Prediction of Acute Kidney Injury in Cardiac Surgery Patients: Interpretation using Local Interpretable Model-agnostic Explanations

Harry Freitas da Cruz, Frederic Schneider and Matthieu-P. Schapranow
Hasso Plattner Institute, Digital Health Center, Rudolf-Breitscheid-Straße 187, Potsdam, Germany

Keywords: Clinical Prediction Models, Supervised Learning, Interpretability, Nephrology, Acute Kidney Injury.

Abstract: Acute kidney injury is a common complication of patients who undergo cardiac surgery and is associated with additional risk of mortality. Being able to predict its post-surgical onset may help clinicians to better target interventions and devise appropriate care plans in advance. Existing predictive models either target general intensive care populations and/or are based on traditional logistic regression approaches. In this paper, we apply decision trees and gradient-boosted decision trees to a cohort of surgical heart patients of the MIMIC-III critical care database and utilize the locally interpretable model agnostic approach to provide interpretability for the otherwise opaque machine learning algorithms employed. We find that while gradient-boosted decision trees performed better than baseline (logistic regression), the interpretability approach used sheds light on potential biases that may hinder adoption in practice. We highlight the importance of providing explanations of the predictions to allow scrutiny of the models by medical experts.

1 INTRODUCTION

Heart patients often have to undergo surgical interventions during the course of the disease. Particularly surgeries utilizing a cardiopulmonary bypass place a significant burden on the patient's kidneys and may lead to Acute Kidney Injury (AKI). This condition occurs in up to 30% of patients following cardiosurgical treatment and is associated with complications such as sepsis, an increase of in-hospital and long-term morbidity, and generally poor patient outcomes (O'Neal, Jason and others, 2016).

Identifying patients at high risk for developing AKI before the surgical intervention can assist care providers in adopting targeted renal-protective strategies, such as increasing renal blood flow and avoidance of nephrotoxins (Rosner and Okusa, 2006). While there is little consensus on drugs that can effectively prevent AKI onset in heart patients, early detection can furthermore be of value for preoperative patient management and clinical trial recruitment (Ng et al., 2014). Therefore, a number of studies have been targeted at developing appropriate risk scores and Clinical Prediction Models (CPM).

Previous work dealing with the task of predicting heart surgery-associated AKI take into account biomarkers and/or clinical data before, during and after the surgical intervention. Particularly concerning biomarker-based approaches, measurements of interest are usually taken after the surgery (Sawhney et al., 2015), thus posing barriers for use prior to the intervention. In this paper, we derive a CPM which utilizes only preoperative variables in order to predict the onset of AKI and compare our results to models which sought to perform the same task. The available models published to date tend to employ preoperative variables are often based on linear approaches. While the

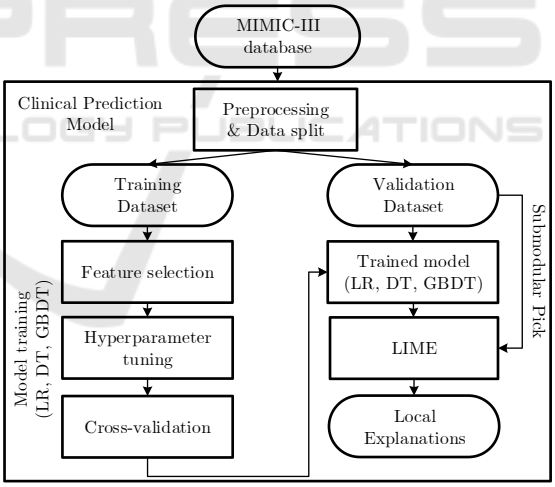

Figure 1: Graphical abstract depicting the set-up of the experiment in this paper as a Fundamental Modeling Concepts block diagram.

Freitas da Cruz, H., Schneider, F. and Schapranow, M.
Prediction of Acute Kidney Injury in Cardiac Surgery Patients: Interpretation using Local Interpretable Model-agnostic Explanations.
DOI: 10.5220/0007399203800387
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 380-387
ISBN: 978-989-758-353-7

results thus achieved are satisfactory, we hypothesize that is possible to improve them even further using machine learning approaches.

Indeed, many complex machine learning and prediction modeling techniques, such as neural networks and ensemble methods, have been shown to outperform linear modeling approaches in terms of accuracy and precision, but they typically lack interpretability. This trade-off is a critical hindering factor for the adoption of such models in high-stakes domains (Valdes et al., 2016; Letham et al., 2015; Katuwal and Chen, 2016). Therefore, we employ the interpretability method Local Model-agnostic Explanations (LIME) to provide intelligible explanations for the prediction results (Ribeiro et al., 2016).

Our CPM is based on pre-surgery clinical data obtained from the MIMIC-III clinical care database. For this task, we compare the performance of three different prediction algorithms: a more readily intelligible Decision Tree (DT) model, a blackbox-type model Gradient Boosted Decision Tree (GBDT) along with a Logistic Regression (LR) baseline model. The models thus trained are internally validated on a held-out dataset. For interpretation, we select specific classification instances to be explained by LIME. This paper's set-up is illustrated by the Fundamental Modeling Concepts block diagram in Figure 1.

The remainder of this work is structured as follows: Section 2 discusses the context of interpretable classification and prediction models specifically in medicine as well as the state-of-the-art algorithms for clinical prediction models. Section 3 details our approach to data acquisition, preprocessing, modeling, and use of LIME. In Section 4 we present an overview of our model's prediction quality to be able to, in Section 5, relate these results to the interpretability of our two employed models and results.

2 RELATED WORK

2.1 Cardiac Surgery-associated AKI

Research regarding AKI and its occurrence after cardiac surgery is typically focused on detecting in-serum and urinary biomarkers, e.g. serum creatinine (Latini et al., 2016; Flynn and Dawnay, 2015) and neutrophil gelatinase-associated lipocalin. Concerning prediction models using electronic health data, an early approach has been the Cleveland score (Thakar et al., 2004) derived from a large cohort of open-heart surgery patients, which was followed by the publication of the AKICS score (Palomba et al., 2007) based on a cohort of Brazilian patients. In a multicentric, multinational study, Mehta et al developed a score using the National Cardiac Surgery Database of the Society of Thoracic Surgeons (STS), achieving satisfactory results. The Simplified Renal Index (SRI) strived to achieve a succinct set of predictors, but ultimately performed worse than the Cleveland score on a validation cohort of the Mayo Clinic (Wijeysundera et al., 2007). Since the publication of those scores, a literature review recommended the Cleveland score, as it is the most frequently validated score (Huen and Parikh, 2012). However, the Cleveland score presented poor discrimination metrics when validated in a Chinese cohort, suggesting limited generalizability for populations not predominantly Caucasian (Jiang et al., 2017).

The models developed until now have been based on logistic regression, with the AKICS score presenting the best performance upon derivation. Even considering possible overfitting effects, the Area Under the Curve (AUC) of most models range from 0.74 to 0.84 (Huen and Parikh, 2012). By using machine learning algorithms as opposed to simple logistic regression, we achieved better discriminative performance for preoperative AKI prediction than extant models (AUC=0.9). An overview of the results achieved in comparison with the literature is provided in Table 2.

2.2 Model Interpretability

Even though blackbox models may promise better results, specifically in high-stakes domains such as medical care, the trade-off between model performance and intelligibility results in domain experts and professionals favoring interpretable prediction models with verifiable outcomes over opaque, machine learning models (Caruana et al., 2015). As such, in addition to applying new algorithms on the problem, we employ the LIME explainer to lend intelligibility to the algorithms' predictions.

Model interpretability is a research topic gaining traction, partially due to ethical concerns and laws regulating the use of machine learning techniques on individual-related data (Goodman and Flaxman, 2016). The notion of interpretability, however, is not yet well-defined and publications often present different characteristics and desiderata for interpretable models. In this paper, we adopt the notion of post-hoc interpretability as defined by Lipton, i.e. the availability of indirect information about a model's behavior or specific results (Lipton, 2016). In effect, LIME provides exactly such post-hoc interpretability on a single-prediction basis using local explanations (Ribeiro et al., 2016).

3 METHODS

In the following, we describe the methodological setup for the development of our CPM, depicted in Figure 1. We provide implementation details, such as relevant software libraries, data used for model training and validation, data preprocessing steps, employed prediction models and the interpretability method LIME.

3.1 Experimental Setup

As depicted in Figure 1, we utilized a cohort of intensive care patients from the MIMIC-III critical care database (Johnson et al., 2016). From this cohort, we extracted an initial feature set based on expert consultation and analysis of literature. Following a number of preprocessing steps and data split to obtain training and validation datasets following the 80:20 ratio, we proceeded to train tree models, DT, GBDT and LR as baseline. For each of these algorithms, we applied a feature selection step using univariate analysis deciding to retain features where $p < .001$. Subsequently, we performed hyperparameter optimization with grid-search using 10-fold cross-validation as score. The models thus trained were then validated on a held-out dataset comprised of 20% of the original dataset. The 10-fold cross-validation discrimination and calibration metrics for each of the algorithms are compared side-by-side. The LIME explainer takes as input the trained classifier and an instance for explanation. LIME enabled us to analyze the results achieved not only in terms of raw performance but also in terms of their medical adequacy.

We implemented all components of the CPM using the Python programming language at version 3.6.1 (Rossum and Drake, 2010). For data handling, loading and storing, as well as preprocessing we have made extensive use of the Python library Pandas at version 0.23.4 (McKinney, 2010). Furthermore, we used the DT, GBDT, and LR classifier implementations of the Python machine learning library scikit-learn at version 0.19.1 for model development and evaluation.

3.2 Patient Cohort

The MIMIC-III critical care database contains close to 59,000 hospital admissions that were recorded over an eleven year period at the Beth Israel Deaconess Medical Center in Boston, MA, USA (Johnson et al., 2016). From this data we selected a cohort of 6,782 admissions of adult patients who underwent cardiac surgery during their hospital stay for use as labeled

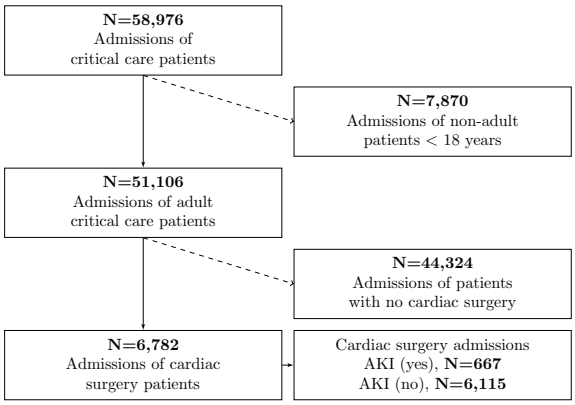

Figure 2: Cohort diagram of relevant cases from the MIMIC-III clinical care database. The data corresponding to the depicted 6,782 patient admissions is used for model training and validation of the proposed prediction method.

training and validation data, as depicted in Figure 2. Surgery cases comprise coronary artery bypass graft and aortic valve repair and/or replacement.

As concerns target variable for the prediction task, we defined AKI according to the Acute Kidney Injury Network (AKIN) classification occurring after the surgery. The AKIN classification is used for diagnosing AKI and ranks patients' stages of AKI from 0 to 3, 0 being no injury and 3 being the most severe, often indicating complete renal failure. The classification is based on patients' measured serum creatinine and urine output, and does consider if patients receive renal replacement therapy (Lopes and Jorge, 2013). In our cohort, the incidence of AKI was approximately 10%, agreeing with general clinical observations (O'Neal, Jason and others, 2016).

As per Figure 2, the classes of patients who did and did not develop AKI following their surgery are not equally distributed. The data exhibit considerable class skew, as the patient cohort comprises ca. 10 times as many negative class cases. Given this class skew, the robustness of training models towards skewed training data must be considered.

3.3 Initial Feature Set

The initial feature set from the MIMIC-III database was derived from consultations with medical experts and analysis of extant literature as laid out in Section 2. The input features include demographical patient data such as patients' age and gender, binary variables indicating the presence of certain comorbidities, such as diabetes, and laboratory values. In addition to the specific comorbidities, we also computed and included the Elixhauser comorbidity score as a further feature. This score is a scalar value that is used to assess a patient's prognosis during their hos-

pital stay based on the presence or absence of 30 comorbidities. A large value indicates a higher risk to the patient and therefore, increased urgency for treatment (Elixhauser et al., 1998). The laboratory values comprised a set of 23 tests, including i.a. serum creatinine, glucose, blood urea nitrogen. For each, we extracted three values for the last three days leading up to the time of the patient's surgery. In total, we extracted 103 features for modeling, which were then submitted to preprocessing and feature selection.

3.4 Preprocessing and Imputation

Data preprocessing is comprised of feature scaling and missing values imputation. The former entails removing the mean of the individual feature dimensions, i.e. centering the data's columns, and scaling to unit variance. The latter is necessary to handle missing values, which often occur in a clinical context (Bai et al., 2015) and to account for the fact that the machine learning models employed for this work do not support missing values. To handle missing values, we applied the k-Nearest Neighbors (kNN) imputation method with k=3, which operates under the assumption that missing values can be approximated by samples that are most similar to it. The utilized the version in the fancyimpute Python library (Rubinsteyn and Feldman, 2018).

For the examined cohort, the amount of missing values depends significantly on the considered feature dimension. For demographical data dimensions such as gender and age, as well as comorbidity data the amount of missing data is understandably low varying from 0% to 1% as the collection of patients' personal information is standard procedure and a non-apparent comorbidity does not result in missing values. Some laboratory results, however, do exhibit a significantly large amount of missing values ranging from 50% for i.a. blood creatinine 1 day before surgery to up to 98% for i.a. blood hematocrit results 3 days before the date of surgery. It must be noted that the amount of missing values for laboratory results decreases as we consider times closer to the time of surgery. That is, the laboratory results for one day before surgery typically present the least missing values compared to values for the same type of test on earlier days.

3.5 Feature Selection

In this processing step, we performed tests using different percentiles of top features, using the full set of features (103), top 50% and top 25%, reporting the results for all the algorithms tested. We chose feature selection based on the mutual information approach, since it can capture non-linear dependencies among among variables, unlike an F-test, which can capture only linear correlations. The list of the top 50% features used in the final model after automatic selection is provided as supplementary material on-line [1].

3.6 Classification Models

We compared different classification techniques that were trained and evaluated on the same data using the same preprocessing pipeline. In the following, we outline the respective hyperparameters' configuration and optimization strategy.

Logistic Regression. Usual parameters to adjust for LR include regularization strength and the type of penalty, L1 or L2. Regularization can improve model performance for unseen data by penalizing large coefficients in an effort to reduce overfitting or learning training data 'peculiarities'. Higher values for λ can lead to more sparse models. The library utilized exposes the parameter C defined as the inverse of regularization strength. The regularization parameters were set to 10^{-6} and the number of iterations before convergence were set to 300.

Decision Tree. We applied the Gini impurity measure to calculate optimal splits. Furthermore, we used class weights of 1:10 (AKI=no; AKI=yes) in order to compensate for the class skew that is evident in our training data. Finally, we determined an optimal maximum tree depth of 6, minimum of 6 samples for each leaf node, and a minimum of 5 samples for each split by using a parameter grid-search over a parameter grid of 3 to 10 for tree depth, 3 to 10 for minimum samples per split, and 1 to 16 for minimum samples per leaf, and 5-fold cross validation.

Gradient-boosted Decision Trees. This ensemble classification approach entails using a large set of decision trees or decision tree stumps as weak learners which are trained iteratively (Gron, 2017). In our work, the ensemble size was determined using an early stopping approach after 10 consecutive performance decreases at 136 indicating that a set of 126 learners provides the best prediction results. Furthermore, we applied a maximum tree depth for the stumps of 3 and learning rate of 0.1, which influences the contribution of each component tree stump.

[1] Supplementary material. Top 50% selected model features. Available at: `https://goo.gl/xnUux2`

3.7 Local Interpretable Model-agnostic Explanations

For practical use in high-stakes domains, not only prediction accuracy is relevant but also the level of trust a model provides. Ensemble models such as GBDT are typically considered black box models as they lack interpretability due to the fact that the models' behavior is determined by a large set of individual classifiers in a voting process (Valdes et al., 2016; Moon et al., 2007).

We use the interpretability method LIME to shed light on the prediction results of the GBDT model. LIME which uses more intelligible models, such as linear regression, to approximate the behavior of a given model in the vicinity of the instance/prediction being explained. The algorithm generates a number of perturbed instances close to the instance of interest, weighing this perturbed input according to a distance measure. After applying the original model on these perturbed instances, a linear function is applied to approximate the thus resulting outputs (Ribeiro et al., 2016). The coefficients of this linear function represent the degree of influence of a given feature for the original prediction we intended to explain. The higher the number of samples, the higher the fidelity of the approximate model, but the higher the algorithm runtime. In this work, we used a sample size of 100.

LIME differs from alternative interpretability methods, such as mimic learning, in that not the entire prediction model's behavior is explained, but rather one single prediction instance (Che et al., 2016). Therefore, the explanations provided are faithful locally but not globally. To make up for this behavior, LIME offers a procedure called submodular pick, which selects a number representative instances that can provide some insight into the model's global behavior (Ribeiro et al., 2016).

4 RESULTS

In this section, we present the performance results achieved using the proposed pipeline along with insights from applying LIME on the best classifier (GBDT). Besides traditional performance metrics for classification tasks, such as precision, recall, and the analysis of Area Under the Receiver Operating Curve (AUROC), we provide the diagnostic odds ratio (DOR), a performance measure for diagnostic tests which is prominently used in the medical domain (Bewick et al., 2004; Glas et al., 2003).

4.1 Discrimination

Table 1 reports the selected metrics across all feature selection configurations and models tested, considering respectively all features, top 50% and top 25% percentiles. DT performed worse than LR and GBDT for most metrics, regardless of feature selection, except for recall, where it presented a substantial advantage against the other two approaches, e.g. recall of 0.66 as opposed to GBDT's 0.48 for the top 50% features.

The different configurations chosen for feature selection demonstrated that the models achieve a similar performance even when only half of the available features are used. Particularly when it comes to the GBDT, the DOR was substantially improved by removing 50% of the features, from 90.74 to 149.92. As more features are removed, though, performance begins to deteriorate perceptibly while not by a large margin, e.g. a drop of approximately 3% in AUROC when only 25% of the features are used in the GBDT.

Overall, the GBDT classification method provides better prediction performance when compared to the results yielded by conventional decision trees or logistic regression, most notably when it comes to precision (40% increase over LR) and AUROC (6% increase over LR), albeit it performs poorly when it comes to recall. Furthermore, GDBT presents substantially better results as it refers to DOR with a 7-fold increase when compared to LR with 50% of features.

4.2 Local Interpretability

The LIME method expects a given prediction sample – or in our case a patient – as along with the trained model an inputs. Therefore, an expert can inquiry the model as to 'why' a given decision was made by the algorithm. However, to obtain an understanding of the model as whole, one would have to explain many instances. Since, this task might too consuming, LIME provide a strategy called submodular pick that provides the instances that are the most representative of the overall model's behavior (Che et al., 2016; Elith et al., 2008).

For GBDT predictions, LIME provides insight into which feature dimensions are most relevant to the results. The exact relevance of dimensions varies depending on the specific data input and outputs, but a small set of dimensions are found to be relevant for the model's decisions quite often across the different instances; the Elixhauser score, cardiac arrhythmia, hemoglobin, hematocrit, serum creatinine, and blood urea nitrogen (BUN) laboratory values before surgery.

Table 1: Precision, recall, diagnostic odds ratio (DOR), and area under receiver operating curve (AUROC) for AKI=1 achieved with the proposed approach employing logistic regression (LR), decision tree classification (DT) and gradient-boosted decision trees (GBDT) respectively for different feature selection configurations (all features, top 50 and 25%). The results were obtained by applying the trained models on a hold-out validation dataset made up of 20% of the original dataset.

Metrics	Precision			Recall			DOR			AUCROC		
	All	Top 50%	Top 25%	All	Top 50%	Top 25%	All	Top 50%	Top 25%	All	Top 50%	Top 25%
LR	0.63	0.63	0.59	0.28	0.25	0.25	19.55	19.14	16.67	0.84	0.84	0.82
DT	0.33	0.35	0.29	0.67	0.66	0.70	10.86	11.22	10.16	0.80	0.80	0.78
GBDT	0.86	0.90	0.62	0.43	0.48	0.32	90.74	149.92	115.50	0.89	0.90	0.87

The results provided by LIME represent the discretized coefficients of the regression applied locally to provide the explanations. We choose to report the top five features that explain the onset of AKI using submodular pick of 6 explanations, as displayed in Figure 3.

5 DISCUSSION

5.1 Discriminative Performance

GBDTs have been found to perform exceptionally well for classification and prediction tasks in multiple domains and the results from our work regarding prediction performance illustrate the expected advantage of the blackbox GBDT model over the simple decision tree or logistic regression prediction models (Che et al., 2016; Elith et al., 2008). In the medical domain, the superior precision and class label discrimination of the GBDT model effectively means more accurately predicting AKI cases after surgery, which might empower doctors to adopt targeted kidney-protective measures.

As exemplified by Table 2, our approach outperforms the Cleveland score by a considerable margin. However, the Cleveland score's authors utilize a substantially larger and more diverse cohort. The same observation applies to the STS score. One could therefore reasonably argue that those two scores potentially present higher generalizability. As such, our model must be subject to a validation study in order to assess its application in different clinical scenarios.

Even though our model performed well across most metrics, it showed significant drawbacks with regards to recall. While this issue can possibly be mitigated by proper model calibration, i.e., by adjusting classification thresholds, it might have critical implications for clinical practice. Since a lower recall means that patients who will develop AKI might incorrectly be deemed as not under risk, calibration a must be conducted, at the expense of possibly harming patients. This fact speaks for the necessity of a holistic evaluation of discrimination metrics.

Our model included laboratory values prior to the surgery. Arguably, these are not always available for surgery patients, particularly when it comes to emergency surgeries. This fact has led to a high degree of missing values in our cohort. Even though imputation has been performed, it is not possible to guarantee that the model has not been biased in some way or another. Missing values are widely-discussed topic in clinical predictive modeling and we intend to compare different approaches for imputation side by side.

5.2 Model Interpretation

Since the GBDT performed best in the given task, we submitted the model to the LIME explainer. In fact, model interpretability is especially important for medical practitioners and patients to promote acceptance for clinical use and build trust in predictive decision support (Katuwal and Chen, 2016).

Upon examination of the instances chosen by the submodular pick by LIME, we can observe that a high Elixhauser score, i.e., over 7, is often implicated with increasing risk of post-surgical AKI. Note that positive coefficients are positively correlated with the outcome and vice-versa. This observation agrees with the medical significance of this co-mordibities score: higher values are in general associated with poorer patient outcomes in general (Austin et al., 2015).

With regards to blood (or serum) creatinine, it is an important marker of kidney function, being present in the definition of AKI itself, with higher values indicating deterioration of kidney function (Lopes and Jorge, 2013). Values between 0.6 and 1.2mg/dL for creatinine are usually considered normal, and LIME correctly shows a protective effect of values below 1.4mg/dL. However, the explanations do not include higher serum creatinine values as a risk factor, possibly casting doubt on the generalizability of the model.

Furthermore, presence of liver disease is generally implicated in poorer outcomes for kidney patients (Targher et al., 2008). One would expect the model explanations to fully reflect medical knowledge. However, the opposite can be verified as per LIME's explanations: absence of liver disease often

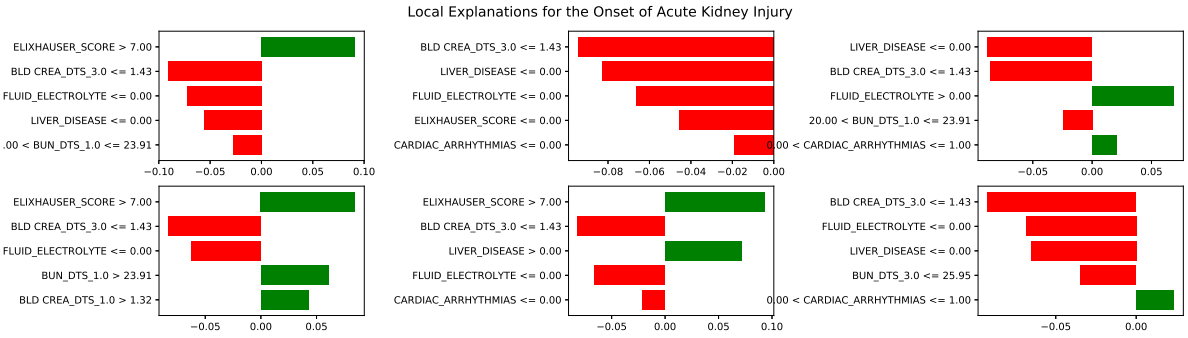

Figure 3: Local explanations provided by LIME. Using submodular pick, LIME chooses the most significant examples for explanation, i.e, the individual subplots. Each shows the top 5 features chosen by LIME as the most meaningful for the local predictions. Note that positive coefficients are correlated with increased likelihood of Acute Kidney Injury. Abbreviations: BLD_CREA_DTS_N=Blood Creatinine N days to surgery; BUN_DTS_N=Blood Urea Nitrogen N days to surgery.

Table 2: Overview of CPMs for cardiac surgery-associated AKI. Abbreviations: CPM=Clinical Prediction Model; N=number of patients; AUC=Area Under the Curve.

CPM	*N*	AUC
Cleveland	33,217	0.81
STS score	86,009	0.83
AKICS score	603	0.84
SRI score	2,566	0.78
Ng score	28,422	0.77
Jiang score	7,233	0.74
Our approach	6,782	0.90

appears as a protective factor, with only one explanation displaying the expected behavior. The use the LIME approach makes it possible to critically analyze model predictions at the single instance level, i.e., patient level, revealing potential bias in the models that might compromise applicability in practice. Correlations that statistically relevant but medically indefensible are not only inconsistent, but potentially dangerous for patients. The application of interpretability approaches can therefore help shed light on the obscure side of black-box models. Finally, it is worth noting that alternative interpretability methods, such as mimic learning, have also been applied in previous research and a thorough comparison of available interpretability approaches highlighting weaknesses and strengths constitutes valuable future work in this field.

6 CONCLUSION

We have devised a clinical prediction model that employs machine learning methodologies on clinical patient data in order to assess the risk of AKI in heart patients before the time of surgery using the MIMIC-III database. This can allow physicians to make better decisions about surgical therapy and plan accordingly for complications in high risk patients, e.g. by readying renal replacement resources in advance and avoiding nephrotoxic agents. By comparing the usage of traditional decision tree models with GBDT prediction models we showed the advantage in prediction quality of a more complex and non-interpretable black box model over an easily understandable white box modeling technique such as logistic regression or decision trees.

Our GBDT model outperformed established clinical scores for post-surgical AKI onset by a significant margin (AUROC of 0.9 vs. 0.83). While external validation with a bigger and more diverse cohort remains to be performed for the model to be considered generally applicable, results suggest GDBT as an appropriate algorithm for this specific prediction task. Future work shall also consider other algorithms such as deep learning and random forests, as well as different strategies for imputation and feature selection.

Despite the promising results, in the light of the importance of model intelligibility in high-risk domains such as medicine, we also utilized the interpretability method LIME on the GBDT model. Using this explainer, we regained a significant amount of meta-information about which features are most relevant for the prediction model's output. It ultimately revealed possible incongruencies between model explanations and medical evidence that must be addressed for such a model to be used in practice.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Alexander Meyer and his team at the German Heart Center Berlin for valuable guidance on the medical challenges involved. Parts of the given work were generously supported by a grant of the German Federal Ministry of

Economic Affairs and Energy (01MD15005).

REFERENCES

Austin, S. R. et al. (2015). Why Summary Comorbidity Measures such as the Charlson Comorbidity Index and Elixhauser Score Work. *Medical care*, 53(9):e65.

Bai, B. M., Mangathayaru, N., and Rani, B. P. (2015). An Approach to Find Missing Values in Medical Datasets. In *Proc. Intl. Conf. on Engineering & MIS*, pages 70:1–70:7, New York, NY, USA. ACM.

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics Review 13: Receiver Operating Characteristic Curves. *Critical Care*, 8(6):508.

Caruana, R. et al. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Intl. Conf. Knowledge Discovery and Data Mining*, pages 1721–1730, New York, NY, USA. ACM.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. (2016). Interpretable Deep Models for ICU Outcome Prediction. *AMIA Symposium*, 2016:371–380.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A Working Guide to Boosted Regression Trees.

Elixhauser, A. et al. (1998). Comorbidity Measures for Use with Administrative Data. *Medical Care*, 36(1):8–27.

Flynn, N. and Dawnay, A. (2015). A Simple Electronic Alert for Acute Kidney Injury. *Ann Clin Biochem*, 52(2):206–212.

Glas, A. S. et al. (2003). The Diagnostic Odds Ratio: A Single Indicator of Test Performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135.

Goodman, B. and Flaxman, S. (2016). EU Regulations on Algorithmic Decision-making and a "Right to Explanation". In *ICML Workshop on Human Interpretability in Machine Learning*.

Gron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc., 1st edition.

Huen, S. C. and Parikh, C. R. (2012). Predicting Acute Kidney Injury after Cardiac Surgery: a Systematic Review. *Ann Thorac Surg*, 93(1):337–47.

Jiang, W. et al. (2016). Dynamic Predictive Scores for Cardiac Surgery–Associated Acute Kidney Injury. *Am Heart J*, 5(8).

Jiang, W. et al. (2017). Validation of Four Prediction Scores for Cardiac Surgery-Associated Acute Kidney Injury in Chinese Patients. *Braz J Cardiovasc Surg*, 32(6):481–486.

Johnson, A. E. et al. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3.

Katuwal, G. J. and Chen, R. (2016). Machine Learning Model Interpretability for Precision Medicine. *ArXiv e-prints*.

Latini, R., Aleksova, A., and Masson, S. (2016). Novel Biomarkers and Therapies in Cardiorenal Syndrome.

Letham, B. et al. (2015). Interpretable Classifiers using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *Annals of Applied Statistics*, 9(3):1350–1371.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ArXiv e-prints*.

Lopes, J. A. and Jorge, S. (2013). The RIFLE and AKIN Classifications for Acute Kidney Injury: a Critical and Comprehensive Review. *Clin Kidney J*, 6(1):8–14.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In van der Walt, S. and Millman, J., editors, *Proc. 9th Python in Science Conference*, pages 51 – 56.

Moon, H. et al. (2007). Ensemble Methods for Classification of Patients for Personalized Medicine with High-Dimensional Data. *Artif Intell Med*, 41(3):197–207.

Ng, S. Y. et al. (2014). Prediction of Acute Kidney Injury within 30 Days of Cardiac Surgery. *J Thorac Cardiovasc Surg*, 147(6):1875–1883.e1.

O'Neal, Jason and others (2016). Acute Kidney Injury Following Cardiac Surgery: Current Understanding and Future Directions. *Critical Care*, 20(1):187.

Palomba, H. et al. (2007). Acute Kidney Injury Prediction following Elective Cardiac Surgery: AKICS Score. *Kidney International*, 72(5):624–631.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. 22nd Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY, USA. ACM.

Rosner, M. H. and Okusa, M. D. (2006). Acute Kidney Injury Associated with Cardiac Surgery. *Clin J Am Soc Nephrol*, 1(1):19–32.

Rossum, G. V. and Drake, F. L. (2010). Python Tutorial. *History*, 42(4):1–122.

Rubinsteyn, A. and Feldman, S. (2018). fancyimpute: A Variety of Matrix Completion and Imputation Algorithms Implemented in Python. https://github.com/iskandr/fancyimpute Accessed: October, 2018.

Sawhney, S. et al. (2015). Acute Kidney Injury - How Does Automated Detection Perform? *Nephrol. Dial. Transplant.*, 30(11):1853–61.

Targher, G. et al. (2008). Non-alcoholic fatty liver disease is independently associated with an increased prevalence of chronic kidney disease and proliferative/laser-treated retinopathy in type 2 diabetic patients. *Diabetologia*, 51(3):444–450.

Thakar, C. V. et al. (2004). A Clinical Score to Predict Acute Renal Failure after Cardiac Surgery. *J Am Soc Nephrol*, 14(8):2176–7.

Valdes, G. et al. (2016). MediBoost: A Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Scientific Reports*, 6.

Wijeysundera, D. N. et al. (2007). Derivation and Validation of a Simplified Predictive Index for Renal Replacement Therapy After Cardiac Surgery. *JAMA*, 297(16):1801.