**Introduction**

The development of new onset postoperative atrial fibrillation (POAF) after coronary artery bypass grafting (CABG) is a common complication with an incidence between 10%-40%1–3. The exact pathogenesis of POAF remains unknown, but has been shown to be associated with a higher risk of both early and late morbidity and death4. This translates to increases in resource utilization, longer ICU stays and an increase in overall hospital length of stay5–7. It is conservatively estimated that the extra cost of POAF after cardiac surgery is approximately 2 billion dollars annually8. Therefore, identifying and predicting patients who will go on to develop POAF would prevent morbidity and mortality but also result in significant hospital savings.

Prediction of POAF after CABG remains difficult. CHADS2 and CHA2DS2-VASc have been shown to be predictive of POAF after cardiac surgery, with modest sensitivity and specificity9,10. Several models have been developed to predict POAF after cardiac surgery and CABG, again with modest accuracy, sensitivity and specificity1,3,11,12. Machine learning algorithms are currently being investigated and tasked with analyzing and predicting outcomes in numerous areas of medicine13. We sought to investigate and assess the ability of common supervised machine learning algorithms in predicting the development of new-onset POAF in patients undergoing CABG at a single institution using the Society of Thoracic Surgeons database.

**Methods**

Study Sample

Data used for this analysis were obtained from the Society of Thoracic Surgeons (STS) database for all patients undergoing CABG at our institution. An IRB was submitted and approved by a multi-disciplinary board at our institution. The sample used for analysis included patients who underwent CABG at Rhode Island Hospital between January 2014 and December 2017. Patients were excluded if a prior history of atrial fibrillation and atrial flutter was noted.

*Candidate variables for Prediction Models*

All available variables that were recorded in the STS database data collection form were exported. This included demographic data, pre-operative data including, and intraoperative data. Other post-operative data were discarded, except for postoperative atrial fibrillation, which was the outcome of interest. A full list and description of variables included in the initial analysis is available in the **Supplemental Table.**

*Training/Test Division and Preprocessing*

The overall schema of training and testing is illustrated in **Figure 1**. Data was imported into Python. Packages utilized in python included: sklearn, missingpy, pandas, numpy, IPython, statistics and matplotlib. Variables that were missing >50% of their data were removed. Data that was missing >20% were either recoded into categorical data if able (with a distinct value for unknown), or was removed. Given class imbalance of the outcome, the data was then split into stratified training (75%) and test sets (25%). Data with <= 10% of missing data was imputed using a *k-*nearest neighbors imputation (k=10) algorithm with which has been shown be effective in moderate sized datasets14,15. Ultimately, 395 variables were identified and used for training. Continuous variables were standardized.

Feature Selection and Hyperparameter Tuning

*Feature Selection*

Recursive Feature Elimination (RFE) is a greedy backwards selection technique that starts by creating a given model on the whole set of predictors, generates an importance score for each predictor, removes the least important predictors, and the process repeats16. RFE works with logistic regression and linear discriminant analysis as long as the number of predictors does not exceed the number of samples17. Random forests tend to pair well with RFE as random forest has an internal method for measuring the important of features17,18. RFE was applied to logistic regression and random forests during each *k* fold to optimize each particular model.

*Hyperparameter Tuning*

All tuning parameters were run in the stratified k-fold analysis, tuned parameters are detailed below. Tuning parameters were searched via an exhaustive grid search unless noted otherwise, through the GridSearchCV function in python.

*k-Nearest Neighbors* (kNN)

Performance of kNN depends on the value of *k* as well as the distance function used (Euclidean distance being the most popular)19. For small to moderate sized data sets, there is no true guideline for choosing the optimum value of *k*19*.* Small values tend to have erratic decision boundaries while larger values are likely to classify a value as the most probable class. For binary classification problems choosing an odd number for *k* may prevent tied votes20. For our experiment, the range of *k* was 5-40, the weight function used for prediction was varied as either uniform (i.e. all points in each neighborhood are weighted equally) or distance (i.e. points were weighted based on the inverse of their distance), and the possible distance functions supplied were the Manhattan distance and the Euclidean distance.

*Logistic Regression*(LR)

Logistic Regression is a common linear classifier algorithm that employs regularization parameters. Python sets LR to L­2 by default. The strength of regularization is governed by the parameter *C*, with higher values of *C* corresponding to less regularization (in python *C* is the inverse of regularization strength which is typically lambda in the literature). The values of 0.001 0.01, 0.1, 1, 10 and 100 were chosen to test which value of *C* would optimize the training set.

*Kernelized Support Vector Machines* (SVM)

SVMs work by constructing an optimal hyperplane that separates the data. SVMs have multiple tunable parameters including: *C* which controls the tradeoff between error minimization and margin maximization, the type of kernel used (linear, polynomial, gaussian, gaussian radial basis function, sigmoid), and gamma which is the kernel coefficient for the sigmoid, gaussian radial basis function and polynomial kernels21,22. For our testing, the kernels provided included: linear, polynomial, gaussian, gaussian radial basis function, and sigmoid.

*Random Forests* (RFs)

RFs work by creating a defined group of decision trees are built to best predict the outcome in the training set. RFs have a number of tunable parameters including the number of trees in the forest, the depth of each tree, the minimum number of samples required to split a node, the minimum number of samples required to be at a leaf node, and the maximum number of features to consider when optimizing a split23. Data has shown that the number of trees should be set sufficiently high, with the highest amount of performance gain occurring at 100 trees24,25. For our analysis, \*\*\*

Oversampling

Due to the class imbalance of the dataset, we expected that the machine learning models may bias towards the majority class (in this case, not developing POAF), and so a oversampling technique called synthetic minority over-sampling technique (SMOTE) was applied to the dataset at each iterative fold. SMOTE uses a kNN algorithm to create new instances of the minority class that exist between true minority values26. This was done at each fold to prevent overfitting.

Model Selection and Evaluation

As illustrated in Figure 1, each algorithm ran through a stratified k-fold cross validation, with hyperparameter tuning and RFE (where applicable) occurring at each *k* fold in order to prevent overfitting. Since there was a class imbalance of the outcome, the model with the highest average area under the receiver operating characteristic curve (ROC) was exported and applied to the test set. The ROC, precision, recall, F1 statistic were calculated to assess the model’s performance. Precision is similar to positive predictive value and is the true positives over the sum of true positives and false positives. Recall is defined as the proportion of positive samples captured by the positive predictions. F1 score is the harmonic mean of precision and recall.

**Results**

*Patient Characteristics*

After excluding patients with a prior history of atrial fibrillation, a total of 1,202 patients were identified. Baseline characteristics of the patients are shown in **Table 1.** Our study population consisted of 22% women and 89% white patients. Over half of the cohort had a history of smoking, and nearly half had diabetes. The median age was 67. 36% of patients underwent CABG electively and ultimately 34% of the cohort developed POAF.

*Model Performances*

**Table 2** summarizes the performances of all the classifier algorithms and **Figure 2** summarizes the confusion matrix of each fitted model against the test set.

*kNN*

The best kNN classifier had a ROC of 0.58 on the training set and the optimized parameters were 36 neighbors, distance weight (points weighted on the inverse of their distance), and Euclidean distance. When applied to the test set, the ROC was 0.53. The precision, recall, and F1 scores for No POAF were 0.83, 0.10 and 0.17 respectively. The precision, recall, and F1 scores for POAF were 0.35, 0.96 and 0.52 respectively.

*LR*

The best ROC during training for LR was 0.63 with a C parameter of 0.001. When applied to the test set, the ROC was 0.63. The precision, recall, and F1 scores for No POAF were 0.76, 0.63 and 0.69 respectively. The precision, recall, and F1 scores for POAF were 0.46, 0.62 and 0.53 respectively.

*LR + RFE*

SVM

**Limitations**

This was a single institutional study, which may limit the generalizability of these results to the overall population. In addition, our study size was small to moderate and it may be that a larger dataset may boost the signal to noise ratio for these models.

**Discussion**

Our study is the first to our knowledge that attempted to study classifier machine learning algorithms in predicting new onset POAF in patients undergoing CABG. The overall incidence of POAF in our cohort was similar to previous reports1,2. Our models performed on par or slightly better than most existing models in the literature1,3,11,12,27,28. It is important to note that some of the models in the literature were not validated and included patients with a prior history of atrial fibrillation. The best model from our cohort was \*\*\*

Our poorest model was kNN with a test ROC of 0.53. This was not entirely unexpected as kNN tends to perform poorly on high dimensional data29. However, our results show a non-insignificant drop in ROC from training to test set for models including kNN and SVM. This typically indicates overfitting of the models during training, however all models were trained with a 5 fold stratified cross validation which attempted to mitigate overfitting as much as possible. Another possibility is that some of the features in the dataset did not correlate well with the outcome of interest. Even after an exhaustive iterative search with RFE, our models \*\*\*. This may be due to the fact that the variables collected in the STS registry do not correlate strongly with POAF. In fact, a recent analysis that compared existing models using the STS registry found that only the ROC area for the CHARGE-AF score (0.68) performed better than a logistic regression model of age model, while other models like CHADS2VASC, the POAF score, and the STS risk score did no better than the age model30.

Our study also highlights the importance of calculating precision, recall and conducting *k*-fold cross validation for in machine learning algorithms, as a recent article on machine learning states31. These measures are important to calculate especially in imbalanced datasets because machine learning models tend to bias towards the majority class32. Given this, the AUC is a better metric as compared to accuracy33.

1. Magee MJ, Herbert MA, Dewey TM, et al. Atrial Fibrillation After Coronary Artery Bypass Grafting Surgery: Development of a Predictive Risk Algorithm. *Ann Thorac Surg*. 2007;83(5):1707-1712. doi:10.1016/J.ATHORACSUR.2006.12.032

2. Hernandez A V., Kaw R, Pasupuleti V, et al. Association Between Obesity and Postoperative Atrial Fibrillation in Patients Undergoing Cardiac Operations: A Systematic Review and Meta-Analysis. *Ann Thorac Surg*. 2013;96(3):1104-1116. doi:10.1016/J.ATHORACSUR.2013.04.029

3. Mariscalco G, Biancari F, Zanobini M, et al. Bedside tool for predicting the risk of postoperative atrial fibrillation after cardiac surgery: the POAF score. *J Am Heart Assoc*. 2014;3(2):e000752. doi:10.1161/JAHA.113.000752

4. Al-Shaar L, Schwann TA, Kabour A, Habib RH. Increased late mortality after coronary artery bypass surgery complicated by isolated new-onset atrial fibrillation: A comprehensive propensity-matched analysis. *J Thorac Cardiovasc Surg*. 2014;148(5):1860-1868.e2. doi:10.1016/J.JTCVS.2014.05.020

5. Shen J, Lall S, Zheng V, et al. The persistent problem of new-onset postoperative atrial fibrillation: a single-institution experience over two decades. *J Thorac Cardiovasc Surg*. 2011;141(2):559-570. doi:10.1016/j.jtcvs.2010.03.011

6. Mathew JP, Parks R, Savino JS, et al. Atrial Fibrillation Following Coronary Artery Bypass Graft Surgery. *JAMA*. 1996;276(4):300. doi:10.1001/jama.1996.03540040044031

7. Aranki SF, Shaw DP, Adams DH, et al. Predictors of atrial fibrillation after coronary artery surgery: Current trends and impact on hospital resources. *Circulation*. 1996;94(3):390-397. doi:10.1161/01.CIR.94.3.390

8. Echahidi N, Pibarot P, O’hara G, Mathieu P. Mechanisms, Prevention, and Treatment of Atrial Fibrillation After Cardiac Surgery. *J Am Coll Cardiol*. 2008;51(8). doi:10.1016/j.jacc.2007.10.043

9. Kashani RG, Sareh S, Genovese B, et al. Predicting postoperative atrial fibrillation using CHA2DS2-VASc scores. *J Surg Res*. 2015;198(2):267-272. doi:10.1016/j.jss.2015.04.047

10. Sareh S, Toppen W, Mukdad L, et al. CHADS2 score predicts atrial fibrillation following cardiac surgery. *J Surg Res*. 2014;190(2):407-412. doi:10.1016/j.jss.2014.02.007

11. Amar D, Shi W, Hogue CW, et al. Clinical prediction rule for atrial fibrillation after coronary artery bypass grafting. *J Am Coll Cardiol*. 2004;44(6):1248-1253. doi:10.1016/j.jacc.2004.05.078

12. Zaman AG, Archbold RA, Helft G, Paul EA, Curzen NP, Mills PG. Atrial fibrillation after coronary artery bypass surgery: a model for preoperative risk stratification. *Circulation*. 2000;101(12):1403-1408. doi:10.1161/01.cir.101.12.1403

13. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132:1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

14. Eskelson BNI, Temesgen H, Lemay V, Barrett TM. The roles of nearest neighbor methods in imputing missing data The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases in forest inventory and monitoring databases. 2009. doi:10.1080/02827580902870490

15. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. *Comput Stat Data Anal*. 2015;90:84-99. doi:10.1016/j.csda.2015.04.009

16. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1-3):389-422. doi:10.1023/A:1012487302797

17. Kuhn M, Johnson K. *Feature Engineering And Selection: A Practical Approach for Predictive Models*. Boca Raton: CRC Press; 2020.

18. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest:  A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf Comput Sci*. 2003;43(6):1947-1958. doi:10.1021/ci034160g

19. Ghosh AK. On optimum choice of k in nearest neighbor classification. *Comput Stat Data Anal*. 2006;50(11):3113-3123. doi:10.1016/j.csda.2005.06.007

20. Supardi NZ, Mashor MY, Harun NH, Bakri FA, Hassan R. Classification of blasts in acute leukemia blood samples using k-nearest neighbour. In: *Proceedings - 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications, CSPA 2012*. ; 2012:461-465. doi:10.1109/CSPA.2012.6194769

21. Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. *Mach Learn*. 2002;46(1-3):131-159. doi:10.1023/A:1012450327387

22. Belousov AI, Verzakov SA, Von Frese J. A flexible classification approach with optimal generalisation performance: Support vector machines. *Chemom Intell Lab Syst*. 2002;64(1):15-25. doi:10.1016/S0169-7439(02)00046-1

23. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(3). doi:10.1002/widm.1301

24. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2012;2(6):493-507. doi:10.1002/widm.1072

25. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: ; 2012:154-168. doi:10.1007/978-3-642-31537-4\_13

26. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. *SMOTE: Synthetic Minority Over-Sampling Technique*. Vol 16.; 2002.

27. Mathew JP, Fontes ML, Tudor IC, et al. A Multicenter Risk Index for Atrial Fibrillation after Cardiac Surgery. *J Am Med Assoc*. 2004;291(14):1720-1729. doi:10.1001/jama.291.14.1720

28. Thorén E, Hellgren L, Jidéus L, Ståhle E. Prediction of postoperative atrial fibrillation in a large coronary artery bypass grafting cohort. *Interact Cardiovasc Thorac Surg*. 2012;14(5):588-593. doi:10.1093/icvts/ivr162

29. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “nearest neighbor” meaningful? In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 1540. Springer Verlag; 1998:217-235. doi:10.1007/3-540-49257-7\_15

30. Pollock BD, Filardo G, da Graca B, et al. Predicting New-Onset Post-Coronary Artery Bypass Graft Atrial Fibrillation With Existing Risk Scores. *Ann Thorac Surg*. 2018;105(1):115-121. doi:10.1016/J.ATHORACSUR.2017.06.075

31. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw open*. 2018;1(4):e181404. doi:10.1001/jamanetworkopen.2018.1404

32. Gu J, Zhou Y, Zuo X. Making class bias useful: A strategy of learning from imbalanced data. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 4881 LNCS. ; 2007:287-295. doi:10.1007/978-3-540-77226-2\_30

33. Ling CX, Huang J, Zhang H. AUC: A better measure than accuracy in comparing learning algorithms. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 2671. ; 2003:329-341. doi:10.1007/3-540-44886-1\_25

Training Set

Test Set

Evaluation of Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Recursive Feature Elimination and Hyperparameter Tuning at each iteration (where applicable)

Best Model Selected

**Figure 1.** Overall scheme of data processing, model training and model testing. Orange squares represent the validation fold, and the black squares represent the training folds.