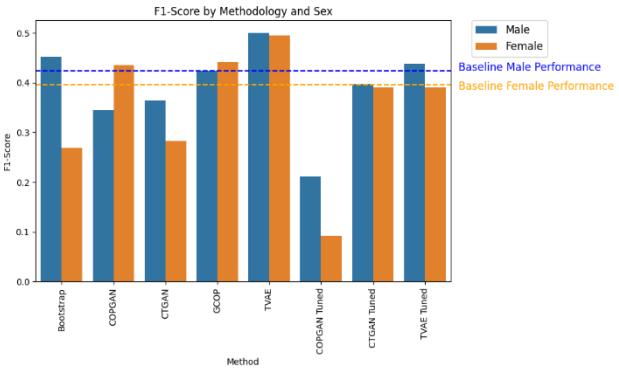
Synthetic Data Generation to Improve Classification Performance for Heart Failure Prediction

Roy S. Zawadzki, Saman Parvaneh

Edwards Lifesciences, Irvine, CA, USA

In cardiology, datasets are often small and contain class imbalance for the target variable, which hinders machine learning (ML) prediction performance. Furthermore, models may have suboptimal performance in certain subpopulations (e.g., sex and race) due to limited data. One solution lies in synthetic data generation (SGD) where models are trained to generate realistic patient observations. We investigate the use of SGD for two aims: (1) to increase overall ML performance by augmenting the training set with generated synthetic data and (2) to mitigate ML performance disparities by only generating data for certain underrepresented subpopulations and augmenting this to the training data.



As a case study, we utilize the University of California, Irvine myocardial infarction dataset (n=1,700) using features available upon hospitalization with the Catboost algorithm to predict chronic heart failure, a binary outcome. First, we partitioned the data into 70% training and 30% test stratified by chronic heart failure and sex. Then, utilizing open-source generators in Python, both with and without hyperparameter tuning, for aim one, we doubled our training set (n=1000), and for aim two, we generated female observations such that the number of males and females were the same (n=500 each). Additionally, we used bootstrap sampling as a simpler data augmentation method. To quantify performance, we looked at the change in accuracy, AUC, and F1-score on the test set before and after augmenting generated synthetic data for training Catboost. Overall, we observed performance gains for both aims. For aim one, bootstrap sampling and Triplet-based Variable AutoEncoders (TVAE) increased performance; in aim two, Copula Generative Adversarial Model (COPGAN), Gaussian Copula (GCOP), and TVAE saw increases in female classification performance. Hyperparameter tuning did not offer notable gains. We believe the overall results could improve with larger sample sizes, which motivates future methodological developments under smaller sample sizes.