International Journal of Nephrology and Renovascular Disease

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery

Evi Diana Omar[1], Hasnah Mat[1], Ainil Zafirah Abd Karim[1], Ridwan Sanaudi[1], Fairol H Ibrahim[2], Mohd Azahadi Omar[1], Muhd Zulfadli Hafiz Ismail[1], Vivek Jason Jayaraj[1], Bak Leong Goh[2]

[1]Sector for Biostatistics and Data Repository, National Institutes of Health, Ministry of Health Malaysia, Shah Alam, Selangor, Malaysia; [2]Hospital Sultan Idris Shah Serdang, Ministry of Health Malaysia, Kajang, Selangor, Malaysia

Correspondence: Ainil Zafirah Abd Karim, Email ainilzafirah@moh.gov.my

Purpose: This study aimed to identify the best-performing algorithm for predicting Acute Kidney Injury (AKI) necessitating dialysis following cardiac surgery.

Patients and Methods: The dataset encompassed patient data from a tertiary cardiothoracic center in Malaysia between 2011 and 2015, sourced from electronic health records. Extensive preprocessing and feature selection ensured data quality and relevance. Four machine learning algorithms were applied: Logistic Regression, Gradient Boosted Trees, Support Vector Machine, and Random Forest. The dataset was split into training and validation sets and the hyperparameters were tuned. Accuracy, Area Under the ROC Curve (AUC), precision, F-measure, sensitivity, and specificity were some of the evaluation criteria. Ethical guidelines for data use and patient privacy were rigorously followed throughout the study.

Results: With the highest accuracy (88.66%), AUC (94.61%), and sensitivity (91.30%), Gradient Boosted Trees emerged as the top performance. Random Forest displayed strong AUC (94.78%) and accuracy (87.39%). In contrast, the Support Vector Machine showed higher sensitivity (98.57%) with lower specificity (59.55%), but lower accuracy (79.02%) and precision (70.81%). Sensitivity (87.70%) and specificity (87.05%) were maintained in balance via Logistic Regression.

Conclusion: These findings imply that Gradient Boosted Trees and Random Forest might be an effective method for identifying patients who would develop AKI following heart surgery. However specific goals, sensitivity/specificity trade-offs, and consideration of the practical ramifications should all be considered when choosing an algorithm.

Keywords: acute kidney injury, cardiac surgery, machine learning, predictive analytics

Introduction

In recent years, the field of medical research has seen exponential growth in the use of machine-learning algorithms to improve clinical decision-making and patient outcomes.[1–3] Amongst the countless applications, predicting postoperative complications after cardiac surgery has emerged as a crucial area of research. Acute kidney injury (AKI) is a common and serious complication that has been shown to have a significant impact on patient morbidity, mortality, and healthcare costs. In particular, cases requiring dialysis due to AKI represent even greater challenges and require rapid and accurate prognostic tools to identify patients at risk.

Generally, clinical risk assessment scores are used to assess the likelihood of postoperative complications, including AKI.[4,5] Many scoring systems fall short of accurately predicting outcomes when dealing with intricate interactions among multiple clinical variables. This is where machine learning algorithms can be utilized to their fullest potential, as they possess the ability to detect intricate patterns in vast data sets, resulting in more accurate predictions. Recent studies have highlighted the efficacy of machine learning approaches in predicting postoperative complications and outcomes.

Received: 4 March 2024
Accepted: 13 June 2024
Published: 24 July 2024

For instance, Zhang et al developed a machine learning model that exhibited superior performance compared to traditional risk scores in forecasting postoperative acute kidney injury.[6] Similarly, Smith et al applied machine learning techniques to enhance the prediction of various postoperative complications, demonstrating improved accuracy and reliability.[7] Furthermore, Qing Qian et al demonstrating the potential of machine learning approaches in improving risk prediction compared to traditional scoring systems.[8]

This study deals with a comprehensive comparative analysis of four well-known machine learning algorithms, namely Logistic Regression, Gradient Boosted Trees, Support Vector Machine, and Random Forest. Each of the algorithms used for this study represents a unique approach to predictive modelling. Logistic Regression is a simple yet interpretable algorithm well-suited for binary classification tasks, while Gradient Boosted Trees iteratively build decision trees, adeptly handling complex non-linear relationships and performing notably well in medical applications. Support Vector Machines, on the other hand, excel at finding the optimal class-separating hyperplane, thus performing robustly on complex datasets with numerous features. Lastly, Random Forests construct an ensemble of decision trees and combine their predictions, effectively reducing overfitting while enhancing generalization performance.

While several studies have explored machine learning techniques for predicting postoperative complications, including AKI, our research offers a comprehensive comparative analysis of four distinct algorithms: Logistic Regression, Gradient Boosted Trees, Support Vector Machine, and Random Forest. This multi-algorithm approach allows for a thorough evaluation of their respective strengths and weaknesses in the context of AKI prediction after cardiac surgery.

Xue et al studied a machine learning algorithm to obtain better predictive power for Cardiac surgery-associated AKI outcomes.[9] This study compared four machine learning algorithms (random forest (RF), logistic regression with LASSO regularization, extreme gradient boosting (Xgboost), and support vector machine (SVM)) for the entire dataset to identify the most important clinical variables for AKI prediction outcomes. Similarly, Jiang et al evaluated whether machine learning algorithms could significantly improve the risk prediction of postoperative AKI.[10] This study utilized conventional logistic regression (LR) alongside five ML algorithms (decision tree, random forest, gradient boosting classifier (GBC), Gaussian Naive Bayes and multilayer perceptron).

Even though previous studies show it was conducted with a broader range of algorithms, our study focuses specifically on the Malaysian population, providing insights into the applicability and performance of these models in a different context. Socioeconomic status and the prevalence of comorbidities may be different from other studies that may play the factors that could influence the effectiveness of AKI prediction models in the Malaysian population.

This research study aims to determine the best-performing algorithm for predicting AKI that requires dialysis after cardiac surgery. By leveraging diverse patient characteristics including demographic, clinical, and laboratory variables, we intend to develop reliable predictive models that can aid physicians in identifying high-risk patients and providing timely interventions. Furthermore, this study provides valuable insight into the advantages and disadvantages of each technique in the context of AKI prediction. The ultimate objective is to contribute to the growing body of knowledge at the intersection of machine learning and healthcare, allowing for more accurate risk assessment and enhanced patient care in cardiac surgery.

Materials and Methods

Data

This study was based on "Validating Cleveland Clinic Score to Predict Acute Kidney Injury Requiring Dialysis After Cardiac Surgery" which included all data collection (at a national level) of patients who underwent cardiac surgery at a tertiary cardiothoracic center in Malaysia between 2011 and 2015. The approach of this data collection process was described in a prior paper and published elsewhere.[11] The electronic health records (EHR), which held detailed patient information, including clinical, surgical, and laboratory data, served as the primary data source for this investigation. The dataset comprised 1741 patients, of which 179 patients were diagnosed with AKI. The Cleveland Clinic Score, a commonly used predictive model for AKI requiring dialysis, served as the foundation for the variables included in this study. The variables were gender, insulin use, history of chronic obstructive pulmonary disease (COPD), prior surgeries, left ventricular ejection fraction (LVEF), intra-aortic balloon pump (IABP) use, congestive heart failure (CHF) diagnosis, type of surgery, emergency status, creatinine levels, and the need for renal replacement therapy as the primary endpoint. Additionally, other variables such as age, hypertension status, discharge status, and length of stay were incorporated due to their potential significance in predicting the outcome.

Preprocessing

The dataset underwent meticulous preprocessing to ensure quality and consistency. Patients with insufficient or missing data were excluded from the analysis. We applied SMOTE (Synthetic Minority Over-sampling Technique) in RapidMiner to handle class imbalance in our dataset. This technique involves increasing the number of instances in the minority class by duplicating or generating synthetic samples. Furthermore, it can improve the model's ability to recognize patterns in the underrepresented class, leading to better overall classification performance and a more robust and fair model evaluation.

Feature Selection

Feature selection was applied to identify the most relevant predictors for predicting AKI requiring dialysis. The primary focus was on assessing data quality, with a critical emphasis on removing predictors that contributed little to the analysis. Hence, in our study, the predictors were identified based on their significant contributions during univariate analysis and by detecting discernible patterns within the data. Predictors of interest included those closely related to the target column (Correlation), those with a high degree of uniqueness (ID-ness), those exhibiting stability (Stability), those with missing values (Missing), and those potentially containing free-text information (Text-ness). Generally, predictors with low Missing, Stability, and ID-ness values were favored, with some cases warranting the retention of text columns.

Analysis

In order to depict various methods of predictive modelling, the four algorithms under consideration; Logistic Regression, Gradient Boosted Trees, Support Vector Machine, and Random Forest were picked. As a linear technique for binary classification, Logistic Regression served as the foundational model. Logistic Regression has been developed for binary classification tasks, making it an important choice because of its simplicity and interpretability. Meanwhile, Gradient Boosted Trees were chosen because of their resistance to complicated interactions and capacity for dealing with non-linearity. This ensemble method iteratively builds successive decision trees, with each tree correcting the shortcomings of the previous one, known for handling complex non-linear relationships, excelling in capturing intricate patterns within data, and achieving particularly notable success in medical applications.

Support Vector Machine was chosen as an example of kernel-based techniques due to its ability to recognize high-dimensional patterns. It is a powerful classification algorithm that emphasizes finding the most efficient class-separator hyperplane and helps to improve performance on complex datasets with large number of features. Finally, random forest is a cluster learning method that builds multiple decision trees and combines their results for improved accuracy and generalization. By using decision trees and pooling their predictions, random forests effectively reduce overfitting, increasing the ability of the model to accurately predict the unobserved data.

The dataset was divided into training and validation sets, ensuring an 80:20 split, to facilitate impartial model evaluation. To improve model performance, the hyperparameters of each algorithm were automatically tuned by RapidMiner's AutoModel feature during the model training process. This automated approach effectively searches the hyperparameter space, leveraging sophisticated optimization techniques to identify the optimal or near-optimal hyper-parameter configurations. These configurations are specifically tailored to maximize the performance metrics of the respective machine learning models. The models were subsequently developed using the training set, and the validation set was used to assess them using the proper metrics, including accuracy, AUC, precision, F-measure, sensitivity, and specificity.

RapidMiner, a powerful data science platform, was used to implement our machine learning workflows.[12] Custom processes were created to handle data preprocessing, feature selection, and model training. RapidMiner's built-in operators were leveraged for algorithm implementation. Figure 1 provides a comprehensive overview of the key stages

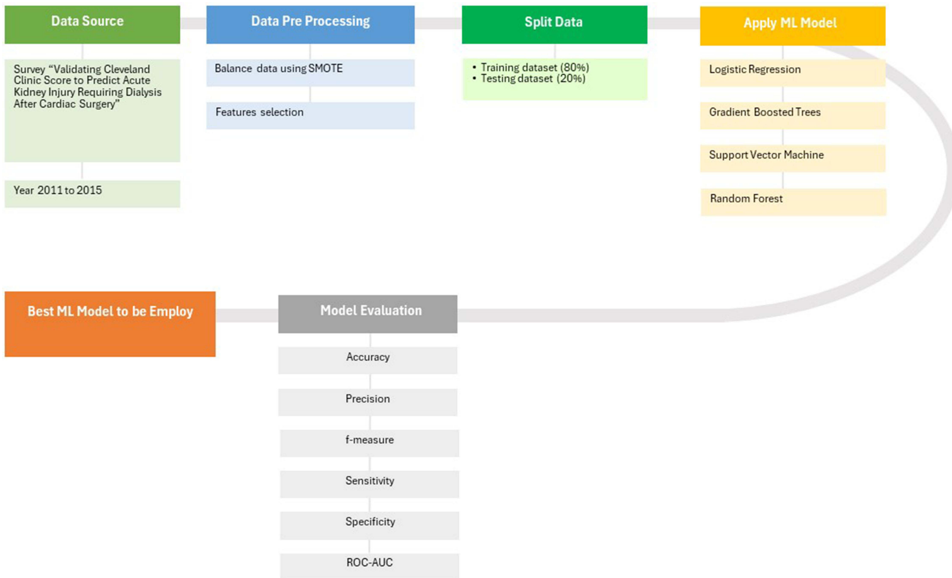

Figure 1 Overview of data preprocessing, model development and model evaluation using RapidMiner.
Abbreviation: ML, machine learning.

involved, namely data preprocessing, model development, and model evaluation, all of which were carried out using the RapidMiner software.

Ethics

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. This study did not require ethical approval as it is an article discussion from a methodological perspective, utilizing secondary data. The requirement for informed consent was waived since all identifiable data was removed, ensuring anonymity and confidentiality. However, this article has received approval for publication from the Director General, Ministry of Health, Malaysia.

Results

In this analysis, we began with a dataset comprising 1741 patients, of which only 179 patients were diagnosed with AKI. This presented an inherent class imbalance issue, as the AKI cases were significantly underrepresented compared to non-AKI cases. To mitigate this imbalance and ensure our dataset's suitability for modeling, we performed a resampling technique known as upsampling. This technique involved increasing the number of AKI patients to 1250 while keeping the number of non-AKI cases at 1250, resulting in a balanced representation of both classes.

Based on univariate statistical analyses, we initially identified a set of variables that showed significant associations with AKI, particularly with emphasis on p-values. The selected variables included age, insulin use, hypertension status, LVEF, IABP use, type of surgery, emergency status, creatinine levels, discharge status, and length of stay. We performed an additional selection step as part of our model refining approach to optimize model stability and performance by detecting discernible patterns within the data as mentioned in the methodology section. As a result, we omitted two variables, IABP and emergency status, because their stability metrics exceeded 90%.

We constructed an analysis of four machine learning models using Logistic Regression, Random Forest, Gradient Boosted Trees, and Support Vector Machine for predicting AKI requiring dialysis after cardiac surgery. The final model incorporated a comprehensive set of variables, including age, insulin use, hypertension status, left ventricular ejection fraction (LVEF), type of surgery, creatinine levels, discharge status, and length of hospital stay to the final model that consist variable age, insulin use, hypertension status, LVEF, type of surgery, creatinine levels, discharge status, and length of stay. Table 1 summarizes the performance metrics achieved by each algorithm in the validation dataset. The

Table 1 Performance of Four Different Machine Learning Algorithms

Model	Accuracy (%)	AUC (%)	Precision (%)	F-measure (%)	Sensitivity (%)	Specificity (%)
Logistic Regression	87.39	93.78	87.53	87.48	87.70	87.05
Random Forest	87.39	94.78	88.73	87.32	86.14	88.70
Gradient Boosted Trees	88.66	94.61	86.78	88.93	91.30	86.06
Support Vector Machine	79.02	93.01	70.81	82.40	98.57	59.55

Abbreviation: AUC, area under the curve.

algorithms were evaluated using various performance metrics including accuracy, AUC, precision, F-measure, sensitivity, and specificity.

The results in Figure 2 show variations in the performance of the various models. Gradient Boosted Trees achieved the best accuracy (88.66%) and AUC (94.61%), indicating its effectiveness at predicting AKI that requires dialysis. It also had the highest sensitivity (91.30%), indicating its ability to appropriately identify patients at risk. Random Forest exhibited high AUC (94.78%) and precision (88.73%) values.

In contrast, the Support Vector Machine showed relatively lower accuracy (79.02%) and precision (70.81%), as well as a higher sensitivity (98.57%) but lower specificity (59.55%). The overall performance of Logistic Regression was consistent, with balanced sensitivity (87.70%) and specificity (87.05%).

In addition to the performance metrics, we analyzed each model's Receiver Operating Characteristic (ROC) curves to assess their discriminative power. The ROC curves graphically illustrate in Figure 3, the trade-off between sensitivity and specificity at various decision thresholds.

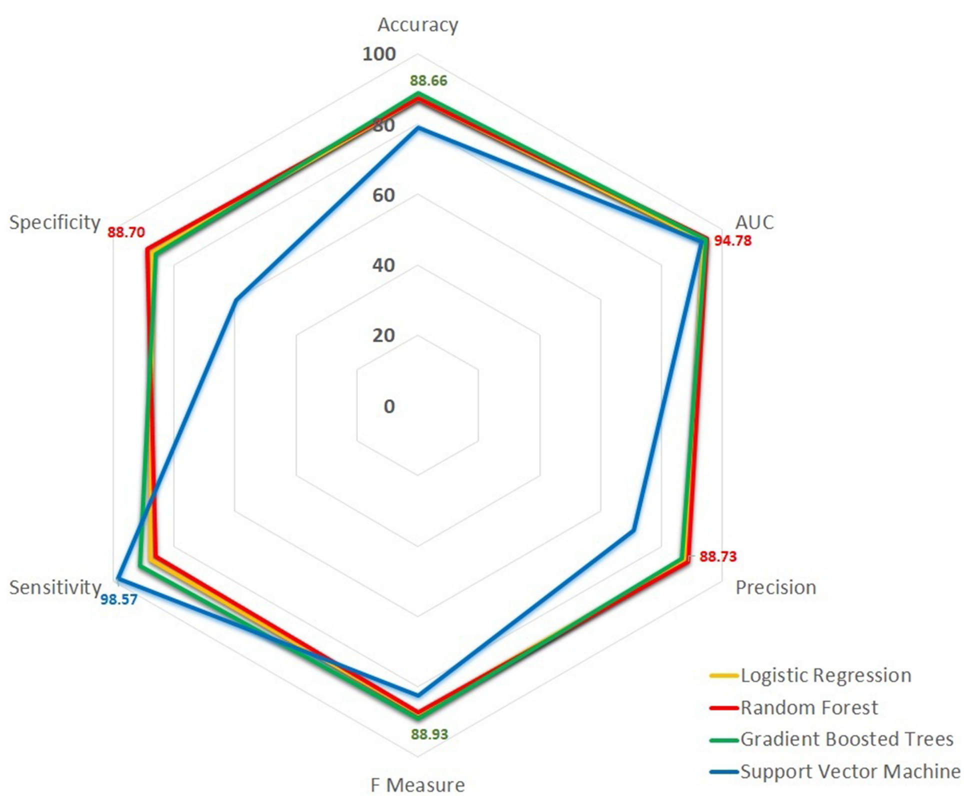

Figure 2 Machine learning performance.
Abbreviation: AUC, area under the curve.

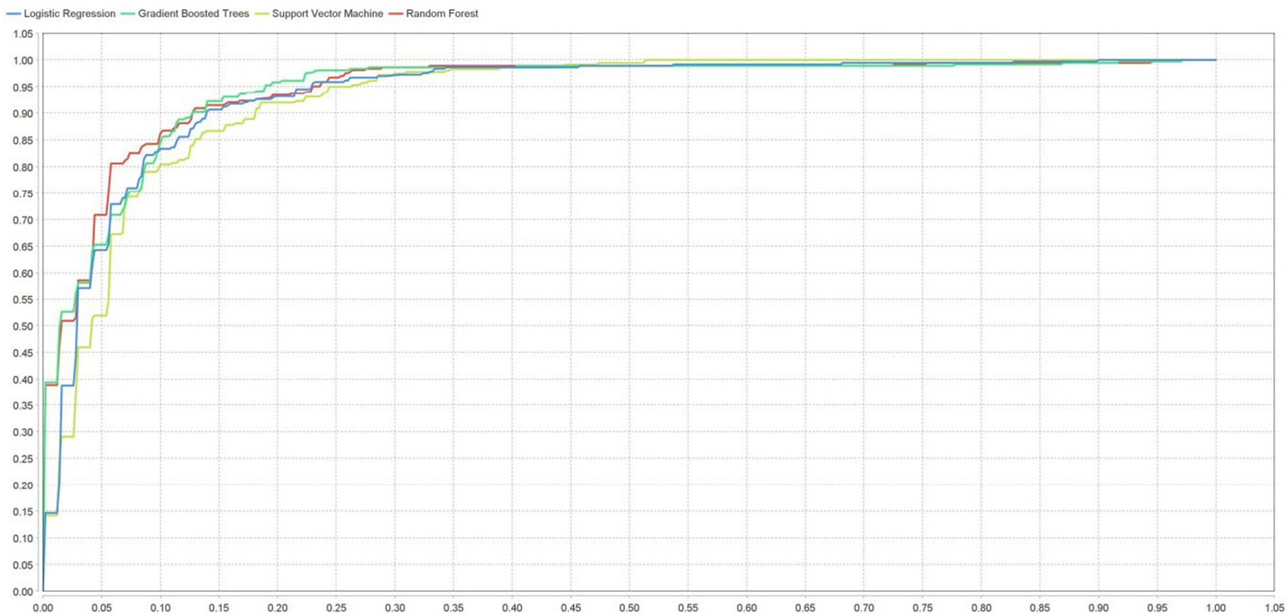

Figure 3 ROC curve for estimating the discrimination power between the machine learning models.

From the ROC curve comparisons, it is evident that Gradient Boosted Trees and Random Forest outperformed the other models, showing a higher AUC value, which indicates their superior ability to discriminate between patients at risk of AKI requiring dialysis and those who are not. Logistic Regression demonstrated a balanced ROC curve, while Random Forest exhibited relatively lower discriminative power. These results indicate that Gradient Boosted Trees and Random Forest may be promising models for predicting AKI requiring dialysis after cardiac surgery.

Discussion

The results from the different machine learning models, such as Logistic Regression, Random Forest, Gradient Boosted Trees, and Support Vector Machine, offer important insights into how well these models perform at identifying patients who are at risk of AKI and need dialysis after cardiac surgery.

The accuracy rates of 87.39% for both the Logistic Regression and Random Forest models show that they are both capable of classifying instances accurately. Random Forest surpassed Logistic Regression in terms of AUC with an AUC of 94.78%, indicating that it is more effective at differentiating between positive and negative situations. Gradient Boosted Trees, on the other hand, had a competitive AUC of 94.61% and the method with the best accuracy, at 88.66%.

In comparison to the other models, Random Forest showed the best precision (88.73%) and a lower false positive rate. The precision ratings of all models are nevertheless reasonably near to one another, showing a reasonable trade-off between recall and precision. Gradient Boosted Trees scored the highest overall balance between precision and recall, with a score of 88.93%, according to the F-measure, which combines precision and recall into a single metric.

Support Vector Machine showed the highest sensitivity (recall) at 98.57%, indicating that it successfully identified true positive situations with a low false negative rate. However, it's essential to consider the specificity as well. The Support Vector Machine had the highest false positive rate with the lowest specificity (59.55%). This shows that the Support Vector Machine might have a propensity to categorize some negative cases as positive.

Gradient Boosted Trees outperform the other models in terms of performance, with strong F-measure, balanced precision, and high sensitivity. Random Forest was a competitive option for this classification assignment because it had the greatest AUC and also did well in terms of precision and F-measure. Its dependability for this work was demonstrated by the constant performance of the Logistic Regression model across a range of parameters. Support Vector Machine performed exceptionally well in terms of sensitivity but had shortcomings in terms of specificity, which could be problematic in situations where false positives are expensive.

Our results are consistent with earlier studies that have shown the capability of machine learning algorithms in foretelling AKI following cardiac surgery.[13–15] However, by highlighting how the choice of the most suitable model depends on the precise goals and limitations of the classification problem, our analysis adds a fresh dimension. Gradient Boosted Trees or Random Forest stands out as strong contenders when looking for a balance between sensitivity and precision. These models fared better than the more conventional Logistic Regression, showing that cutting-edge machine-learning techniques can greatly improve forecast accuracy. However, Logistic Regression continues to be a solid option for people who value simplicity and interpretability.

Despite the fact that our study provided insightful findings, it is important to recognize its shortcomings. The generalizability of our findings to larger patient populations may be constrained because our analysis was based on a retrospective dataset from a single cardiac surgery unit. The dataset also included a few missing values and other source-related biases. Future studies should investigate techniques to deal with data restrictions as well as the validation of these models in various clinical situations.

Future studies in this field might concentrate on improving the predictive models by including more clinical factors, like biomarkers or genetic data, to increase prediction precision. It is necessary to conduct prospective studies in clinical settings to evaluate the models' practicality and validate their effectiveness. Additionally, investigating interpretable machine-learning methods can help to improve our comprehension of the causes of AKI following heart surgery.

Conclusion

In conclusion, the precise goals of the classification task and the respective weights of precision, recall, and other performance indicators should serve as the basis for choosing the best machine-learning model. It could be necessary to conduct further validation and fine-tuning studies to guarantee reliable model performance in real-world scenarios.

Acknowledgments

The author would like to thank the Director-General of Health Malaysia for granting permission to publish this report. The author is also very thankful for the reviewers' insightful comments and constructive suggestions. Additionally, we extend our appreciation to Dr Goh Kheng Wee for providing valuable data that significantly contributed to the completion of this research project.

Disclosure

The author(s) report no conflicts of interest in this work.

References

1. Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applications. *Int J Intell Netw*. 2022;3:58–73. doi:10.1016/j.ijin.2022.05.002
2. Russak AJ, Chaudhry F, De Freitas JK, et al. Machine learning in cardiology—ensuring clinical impact lives up to the hype. *J Cardiovasc Pharmacol Therap*. 2020;25(5):379–390. doi:10.1177/1074248420928651
3. Delen D, Davazdahemami B, Eryarsoy E, Tomak L, Valluru A. Using predictive analytics to identify drug-resistant epilepsy patients. *Health Inform J*. 2019;12:146045821983312. doi:10.1177/1460458219833120
4. Thakar CV, Worley S, Arrigain S, Yared JP, Paganini EP. Improved Survival in Acute Kidney Injury After Cardiac Surgery. *Am J Kidney Dis*. 2007;50(5):703–711. doi:10.1053/j.ajkd.2007.07.021
5. Che M, Wang X, Liu S, et al. A clinical score to predict severe acute kidney injury in Chinese patients after cardiac surgery. *Nephron*. 2019;142(4):291–300. doi:10.1159/000499345
6. Zhang Y, Yang D, Liu Z, et al. An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation. *J Transl Med*. 2021;19(1):1–15. doi:10.1186/s12967-021-02990-4
7. Smith MT, Bruce JL, Clarke DL. Using machine learning to establish predictors of mortality in patients undergoing laparotomy for emergency general surgical conditions. *World J Surg*. 2022;46(2):339–346. doi:10.1007/s00268-021-06360-5
8. Qian Q, Wu J, Wang J, Sun H, Yang L. Prediction models for AKI in ICU: a comparative study. *Int J Gene Med*. 2021;14:623–632. doi:10.2147/IJGM.S289671
9. Xue X, Liu Z, Xue T, Chen W, Chen X. Machine learning for the prediction of acute kidney injury in patients after cardiac surgery. *Front Surg*. 2022;9:946610. doi:10.3389/fsurg.2022.946610
10. Jiang J, Liu X, Cheng Z, et al. Interpretable machine learning models for early prediction of acute kidney injury after cardiac surgery. *BMC Nephrol*. 2023;24(1):326. doi:10.1186/s12882-023-03324-w

11. Goh KW, Goh BL, Ibrahim FH, et al. 126 validating Cleveland clinic score to predict acute kidney injury requiring dialysis after cardiac surgery. *Kidney Int Rep*. 2017;2(4):S18. doi:10.1016/j.ekir.2017.06.066
12. Altair® RapidMiner® Data Analytics and AI Platform. Available from: https://rapidminer.com/. Accessed May 22, 2023.
13. Penny-Dimri JC, Bergmeir C, Reid CM, Williams-Spence J, Cochrane AD, Smith JA. Machine learning algorithms for predicting and risk profiling of cardiac surgery-associated acute kidney injury. *Semin Thor Cardiovasc Surg*. 2020;33(3):735–745. doi:10.1053/j.semtcvs.2020.09.028
14. Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care*. 2020;24(1). doi:10.1186/s13054-020-03179-9
15. Demirjian S, Schold JD, Navia J, et al. Predictive models for acute kidney injury following cardiac surgery. *Am J Kidney Dis*. 2012;59(3):382–389. doi:10.1053/j.ajkd.2011.10.046