Health. Innovation. Community: It Starts With Us
J. Bichel-Findlay (Ed.)

doi:10.3233/SHTI240908

Predicting Unplanned Return to Operating Room Following Primary Total Shoulder Arthroplasty: Insights from Fair and Explainable Ensemble Machine Learning

Annie KIM [a], Hongtao WANG [b], Nicole MYERS [a], Puneet GUPTA [c], Fritz STEUER [d], Michael R KANN [d], Ting CONG [e], Hongfang LIU [f] and Ahmad P TAFTI [a, b1]
[a] *School of Health and Rehabilitation Sciences, University of Pittsburgh, PA, USA*
[b] *School of Computing and Information, University of Pittsburgh, PA, USA*
[c] *School of Medicine, George Washington University, DC, USA*
[d] *School of Medicine, University of Pittsburgh, PA, USA*
[e] *University of Pittsburgh Medical Center, PA, USA*
[f] *University of Texas Health Science Center at Houston, TX, USA*

ORCiD ID: Annie Kim https://orcid.org/0000-0001-7887-0872, Hongtao Wang https://orcid.org/0009-0008-4333-6440, Nicole Myers https://orcid.org/ 0009-0003-5555-2005, Puneet Gupta https://orcid.org/0000-0001-8274-6970, Fritz Steuer https://orcid.org/0009-0003-4155-2979, Michael Kann https://orcid.org/0009-0000-6362-2083, Hongfang Liu https://orcid.org/0000-0003-2570-3741, Ahmad P. Tafti https://orcid.org/0000-0001-9650-2862

Abstract. Reoperation is the most significant complication following any surgical procedure. Developing machine learning methods that predict the need for reoperation will allow for improved shared surgical decision making and patient-specific and preoperative optimisation. Yet, no precise machine learning models have been published to perform well in predicting the need for reoperation within 30 days following primary total shoulder arthroplasty (TSA). This study aimed to build, train, and evaluate a fair (unbiased) and explainable ensemble machine learning method that predicts return to the operating room following primary TSA with an accuracy of 0.852 and AUC of 0.91.

Keywords. total shoulder arthroplasty, TSA, fair and explainable machine learning

1. Introduction

Total shoulder arthroplasty (TSA) is a well-established surgical procedure that aims to replace impaired parts of the shoulder joint with implants. While TSA has already demonstrated substantial improvements within the last few years, reoperation might still be needed following primary TSA [1]. The current study aimed to (1) computationally

[1] Corresponding Author: Ahmad P. Tafti, tafti.ahmad@pitt.edu.

assemble a TSA cohort using the American College of Surgeons National Surgical Quality Improvement Database (NSQIP), (2) train, test, and evaluate an ensemble machine learning method equipped with fairness and explainability using the argmax of the sums of the predicted probabilities provided by Random Forest, Bagged Logistic Regression as the base model, Support Vector Machine, and AdaBoost.

The present study is expected to open several research avenues to incorporate multi-modal data to inquire about surgical outcome(s) following primary TSA.

2. Methods

A cohort of 19,055 patients in the American College of Surgeons National Surgical Quality Improvement Database (NSQIP) who underwent primary TSA between 2016 and 2020 were computationally assembled. Thirty-seven predictors, including basic demographics, preoperative and intraoperative variables, comorbidities, and laboratory results, were employed to predict 30-day unplanned reoperation following primary TSA. Utilising pre- and post-processing steps (e.g., eliminating missing values) resulted in 525 TSA patients, of which 153 patients experienced reoperation within 30 days and 372 did not. Bootstrapping resampling was then utilised to tackle the problem of the imbalanced dataset, which resulted in 372 patients with reoperation within 30 days and 372 without.

The next step involved training, testing, and evaluating a fair and explainable ensemble machine learning (ML) model to predict unplanned reoperation following primary TSA using the balanced dataset. The argmax of the sums of the predicted probabilities generated by Random Forest, Bagged Logistic Regression as the base model, Support Vector Machine, and AdaBoost were used to build an ensemble ML model.

3. Results

3.1 Fairness Visualisation

A difference plot is presented (Figure 1) that visualises the impact of the fairness interventions, specifically using the exponentiated gradient reduction (EGR) algorithm [2], on the accuracy of a range of ML models. This comparison is made within the equalised odds fairness constraint framework, which aims to achieve fair predictions for protected groups (race and sex). The initial accuracy (blue) in Figure 1 reflects the model's performance without any fairness correction, while the subsequent accuracy (orange) shows the performance after recalibration following the equilibrium odds constraint.

Notably, the reduction in accuracy after fairness interventions reflects trade-offs often encountered when enhancing the fairness of the predictive models. For example, the significant decrease in accuracy for the Bagging model, which indicates a significant change in prediction results after the intervention, suggests that the fairness algorithm significantly impacts the model's decision boundaries. In contrast, models such as Random Forest and Voting showed less pronounced changes, suggesting a lower level of bias or a more subtle effect of the EGR algorithm. The XGBoost model also maintained relatively stable performance, implying that fairness interventions had a negligible impact on predictive accuracy.

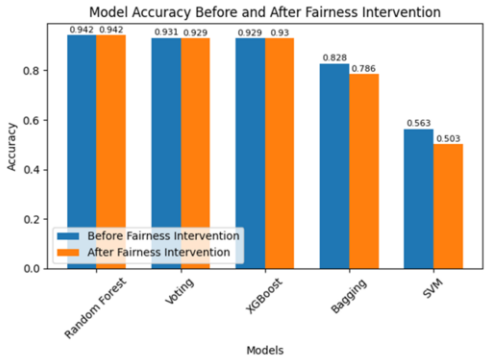

Figure 1. Disparity map using the EGR algorithm

3.2 Explainability Visualisation

The distribution of SHAP values for each feature across all predictions made by the model is illustrated in Figure 2. Features are ranked in descending order of importance, as determined by the magnitude of their SHAP values. The colour representation indicates the feature value: high (in red) or low (in blue). This colour coding provides additional context, highlighting the relationship between the value of a feature and its impact on the model's output.

The AGE is the most important feature with consistently negative SHAP values, indicating that regardless of age, the predicted probability of reoperation decreases uniformly across all ages. This uniform negative impact suggests that the model does not differentiate between different age ranges and applies a blanket decrease in probability, highlighting a potential area for model refinement to capture more nuanced patterns. DIABETES, SEX, and BMI substantially influence the model's output considerably. BMI predominantly has positive SHAP values, indicating it generally increases the predicted probability of reoperation. SEX shows a mix of positive and negative SHAP values but with a more concentrated distribution, suggesting specific gender differences influence the model in different directions.

In addition, LIME was employed to deconstruct the decision-making process of the bagging model for individual predictions. The LIME explanation plot (Figure 3) provides an instance-specific explanation for a prediction made on the need for reoperation within 30 days following primary TSA.

The left side of the plot displays the prediction probabilities, with the model estimating an 87% probability of the outcome being "Yes" and a 13% probability of being "No." The central bar chart shows the contribution of each feature to the predicted outcome. Blue indicates a positive contribution to the "No" classification and orange indicates a contribution to the "Yes" classification. For example, the higher feature value "BMI" significantly increases the probability of a "Yes" prediction, reflecting its strong impact on model evaluation. The right side of the bar chart details the actual values of the patient's features, allowing a direct comparison of the impact of each feature value on the prediction.

Figure 2. SHAP Summary Plot

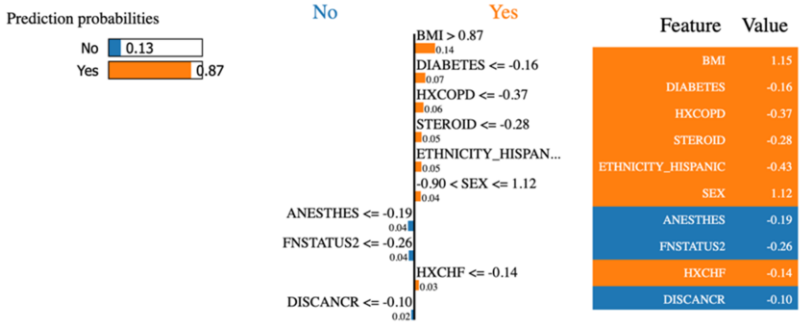

Figure 3. LIME explanation plot

4. Discussion

The balanced dataset was randomly grouped into 80% train and 20% test sets. The proposed pipeline resulted in an accuracy of 0.852 and an AUC of 0.91 (Figure 4). The weighted average of F1 score was 0.85. The proposed ensemble model outperformed each of those four individual classifiers. While the proposed pipeline demonstrates the successful application of an ensemble ML method, future works will focus on analysing

the fairness of the predictive models by integrating any available social determinants of health.

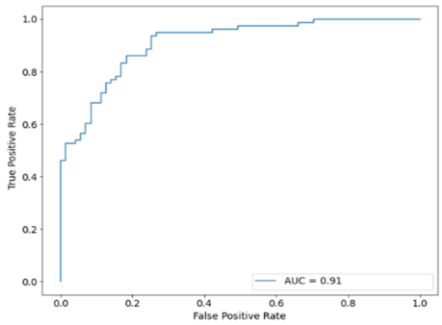

Figure 4. Proposed ensemble machine learning pipeline resulted in an AUC of 0.91.

5. Conclusions

The development and validation of an unbiased and explainable ensemble machine learning model for predicting 30-day unplanned reoperation following primary total shoulder arthroplasty (TSA) represent a significant advancement in surgical outcome prediction. This study successfully utilised a balanced dataset, comprehensive feature selection, and rigorous model evaluation, integrating fairness constraints and explainability techniques such as SHAP and LIME. The proposed AI model(s) performs well and offers transparent insights into the factors influencing reoperation risks. Future research should incorporate additional data sources, including social determinants of health, to further refine the model's fairness and predictive power. Moreover, implementing this model in clinical settings can enhance preoperative planning and patient-specific risk assessment, ultimately improving surgical decision-making and patient outcomes. The promising results of this study lay the groundwork for ongoing advancements in the application of machine learning to orthopedic surgery and beyond.

References

[1] Arvind V, London DA, Cirino C, Keswani A, Cagle PJ. Comparison of machine learning techniques to predict unplanned readmission following total shoulder arthroplasty. J Shoulder Elbow Surg. 2021 Feb;30(2):e50-e59, doi: 10.1016/j.jse.2020.05.013.

[2] Li C, Ding S, Zou N, Hu X, Jiang Z, Zhang K. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. J Biomed Inform. 2023 Jul;143:104399, doi: 10.1016/j.jbi.2023.104399.