Cardiothoracic Surgery Analysis for Predicting Acute Renal Failure Outcomes

Willa Potosnak[1], Anthony Wertz, MS[2], James K. Miller, PhD[2], Arman Kilic, MD[3], Keith A. Dufendach, MD[3] and Artur Dubrawski, PhD[2]

***Abstract*— Acute renal failure is a serious medical complication that can occur following coronary artery bypass grafting (CABG) surgery and can pose other serious medical complications if left undiagnosed and untreated. Risk models based on logistic regression were developed by the Society of Thoracic Surgeons (STS) to provide information on the potential mortality and morbidity outcomes of patients for cardiac surgeries. Previous work strove to improve the STS risk models with machine learning algorithms using pre-operative data similar to that used to develop the STS risk models. In this research, an intra-operative dataset with data obtained during CABG surgery was analyzed that is separate from that used to develop the STS risk models and previous work. The focus of this research was to determine through intra-operative data analysis whether surgical procedures and/or patient condition changes during surgery are associated with acute renal failure outcomes. Information from the analysis was used to generate a binary classification model for the purpose of assisting the pre-operative model in identifying patients' risk of developing renal failure following CABG surgery. The research identified 20 features of interest with significant (p-value < 0.05) deviations between renal failure (RF) and non-renal failure (NRF) patients during CABG surgery. The model accurately identifies approximately 10 percent of RF patients at a false positive rate (FPR) of 1 percent and approximately 22 percent at a higher FPR of 10 percent based on their surgical parameter and patient condition measurements. The model has the potential to be used as an overlay to the pre-operative model and current practices to help identify patients with higher risk of RF, thereby allowing clinicians to increase preventative care measures for these patients.**

I. INTRODUCTION

Coronary artery bypass grafting (CABG) is the most common type of heart surgery in the U.S. [1]. Approximately 340,000 procedures are performed each year [2]. Serious medical complications from CABG surgery can occur, including stroke, heart attack, acute renal failure, and death. This project focuses on acute renal failure, which is defined as a significant post-operative increase in serum creatinine or the post-operative requirement for dialysis [3, 4]. The complication of acute renal failure was chosen to analyze for this project due to a request from cardiothoracic surgeons for a predictive model that could provide additional insight for improving surgical parameters during and following CABG surgery for patients classified as likely to develop renal failure. Acute renal failure reduces the kidneys' ability to filter waste products and balance fluid and electrolytes. It also increases risks associated with other serious health complications, such as permanent kidney damage, if not diagnosed and treated immediately [5]. Because of the importance of renal function in maintaining homeostasis in the body, acute renal failure is an independent risk factor for post-operative mortality for patients requiring dialysis or other renal replacement therapies [3, 4].

Risk models based on logistic regression were developed by the Society of Thoracic Surgeons (STS) to provide information on the potential mortality and morbidity outcomes of cardiac surgery patients. The STS risk models provide a risk assessment capability based on patients' characteristics that enables doctors to better judge patients' fitness for specific types of cardiac surgeries. Previous work [3, 6] strove to improve predictions for 7 outcomes of the STS risk models, including post-operative acute renal failure and mortality, using pre-operative data similar to that used to develop the STS risk models. This previous work used the machine learning algorithm Extreme Gradient Boosting (XGBoost) to develop models which showed improved classification results for acute renal failure and modest improvement for mortality compared with results based on the existing STS models [3, 6].

The focus of this research was to determine through intra-operative data analysis whether surgical procedures and/or patient condition changes during surgery are associated with acute renal failure outcomes. Intra-operative data, which consist of data collected during CABG surgeries separate from the data used in developing the STS risk models and previous work, are used in the analysis. The intra-operative data employed in this project consist of patient demographics as well as surgical medications and surgical parameters recorded during CABG surgery for 362 patients.

This work is novel because it focused on identifying intra-operative feature differences between the renal failure (RF) and non-renal failure (NRF) patient classes for use in generating a binary classification model to predict post-operative acute renal failure outcomes. A binary classification model using solely intra-operative features has the potential to be used as an overlay to the pre-operative

[1]Willa Potosnak is a student in her 3rd year in the Biomedical Engineering Department at Duquesne University, Pittsburgh, PA, USA potosnakw@duq.edu

[2]Anthony Wertz, MS, James K. Miller, PhD, and Artur Dubrawski, PhD are with the Auton Lab, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA awertz@cmu.edu, mille856@andrew.cmu.edu, awd@cs.cmu.edu

[3]Arman Kilic, MD and Keith A. Dufendach, MD are with the Division of Cardiac Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA, USA kilica2@upmc.edu, dufendachka@upmc.edu

model and current practices to help identify potential RF patients. The intra-operative data used in this project contain time-series features (i.e., surgical/patient measurements), many of which are different from those collected prior to surgery. Information on the minute discrepancies in surgical parameters and patient condition changes between RF and NRF patients during surgery can be discerned through time-series analysis. Time-series analysis results can influence medication and procedures applied during and following CABG surgery to mitigate the risk of RF. Patient condition changes, especially those recorded just prior to, during and directly after cardiopulmonary bypass time, are of special interest for identifying the best surgical parameters.

The objectives of this research were to: 1) analyze intra-operative data from patients undergoing CABG surgery; 2) determine whether surgical procedures and patient condition changes are associated with renal failure outcomes; and 3) model determined features related to renal failure outcomes in an explainable format for better prediction and prevention of renal failure.

II. Data

The data used in this research consist of 362 patients who underwent isolated CABG surgery, meaning CABG is the only performed procedure. There is 4:1 propensity matching (4 NRF patients are present for every 1 RF patient) based on pre-operative risk scores generated from the STS risk model. Propensity matching is used to help remove potential bias that could cause a classification model to overfit the data and perform sub-optimally when applied to patients external to the project data. 75 patients developed acute renal failure following surgery and comprise the test group while the other 287 patients did not develop acute renal failure and comprise the control group. It is important to note that surgery duration varies for each patient and only half of the patients in the data have surgery durations that exceed approximately 4.5 hours as shown in Figure 4. Furthermore, CABG surgical procedures are not time-specific, but depend on the condition of the individual patient. The data subset used in the analysis consists of patient condition measurements, surgical parameters, and the top 5 medication features that affect heart rate and blood pressure. Negative time values indicate that the measurement was recorded prior to the first incision which occurs at 0 minutes and is specified in Figures 3 and 4. Patient condition measurement and surgical parameter data were forward filled for each patient by propagating the last valid measurement forward to avoid data sparsity given these features are continuous [7]. Medication features were incorporated into the dataset for the specified infusion start and stop times.

III. Analysis Methods

Patient condition, surgical parameter and medication measurements were incorporated into a dataset as individual features. An analysis of these features was conducted to discern if feature values differed between RF and NRF patients. Several methods were used for the analysis. Median value time-series plots were generated for features that showed distinct value separation for RF and NRF patients. P-values were generated for each feature using the Kolmogorov-Smirnov test on the distributions for RF and NRF patients. This tests the null hypothesis that two independent samples are drawn from the same continuous distribution. Features with significant (p-value < 0.05) deviations between classes indicate that only less than 5 percent of the time would the same distribution generate the two class samples. P-values less than 0.05 were used to help confirm features of interest and are shown in Table I. Time-series p-value plots were also generated for each feature using the Kolmogorov-Smirnov test for the two patient class distributions at each minute. These provide more minute visualizations of significant deviations between classes throughout surgery duration.

A separate analysis of medication features was conducted using the Fisher Exact test to determine if there is 1) a statistically significant association between patient class and the presence of specific medication features; and/or 2) a statistically significant association between patient class and the 20 percent highest and lowest total medication amounts administered. The p-values for these tests are shown in Table II. Medication features were not used to train the model due to their sparsity and poor effect on model performance.

6 classifiers were tested with various dataset adjustments. The classifiers tested include Logistic Regression, Random Forest, Extremely Randomized Trees (Extra Trees), Gaussian Naïve Bayes, K-Nearest Neighbors (KNN), and Quadratic Discriminant Analysis (QDA). Logistic Regression has been the model of choice for cardiac surgery risk modeling, such as the STS models [7, 8]. Even with the high calibration seen with the STS models using logistic regression [14], there are downsides to logistic regression as it requires a linear relationship between covariates and is prone to overfitting for multicollinear and large datasets like the one used in this research [9]. For these reasons, additional classifiers were tested to determine the most optimal classifier for this dataset. 10-fold cross-validation was applied to the data when training and testing the model: the dataset was split into 10 groups, the model was trained on 9 groups, and then tested on 1 group with this this process repeating a total of 10 times.

The performance of each classifier in predicting the RF class was assessed based on the true positive rate (TPR) percent values at false positive rates (FPR) of 0.01 and 0.10 percent as this indicates how well the model classifies true positives (RF patients) while still having a low FPR (i.e., the rate of classifying NRF patients as RF patients). The TPR at a low FPR can be seen clearly on an ROC Curve with the x-axis set to the $\log_{10}$ scale as shown in Figure 2 for the classifier with the best performance. The true negative rate (TNR) percent value at a false negative rate (FNR) of 0.01 percent was also taken into consideration as to how well the model classifies true negatives (NRF patients) while still having a low FNR (i.e., the rate of classifying RF patients as NRF patients). While the the main objective was to classify RF patients, a second application

of the model for helping to clear patients from consideration of developing RF would also be useful for this objective. Various dataset adjustments were tested with these classifiers. The main dataset adjustments that were used to determine the best model are:

1) Computing a rolling standard deviation for each of the original numeric attributes to generate new features for model training, referred to as featurization. This serves to generate features that indicate patient condition measurements that deviate from the mean feature values. This technique is used as classifiers may discern class differences better for certain features in this format.

2) Using scikit-learn [10] Robust Scaler with pre-operative data, or data collected in the 10-minute window prior to each patient's surgery start time. Robust Scaler removes the median and scales the data to the quantile range making it robust to outliers present in the dataset. The values from the 10-minute windows serve as patient baseline values that when applied to scale the dataset, result in features that indicate patient conditions that deviate from baseline values during surgery.

3) Applying the scikit-learn [10] Recursive Feature Elimination (RFE) tool for specific classifiers and using only the RFE-selected features when training and testing the model. When applied to the training set of data, RFE eliminates features that are least important to the model and returns features with the highest importance for the specified estimator (classifier). This tool can help improve model performance by reducing noise in the dataset due to sparse and/or irrelevant features.

Additional adjustments include training the model at certain time intervals that showed improved class separation. Time intervals with improved class separation were determined through an analysis of the median predicted positive probabilities of the patients: predicted positive probabilities for RF patients close to 1 and a predicted positive probabilities for NRF patients close to 0 is ideal. The median positive probability plot which assessed all patients' probabilities of developing renal failure throughout the surgery duration for the classifier with the best performance is shown in Figure 3.

IV. Results

Features with significant (p-value < 0.05) deviations between classes, meaning that only less than 5 percent of the time would the same distribution generate the two class samples, are listed in Table I with their respective p-values and difference between mean values. Classifiers trained and tested on only features with p-values considered significant showed poor classification performance and a low percentage of correctly identified RF patients.

A specific analysis of medication features using p-values generated from the Fisher Exact Test showed that the presence of medication features for either class is not statistically significant, nor is the number of RF patients in the groups of patients administered the 20 percent highest and lowest total dosages/volumes. The medication features and their p-values are shown in Table II. P-value 1 indicates statistical significance between patient class and the presence of specific medication features. P-values 2 and 3 indicate statistical significance between patient class and the 20 percent highest and lowest total medication amounts administered, respectfully.

The classifier that showed the best model performance is Extra Trees with additional standard deviation features for each feature in the dataset, no scaling of the dataset with pre-operative data and the dataset containing only RFE-selected features. This model is shown in Figure 1. The top 30 RFE-selected features and their feature importances as determined by Extra Trees are shown in Figure 5. Extra Trees shows the best model performance in terms of having the highest TPR at an FPR of 0.01 percent out of all classifiers tested for this combination of dataset adjustments. This classifier accurately identifies approximately 10 percent of RF patients at an FPR of 1 percent and approximately 22 percent at a higher FPR of 10 percent as shown in the first plot in Figure 1, which is enlarged in Figure 2. In addition, it accurately identifies approximately 10 percent of NRF patients at an FNR of 1 percent as shown in the third plot of Figure 1. The approximation in identification is due to the randomized nature of Extra Trees in terms of selected features and cut-point choice that is explained further in [11]. Class separation based on the probability estimates for the positive (RF) class as determined throughout surgery duration by the Extra Trees classifier is shown in Figure 3. Distinct class separation, especially in the 3-7 hour time interval, indicates that the classifier can discern a distinction between patients, and that other machine learning methods may improve model performance.

Significant features were used to assess credibility of the Extra Trees model in classifying patients based on its ranked feature importances as shown in Figure 5. Of the top 30 features with the largest importances out of all 54 RFE-selected features, 9 features with significant deviations between classes were present.

V. Discussion

The model has the potential to be used as an overlay to the pre-operative model and current practices to help identify patients with higher risk of RF, thereby allowing clinicians to increase preventative care measures for these patients. In addition, this classifier can help identify NRF patients, which can allow clinicians to better allocate preventive measures to patients who show higher risk of RF as well as those not identified by the classifier. Moreover, the model performance indicates that an entire intra-operative time series analysis may not be the best approach, and that the surgery time-series analysis should be partitioned into sections based on surgical procedures. Because there is a large variation in patient surgery durations as shown in Figure 4, there could be distinct variations in patient conditions at their respective stages. Including the entire intra-operative time series in the analysis could be convoluting the data. So,

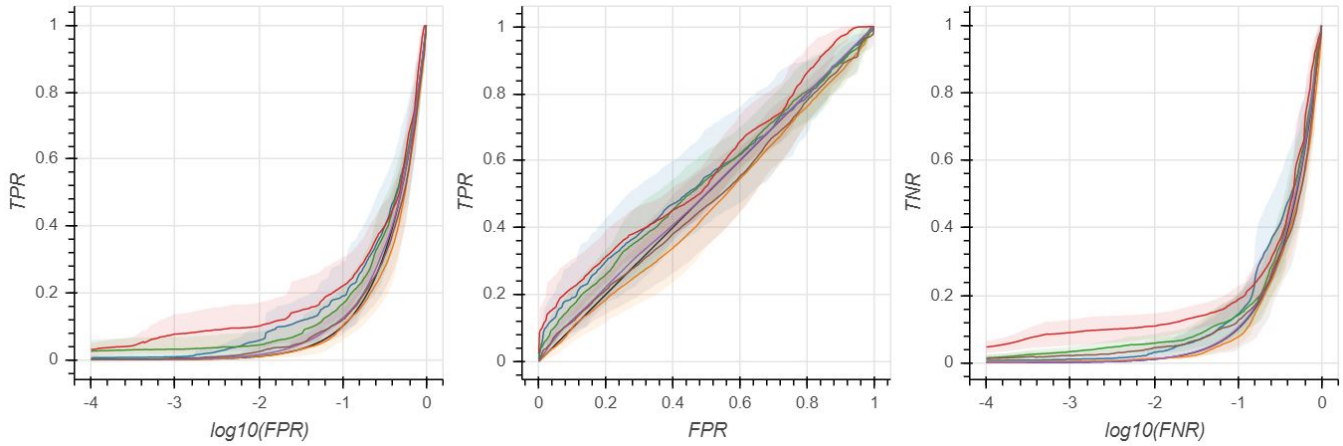


Fig. 1: ROC curves for all 6 classifiers. Shaded region is the standard error.

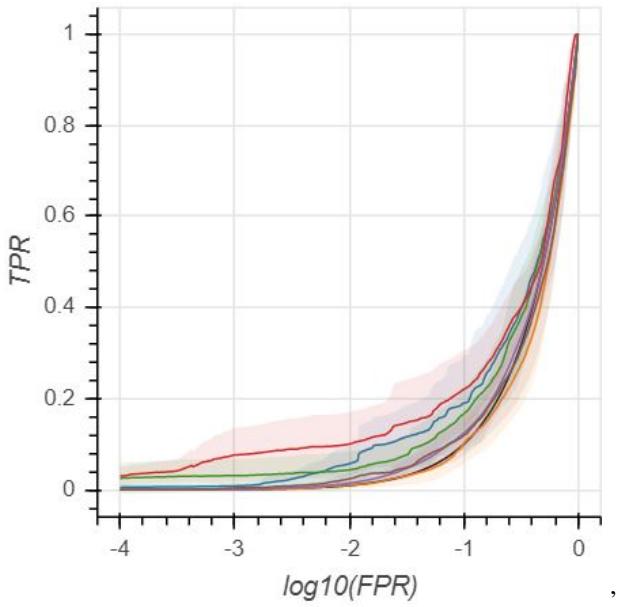
Feature	P-Value	Mean Difference
Stroke Volume	-7.26	0.62
RV End Diastolic Volume	-6.88	2.04
Pulse Pressure-Blood	-5.49	2.14
Central Venous Pressure	-4.54	0.76
NIRS Cerebral Oxygenation-L	-3.96	2.37
Mean Blood Pressure	-3.96	0.76
SvO2	-3.69	1.12
Systolic Blood Pressure	-3.69	1.45
Oxygen Percent (FiO2)	-3.69	1.43
RV Ejection Fraction	-3.42	0.50
Arterial Diastolic Pressure	-3.17	0.70
BIS Value	-3.17	1.28
Heart Rate-Pleth	-2.92	3.41
Diastolic Blood Pressure	-2.92	0.68
NIRS Cerebral Oxygenation-R	-2.46	1.90
Heart Rate	-2.24	1.78
Mean Arterial Pressure	-2.24	0.55
Pulse Pressure-Arterial	-2.03	1.10
Pulmonary Artery Mean	-1.83	0.23
Epinephrine 64 Dose	-1.64	3.13

TABLE I: $\log_{10}$(p-values) and mean differences between RF and NRF patients for significant features

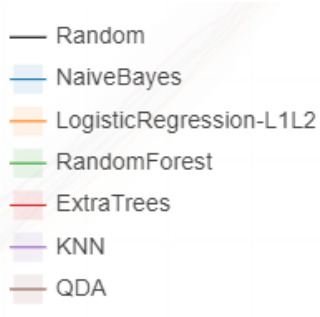
Feature	P-value 1	P-value 2	P-value 3
Phenylephrine Volume	1.00	0.712	0.856
Phenyleprhine Dosage	1.00	0.461	0.856
Vasopressin Volume	0.007	0.545	0.360
Vasopressin Dosage	0.007	0.545	0.360
Epinephrine 10 mcg/mL Volume	0.018	0.526	1.00
Epinephrine 10 mcg/mL Dosage	0.018	0.526	1.00
Epinephrine 64 mcg/mL Volume	0.019	0.377	0.517
Epinephrine 64 mcg/mL Dosage	0.027	0.828	1.00
Norepinephrine Volume	0.005	0.450	0.450
Norepinephrine Dose	0.005	0.205	0.405
Albumin 5 percent Volume	0.007	1.00	0.819

TABLE II: P-values for medication features. None are considered statistically significant after Bonferroni correction (p-value = 0.0045).

while the model correctly identifies approximately 10 percent of RF patients at a FPR of 1 percent and approximately 22 percent at a higher FPR of 10 percent, partitioning the data based on procedures within the surgery could improve



(a) Plot 1 of Figure 1 on a larger scale. Approximately 10 percent accurate identification of RF patients at a FPR of 0.01 percent and 22 percent at an FPR of 0.10 percent for Extra Trees Classifier



(b) ROC curve legend

Fig. 2: Extra Trees classifier has the best ROC curve results out of the 6 tested classifiers

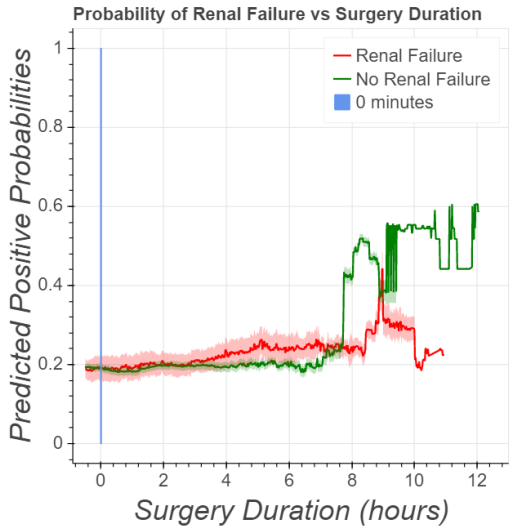


Fig. 3: Positive probability estimates for the optimal model with Extra Trees classifier. 0 minutes indicates first incision of surgery.

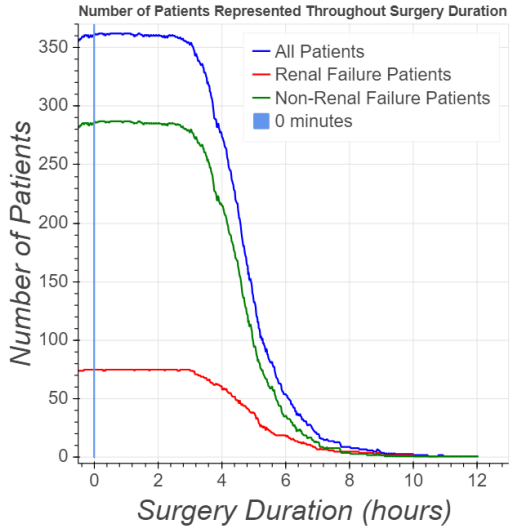


Fig. 4: The number of patients present throughout surgery. 0 minutes indicates first incision of surgery.

model performance. Partitioning may also provide results that clinicians can more easily use to discern what and when surgical parameters should be adjusted.

The model performance could also indicate that features currently collected during CABG surgery may not be representative of renal function. Collecting and evaluating features more directly correlated to renal function during CABG surgery could improve the intra-operative analysis and, ultimately, the model performance. For example, serum creatinine level directly indicates glomerular filtration rate, which directly correlates to renal function [9, 13, 15]. Glomerular filtration rate decreases as renal function decreases, leading to an increased serum creatinine level, which is a strong risk factor for RF [16]. Creatinine level is

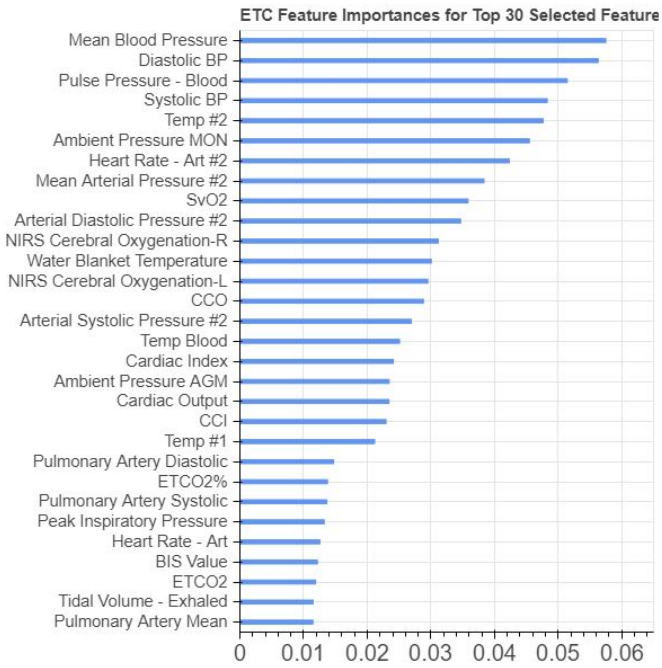


Fig. 5: Extra Trees RFE selected features ranked by importance determined by the classifier with respect to all features

measured pre-operatively and used in pre-operative models to predict RF outcomes [3, 9, 12, 14, 15]; it is also measured post-operatively to monitor RF, so its inclusion in time-series intra-operative data could potentially improve model performance.

VI. Future Work

Future work will involve partitioning the data for each patient into 5 stages based on CABG surgical procedures. Recorded features and their values will indicate the start and end points of the 5 stages for each patient. An analysis of patient condition, surgical parameter and medication features will be performed for each stage to better assess patient changes regarding specific events during CABG surgery. Classifiers will be trained on each of the surgical stages to refine the model and minimize possible data convolution. The XGBoost algorithm will also be applied as it has improved model performance as shown for work in [3, 6, 9].

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1659774. Thank you to Carnegie Mellon Robotics Institute for funding this research. A special thanks to Rachel Burcin and Dr. John Dolan for their efforts to help make the RISS program possible this summer via a virtual approach.

References

[1] "Bypass Surgery Shows Advantage," *nih.gov*, 2012.
[2] "New Study Shows Approximately 340,000 CABG Procedures per Year in the United States," *idataresearch.com*, 2018.

[3] A. Kilic, A. Goyal, J. K. Miller, T. G. Gleason, and A. Dubrawski, "Performance of a Machine Learning Algorithm in Predicting Outcomes of Aortic Valve Replacement," *The Annals of Thoracic Surgery*, 2020.

[4] Duca, S. Iqbal, E. Rahme, P. Goldberg, and B. Varennes, "Renal Failure After Cardiac Surgery: Timing of Cardiac Catheterization and Other Perioperative Risk Factors," *The Annals of Thoracic Surgery*, vol. 84, no. 4, p. 1264-1271, 2007.

[5] "Acute kidney failure," *mayoclinic.org*.

[6] A. Kilic, A. Goyal, J. K. Miller, E. Gjekmarkaj, W. Lam Tam, T. G. Gleason, I. Sultan, and A. Dubrawski, "Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery," *The Annals of Thoracic Surgery*, vol. 109, no. 6, p. 1811-1819, 2020.

[7] D. M. Shahian, E. H. Blackstone, F. H. Edwards, F. L. Grover, G. L. Grunkemeier, D. C. Naftel, S. A.M. Nashef, W. C. Nugent, and E. D. Peterson, "Cardiac Surgery Risk Models: A Position Article," *The Annals of Thoracic Surgery*, vol. 78. no. 5, p. 1868-1877, 2004.

[8] D. M. Shahian, J. P. Jacobs, V. Badhwar, L. Feng, X. He, S. M. O'Brien, et al., "The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development," *The Annals of Thoracic Surgery*, vol. 105, no. 5, p. 1411-1418, 2018.

[9] Lee, H. Yoon, K. Nam, Y. J. Cho, T. K. Kim, W. H. Kim, and J. Bahk, "Derivation and Validation of Machine Learning Approaches to Predict Acute Kidney Injury after Cardiac Sugery," *Journal of Clinical Medicine*, vol. 7, no. 322, 2018.

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

[11] P.Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, p. 3-42, 2006.

[12] C.V. Thakar, S. Arrigain, S. Worley, J. Yared, and E. Paganini, "A clinical score to predict acute renal failure after cardiac surgery," *Journal of American Society of Nephrology: JASN*, vol. 16, no. 1, 2005.

[13] T. Coulson, M. Bailey, D. Pilcher, C. M. Redi, S. Seevanayagam, J. Williams-Spence, and R. Bellomo, "Predicting Acute Kidney Injury After Cardiac Surgery Using a Simpler Model," *Journal of Cardiothoracic and Vascular Anesthesia*, 2020.

[14] S. M. O'Brien, L. Feng, X. He, N. D. Desai, F. H. Edwards, D. M. Shahian, et al., "The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2-Statistical Methods and Results," *The Annals of Thoracic Surgery*, vol. 105, no. 5, p.1419-1428, 2018.

[15] J. Chikwe, J.G. Castillo, P.B. Rahmanian, A. Akujuo, D. H. Adams, and F. Filsoufi, "The impact of moderate-to-end-stage renal failure on outcomes after coronary artery bypass graft surgery," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 24, no. 4, p. 574-579, 2010.

[16] R. Bellomo, C. Ronco, J. A. Kellum, R. L. Mehta, and P. Palevsky, "Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group," *Critical Care*, vol. 8, no. 4, 2004.