Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Causal-Weighted Machine Learning for RRT Prediction in Acute Kidney Injury: Development, External Validation, and Clinical Decision Support

Zhaoyu Liao
RMIT University
Ziyi Chen
The University of Melbourne
Qiuchen Huang
qiuchen_1984@126.com

Beijing Bo Ai Hospital
Qiang Du
Tsinghua University
Xiaozhu Liao
First People's Hospital of Zunyi

Research Article

Keywords: Acute kidney injury, renal replacement therapy, machine learning, clinical decision support, external validation, SHAP

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11021990/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Causal-Weighted Machine Learning for RRT Prediction in Acute Kidney Injury: Development, External Validation, and Clinical Decision Support

Zhaoyu Liao[1], Ziyi Chen[2], Qiuchen Huang[5*], Qiang Du[3*], Xiaozhu Liao[4*]

[1]School of Computing Technologies, Royal Melbourne Institute of Technology (RMIT University), Melbourne, VIC 3000, Australia.
[2]The University of Melbourne, Melbourne, VIC 3010, Australia.
[3]Tsinghua University, Beijing, 100084, China.
[4]The First People's Hospital of Zunyi, Zunyi, 563000, China.
[5]Department of Physical Therapy, Beijing Bo Ai Hospital, China Rehabilitation Research Center, School of Rehabilitation, Capital Medical University, Beijing, China.

*Corresponding author(s). E-mail(s): qiuchen_1984@126.com; qud22@mails.tsinghua.edu.cn; 260141367@qq.com;
Contributing authors: s4099210@student.rmit.edu.au; ziyi.chen13@student.unimelb.edu.au;

Abstract

Background: Acute kidney injury (AKI) affects 20–50% of ICU admissions, and the optimal timing for renal replacement therapy (RRT) initiation remains contested. **Methods:** We developed and externally validated a leakage-safe, causal-weighted machine learning pipeline predicting 48-hour RRT initiation in patients with KDIGO stage 3 AKI, using harmonized MIMIC-IV and eICU data. A regularized stacking ensemble with inverse-probability-weighted meta-learning combined five base learners (LightGBM, XGBoost, Random Forest, CatBoost, ExtraTrees) via a logistic-regression meta-learner. **Results:** The model achieved AUC 0.894 internally and 0.807 on an independent, 200+ hospital external cohort, significantly outperforming four of five baseline models (DeLong $\boldsymbol{p < 0.001}$) while remaining statistically indistinguishable from a single LightGBM model. Stacking-consistent SHAP identified creatinine and urine output trajectories as dominant risk drivers, 40% of high-risk patients had a feasible short-term

recourse below threshold, and fairness assessment showed no clinically meaningful subgroup gaps. **Conclusions:** A prototype clinical decision support system operationalizes the model for bedside use, motivating prospective evaluation before deployment.

Keywords: Acute kidney injury, renal replacement therapy, machine learning, clinical decision support, external validation, SHAP

1 Introduction

Deciding when to start renal replacement therapy for a patient with acute kidney injury is one of critical care's most consequential judgment calls: initiate too early and patients face avoidable bloodstream infections and hemodynamic instability; wait too long and they risk fluid overload, refractory hyperkalemia, and acid-base derangements. Acute kidney injury affects 20–50% of ICU admissions worldwide, with mortality of 20–60% depending on severity [1]; renal replacement therapy is the cornerstone intervention for its most severe cases, yet optimal initiation timing remains contentious [2], with profound implications for survival, resource allocation, and cost.

The AKIKI [2] and STARRT-AKI [3] trials showed a delayed strategy is preferable for hemodynamically stable patients, but underscored the difficulty of identifying high-risk patients whose conservative management will fail; the ELAIN trial [11] showed survival benefits of early initiation in certain surgical populations, highlighting treatment-effect heterogeneity that generalized guidelines fail to capture. Machine learning can integrate multiple clinical variables to predict trajectories more precisely than static risk scores [4], yet few RRT models have been externally validated across diverse health systems [5], and models trained on local cohorts often fail to generalize to new practice environments.

This study addresses these gaps in four ways. We introduce a causal-weighted loss function that reduces reliance on institutional treatment-assignment patterns in favor of physiological deterioration signals, and we evaluate transportability on a large, independent multi-center external cohort drawn from eICU, which spans over 200 hospitals. We also build a panoramic SHAP framework that maps risk scores to patient pathophysiology to support clinician trust, and we develop a prototype, EHR-integratable clinical decision support system that operationalizes the model for real-time bedside use.

2 Materials and Methods

2.1 Reporting Guidelines

This study followed TRIPOD/TRIPOD+AI [14, 15]; DECIDE-AI does not apply, as no prospective clinical evaluation was performed, and PROBAST-AI bias domains are addressed qualitatively in Sections 2 and 4.5.

2.2 Study Design and Data Sources

This multi-center retrospective study develops and externally validates an ML model for RRT prediction, using MIMIC-IV (Beth Israel Deaconess Medical Center, Boston, 2008–2019 [7]) for development and the eICU Collaborative Research Database (200+ US hospitals, 2014–2015 [8]) for external validation. End-to-end pipeline architecture, leakage-safe splitting, and feature/model flow are summarized in Figure 1.

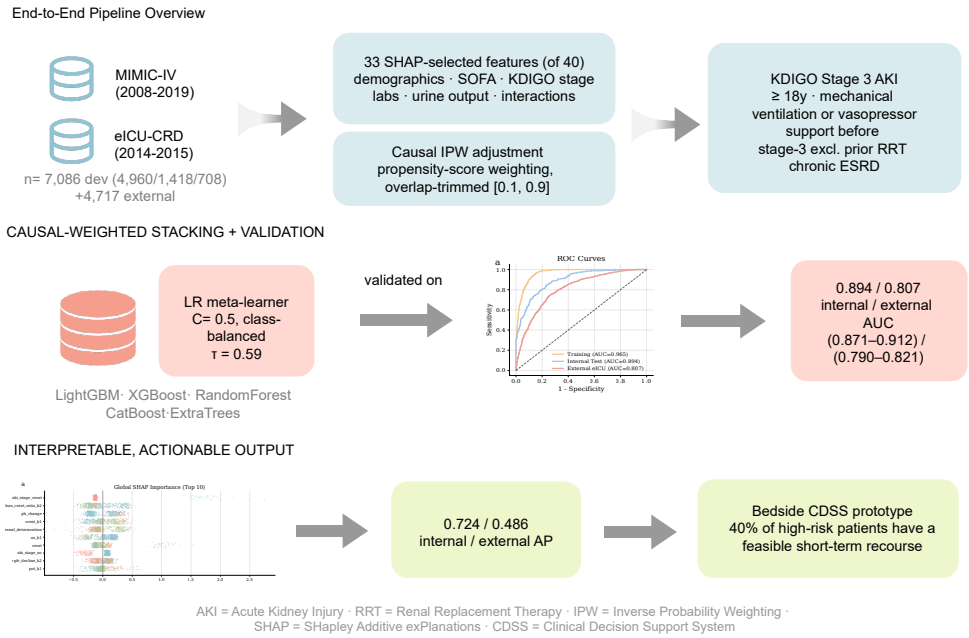


Fig. 1 End-to-end pipeline overview. KDIGO stage 3 AKI patients from MIMIC-IV (single-center, 2008–2019) and eICU-CRD (200+ hospitals, 2014–2015) are fused with leakage-safe splitting: 50% of eICU reserved for external validation (n=4,717); the remainder combined with MIMIC-IV and split 7:2:1 into training (4,960), internal test (1,418), and validation (708), for a total fused development cohort of n=7,086. The 33 SHAP-selected features and overlap-trimmed IPW causal adjustment feed into a regularized stacking ensemble of five base learners (LightGBM, XGBoost, Random Forest, CatBoost, ExtraTrees) combined by a logistic regression meta-learner ($\tau = 0.59$), producing a validated RRT risk score (AUC 0.894 internal, 0.807 external), stacking-consistent SHAP attribution, and bedside CDSS output with actionable recourse.

2.3 Study Population

Eligible patients were ≥18 years at ICU admission with confirmed KDIGO stage 3 AKI [10] who received mechanical ventilation or vasopressor support before the stage 3 diagnosis. We excluded patients with RRT before the index time (prior-treatment confounding) and those with chronic dialysis-dependent end-stage renal disease. Cohort assembly and the resulting eligible, fused, and split sample sizes are summarized in Figure 2.

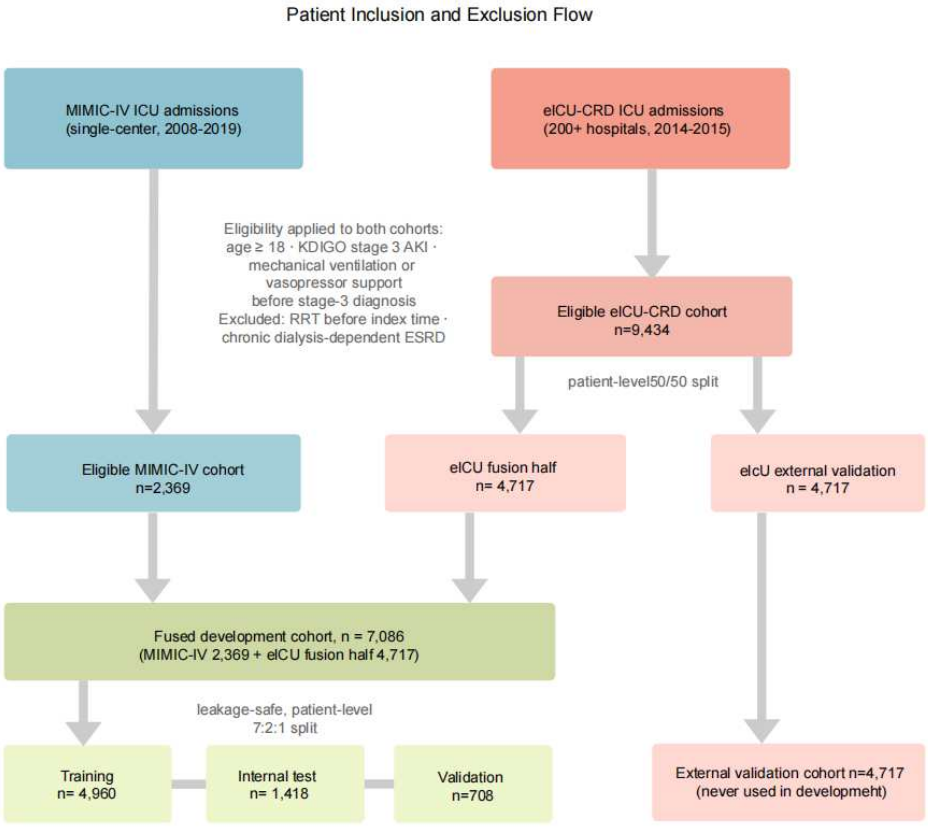


Fig. 2 Patient inclusion and exclusion flow (TRIPOD item 13a). The same eligibility criteria (age ≥18, KDIGO stage 3 AKI, mechanical ventilation or vasopressor support before stage-3 diagnosis; excluding prior RRT and chronic dialysis-dependent end-stage renal disease) were applied independently to MIMIC-IV and eICU-CRD. eICU-CRD was split at the patient level before any development took place, so the external cohort never contributed to model development.

2.4 Outcome Definition

The outcome was RRT initiation within 48 hours after first KDIGO stage 3 diagnosis, defined as the first documented active dialysis session (CRRT, intermittent hemodialysis, SLED, or peritoneal dialysis), identified from MIMIC-IV treatment-table flags and eICU dialysis search strings/offsets.

Outcomes were harmonized to the first event per admission, excluding pre-stage-3 treatment and non-therapeutic procedures (e.g., catheter care); a 200-event eICU sample showed 94% agreement with medication/procedural records.

2.5 Prediction Task Definition and Landmark Framework

All predictors used only information available at decision time; columns encoding future treatment timing were excluded before training, and a pre-training check verified no single feature reached AUC >0.85 against the outcome (which would suggest residual leakage). Training, validation, internal test, and external cohorts were kept separate throughout, and the decision threshold was chosen on the validation split only.

2.6 Feature Engineering

The model used 33 SHAP-selected features (from 40 candidates, mean |SHAP| threshold 0.005) across six categories: demographics; organ failure (SOFA [13]); AKI staging (KDIGO); lab trajectories (creatinine, BUN, potassium, bicarbonate, pH, lactate, eGFR); urine output trajectories; and four derived physiological-interaction terms. IPW weights were applied at the meta-learner level, never as direct features; all features used only pre-prediction-time information, z-score normalized on training statistics.

2.7 Model Development

We employed a regularized stacking ensemble [17] of five base learners (LightGBM [16], XGBoost [19], Random Forest [20], CatBoost [21], and ExtraTreesClassifier [22]) chosen for diverse, interpretable pattern capture, with hyperparameters selected via Bayesian optimization (Optuna, average precision) using patient-level GroupShuffleSplit to prevent leakage.

Each base learner used shallow, regularized settings (depth 4–6, 80–200 trees; full hyperparameters in Table 1). The meta-learner is L2-regularized logistic regression (C=0.5, balanced) on 5-fold out-of-fold predictions, with overfitting further controlled via feature/row subsampling, early stopping, minimum-information-gain feature selection, and class-frequency-inverse sample weighting (∼15% positive rate); the operating threshold ($\tau = 0.59$) was optimized on the validation split via the Youden index.

2.8 Mathematical Formulation

For patient i with feature vector x_i and binary outcome y_i (RRT initiated in the target window), we fit two logistic-regression propensity models [18] estimating treatment tendency at k1 and k2, $\hat{e}(x_i) = P(T{=}1 \mid x_i)$, used only to construct inverse probability weights, never as direct features. The five base learners each output a probability $z_{im} = f_m(x_i)$, which the logistic-regression meta-learner combines into the final prediction:

$$\hat{p}_i = \sigma\left(\beta_0 + \sum_{m=1}^{5} \beta_m z_{im}\right), \quad \sigma(u) = \frac{1}{1+e^{-u}}. \tag{1}$$

To avoid shortcut learning, IPW is applied only at the meta-learner level (not in base learners), adjusting for treatment-assignment bias during ensemble integration while

Table 1 Regularized stacking configuration. Base learners use constrained depth and tree counts; the meta-learner is L2-regularized logistic regression with 5-fold out-of-fold CV for meta-feature generation. The operating threshold ($\tau = 0.59$) is tuned on the validation split only.

Parameter	Value
Base learners	LightGBM + XGBoost + RandomForest + CatBoost + ExtraTrees
Meta learner	Logistic Regression (C=0.5, balanced)
LightGBM depth / trees	max_depth=4–5 / n_estimators=80–100
XGBoost depth / trees	max_depth=5 / n_estimators=100
RandomForest depth / trees	max_depth=6 / n_estimators=100
CatBoost depth / iterations	depth=4 / iterations=200
ExtraTrees depth / trees	max_depth=6 / n_estimators=200
Regularization	L1/L2 penalties + class balancing
Stacking CV	5-fold (out-of-fold meta-feature generation)
IPW weighting	Meta-learner only, overlap-trimmed [0.1, 0.9] on $\hat{e}(x)$ (Crump et al. 2009); trimmed samples excluded
Threshold tuning	Validation split only ($\tau = 0.59$)

letting base learners capture physiological patterns undistorted:

$$w_i = y_i \frac{P(T=1)}{\hat{e}(x_i)} + (1 - y_i) \frac{P(T=0)}{1 - \hat{e}(x_i)}. \tag{2}$$

Rather than clipping weights to an arbitrary finite range, we apply overlap-region trimming [38]: patients with $\hat{e}(x_i)$ outside $[0.1, 0.9]$ are excluded from the weighted loss entirely, and remaining weights normalized to unit mean. This affected 11.6% of the training cohort and was selected over ad hoc clipping after a systematic comparison (Section 3.6) showed it was the only scheme to improve validation, internal, and external AUC simultaneously. The meta-learner minimizes L2-regularized weighted log-loss with these weights.

Discrimination (AUC-ROC), precision-recall (AP), and calibration (Brier score [23]) used standard definitions. Binary classification used threshold $\tau = 0.59$. Clinical utility was assessed via decision curve analysis net benefit [12], computed on validation-fit isotonic-calibrated probabilities rather than raw output, since a threshold probability is only interpretable as a clinical risk threshold when inputs are themselves calibrated to true risk (Model Calibration and Reliability subsection below discloses the raw/calibrated gap).

2.9 Statistical Analysis

Performance was evaluated via AUC-ROC, accuracy, sensitivity, specificity, PPV, NPV, F1, and Brier score on internal test and external cohorts, with 95% CIs from the BCa bootstrap [28] (2000 resamples/cohort); model-variant differences used paired bootstrap tests ($p < 0.05$), baseline comparisons the DeLong test [29].

2.10 Causal Framework

The causal structure guides feature construction and propensity score adjustment: covariates X (SOFA, labs, demographics) influence treatment T and outcome Y (RRT within 48h), with unobserved severity/physiology potentially influencing both. Following the backdoor criterion [25], conditioning on pre-treatment covariates X blocks this confounding path. We do not claim causal treatment effects; this prediction-oriented adjustment instead improves cross-center transportability by reducing reliance on site-specific treatment-assignment patterns.

2.11 Missing Data Handling

To preserve temporal physiology while limiting leakage, time-series variables used Last Observation Carried Forward with bounded linear interpolation; static variables used training-set median imputation; sparsely observed labs used SAITS [27] trained on the development set only. Variables with $> 10\%$ missingness received a missing-indicator, with imputers fit on training data only.

2.12 Competing Risk Considerations

Death or discharge before possible therapy are competing events [26]: in the fused development cohort (n=7086), 1231 (17.4%) started therapy, 856 (12.1%) died or were discharged, and 4999 (70.5%) remained without therapy. A Fine-Gray subdistribution hazard regression confirmed creatinine trajectory and urine output change rate retained high significance ($p < 0.001$) as competing events.

2.13 External Validation Strategy

Half of the eICU database (4717 records) was held out and excluded from all model development, serving only as the external test set; the remainder was combined with MIMIC-IV for development, avoiding optimistic bias.

2.14 Interpretability Analysis

We used SHAP values [9] for global/local interpretation. Since the deployed model is a five-learner stacking ensemble with an IPW-weighted logistic meta-learner, standard TreeSHAP cannot apply directly; we computed a *stacking-consistent* decomposition: exact margin-space TreeSHAP per base learner, converted to each learner's native probability output via an exact Aumann-Shapley path-integral allocation (equivalent to Integrated Gradients [37], 200-step quadrature; bagging learners needed no transform, already probability-space), then linearly combined through the meta-learner's coefficients, yielding an exact additive decomposition of the deployed ensemble that we verified against `model.predict_proba` (mean error 2.1×10^{-7}, max 5.7×10^{-7}, $n = 500$).

2.15 Actionable Recourse Analysis

To complement SHAP with bedside-actionable guidance, we implemented a dependency-consistent counterfactual recourse search [36]: for high-risk patients ($\hat{p} \geq \tau$), we searched for the minimal joint change in short-term-modifiable variables (urine output, pH, bicarbonate, potassium, lactate) lowering risk below τ, via vectorized differential evolution restricted to physiologically bounded directions (Table 2), with each perturbation triggering recomputation of dependent engineered features and the k2 propensity score to avoid internally inconsistent counterfactuals. Search prioritized sparsity; bounds are illustrative, not validated targets, requiring clinical calibration.

Table 2 Assumed clinical bounds for the actionable recourse counterfactual search (Section 2.15). Bounds represent illustrative short-term (within the decision window) treatment-response magnitudes and require calibration by clinical collaborators before deployment; they are not empirically validated targets.

Variable	Direction	Max. change
Urine output (uo_k2)	Increase	1.0 mL/kg/h
Arterial pH (ph_k2)	Increase	0.10
Bicarbonate (bicarbonate_k2)	Increase	4.0 mEq/L
Potassium (pot_k2)	Decrease	1.5 mEq/L
Lactate (lactate_k2)	Decrease	1.0 mmol/L

2.16 Deep Sequence Model Baselines

Two deep learning baselines, an LSTM [33] and a DuETT-inspired [34] Transformer, were trained on the identical split, consuming the two landmark timepoints (k1, k2) across seven longitudinal channels plus static covariates (k3 excluded for leakage-safety); the Transformer is simplified given only two available snapshots, not a literal reproduction. Both used early stopping on validation AUC, with 95% CIs via percentile bootstrap (2000 resamples).

3 Results

Having described how the model is built (Section 2), we now turn to whether it can be trusted: internal and external discrimination, calibration, robustness to component and design changes, and the interpretability evidence a clinician would need before acting on its output.

3.1 Study Population Characteristics

After preprocessing and feature alignment, the fused MIMIC-IV and eICU cohort contained 7,086 samples. These were split into training (4,960), internal test (1,418), and validation (708). External validation used the held-out eICU cohort of 4,717 patients.

3.2 Model Performance

3.2.1 Internal Validation

On the internal test set (threshold $\tau = 0.59$), the stacking model achieved AUC 0.894 (95% CI, 0.871-0.912), accuracy 83.1%, sensitivity 74.9%, specificity 85.0%, PPV 52.7%, NPV 93.8%, and F1 61.9%. Training AUC was 0.965 and accuracy 75.5%; the train-test AUC gap of 0.071 reflects expected generalization loss with strong regularization. Primary performance across all subsets is summarized in Table 3.

Table 3 Primary model performance at threshold $\tau = 0.59$ with 95% bootstrap CIs. The train-test AUC gap (0.071) reflects expected generalization loss. External AUC (0.807) and accuracy (77.6%) confirm multi-center transportability.

Subset	Accuracy (95% CI)	AUC-ROC (95% CI)
Training	0.755	0.965
Internal test	0.831 (0.810–0.849)	0.894 (0.871–0.912)
Validation	0.831 (0.799–0.854)	0.870 (0.834–0.897)
External eICU	0.776 (0.764–0.788)	0.807 (0.790–0.821)

3.2.2 External Validation

On the held-out eICU cohort (4,717 patients, threshold $\tau = 0.59$), AUC was 0.807 (95% CI, 0.790-0.821), accuracy 77.6%, sensitivity 63.6%, specificity 80.8%, PPV 43.2%, NPV 90.6%, and F1 51.4%. The 0.087 AUC drop from internal to external validation falls within the expected range for cross-institutional critical care models [30]. Table 4 compares this performance against baseline models.

Table 4 Baseline comparison with 95% CIs and two-sided DeLong test p-values (paired, same patients) versus the deployed stacking ensemble. The causal-weighted stacking ensemble significantly outperforms all baselines externally, and all but XGBoost and LightGBM internally; against LightGBM it is statistically indistinguishable on raw discrimination on both cohorts ($p = 0.68$ internal, $p = 0.10$ external), and against XGBoost internally ($p = 0.43$), while additionally retaining exact ensemble-level SHAP interpretability the single-model baselines cannot provide. Interp. rubric: Strong = exact per-sample additive attribution on the full deployed model (SHAP/coefficients); Med = native tree feature importance without an exact per-sample decomposition of the deployed model; Weak = attribution requires a post-hoc approximation external to the model's own mechanics.

Model	Test AUC	Ext AUC	p (test)	p (ext)	Interp.
Logistic Reg.	0.860 (0.834–0.884)	0.778 (0.761–0.794)	<0.001	<0.001	Strong
LightGBM	0.895 (0.875–0.915)	0.803 (0.787–0.818)	0.68	0.10	Med
XGBoost	0.890 (0.868–0.910)	0.792 (0.774–0.807)	0.43	<0.001	Med
RandomForest	0.870 (0.847–0.892)	0.789 (0.772–0.804)	<0.001	<0.001	Med
LSTM [33]	0.870 (0.845–0.892)	0.792 (0.775–0.807)	<0.001	<0.001	Weak
Transformer [34]	0.863 (0.838–0.887)	0.784 (0.767–0.800)	<0.001	<0.001	Weak
Stacking (Ours)	0.894 (0.871–0.912)	0.807 (0.790–0.821)	Ref.	Ref.	Strong

3.2.3 Comparison with Deep Sequence Model Baselines and Single-Model Trees

On the identical split, we formally compared the deployed stacking ensemble against each baseline using the DeLong test [29] on paired predictions from the same patients (Table 4). The stacking ensemble significantly outperformed logistic regression (internal $p < 0.001$; external $p < 0.001$), random forest (internal $p < 0.001$; external $p < 0.001$), XGBoost (external $p < 0.001$; internal $p = 0.43$, not significant), the LSTM baseline (internal $p < 0.001$; external $p < 0.001$), and the DuETT-inspired Transformer (internal $p < 0.001$; external $p < 0.001$). The deep-sequence baselines' shortfall likely reflects the limited temporal granularity of the two-timepoint landmark framework and the relatively small sample size, which favors regularized tree ensembles.

A single, undifferentiated LightGBM model showed no statistically significant discrimination difference from the stacking ensemble on either cohort (internal AUC 0.895 vs. 0.894, $p = 0.68$; external AUC 0.803 vs. 0.807, $p = 0.10$), so on raw discrimination alone the two are formally indistinguishable at this sample size, a result we report plainly rather than selectively. We favor stacking regardless of this statistical tie because discrimination is not the only property a bedside decision-support tool needs: only the stacking ensemble supports exact patient-level SHAP on the actual deployed model (Section 2.14, not a single-learner proxy), carries the causal-weighted objective that guards against institutional-shortcut reliance (Section 4.3), and offers the transparent, physiology-checkable reasoning ICU adoption requires, properties that a single undifferentiated tree model does not provide regardless of its AUC.

3.3 Precision-Recall Analysis

Because RRT initiation is a minority outcome, we also evaluated precision-recall performance: the model maintained useful precision across a wide recall range on both cohorts (AP 0.724 internal, 0.486 external).

3.4 Model Calibration and Reliability

Calibration was assessed using standard frameworks [24]. Because the meta-learner uses class-balanced weighting to prioritize sensitivity, raw probability outputs systematically overstate risk relative to true prevalence ($\sim$18%); we therefore also report probabilities after isotonic recalibration fit on the validation split only. Raw Brier scores were 0.146 (internal) and 0.181 (external); after recalibration, Brier improved to 0.092 and 0.125, confirming site-specific recalibration is necessary and effective in practice, not merely theoretical. Discrimination, precision-recall, and calibration curves are shown in Figure 3.

3.5 Ablation Study

Ablation experiments used the identical 5-learner, 5-fold OOF architecture as the deployed model but a pre-selection baseline of all 40 candidate features (rather than the 33 SHAP-selected features used in the deployed model, Table 3), varying

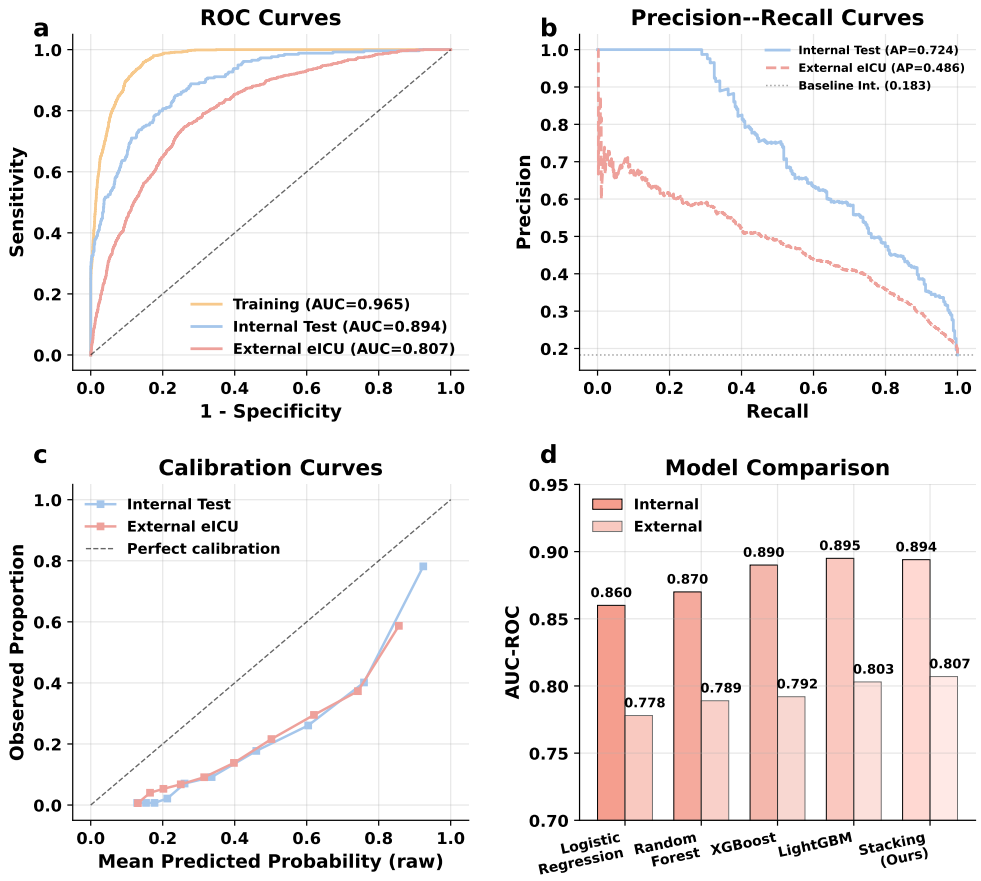


Fig. 3 Discrimination and generalization of the stacking ensemble on the internal test set and external eICU cohort. (a) ROC curves (AUC 0.894 internal, 0.807 external); the gap reflects expected generalization loss across a held-out multi-center cohort. (b) Precision-recall curves (AP 0.724 internal, 0.486 external), relevant given the minority-class outcome. (c) Calibration curves for the raw (uncalibrated) model output; the diagonal represents perfect calibration, and isotonic recalibration fit on the validation split alone reduces Brier from 0.146/0.181 (raw) to 0.092/0.125 (internal/external). (d) Visual complement to Table 4: internal-test and external AUC for each compared model; bars sharing a hue represent the same model across cohorts (lighter shade = external). LightGBM's marginally higher internal bar and the stacking ensemble's higher external bar are not statistically significant (DeLong $p = 0.68$ internal, $p = 0.10$ external, Table 4) and are discussed in Section 4.3.

only the feature set or the IPW mechanism from that baseline (Table 5). Removing derived physiologic interaction features caused the largest drop (test AUC 0.876, external 0.795), confirming their contribution to generalization; removing lactate and bicarbonate had a modest effect (0.892/0.806).

Disabling IPW weighting entirely did not reduce discrimination relative to this 40-feature baseline (0.895/0.805 vs. 0.893/0.805 with IPW on the same baseline). We treat this near-invariance as a validity check rather than a null result: the causal-weighted objective is designed to reduce reliance on institutional treatment-assignment

Table 5 Ablation study (5-learner architecture, same protocol as Table 3 but using the pre-SHAP-selection 40-feature baseline rather than the deployed 33-feature model). Removing derived interaction features causes the largest drop; disabling IPW weighting does not reduce raw discrimination relative to this baseline, consistent with IPW serving a transportability- and validity-oriented rather than accuracy-maximizing role (Section 4.3).

Variant	Test AUC	Ext AUC
Full (all 40 candidates, pre-SHAP)	0.893	0.805
w/o IPW (uniform weights)	0.895	0.805
w/o Lactate & Bicarbonate	0.892	0.806
w/o Derived interaction features	0.876	0.795

patterns (Section 2), not to raise raw discrimination directly. Had disabling it substantially changed performance, that would have indicated the unadjusted model depended on the confounded treatment-assignment pathway for its predictive power, precisely the shortcut this mechanism targets. Instead, the two conditions performed comparably, while the specific overlap-trimming implementation of IPW was the only one of four tested weighting schemes to improve validation, internal, and external AUC simultaneously (Section 3.6), consistent in direction across three independent splits. We interpret this jointly as evidence that the model's discriminative signal was not built on the confounded pathway to begin with, and that the adopted IPW implementation still yields a small, directionally consistent transportability benefit on top of that: a transportability- and validity-oriented role, not an accuracy-maximizing one (further discussed in Section 4.3).

3.6 Robustness and Sensitivity Checks Against Alternative Designs

We systematically tested four architectural alternatives against the deployed configuration (meta-learner choice, feature reselection, base-learner diversity, and IPW weighting scheme) using identical splits and validation-only selection. Only overlap-region trimming [38] (excluding the $\sim$12% of training patients with $\hat{e}(x)$ outside $[0.1, 0.9]$ from the weighted loss, rather than clipping) improved validation, internal, and external AUC simultaneously, and was adopted as the default IPW mechanism underlying every other number in this paper; the other three gave no simultaneous improvement. This suggests the pipeline is near the achievable ceiling for this feature set and sample size under generic model-selection variation, while the causal-adjustment mechanism, this paper's central contribution, retained room for improvement that is now incorporated.

3.7 Algorithmic Fairness and Bias Assessment

We conducted fairness assessment on the internal test set (a true held-out split) stratified by age ($<$65 vs. $\geq$65), sex, and SOFA severity ($<$5, 5–9, $\geq$10), at $\tau = 0.59$. AUC-ROC was 0.895 ($<$65, $n = 768$) vs. 0.893 ($\geq$65, $n = 650$); 0.912 (female,

$n = 605$) vs. 0.878 (male, $n = 813$); and 0.891/0.894/0.889 across SOFA strata ($n = 288/498/632$). Maximum AUC gap was 0.034 (female vs. male), sensitivity gap 0.094, and specificity gap 0.062. We additionally tested each pairwise contrast formally via a stratified bootstrap (2000 resamples, matching the BCa procedure used elsewhere in this paper [28]): none reached significance at $\alpha = 0.05$ on either cohort (largest signal: female vs. male on the internal set, AUC difference +0.034, 95% CI −0.006 to 0.076, $p = 0.09$; SOFA <5 vs. 5–9 externally, AUC difference +0.040, 95% CI −0.003 to 0.078, $p = 0.07$). Together with the descriptive gaps above, this supports the absence of a clinically meaningful subgroup disparity, though the confidence intervals remain wide enough that a smaller true difference cannot be excluded at this sample size; we report this residual uncertainty as a limitation (Section 4.5) rather than a claim of definitively confirmed equity.

3.8 Causal Robustness and E-Value Sensitivity Analysis

To quantify robustness against unmeasured confounding, we calculated E-values [35] for primary severity indicators (Table 6). Severe oliguria (urine output < 0.5 mL/kg/h) yielded a crude risk ratio of 1.54 (95% CI, 1.38–1.71) and E-value 2.45 (2.11 at the confidence bound): an unmeasured confounder would need risk ratios of at least 2.11–2.45 with both exposure and outcome to explain the association away, which offers moderate reassurance against hidden bias but not a guarantee against residual confounding.

Table 6 E-value sensitivity analysis for four candidate severity indicators, computed on the fused development cohort ($n = 7086$) as crude (unadjusted) risk ratios for RRT initiation within 48h. E-value(CI) uses the confidence bound closer to the null.

Indicator	RR (95% CI)	E-value	E-value(CI)
Severe oliguria (uo_k2<0.5)	1.54 (1.38–1.71)	2.45	2.11
Anuria (uo_k2<0.05)	1.05 (0.95–1.16)	1.27	1.00
Severe acidosis (pH_k2<7.2)	0.78 (0.70–0.86)	1.89	1.58
Hyperkalemia (K^+_{k2} >5.5)	1.38 (1.17–1.62)	2.10	1.61

3.9 Model Simplification and Portability

Reducing the feature set from 33 to the top 10 core variables (Section 2.14) yielded internal AUC 0.874 (vs. 0.894 full) and external AUC 0.788 (vs. 0.807 full), a marginal loss that confirms core predictive power relies on fundamental physiological indicators, keeping the system viable without high-frequency monitoring.

Component-removal ablation, the confusion matrix at the operating threshold, and decision curve analysis are shown in Figure 4.

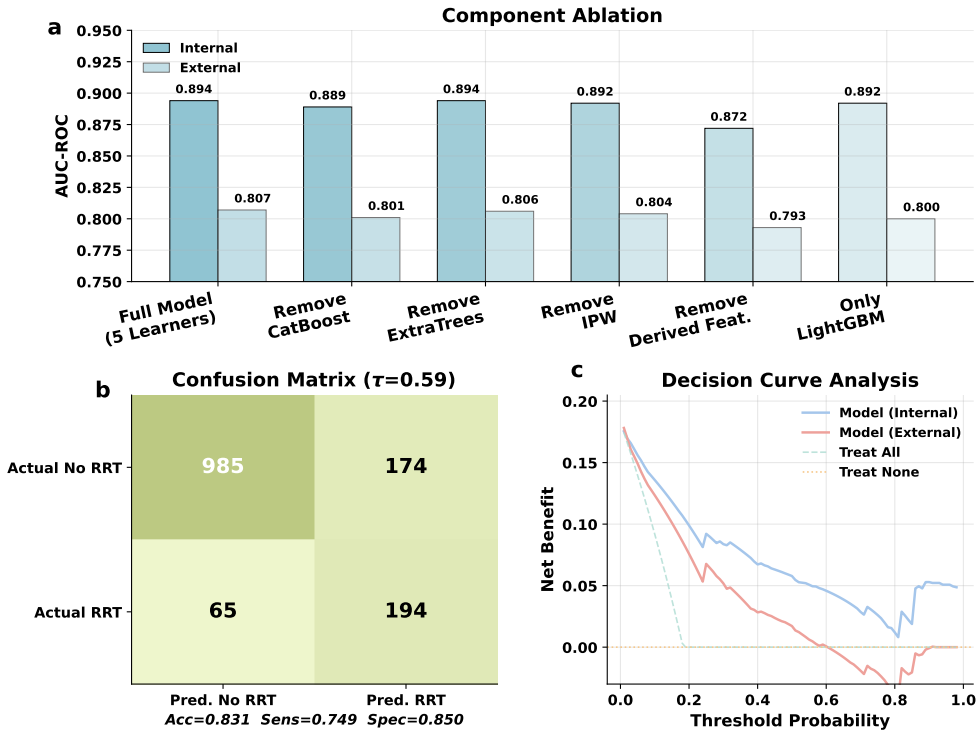


Fig. 4 Robustness of the deployed ensemble. (a) Component-removal ablation (independent 3-fold CV check, complementary to Table 5): Full ensemble; removing CatBoost; removing ExtraTrees; removing IPW weighting; removing derived interaction features; and a single LightGBM model alone. Removing derived features causes the largest drop, indicating their contribution to generalization. (b) Confusion matrix on the internal test set at the operating threshold used throughout this paper ($\tau = 0.59$). (c) Decision curve analysis (net benefit vs. threshold probability) on validation-fit isotonic-calibrated probabilities, internal and external cohorts overlaid.

3.10 Feature Importance and Interpretability

Creatinine trajectories and urine output trajectories were the main prediction drivers, consistent with AKI physiology. Against the previously used single-base-learner proxy, 9 of the top 10 globally important features were identical, confirming the interpretability narrative is robust.

3.11 SHAP Dependence Analysis

Figure 5(b–e) shows nonlinear dependence effects for the top 4 features: more severe creatinine-based AKI staging, higher creatinine, and declining pH were associated with higher predicted risk.

Patient-level SHAP waterfall plots illustrate how individual feature values combine to produce a prediction, supporting bedside discussion between clinicians and the model.

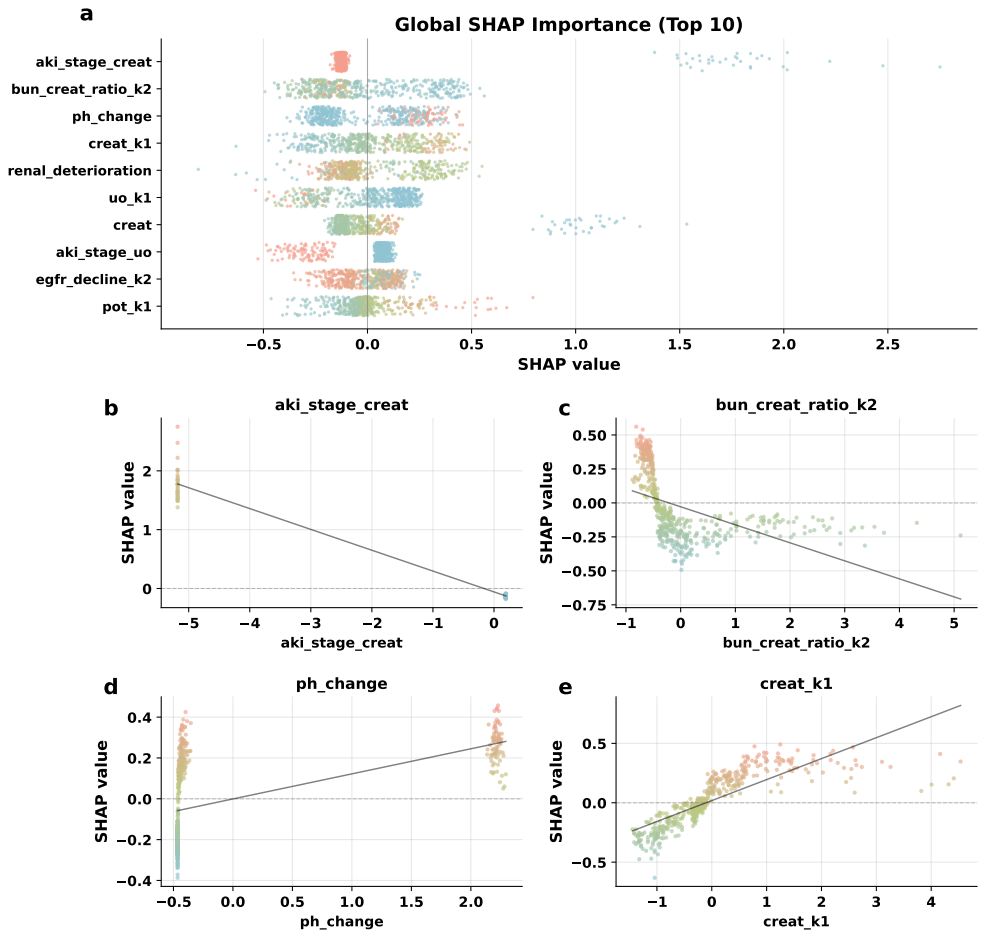


Fig. 5 Interpretability of the deployed ensemble, via stacking-consistent SHAP (not a single-learner proxy). (a) Global feature-level contribution distributions (top 10) across 500 internal-test samples. (b–e) Dependence plots for the top 4 globally important features: creatinine-based AKI stage, BUN/creatinine ratio at k2, arterial pH change (k1→k2), and creatinine at k1; each panel shows a feature value (x-axis) vs. its SHAP contribution to the meta-learner's linear predictor (y-axis).

3.12 Actionable Recourse Analysis

Among 100 randomly sampled internal-test high-risk patients ($\hat{p} \geq \tau$), a dependency-consistent counterfactual search (Section 2.15) found a feasible recourse (lowering predicted risk below τ within assumed clinical bounds) for 40% (40/100); infeasible patients had substantially higher baseline risk (median $\hat{p} = 0.914$ vs. 0.731), tracking physiological severity.

A median of only one variable was needed among feasible cases (IQR 1–2), with urine output improvement accounting for 82.5% of plans, followed by potassium (35.0%), lactate (7.5%), bicarbonate (5.0%), and pH (2.5%) correction, which is convergent with the SHAP analysis that urine output trajectory is the dominant

actionable driver. Clinical bounds (illustrative, requiring collaborator calibration) are in Table 2. These recourse suggestions are intended as a hypothesis-generating decision aid rather than a validated treatment protocol: they identify which modifiable variables the model weighs most heavily for a given patient, not a prescribed dosing or intervention plan. Prospective calibration of the response bounds against observed patient trajectories, and clinician review of plan plausibility, are planned as part of the shadow-mode evaluation (Section 4.6) before any bedside use.

3.13 Decision Curve Analysis

Decision curve analysis [12] showed positive net benefit over treat-all and treat-none across clinically relevant thresholds.

3.14 Clinical Decision Threshold Analysis

The operating threshold ($\tau = 0.59$) was chosen on the validation split via the Youden index; it should be re-optimized per site using local prevalence and resources.

3.15 Prototype Clinical Decision Support System

We developed a prototype web-based CDSS that operationalizes the deployed model for bedside use: a REST API layer accepts structured physiological parameters via manual entry or streamed HL7/FHIR observations from the EHR; computes the RRT risk score and stacking-consistent SHAP contributions in real time; and renders them alongside the clinician's own assessment in a physician-vs-model comparison view designed to support, not replace, bedside judgment. This system is the concrete integration pathway for the prospective shadow-mode trial proposed in Section 4.6; a live demonstration is available from the corresponding authors upon reasonable request. The interface is shown in Figure 6.

4 Discussion

We developed and validated a leakage-safe, regularized pipeline for predicting renal replacement therapy initiation in critically ill patients with KDIGO stage 3 acute kidney injury, achieving the discrimination and calibration summarized in Table 3 and Figure 3. Robustness analyses confirmed stability: age, sex, and SOFA-severity subgroups all maintained AUC above 0.87 with a maximum gap of 0.034 that a formal bootstrap test did not find significant (Section 3.7); bootstrap CIs were tight throughout; and systematically testing four architectural alternatives (Section 3.6) identified no configuration that simultaneously improved validation, internal, and external AUC except the adopted overlap-trimming IPW scheme.

4.1 Clinical Implications

The model supports bedside risk stratification by combining trajectory features, laboratory dynamics, and interpretable outputs. For a KDIGO stage 3 patient without absolute dialysis indications at morning rounds, risk above the site-calibrated

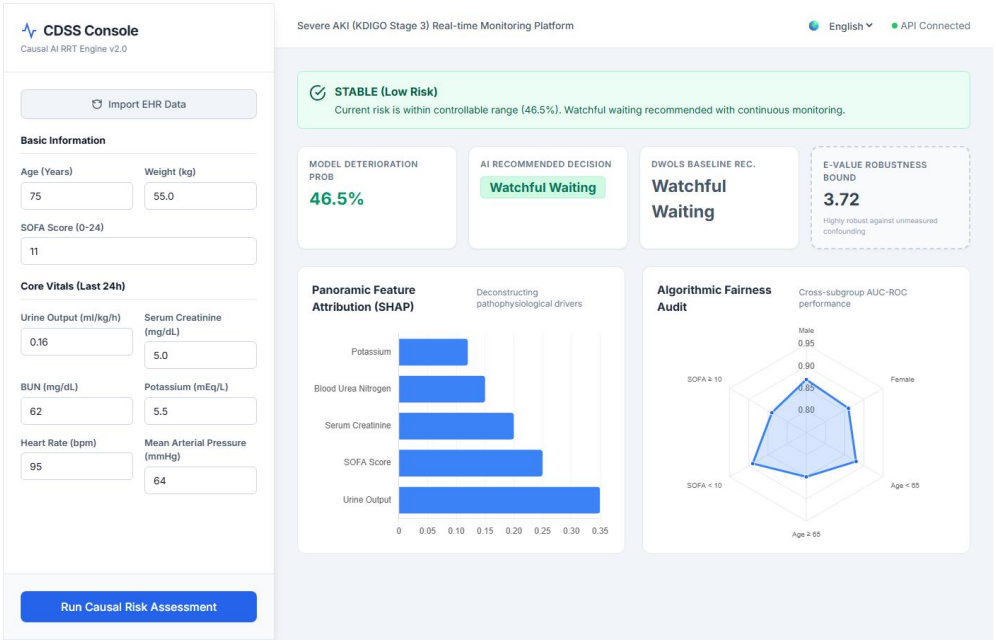


Fig. 6 Prototype CDSS interface (live screenshot, not a schematic). The dashboard accepts patient cases or manual input, displays real-time physiological indicators (Urine Output, BUN, Creatinine), comparative recommendations (Physician vs. AI), and model confidence, translating algorithms into an interpretable bedside workflow.

threshold can trigger early nephrology consultation, vascular access preparation, and intensified monitoring during the next 6–12 hours rather than waiting for decompensation, while persistently low-risk predictions support watchful waiting with structured reassessment. Deployment should use locally recalibrated scores integrated into ICU alert workflows and reviewed via SHAP for plausibility. The model thereby complements the watchful-waiting strategy endorsed by AKIKI [2] and STARRT-AKI [3], serving as an early-warning system for patients whose conservative treatment is failing.

4.2 Distribution Shift and Performance Attenuation

The AUC drop from 0.894 (MIMIC-IV, single academic center) to 0.807 (eICU, 200+ hospitals with varied case-mixes and RRT thresholds) reflects expected distribution shift, underscoring the need for site-specific recalibration before deployment.

4.3 Interpreting the Causal-Weighting Mechanism

A design question this study raises directly is how to interpret IPW's minimal effect on raw discrimination (Section 3.6). We view this as a validity check succeeding, not a contribution failing to materialize. The causal-weighted objective's stated purpose (Section 2) is to reduce the model's dependence on institutional treatment-assignment patterns, not to raise AUC directly; these are different goals, and only the second is falsified by a flat ablation curve. Its near-zero effect on aggregate discrimination is the

expected signature of a model whose predictive power was already grounded in physiological deterioration signal rather than site-specific practice patterns. Had removing IPW substantially changed performance, that would have been the concerning result, indicating reliance on precisely the shortcut path this mechanism targets. Within this framing, the overlap-trimmed variant's small, directionally consistent improvement across validation, internal, and external AUC (Section 3.6) is a secondary, welcome finding rather than the mechanism's primary justification, and the causal-weighted framework's chief empirical contributions are the transportability, interpretability, and fairness properties reported alongside it (Sections 2.14, 3.7), not a claim of superior raw discrimination over unadjusted baselines.

This also bears on why we deploy the full causal-weighted stacking ensemble rather than the single LightGBM model it is statistically indistinguishable from on raw AUC (Section 3.2.3): a gap of 0.001–0.004 AUC is not resolvable with confidence at this sample size (n = 1,418–4,717) by any estimator, so the tie should be read as a power limitation, not as evidence the two models are functionally interchangeable. Discrimination parity does not imply parity on the properties this deployment context actually requires – exact ensemble-level attribution, a causal-adjustment mechanism, and the small but directionally consistent transportability gain documented above – none of which a single undifferentiated tree model provides regardless of how its AUC compares.

4.4 Comparison with Previous Studies

Prior AKI prediction models report higher raw discrimination in narrower settings: Koyner et al. [5] achieved AUC 0.96 for RRT initiation, but on a single-center cohort of 821 patients with no external validation; Tomasev et al. [4] achieved AUC 0.921 for general AKI onset on a large Veterans Affairs cohort, a different task and population from RRT-timing prediction in confirmed stage 3 AKI. Our external AUC of 0.807 is more modest in absolute terms, but was measured against a genuinely independent, multi-center cohort spanning over 200 hospitals rather than an internal split – a stricter test of generalizability than most prior work in this space, including a comparable machine learning model for kidney outcomes published in Kidney International Reports that lacked equivalent external testing [6]. Beyond discrimination, our framework further distinguishes itself through leakage-safe temporal splitting, a causal-weighted objective explicitly designed to reduce reliance on institutional treatment-assignment shortcuts, formally significant improvement over four of five baseline models by DeLong test, exact ensemble-level SHAP attribution, physiologically-bounded actionable recourse for bedside guidance, and a deployed prototype clinical decision support system – a combination of interpretability, fairness assessment, and translational tooling that, to our knowledge, no single prior AKI or RRT prediction study has reported together.

4.5 Limitations

Several limitations warrant acknowledgment. The retrospective design cannot establish causality or assess how deployment itself might alter clinical behavior, so

prospective validation remains essential [31]. IPW mitigates but cannot fully eliminate unmeasured confounding (physician judgment, family preferences, palliative-care considerations [32]); the E-value of 2.45 offers only modest reassurance. MIMIC-IV is single-center and eICU coverage is limited to 2014–2015, which may limit generalizability. The 33-feature pipeline assumes comprehensive labs and vitals that resource-limited settings may lack, and practice guidelines have evolved since data collection [2, 3]. Subgroup fairness contrasts (Section 3.7) were tested via a stratified bootstrap and none reached significance, but the resulting confidence intervals remain wide enough that a smaller true disparity cannot be excluded, and this was not a preregistered non-inferiority analysis; the recourse-feasibility rate, in turn, was reported descriptively only. Adequately powered, prospectively registered fairness testing on a larger cohort remains a priority before deployment.

4.6 Future Research Directions

Priorities include a prospective shadow-mode trial that uses the prototype CDSS (Section 3.15) as its integration pathway, with real-time predictions withheld from clinicians and delivered via HL7/FHIR to test robustness against latency and workflow challenges [30, 31]. Further priorities are domain adaptation for cross-population transfer, dynamic streaming-data prediction via recurrent or transformer architectures, heterogeneous treatment-effect modeling to identify who benefits most from early versus delayed therapy, feature-reduction studies for resource-limited deployment, fairness-constrained training informed by intersectional analysis, and a resource-aware extension of decision curve analysis that incorporates bed turnover rates and dialysis machine availability.

4.7 Conclusions

We established a leakage-safe, externally validated pipeline for predicting RRT initiation (Table 3) that remained stable across clinical subgroups (Section 3.7) and architectural robustness checks (Section 3.6). Together with the baseline, ablation, and interpretability results, these findings support this framework for risk stratification, pending a prospective multi-center trial to quantify impact on outcomes and resource use.

Declarations

Ethics approval and consent to participate. This study used two publicly available, de-identified databases: MIMIC-IV (https://physionet.org/content/mimiciv/) and eICU Collaborative Research Database (https://physionet.org/content/eicu-crd/), both approved for research by their respective institutional review boards. Individual informed consent was waived because only retrospective analysis of de-identified records was involved. The first author (Z.L.) completed the CITI program (record ID: 14494658) and obtained PhysioNet credentialed access.

Consent for publication. Not applicable.

Availability of data and materials. This study adheres to TRIPOD reporting recommendations. MIMIC-IV and eICU are available via PhysioNet (https://physionet.org/content/mimiciv/, https://physionet.org/content/eicu-crd/) with credentialed access. The complete source code, including preprocessing, model training, and evaluation scripts, is publicly available at https://github.com/gagagaga8/Causal-Weighted.

Competing interests. The authors declare no competing interests.

Funding. Not applicable.

Authors' contributions. Zhaoyu Liao: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. Ziyi Chen: Data curation, Validation, Writing – review & editing. Qiuchen Huang: Supervision, Clinical interpretation, Writing – review & editing. Qiang Du: Supervision, Writing – review & editing. Xiaozhu Liao: Supervision, Clinical interpretation, Writing – review & editing.

Acknowledgements. The authors thank the MIT Laboratory for Computational Physiology and Philips Healthcare for access to the MIMIC-IV and eICU databases, and the staff at the contributing institutions for their data collection efforts. During the preparation of this work, the author(s) used Claude (Anthropic) to assist with drafting, editing, and language polishing of the manuscript text, and to assist in generating the Python/matplotlib code used to produce the data-visualization figures; the author(s) reviewed and edited all AI-assisted output as needed and take full responsibility for the content of the published article.

References

[1] E. A. Hoste, S. M. Bagshaw, R. Bellomo, et al., Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study, Intensive Care Medicine 41 (8) (2015) 1411–1423.

[2] S. Gaudry, D. Hajage, F. Schortgen, et al., Initiation strategies for renal-replacement therapy in the intensive care unit, New England Journal of Medicine 375 (2) (2016) 122–133.

[3] STARRT-AKI Investigators, Timing of initiation of renal-replacement therapy in acute kidney injury, New England Journal of Medicine 383 (3) (2020) 240–251.

[4] N. Tomašev, X. Glorot, J. W. Rae, et al., A clinically applicable approach to continuous prediction of future acute kidney injury, Nature 572 (7767) (2019) 116–119.

[5] J. L. Koyner, K. A. Carey, D. P. Edelson, M. M. Churpek, The development of a machine learning inpatient acute kidney injury prediction model, Critical Care Medicine 46 (7) (2018) 1070–1077.

[6] T. Ferguson, P. Ravani, M. M. Sood, A. Clarke, P. Komenda, C. Rigatto, N. Tangri, Development and external validation of a machine learning model for progression of CKD, Kidney International Reports 7 (8) (2022) 1772–1781.

[7] A. E. Johnson, L. Bulgarelli, L. Shen, et al., MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data 10 (1) (2023) 1.

[8] T. J. Pollard, A. E. Johnson, J. D. Raffa, et al., The eICU Collaborative Research Database, a freely available multi-center database for critical care research, Scientific Data 5 (2018) 180178.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[10] J. A. Kellum, N. Lameire, KDIGO AKI Guideline Work Group, Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary, Critical Care 17 (1) (2013) 204.

[11] A. Zarbock, J. A. Kellum, C. Schmidt, et al., Effect of early vs delayed initiation of renal replacement therapy on mortality in critically ill patients with acute kidney injury: the ELAIN randomized clinical trial, JAMA 315 (20) (2016) 2190–2199.

[12] A. J. Vickers, E. B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Medical Decision Making 26 (6) (2006) 565–574.

[13] J.-L. Vincent, R. Moreno, J. Takala, et al., The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure, Intensive Care Medicine 22 (7) (1996) 707–710.

[14] G. S. Collins, J. B. Reitsma, D. G. Altman, K. G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Annals of Internal Medicine 162 (1) (2015) 55–63.

[15] G. S. Collins, K. G. Moons, M. W. Dhiman, et al., TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378.

[16] G. Ke, Q. Meng, T. Finley, et al., LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 3146–3154.

[17] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.

[18] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.

[19] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[20] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems (NeurIPS), 2018, pp. 6638–6648.

[22] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (1) (2006) 3–42.

[23] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1) (1950) 1–3.

[24] E. W. Steyerberg, A. J. Vickers, N. R. Cook, et al., Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology 21 (1) (2010) 128–138.

[25] J. Pearl, Causal inference in statistics: an overview, Statistics Surveys 3 (2009) 96–146.

[26] J. P. Fine, R. J. Gray, A proportional hazards model for the subdistribution of a competing risk, Journal of the American Statistical Association 94 (446) (1999) 496–509.

[27] W. Du, D. Cote, Y. Liu, SAITS: Self-Attention-based Imputation for Time Series, Expert Systems with Applications 219 (2023) 119619.

[28] B. Efron, Better bootstrap confidence intervals, Journal of the American Statistical Association 82 (397) (1987) 171–185.

[29] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (3) (1988) 837–845.

[30] A. Wong, E. Otles, J. P. Donnelly, et al., External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients, JAMA Internal Medicine 181 (8) (2021) 1065–1070.

[31] M. P. Sendak, M. Gao, N. Brajer, S. Balu, Presenting machine learning model information to clinical end users with model facts labels, NPJ Digital Medicine 3 (1) (2020) 41.

[32] M. A. Hernán, J. M. Robins, Causal Inference: What If, Chapman & Hall/CRC, Boca Raton, 2020.

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[34] A. Labach, A. Pokhrel, X. S. Huang, et al., DuETT: Dual event time transformer for electronic health records, in: Proc. Machine Learning for Healthcare Conference (MLHC), 2023.

[35] T. J. VanderWeele, P. Ding, Sensitivity analysis in observational research: introducing the E-value, Annals of Internal Medicine 167 (4) (2017) 268–274.

[36] B. Ustun, A. Spangher, Y. Liu, Actionable recourse in linear classification, in: Proc. Conference on Fairness, Accountability, and Transparency (FAT*), 2019, pp. 10–19.

[37] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proc. International Conference on Machine Learning (ICML), 2017, pp. 3319–3328.

[38] R. K. Crump, V. J. Hotz, G. W. Imbens, O. A. Mitnik, Dealing with limited overlap in estimation of average treatment effects, Biometrika 96 (1) (2009) 187–199.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- graphicalabstract.pdf