REVIEW

Journal of Indonesian Thoracic Cardiac and Vascular Surgery *(JINATCVS)* 2026,
Volume 3, Number 2: 51-59

Machine learning models in predicting acute kidney injury following acute aortic dissection surgery: Meta-analysis of diagnostic accuracy

Published By :
The Indonesian Association of Thoracic and Vascular Surgeons

Bernardinus Prima Ardjie Pradipta[1*], Bernadus Bernardino Bramantyo[2], Muhammad Reva Aditya[3], Paulina Lusty Artanti[4], Koernia Kusuma Wardhana[5]

[1]Department of Cardio-Thoracic and Vascular Surgery, Universitas Airlangga, Surabaya, Indonesia
[2]Department of Cardiology and Vascular Medicine, Sardjito General Hospital, Yogyakarta, Indonesia
[3]Faculty of Medicine, Universitas Brawijaya, Malang, Indonesia
[4]Oen Kandang Sapi Hospital, Solo, Indonesia
[5]Department of Cardio-Thoracic and Vascular Surgery, Saiful Anwar General Hospital, Malang, Indonesia

*Corresponding email:
Bernardinus Prima Ardjie Pradipta;
Department of Cardio-Thoracic and Vascular Surgery, Universitas Airlangga, Surabaya, Indonesia;
primapradipta7@gmail.com

Received: 2026-08-26
Accepted: 2026-09-23
Published: 2026-10-25

ABSTRACT

Introduction: Acute kidney injury (AKI) continues to represent a major complication following surgical repair of acute aortic dissection (AAD). Recent advancements in artificial intelligence (AI) demonstrate significant potential by leveraging extensive electronic health record data to forecast AKI. This meta-analysis seeks to elucidate the diagnostic accuracy and comparative performance of various machine learning (ML) models in predicting AKI occurrence in patients with AAD repair.

Methods: A comprehensive systematic literature search was conducted across PubMed, EuropePMC, and EBSCO for studies evaluating ML algorithms (including Neural Network [NN], Gradient Boosting Machine [GBM], K-Nearest Neighbors [KNN], LightGBM [LGB], Logistic Regression [LR], Random Forest [RF], Support Vector Machine [SVM], and eXtreme Gradient Boosting [XGB]) in predicting post-AAD repair AKI. Study quality and risk of bias were systematically evaluated using the Quality Assessment of Prognostic Accuracy Studies (QUAPAS) tool. Pooled sensitivity, specificity, diagnostic odds ratios (DOR), and summary receiver operating characteristic (SROC) curves with corresponding area under the curve (AUC) were calculated using a bivariate random-effects model. Heterogeneity was evaluated using I2 statistics.

Results: Six studies comprising 1,713 participants and evaluating 22 distinct ML models met the inclusion criteria. Predictors evaluated to predict AKI comprised demographic parameters; pre-existing diseases; baseline, intraoperative, and postoperative hemodynamic parameters; laboratory findings; intraoperative or postoperative events; medication usage; and postoperative management. Meta-analysis revealed acceptable discrimination power for all models (sensitivity=0.655, specificity=0.685, AUC=0.709). Subgroup analysis revealed that NN showed the best discrimination power (sensitivity=0.771, specificity=0.882, AUC=0.907). Decent discrimination power was demonstrated by KNN (sensitivity=0.563, specificity=0.766, AUC=0.748), LGB (sensitivity=0.667, specificity=0.788, AUC=0.793), LR (sensitivity=0.755, specificity=0.660, AUC=0.723), and RF (sensitivity=0.750, specificity=0.499, AUC=0.737).

Conclusion: ML models exhibit acceptable overall diagnostic capability for predicting post-AAD repair AKI, with NN demonstrating superior discrimination over other architectures. External validation in broader international cohorts is warranted.

Keywords:acute kidney injury, aortic dissection, artificial intelligence, machine learning, surgery.
Cite This Article: Pradipta, B.P.A., Bramantyo, B.B., Aditya, M.R., Artanti, P.L., Wardhana, K.K. 2026. Machine learning models in predicting acute kidney injury following acute aortic dissection surgery: Meta-analysis of diagnostic accuracy. *Journal of Indonesian Thoracic Cardiac and Vascular Surgery* 3(2): 51-59.

INTRODUCTION

Rapid renal function decline is a hallmark of acute kidney injury (AKI), which is frequently treatable with the right care. About 7% of all hospital admissions and up to 30% of patients admitted to intensive care units have AKI, which is a prevalent illness, especially among hospitalized people. RIFLE, AKIN (Acute Kidney Injury Network), and KDIGO (Kidney Disease: Improving Global Outcomes) are some of the criteria that have been created to diagnose AKI.[1–5] The second most common cause of AKI is known to be cardiac surgery-associated AKI. Compared to other cardiac procedures, the frequency of AKI after acute aortic dissection (AAD) surgery is significantly greater, with reported rates ranging from 23% to 67%. Postoperative AKI is still a dangerous and difficult complication in AAD cases, despite improvements in surgical methods. Postoperative AKI is still a dangerous and difficult complication in AAD cases despite advancements in surgical techniques. According to recent registries, 22–38% of patients need dialysis, a condition associated with a 3.1-fold increase in mortality risk and a significant impact on overall morbidity.[6,7]

Consequently, interest in developing reliable predictive models for post-AAD AKI has grown significantly. Conventional risk-prediction frameworks rely predominantly on logistic regression (LR), capturing simple linear relationships

among clinical factors. However, LR-based models frequently underperform relative to modern machine learning (ML) paradigms, as standard regression fails to resolve complex non-linear dynamics, higher-order feature interactions, and real-time hemodynamic swings inherent to perioperative acute care.[8]

While previous broader meta-analyses have evaluated ML models for general AKI or mixed cardiac surgery populations, their findings cannot be directly extrapolated to AAD repair. Prior generic AKI meta-analyses consistently identified tree-based ensemble methods (e.g., gradient boosting) as top performers; however, whether these findings apply to the distinct hemodynamic context of AAD repair remains unexamined.[8,9] To date, no systematic review or meta-analysis has evaluated the diagnostic performance of ML models specifically within the post-AAD surgical cohort. To date, no meta-analytic investigation has systematically synthesized ML model performance strictly within post-AAD surgical cohorts. To address this evidence gap, the present study was conducted to evaluate the overall prognostic accuracy of diverse ML algorithms and determine the superior model architecture for predicting AKI following AAD repair.

METHODS

Study Protocol

This systematic review and meta-analysis were carried out according to the updated guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement.[10] The protocol for this review was not prospectively registered, as PROSPERO does not currently provide registration for prognostic reviews.

Search Strategy

A systematic literature search was conducted by three independent investigators (BPAP, BBB, and PLA) to identify studies assessing exclusively ML models, a subset of AI, in predicting the occurrence of AKI in patients with AAD repair surgery. Several search engines and databases were employed for the literature search, including PubMed, EuropePMC, and EBSCO. The studies were restricted to being published from January 2020 to July 2025. We meticulously constructed a specific Boolean search string using a combination of free words for the sake of the literature search: (((artificial intelligence) OR (machine learning)) AND ((acute kidney injury) OR (AKI))) AND ((prediction) OR (predicting) OR (predict))) AND ((aortic dissection) OR (acute aortic dissection) OR (AAD)).

Eligibility Criteria

Inclusion and exclusion criteria for the included studies were predetermined before the literature search. Only studies written and published in English were eligible for this meta-analysis study. The studies were deemed eligible if the following inclusion criteria were satisfied: 1) the studies utilized observational or retrospective cohort study design; 2) the studies recruited adult patients aged 18 years or above with the diagnosis of AAD without any restriction on the type of AAD; 3) the patients in the included studies underwent surgical repair or correction of the AAD; 4) the studies assessed the performance of ML algorithms, not limited to certain models, in predicting AKI occurrence following AAD repair surgery; and 5) studies were required to report data regarding the prevalence of the condition of interest and/or the number of true positive, false negative, false positive, and true negative cases.

AKI was defined according to Kidney Disease: Improving Global Outcomes (KDIGO) criteria, which incorporate serum creatinine (SCr) and urine output. The diagnosis of AKI was confirmed when there was an increase in SCr by at least 1.5 times from the baseline or ≥0.3 mg/dl (≥26.5 mmol/l) and the urine output was less than 0.5 ml/kg/h for 6-12 hours.[11] Studies with incomplete data were excluded. Reviews, research letters, editorials, correspondence, or abstracts were excluded.

Study Selection and Data Extraction

Studies identified from the literature search were subsequently collected in the Rayyan intelligent systematic review website (https://www.rayyan.ai) to be screened. A PRISMA flow diagram was then constructed to depict the process of study selection. Four investigators (BPAP, BBB, MRA, and PLA) screened the titles and abstracts of all identified studies. The full papers of all relevant studies were subsequently obtained and assessed for the inclusion and exclusion criteria. The supporting data from the included studies were extracted into a Google Sheet. The data that were extracted encompassed the name of the first author, year of publication, study center (the country where the study was conducted), sample size, type of AAD, AAD repair technique, ML models evaluated in each study, and predictors of AKI occurrence. The main data included the number of true positives, false negatives, false positives, and true negatives for each ML model. If the studies did not report the data of the confusion matrix, the true positive, false negative, false positive, and true negative cases were calculated using the sensitivity/recall and specificity values. When missing data were encountered, the corresponding author of the associated study was contacted, and if there was no response, the study was excluded. Any discrepancy in the data extraction was discussed until a consensus was reached, with guidance from the senior authors (DWA and KKW).

Risk of Bias Assessment

The quality of each included study was assessed using the Quality Assessment of Prognostic Accuracy Studies (QUAPAS) tool. QUAPAS is an extension of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2)-2 tool, which is adapted to assess the quality of studies regarding prognostic accuracy. This checklist consists of five main domains, namely participant selection, index test, outcome, flow and timing, and analysis. A study is deemed low risk when all domains reveal low risk of bias. When there is an unclear risk of bias in one or more domains, the study is considered to demonstrate some concerns. However, if one or more domains exhibit a high risk of bias, the study is subject to being of high risk.[12]

Statistical Analysis

Statistical analysis was conducted by two investigators (BBB and MRA). We employed RStudio for Windows 11 (ver 2024.12.0+467) and applied the 'mada' package to evaluate the pooled

discrimination power. A bivariate random-effects model with the restricted maximum likelihood (REML) method was used to measure the pooled sensitivity, specificity, and diagnostic odds ratio (DOR) with 95% confidence intervals, which were represented using forest plots. The performance of the ML models in terms of prognostic accuracy was represented using summary receiver operating characteristic (SROC) curves with the corresponding area under the curve (AUC) value. AUC values of 0.9-1.0, 0.8-0.9, 0.7-0.8, and 0.5-0.7 indicated excellent, good, acceptable, and poor prognostic accuracy, respectively. I squared (I^2) statistics value by Zhou and Dengtukuri was used to quantify the heterogeneity of the statistical model.[13] (4) I^2 values of 0-25%, 26-50%, 51-75%, and 76-100% represented no, low, moderate, and high heterogeneity, respectively. Subgroup analysis was performed to assess the prognostic accuracy of various ML models, if at least two studies evaluated their performance.

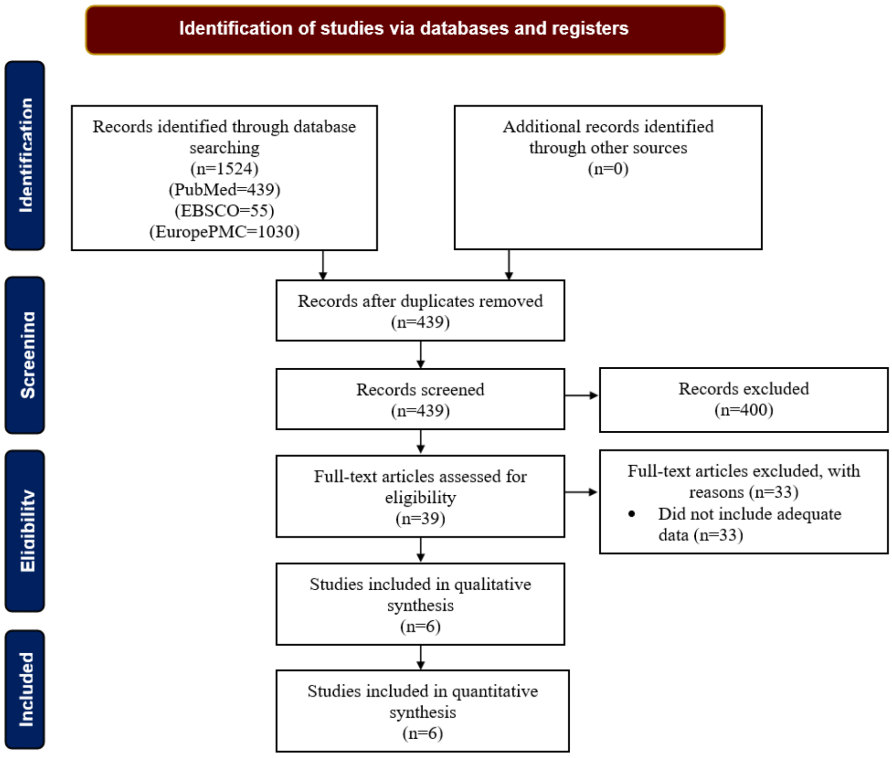


Figure 1. PRISMA flow diagram illustrating the study selection process

RESULTS

Characteristics of Eligible Studies

A total of 1524 studies were identified, and after auto-exclusion of duplicates, non-human studies, and non-English studies, a total of 439 studies remained (**Figure 1**). Of the screened studies, 400 were excluded, and 39 studies were selected for full-text review (PubMed & Europe PMC). Studies that failed to provide prevalence and/or the number of true positives, false negatives, false positives, and true negatives were not included. Four studies were excluded as they did not include data. Among these 6 studies, there were a total of 22 models described, of which 22.7 % (5/22) were categorized as logistic regression, with 45,5 % (10/22) categorized as extreme gradient boosting, support vector machine, gradient boosting machine, light gradient boosting machine, and k-nearest neighbors. An ML model was defined as any method that utilizes an algorithm to learn from the data to improve performance (i.e., AKI prediction). AUC data were provided in 100% (22/22) of the models and were included in subsequent model performan cecomparisons (**Table 1**).

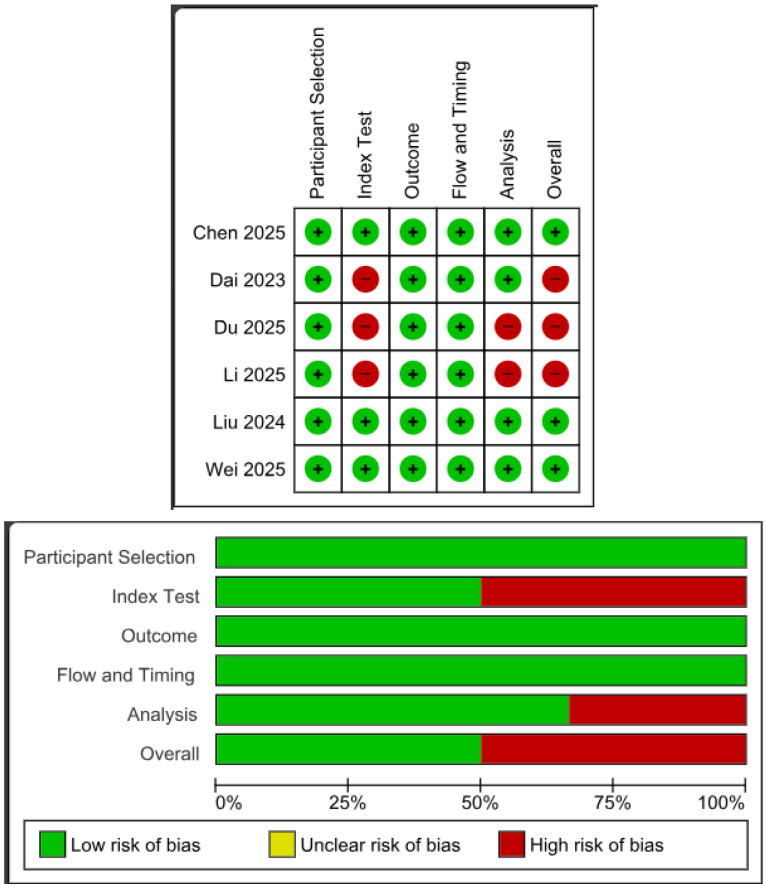


Figure 2. Risk of bias assessment of included studies

Table 1. Study characteristics

Study ID	Study Design	Participants	Patient Enrollment Period	Surgical Procedure	AKI Incidence (n, %)	Model Development Strategy	Model Type	Validation Approach	Outcomes (top models and predictors)
Chen 2025[2]	Retrospective Cohort	202 patients with TA-AAD	November 2018 – October 2023	Thoracotomy and vascular replacement	88 (43.56)	The model was developed using SHAP analysis, which assigns each predictor a value that quantifies its contribution to the model's prediction.	LightGBM, RF, KNN, MLP-NN, NB, LR	DCA and calibration curve	LightGBM model demonstrating the best performance. Ventilation time, Urine output, Diuretics, SCr, Heart rate, Urea, Natriuretic peptide, Urapidil, MCHC, Blood glucose, Circulatory arrest time, BNP, CPB time, AST, ICU VIS, HB, LDH, Creatine Kinase Myoglobin, Lymphocyte, Mean Platelet Volume
Dai 2023[6]	Retrospective Cohort	53 patients with TA-AAD	January 2019 – August 2022	NR	38 (71.70)	The overfitting of the ML models was avoided through a grid search algorithm with 10-fold cross-validation, with the aim of identifying the optimal hyperparameters in the training set.	LR, RFC, SVM, XGB, GBM	DCA	XGBoost model achieves high accuracy. Intraoperative urine output, Cumulative time of mean arterial pressure lower than 65 mmHg outside cardiopulmonary bypass (OUT_CPB_MAP_65 time), Autologous blood transfusion, Smoking, Blood platelet transfusion, Intraoperative diuretics use, Norepinephrine dose, Total area under curve below threshold mean arterial blood pressure 65 mmHg outside cardiopulmonary bypass (OUT_CPB_MAP_AUT_65), Fraction of intraoperative hypotension duration during CPB.
Du 2025[3]	Retrospective Cohort	383 patients with TA-AAD	2009-2015 (phase I), 2016-2025 (phase II)	NR	115 (30.03)	LR with 10-fold cross-validation was used to determine the predictors, and the final independent predictors were selected through backward elimination.	LR	DCA	The nomogram demonstrates superior preoperative discriminative accuracy. Cardiac dysfunction, Microvascular complications of diabetes, Baseline renal impairment, Peripheral vasculopathy, BUN, GFR, Serum creatinine
Li[4] 2025	Retrospective Cohort	581 patients with TA-AAD	February 2010 – July 2023	NR	301 (51.81)	LR was used to remove less-important predictors, and the model's effectiveness was assessed through the ROC curve and Harrel's C-index.	LR	Bootstrap resampling and adjusted relative C-index calculation	The nomogram exhibits robust predictive capabilities. Demographic characteristics (age, sex, body mass index [BMI]), Existing health conditions (hypertension, diabetes, history of cardiovascular diseases), Initial laboratory measurements (serum creatinine, hemoglobin, white blood cell count), Historical medication usage (angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers, diuretics, nonsteroidal anti-inflammatory drugs), Intraoperative factors (blood loss, fluid replacement, use of vasopressors), Postoperative clinical progressions (complications such as bleeding, infection), Subsequent laboratory findings (alterations in serum creatinine, urine output), Postoperative medical management (dialysis requirement, use of nephrotoxic medications).

Table 1. Continued

Study ID	Study Design	Participants	Patient Enrollment Period	Surgical Procedure	AKI Incidence (n, %)	Model Development Strategy	Model Type	Validation Approach	Outcomes (top models and predictors)
Liu[5] 2025	Retrospective Cohort	397 patients with TA-AAD	January 2018 – December 2021	Total Aortic Replacement	93 (23.43)	Intersection predictors identified through 3 ML models (RF, SVM-RFE, and LR) were selected to develop a more stable model.	LR, ANN	Bootstrap resampling	The ANN model showed a higher predictive ability for severe AKI than logistic regression. Baseline serum creatinine (sCr), ICU admission variables: Serum cystatin C (sCysC), Procalcitonin (PCT), Aspartate transaminase (AST), Platelet count, Lactate dehydrogenase (LDH), Urine N-acetyl-β-D-glucosaminidase (uNAG), APACHE II score.
Wei 2025[1]	Retrospective Cohort	97 patients with AAD	2001-2019	NR	25 (25.77)	Model performance was enhanced through 10-fold cross-validation, and SHAP was utilized to identify the optimal model.	SVM, GBM, NNET, KNN, XGB, LGB, CAT	External validation using 179 participants from the MIMIC-III database	The CatBoost model exhibited superior performance. Body weight, BMI, APSIII score, Minimum and maximum values of BUN, Minimum and maximum values of creatinine, Maximum glucose levels, Urine output

AKI = Acute Kidney Injury; ANN = Artificial Neural Network; APACHE = Acute Physiology and Chronic Health Evaluation; APS = Acute Physiology Score; AST = Aspartate Aminotransferase; BMI = Body Mass Index; BNP = Brain Natriuretic Peptide; BUN = Blood Urea Nitrogen; CAT = Categorical Boosting; CPB = Cardiopulmonary Bypass; DCA = Decision Curve Analysis; GBM = Gradient Boosting Machine; GFR = Glomerular Filtration Rate; HB = Hemoglobin; ICU = Intensive Care Unit; KDIGO = Kidney Disease: Improving Global Outcomes; KNN = K-Nearest Neighbors; LDH = Lactate Dehydrogenase; LightGBM/LGB = Light Gradient Boosting Machine; LR = Logistic Regression; MCHC = Mean Corpuscular Hemoglobin Concentration; MLP-NN = Multilayer Perceptron Neural Network; NB = Naive Bayes; NNET = Neural Network; OUT_CPB_MAP < 50/55/60/65 time = Cumulative Time with Mean Arterial Pressure Below 50/55/60/65 mmHg Outside Cardiopulmonary Bypass; OUT_CPB_MAP_AUT_50/55/60/65 = Total Area Under the Curve Below Mean Arterial Pressure Thresholds of 50/55/60/65 mmHg Outside Cardiopulmonary Bypass; RF = Random Forest; RFC = Random Forest Classifier; SCr = Serum Creatinine; SHAP = SHapley Additive exPlanations; SVM = Support Vector Machine; VIS = Vasoactive-Inotropic Score; XGB = Extreme Gradient Boosting.

Risk of Bias Assessment of Included Studies

Risk of bias assessment using the QUAPAS tool revealed that half of the included studies were judged to have a low risk of bias, whereas the remaining three studies were considered to have a high risk of bias (**Figure 2**). The high risk of bias was primarily attributable to the index test and analysis domains. Therefore, the findings of this review should be interpreted with caution, given the potential influence of methodological limitations in the included studies.

Predictors

To determine the variables with the highest predictive value in machine learning models, a comparative synthesis of six studies was conducted. These studies collectively incorporated a diverse array of predictors, including demographic characteristics; pre-existing comorbidities; baseline, intraoperative, and postoperative hemodynamic parameters; laboratory findings; intra- and postoperative events; medication usage; and postoperative management strategies.

Serum creatinine (Scr) emerged as the most consistently cited predictor, identified in four of the six studies (66.7%), underscoring its pivotal role in the early detection of AKI. Blood urea nitrogen (BUN) and blood glucose were also notable predictors, appearing in three studies (50%). Meanwhile, urine output, diuretic use, aspartate aminotransferase (AST), lactate dehydrogenase (LDH), baseline renal impairment, and body mass index (BMI) were reported in two studies each (33.3%). This variation in predictive variables across studies highlights the heterogeneity of AKI risk profiles. It reinforces the importance of developing context-specific, tailored predictive models to enhance early identification and intervention in high-risk patients.

Performance of ML Models

The diagnostic performance of machine learning (ML) models in predicting acute kidney injury (AKI) following acute aortic dissection (AAD) surgery was evaluated using pooled sensitivity, specificity, diagnostic odds ratio (DOR), and area under the curve (AUC). As presented in

Table 2. Performance of ML models

ML Models	Number of Included Studies	Number of Participants	Sensitivity	Specificity	DOR	AUC	I^2
All Model	6	1713	0.655	0.685	4.121	0.709	38%
NN	2	669	0.771	0.882	26.875	0.907	0%
GBM	2	299	0.559	0.827	6.156	0.647	0%
KNN	2	299	0.563	0.766	4.335	0.748	0%
LGB	2	299	0.667	0.788	7.489	0.793	0%
LR	5	3437	0.755	0.660	1.565	0.723	67%
RF	2	467	0.750	0.499	2.991	0.737	0%
SVM	2	362	0.511	0.746	3.013	0.568	0%
XGB	2	362	0.673	0.591	3.017	0.680	0%

Table 2, the meta-analysis included six studies comprising 1,713 participants and 22 model evaluations. The overall analysis demonstrated acceptable discriminative performance, with an AUC of 0.709, pooled sensitivity of 0.655 (95% CI: 0.553–0.744), specificity of 0.685 (95% CI: 0.521–0.814), and DOR of 4.121. The overall heterogeneity reported in **Table 2** was 37.7%.

Subgroup analysis demonstrated differences in diagnostic performance across ML algorithms. Neural network (NN) models achieved the highest AUC of 0.907, with a sensitivity of 0.771 (95% CI: 0.161–0.983), specificity of 0.882 (95% CI: 0.704–0.959), and DOR of 26.875. Light Gradient Boosting Machine (LGB) demonstrated an AUC of 0.793, sensitivity of 0.667, and specificity of 0.788, followed by k-nearest neighbors (KNN; AUC=0.748, sensitivity=0.563, specificity=0.766), random forest (RF; AUC=0.737, sensitivity=0.750, specificity=0.499), and logistic regression (LR; AUC=0.723, sensitivity=0.755, specificity=0.660). In contrast, extreme gradient boosting (XGB; AUC=0.680), gradient boosting machine (GBM; AUC=0.647), and support vector machine (SVM; AUC=0.568) exhibited lower discriminative performance. Heterogeneity was reported as 0% for all model-specific subgroups except LR, which demonstrated an I^2 of approximately 67% (**Table 2**).

The forest plot of sensitivity estimates is presented in **Figure 3**. Individual model evaluations demonstrated considerable variation in sensitivity, with a random-effects pooled estimate of approximately 0.66 (95% CI: 0.55–0.75). Substantial heterogeneity was observed across

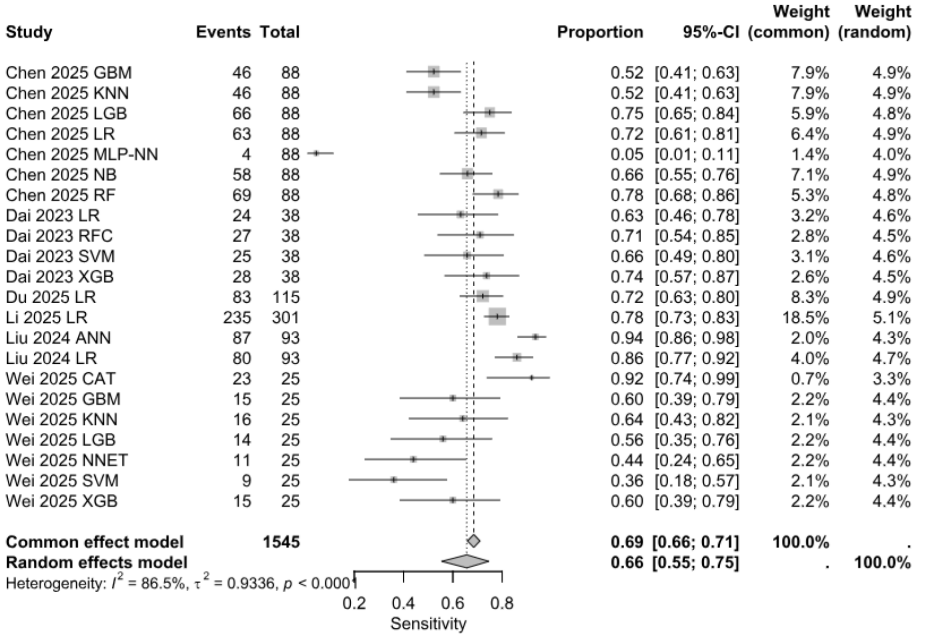


Figure 3. Sensitivity analysis of machine learning models

the individual evaluations (I^2=86.5%). The corresponding specificity analysis, presented in **Figure 4**, demonstrated a random-effects pooled estimate of approximately 0.68 (95% CI: 0.52–0.81), with substantial heterogeneity (I^2=96.6%). These forest plots illustrate the variability in the ability of individual models to correctly identify patients with and without postoperative AKI.

The overall and model-specific summary receiver operating characteristic (SROC) curves are illustrated in **Figure 5**. The overall SROC analysis yielded an AUC of 0.709, while the NN subgroup demonstrated the highest AUC of 0.907. The remaining subgroup AUC values were 0.793 for LGB, 0.748 for KNN, 0.737 for RF, 0.723 for LR, 0.680 for XGB, 0.647 for GBM, and 0.568 for SVM. These results demonstrate differences in discriminative performance across the evaluated algorithms. Subgroup meta-analysis was not performed for categorical boosting (CAT), multilayer perceptron neural network (MLP-NN), and naïve Bayes (NB), because each model was evaluated in only one study.

DISCUSSION

To the best of our knowledge, this study is the first systematic review and meta-analysis to assess ML performance in predicting AKI following AAD repair surgery. On average, all ML models performed similarly at predicting AKI, with substantial variability across models.

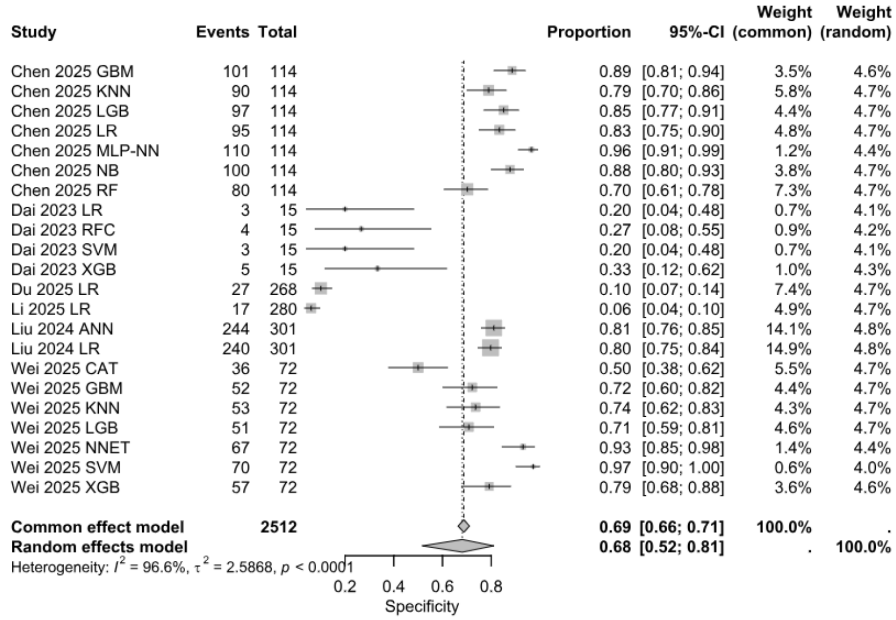


Figure 4. Specificity analysis of machine learning models

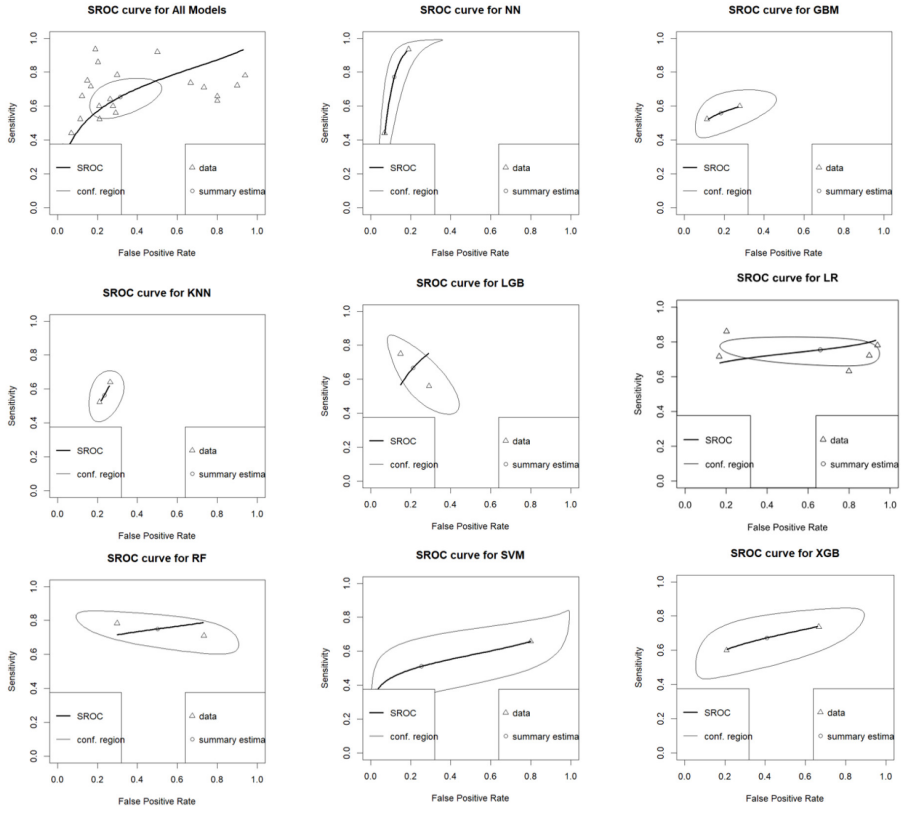


Figure 5. SROC curve for all models and each model, respectively

Subgroup analysis revealed that a neural network consisting of a neural network and an artificial neural network possessed superior performance at predicting AKI compared to logistic regression, k-nearest neighbors, random forest, gradient boosting, and support vector machine.[1–3] Serum creatinine-related metrics were the most frequently identified significant predictors (featured in 5 of 6 studies), followed by urine output measurements (present in 4 of 6 studies). These findings confirm that while ML strategies are generally effective for forecasting post-AAD AKI, predictive accuracy is heavily governed by the choice of algorithmic architecture and feature selection. Notably, our findings contrast with earlier meta-analyses; for example, Song et al. reported that gradient boosting models achieved top performance over artificial neural networks, Bayesian networks, random forest, and logistic regression. However, their synthesis included broad clinical populations without focusing on the specialized context of acute aortic dissection.[7]

Our study showed numerous predictors for calculating acute kidney injury among all models. Although ML algorithms have inherent structural differences, the variables included in the model can drastically alter the quality of prediction. Creatinine and urine output, two variables that are utilized in AKI staging (KDIGO), were noted to be significant predictors in multiple studies that utilized various ML models. Other variables related to kidney function, such as BUN, were noted to be significant predictors in multiple studies that utilized various ML models. Other variables related to kidney function, such as BUN, were also noted in multiple studies. Baseline variables as well as certain medications (e.g., vasopressors, inotropics, and diuretics) in acute aortic dissection should be examined in more detail to determine their value in improving ML performance.[1–6]

In the study by Wei et al., categorical boosting (CatBoost) yielded superior predictive accuracy compared to SVM, GBM, NNET, KNN, XGB, and LightGBM using feature sets that included body weight, BMI, APS III score, extreme BUN/creatinine values, peak glucose, and urine output. Nevertheless, no parallel studies were available to corroborate this specific model's advantage further.[5]

Whether ML algorithms are currently utilized in active point-of-care settings for AKI prediction remains uncertain. The wide diversity of model structures, along with variations in data processing strategies, leads to considerable variability in predictive output, as demonstrated by our results. In clinical workflows, the primary value of an ML model lies in

serving as an automated early warning system, offering dynamic, real-time risk stratification to assist clinicians with timely medical decisions and resource planning.

For any ML framework to achieve clinical utility, rigorous model training, validation, testing, and calibration are mandatory. All six included studies explicitly outlined appropriate training protocols (such as cross-validation), with no major methodological discrepancies identified in model development. However, none of the primary studies evaluated external validation in outside clinical cohorts, and the area under the curve (AUC) remained the predominant reported metric for model evaluation.[1,2,4]

A critical limitation identified across the included literature is the primary reliance on single-center, internal validation cohorts. Predictive performance often drops significantly when ML models are deployed in outside healthcare settings due to site-specific variations in surgical protocols, intraoperative cardiopulmonary bypass (CPB) management, and regional patient demographics. Without rigorous external validation across diverse, multi-institutional cohorts, models risk failing to generalize to broader clinical populations.

Compounding this issue is the high risk of algorithmic overfitting. Several evaluated studies analyzed high-dimensional feature sets relative to their limited sample sizes. When training complex algorithms—such as Neural Networks or Gradient Boosting Machines—on small patient samples with dozens of predictor variables, models risk memorizing noise and center-specific idiosyncrasies rather than learning true physiological signals. Overfitted models demonstrate artificially inflated discrimination during internal validation but experience steep performance degradation in real-world deployment. Future research must prioritize robust cross-validation techniques, feature dimensionality reduction, and prospective multi-center external validation before these predictive tools can be safely implemented in clinical workflows. Finally, seamless integration into existing Electronic Health Record (EHR) systems represents the operational bridge between theoretical models and real-time bedside application. Modern EHR-integrated algorithms should automatically extract dynamic intraoperative and postoperative variables, such as blood pressure fluctuations, urine output, and laboratory trends, to generate automated AKI risk scores in real time. Such integration enables early target interventions, such as optimizing hemodynamic targets and avoiding nephrotoxic agents in high-risk patients before overt renal failure occurs. Addressing the triad of external validity, model interpretability, and automated EHR integration will be key to transitioning ML algorithms from retroactively validated concepts to transformative clinical tools in AAD perioperative care.

In addition, half of the included studies were considered to have a high risk of bias, which limits the ability to draw robust and high-quality conclusions. Furthermore, prospective registration of the study protocol was not possible due to the current limitations of the PROSPERO platform regarding the registration of prognostic accuracy reviews. Further high-quality studies with more homogeneous study protocols, particularly regarding the identification and assessment of predictors of AKI, are warranted to validate and strengthen our findings.

CONCLUSION

In summary, machine learning algorithms demonstrate promising potential as advanced tools for predicting acute kidney injury following acute aortic dissection repair, with Neural Network architectures offering superior discrimination over traditional logistic regression. However, before these algorithms can be safely transitioned into routine perioperative practice, future research must prioritize rigorous multicenter validation, prospective real-world evaluation, and formal assessments of their true clinical impact on patient outcomes. Crucially, machine learning models should serve to complement—rather than replace—established clinical risk assessment approaches and physician judgment. By integrating data-driven predictions with clinical expertise, health systems can enhance early risk identification while maintaining holistic, patient-centered care.

DISCLOSURES

ETHICAL CONSIDERATION

Ethics approval and informed consent were not required for this systematic review and meta-analysis of previously published studies.

AUTHOR CONTRIBUTION

BPAP contributed to the study conceptualization, research design, methodology development, supervision, and critical revision of the manuscript. BBB contributed to the literature search strategy, study selection, data interpretation, and manuscript review. MRA performed the systematic literature search, data extraction, quality assessment, and preparation of the initial manuscript draft. PLA contributed to data extraction, risk of bias assessment, statistical analysis, and interpretation of the meta-analytic findings. KKW contributed to methodological validation, interpretation of diagnostic accuracy outcomes, supervision, and critical revision of the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

FUNDING

No external funding was received for conducting this study.

DECLARATION OF USE OF ANY GENERATIVE AI TOOLS

The authors used QuillBot for language editing only. No AI tools were used for data analysis, interpretation, or generation of scientific content.

REFERENCES

1. Wei Z, Liu S, Chen Y, Liu H, Liu G, Hu Y, et al. Machine Learning Model-Based Prediction of In-Hospital Acute Kidney Injury Risk in Acute Aortic Dissection Patients. Rev Cardiovasc Med. 2025;26(2):25768. Available from: http://dx.doi.org/10.31083/rcm25768

2. Chen Z, Lu X, Liu M, Yu Y, Yu L, Cheng S, et al. Artificial Intelligence–Driven Prediction of Acute Kidney Injury Following Acute Type A Aortic Dissection Surgery in a Chinese Population. J Cardiothorac Vasc Anesth. 2025;39(10):2729–38. Available from: http://dx.doi.org/10.1053/j.jvca.2025.05.047
3. Du R, Wang L, Wang Y, Zhao Z, Zhang D, Zuo S. AKI prediction model in acute aortic dissection surgery: nomogram development and validation. Front Med. 2025;May 15(12):1562956. Available from: http://dx.doi.org/10.3389/fmed.2025.1562956
4. Li W, Yu W, Chen Y, Tan W, Zhang F, Zhang Y. Development and validation of a nomogram for predicting acute kidney injury risks in patients undergoing acute Stanford type A aortic dissection repair surgery. BMC Nephrol. 2025;26(1):257. Available from: http://dx.doi.org/10.1186/s12882-025-04150-y
5. Liu X, Fang M, Wang K, Zhu J, Chen Z, He L, et al. Machine learning-based model to predict severe acute kidney injury after total aortic arch replacement for acute type A aortic dissection. Heliyon. 2024;10(13):e34171. Available from: http://dx.doi.org/10.1016/j.heliyon.2024.e34171
6. Dai A, Zhou Z, Jiang F, Guo Y, Asante DO, Feng Y, et al. Incorporating intraoperative blood pressure time-series variables to assist in prediction of acute kidney injury after type A acute aortic dissection repair: an interpretable machine learning model. Ann Med. 2023;55(2):2266458. Available from: http://dx.doi.org/10.1080/07853890.2023.2266458
7. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int J Med Inform. 2021;151:104484. Available from: http://dx.doi.org/10.1016/j.ijmedinf.2021.104484
8. Goyal A, Daneshpajouhnejad P, Bashir K. Acute Kidney Injury. In Treasure Island (FL); 2026. Available from: https://pubmed.ncbi.nlm.nih.gov/28722925/
9. Wang Z, Xu J, Zhang Y, Chen C, Kong C, Tang L, et al. Prediction of acute kidney injury incidence following acute type A aortic dissection surgery with novel biomarkers: a prospective observational study. BMC Med. 2023;21(1):503. Available from: http://dx.doi.org/10.1186/s12916-023-03215-9
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. J Clin Epidemiol. 2021;134:103–12. Available from: http://dx.doi.org/10.1016/j.jclinepi.2021.02.003
11. Lameire N, Kellum JA, Aspelin P. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl. 2012;2:1–138. Available from: http://dx.doi.org/10.1038/kisup.2012.1
12. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. Ann Intern Med. 2022;175(7):1010–8. Available from: http://dx.doi.org/10.7326/m22-0276
13. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: The case of meta-analyses of diagnostic accuracy. Stat Med. 2014;33(16):2701–17. Available from: http://dx.doi.org/10.1002/sim.6115

This work is licensed under a Creative Commons Attribution