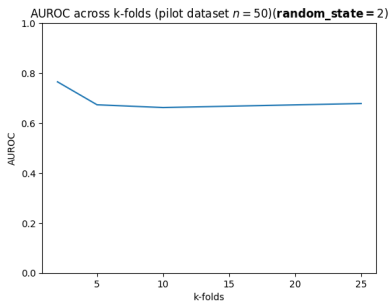


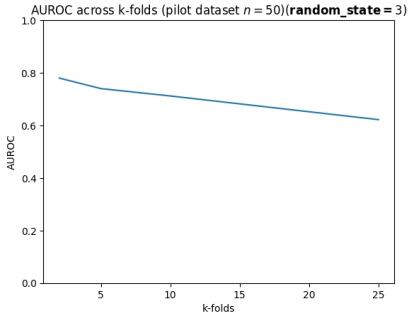




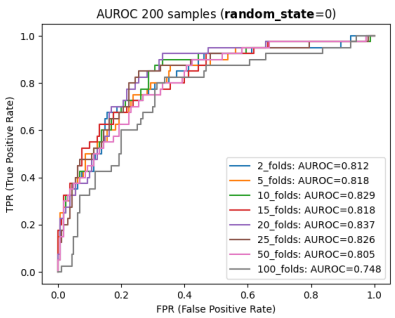


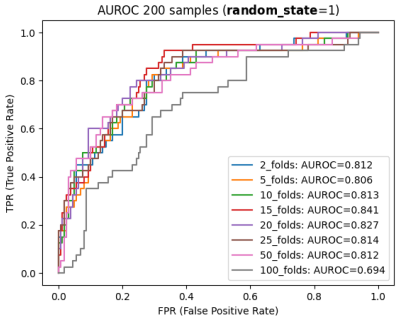


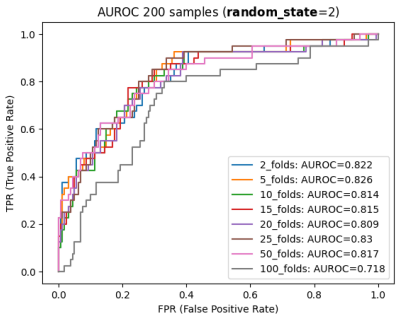


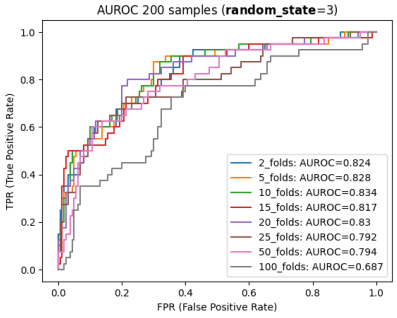


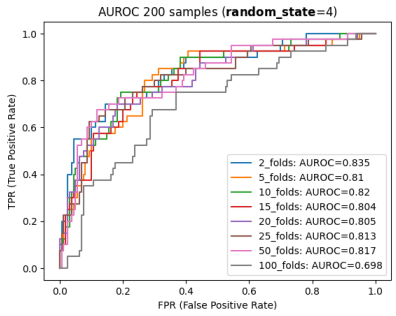
200 samples performance across multiple random shuffles

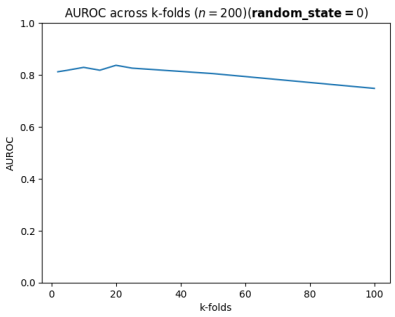


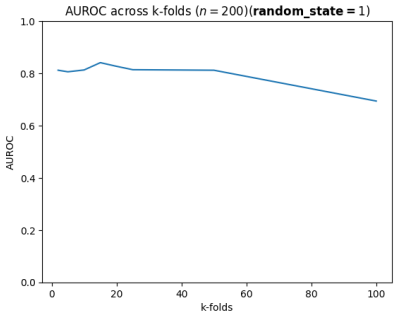


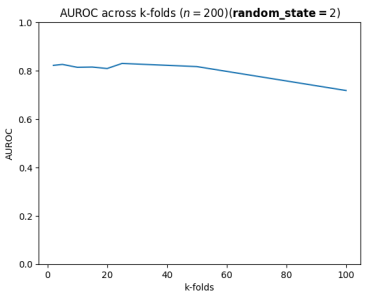


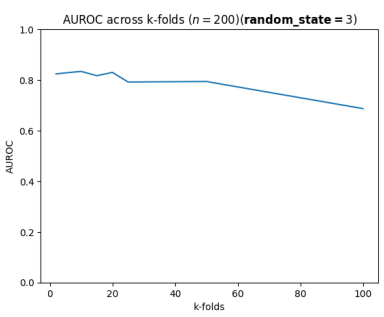


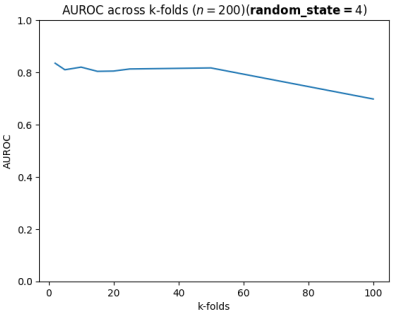












Jacob Krol BIOS 7747 2023-10-04

200 samples (weighted) performance across various random shuffles