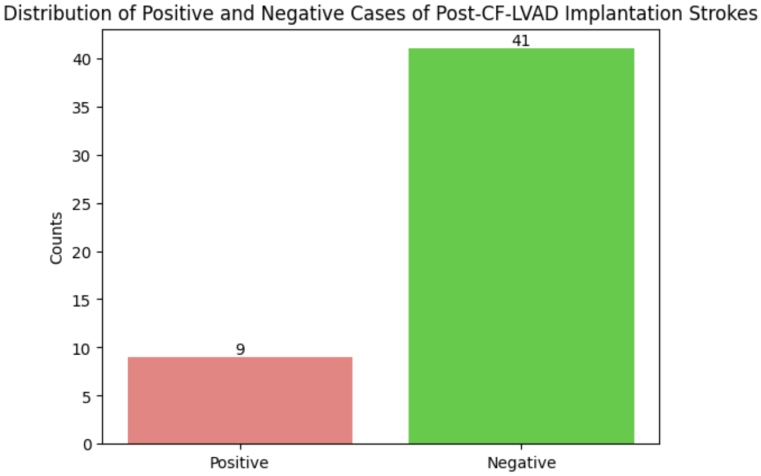
Methodology:

- CTGAN:

Since we have a small dataset with only 50 patients, we use the Conditional Tabular GAN (CTGAN) [1] to generate synthetic tabular data that are similar to the 50 actual patients' data. The CTGAN is a generative adversarial network (GAN) that is explicitly designed for tabular data. It is composed of a generator, a discriminator, and an embedding network. The generator learns from the actual patient data's distribution and provides the synthetic data, while the discriminator is trained to distinguish the real and synthetic data. The output synthetic data will be indistinguishable based on the judgments of the discriminator. The synthetic data serves as extra data points for deep learning models to provide a more accurate result.

- SMOTE:

We also have a class-imbalanced dataset with only 9 patients of the 50 patients developing new stroke within 3 months after CF-LVAD implantation. Since the CTGAN learns from the distribution of the real patients' data, we still face an imbalanced dataset after adding the synthetic data from the CTGAN. In this case, we used the Synthetic Minority Over-sampling Technique (SMOTE) [2] to balance the data distribution through oversampling. It finds the k-nearest neighbors of the minority class and creates new data points. Oversampling can avoid the model from always predicting the majority class when performing a classification.



- TabNet:

We used an attention-based neural network model, TabNet [3], for binary classification and prediction of the post-CF-LVAD implantation strokes. The TabNet uses an encoder-decoder structure. The encoder is composed of a feature transformer, an attentive transformer, and feature masking, while the decoder is composed of a feature transformer. The TabNet is designed to work well on tabular data with columns as features and rows as samples. This matches the structure of our patient data.

- Majority Vote:

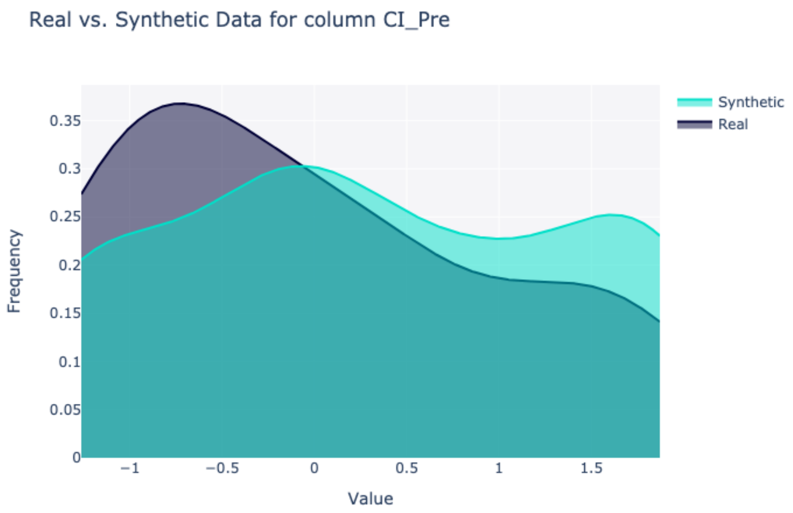
Due to a small dataset, our result from SMOTE can be variating. Therefore, to obtain a more accurate and stable result, we will conduct 100 trials with different SMOTE random states as inputs. After conducting the 100 trials, we will gather the probability vector of each trial on the test data. For each patient in the testing data, we look over the 100 trials and calculate the majority vote. If more than 50 trials have determined this patient to have a higher than 0.5 possibility to have stroke, we will predict this patient to have post-CF-LVAD implantation strokes. Otherwise, we predict this patient to be normal.

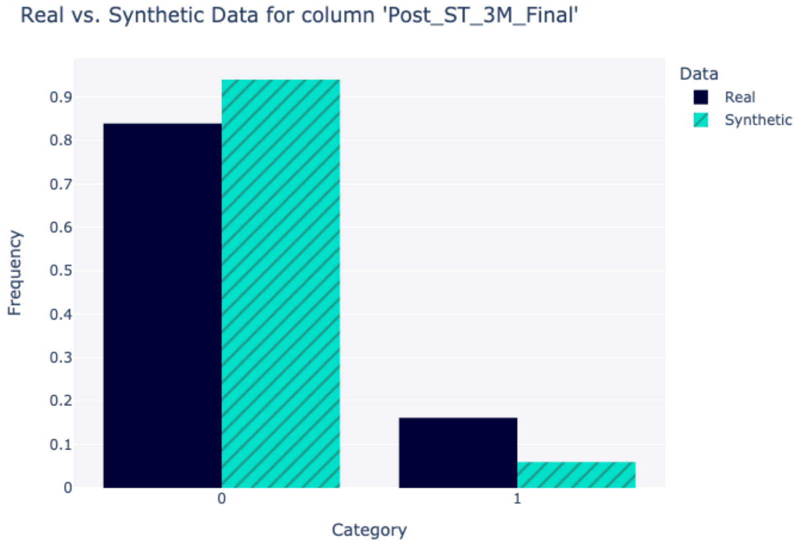
Experiment:

- Procedure:

We are given a data frame with 50 patient entries and 41 valid features to consider. We perform a label encoding for all the categorical string variables. For the rest of the continuous variables, we standardize them by centering around the mean and scaling them to unit variance using the StandardScaler [2]. We extract the target label representing the presence of post-CF-LVAD implantation strokes and split the 50 patients into 85% training and validation data and 15% testing data. There are 2 positive cases in the testing set and 4 negative cases.

After obtaining the original real training data, we add the training data frame to a metadata object [5] and indicate categorical and continuous variable columns. The metadata is then fed into a CTGAN synthesizer [4] with an epoch of 20. We generate 100 more samples and add them to the original training data. The synthetic data obtained an around 75% similarity quality score.





Then, the SMOTE over-sampler takes in this extended training data to balance the data. We instantiate a TabNet classifier [3] with a learning rate of 0.02, an Adam optimizer, and a binary cross entropy loss function. The training and validation data are provided for the classifier to train with an epoch of 150, a patience level of 20, and a batch size of 4. For each epoch, the best weights that give the best validation accuracy are automatically used. After training, the model is tested with the test data and outputs the predictions. An Area Under the ROC Curve (AUC) score is calculated by comparing the predictions to the actual testing data labels to evaluate the performance of the data.

Through our research, we found there is randomness of how SMOTE balances the training data. Therefore, we tested the model over 100 trials with different random states of the SMOTE on the same dataset for an overview of the model’s performance. All other variables were held constant throughout the 100 trials. After the 100 trials, we conducted a majority vote to obtain a finalized result on the test data. We compare the test data with the ground truth to achieve an accuracy score for our model.

Multimodal approach (Jason):

In this work we used machine learning and patient metadata to predict future strokes before they happen. Another approach we believe will strengthen our results, is leveraging CT images of the brain associated with each subject to make predictions. In this section we will discuss our plan to use both tabular and imaging data of all subjects, and the possible impact this could have on predicting future strokes.

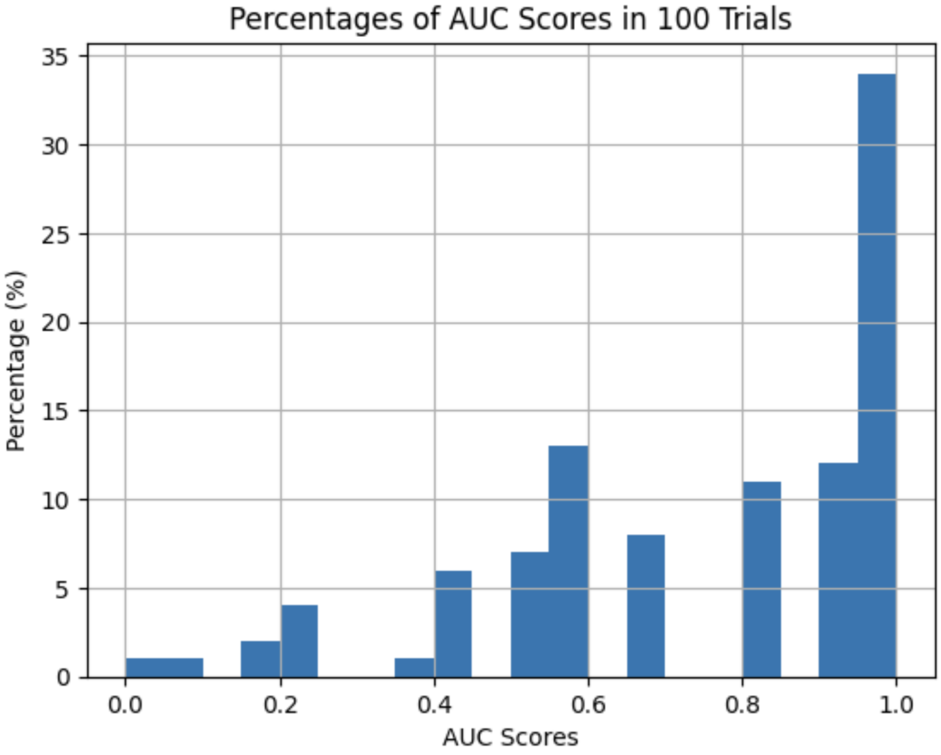
Radiomics is the collection of unique features from medical images, such as shape and texture. In this case, we'd extract a high number of features from CT images of each subject and append the data to our tabular dataset, reducing the high dimensional images to a set of radiomic features. We hypothesize that the inclusion of patient features gained from their respective set of CT images will create a more ML algorithm and lead to the improvement of its accuracy.

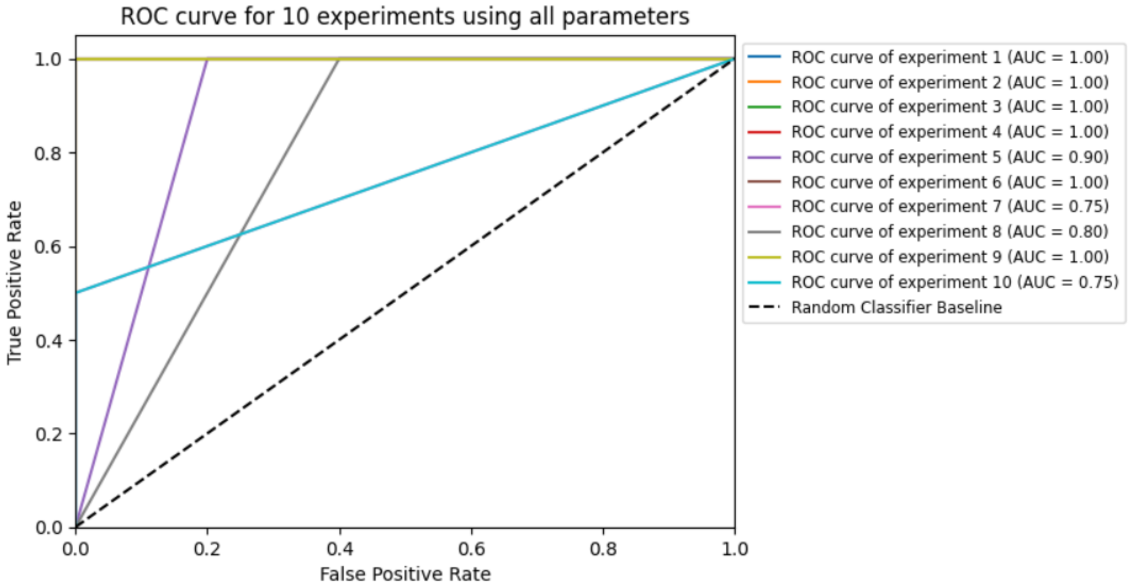
Multimodal ML algorithms are a great solution to solving problems where more than one modality is present. In our case tabular data and CT images are two different modalities, usually used in separate environments or used as inputs to different machine learning algorithms. An example of this is convolutional neural network (CNN)[1] algorithm which only takes images as input, or how TabNet[2] only takes tabular data as input. Both CNN and Tabnet can be considered as unimodal ML algorithms, where they only received one type of data as input. We will use a multimodal approach to predict future strokes and perform an assessment of both multimodal and unimodal approaches to predicting future strokes. We believe that the benefit using prior CT images of all subjects alongside with the existing tabular data will outperform our current solution, which only employs tabular metadata of each subject.

- Results:

The average AUC score of each trial on the test data out of 100 trials is 76%. Around 40 trials received an equal or better AUC score than 90%. We conducted 10 times of these 100 trials to analyze the accuracy result. After the majority vote, we were able to achieve an average of 93% accuracy and an average AUC score of 92%. For the six test patients, the model was able to classify all four negative cases correctly. On average, the model’s prediction will give us a false negative case out of the six patients.

The current limitation of this project is the small dataset. However, this preliminary study shows promising results of deep learning models in predicting post-CF-LVAD implantation strokes. We believe that with more funding to extract more post-CF-LVAD implantation stroke patient data, we will be able to have a much better result.





Related Works:

Several studies have tried machine learning models to predict strokes, such as using Deep Neural Networks (DNN) to predict stroke patient mortality [4]. Although their results were good with an 83.48% AUC score, they used a dataset of 15,099 patients. We have also done a comparative study using SVM and DNN models on our dataset and none achieved a more stable or better result than our proposed method. Another research shows that the DNN network only gives a 71.6% accuracy score when predicting with a dataset of 43,300 patient data [5]. Our study has shown that TabNet is a more suitable model for tabular data stroke prediction than DNN. They have also pointed out that the current limitation includes better data preprocessing methods to balance the imbalance of medical data. Synthetic data produced by CTGAN in our research shows that a tabular GAN network is a solution. Given our current result and our limited data, we believe that with more data points we will be able to produce an even better result. Furthermore, many other researchers have deep learning models such as U-Net and ResNet [6][7] on stroke image data. In our later studies, we expect to use the latest multimodal contrastive learning that can train on both tabular and image data. With our current approach, we believe that having more data will enable us to predict post-CF-LVAD implantation strokes with a stable and high accuracy.

References:

[1] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular data using Conditional GAN,” NeurIPS 2019, Jun. 2019, doi: https://doi.org/10.48550/arxiv.1907.00503

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] S. O. Arik and T. Pfister, “TabNet: Attentive Interpretable Tabular Learning,” arXiv (Cornell University), Aug. 2019, doi: https://doi.org/10.48550/arxiv.1908.07442

[4] S. Cheon, J. Kim, and J. Lim, “The Use of Deep Learning to Predict Stroke Patient Mortality,” International Journal of Environmental Research and Public Health, vol. 16, no. 11, May 2019, doi: https://doi.org/10.3390/ijerph16111876

[5] T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” Artificial Intelligence in Medicine, vol. 101, p. 101723, Nov. 2019, doi: https://doi.org/10.1016/j.artmed.2019.101723

[6] A. Hilbert et al., “Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke,” Computers in Biology and Medicine, vol. 115, pp. 103516–103516, Dec. 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.103516

[7]W. Qiu et al., “Machine Learning for Detecting Early Infarction in Acute Stroke with Non–Contrast-enhanced CT,” Radiology, vol. 294, no. 3, pp. 638–644, Mar. 2020, doi: https://doi.org/10.1148/radiol.2020191193