

Supporting Hemodialysis Decision-Making in Lithium Poisoning: An Explainable and Clinically Interpretable Machine Learning and Nomogram Development

Kamran Rezaei

Shahid Beheshti University of Medical Sciences

Shahin Shadnia

Shahid Beheshti University of Medical Sciences

Babak Mostafazadeh

Shahid Beheshti University of Medical Sciences

Mitra Rahimi

Shahid Beheshti University of Medical Sciences

Peyman Erfan Talab Evini

peyman1346erfan@gmail.com

Shahid Beheshti University of Medical Sciences

Sayed Masoud Hosseini

Shahid Beheshti University of Medical Sciences

Sarina Sadat Abouei Mehrizi

Shahid Beheshti University of Medical Sciences

Fatemeh Saber

Tehran University of Medical Sciences

Pooya Eini

Shahid Beheshti University of Medical Sciences

Joshua King

University of Maryland Medical System

Article

Keywords: Artificial intelligence, Machine Learning, Poisoning, Lithium, Toxicology, Hemodialysis

Posted Date: March 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8680752/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Lithium poisoning (LP) poses a critical management challenge due to its narrow therapeutic index and unpredictable toxicodynamics. Present paper aims to use machine learning capabilities to develop a decision-support system for risk assessment of hemodialysis in LP. We analyzed 207 patients with LP admitted to a referral toxicology center, of whom 27 (13.0%) required hemodialysis. Following a feature selection strategy, four algorithms, logistic regression, support vector machine, artificial neural network, and random forest, were trained. with 5-fold cross-validation. The approach was focused on preventing data leakage into the validation and test. The random forest model outperformed other models with a test-set AUROC of 0.83, sensitivity of 80.0%, specificity of 83.0%, and F1-score of 0.53. Mean cross-validation AUROC was 0.90. SHAP analysis identified serum lithium level, neurological symptoms, alkaline phosphatase, and age as key predictors. Platt recalibration improved the Brier score from 0.126 to 0.086 and calibration slope to 1.06. Decision curve analysis showed net clinical benefit across a wide threshold range. Bedside nomogram increased clinical utility (AUROC = 0.79) by classifying patients into low- moderate- and high-risk for hemodialysis. Although this decision-support system can significantly help the clinicians, an external validation and future studies using a bigger development set is required.

INTRODUCTION

Lithium is a cornerstone therapy for bipolar disorder, but its use is constrained by an exceedingly narrow therapeutic index and its propensity to impair renal function, factors which make toxicity relatively common ¹. The clinical management of lithium poisoning (LP) represents one of the most enduring challenges in critical care and medical toxicology, characterized by a persistent decoupling between serum drug concentrations and the severity of neurotoxic manifestations ^{1,2}. A major clinical challenge is that serum lithium concentrations often do not correlate closely with the severity of intoxication, and this issue, combined with the risk of permanent neurological sequelae in severe cases, complicates management decisions and necessitates a cautious approach to intervention ^{1,3}.

Clinicians have developed expert guidelines to aid in determining when to initiate hemodialysis for LP. Notably, the Extracorporeal Treatments in Poisoning (EXTRIP) workgroup issued consensus criteria that consider serum lithium levels alongside clinical factors ³. However, these established criteria are intentionally broad (favoring sensitivity over specificity) and may over-recommend dialysis. For instance, a study found that strictly applying EXTRIP recommendations would have indicated hemodialysis in 58% of LP cases in their series, far exceeding the proportion of patients who actually developed serious toxicity ³. This underscores a pressing need for more precise risk-stratification tools that improve specificity for hemodialysis in LP, without compromising patient safety.

Machine learning (ML) and artificial intelligence (AI) offer a promising way to fulfill this need by uncovering complex patterns in clinical data that traditional criteria might miss ^{4,5}. These techniques have rapidly expanded in medical use, with growing interest in leveraging them to improve clinical decision-making and outcome prediction. Indeed, the first ML-driven models for LP have only begun to

appear; for instance, a recent ML model using United States National Poison Center data achieved exceptionally high accuracy in distinguishing severe outcomes from mild cases of acute lithium overdose ⁶. Meanwhile, ML models in medical settings are required to be clinically transparent and reliable, in addition to simpler metrics such as accuracy. Explainable AI methods can highlight which features most influence the predictions ⁷, helping clinicians understand and trust an algorithm's reasoning. Likewise, careful calibration of the model's probability estimates is vital to ensure that predicted risks align with real-world outcomes, allowing an ML-based tool to serve as an effective decision-support aid for hemodialysis in LP rather than an inscrutable black box model ^{7,8}.

The present paper aims to capture these capabilities of AI by integrating a supervised ML model into the clinical decision-support systems for the prediction of hemodialysis in LP, while increasing its explainability and post-hoc calibration to enhance an ML model's clinical utility.

RESULTS

Study Population

A total of 207 patients with LP were included in the analysis, of whom 27 (13.0%) required hemodialysis during hospitalization. Baseline demographic, clinical, and laboratory characteristics are summarized in **Table 1**. The hemodialysis group showed a significantly higher mean for age of subjects (40.92 (15.42) vs. 28.78 (12.99), p -value<0.001), and female to male ratio was 2.0 and 4.0 for hemodialysis and no hemodialysis groups, respectively (p -value<0.11). The trend for going from acute to chronic ingestion was significantly associated with hemodialysis ($z = 0.19$, $p < 0.001$) with no evidence of departure from linearity ($\chi^2 = 2.71$, $p = 0.09$). Among the signs and symptoms on admission, significant differences were observed regarding the occurrence of neurological manifestations (Abnormal motor signs such as rigidity, ataxia, or tremor, and myoclonus or seizure) between no hemodialysis (17.2%) and hemodialysis (55.6%) groups ($p < 0.001$). No significant differences between groups were observed regarding the past medical history. No vital sign differences were observed; however, for Glasgow Coma Scale (GCS) on admission, significant monotonic decreasing hemodialysis trend across the higher GCS ($z = -5.01$, $p < 0.001$), with no evidence of departure from linearity ($\chi^2 = 14.41$, $p = 0.11$). Temperature was dropped from further analyses because of higher than 30% missing values, and no significant between group differences. Finally, higher white blood cell count ($p = 0.04$), serum creatinine ($p < 0.001$) and urea ($p = 0.02$), serum potassium ($p = 0.03$), and first serum lithium concentration ($p < 0.001$) were the only significant different lab findings between groups.

Table 1 also describes the outcome details of both hemodialysis and no hemodialysis groups. A significantly higher measure for serum lithium level, confusion, refractory shock or dysrhythmia, serum lithium concentration over 4 mmol/L or no decrease to 1 after 36 hours of acute intoxication, and length of hospital stay were observed in hemodialysis group (p -value<0.001)

Feature Selection

In univariate logistic regression (LR) analyses, multiple clinical and laboratory variables demonstrated associations with the need for hemodialysis at a screening threshold of $p\text{-value} < 0.20$. Multivariable LR using standard maximum likelihood estimation was unstable due to the low event rate and evidence of complete or quasi-complete separation. Firth penalized logistic regression was therefore applied; however, no predictor retained statistical significance at $p\text{-value} < 0.05$ in the fully adjusted model (**Supplementary Tables**).

Given these limitations and the study's predictive objective, feature selection proceeded using a hybrid approach incorporating clinical relevance and stability across ML-based feature importance analyses. Seven predictors according to their mean importance across 5 folds of cross-validation (CV) were retained for downstream modeling: serum lithium level at presentation, presence of severe neurological signs and symptoms (rigidity, ataxia, tremor, myoclonus, or seizure), GCS on admission, age, time elapsed from ingestion to emergency department (ED) admission (minutes), hemoglobin level, and serum alkaline phosphatase. Additionally, the clinical importance of type of toxicity⁹ made it logical to add chronic LP to our features, making a set of eight selected features for model development (**Supplementary Materials-Nestedcv RF top features**).

Model Performance

Four ML models were developed: Elastic-Net LR, linear support vector machine (SVM), shallow artificial neural network (ANN), and constrained RF. Performance metrics for all models are summarized in **Table 2**. Moreover, confusion matrices for all four models are described in **Table 3**. Across models, discriminative performance on the test set was generally acceptable; however, the constrained RF model demonstrated the most favorable overall performance (accuracy=83%), achieving the highest test-set area under the receiver operating characteristic curve (AUROC) on test set (0.83, 95% confidence interval (CI) = 0.65-0.97, **Figure 2**) while maintaining a balanced trade-off between sensitivity (80%) and specificity (83%, F1-score=0.53). Mean AUROC on CV folds was 0.90 (95% CI = 0.80-0.99), showing a good prevention of overfitting for this model. Consequently, the constrained RF was selected as the final model for further evaluation. The structure of each model is presented in supplementary materials, and **Table 2** shows the best set of hyperparameters following using GridsearchCV method.

Model Explainability

SHapley Additive exPlanations (SHAP) analysis (**Figure 3**) of the final RF model identified serum lithium level at presentation, age, presence of severe neurological findings, and alkaline phosphatase as the most influential predictors of hemodialysis requirement. Higher lithium levels, presence of severe neurological findings, higher alkaline phosphatase, and older age were consistently associated with increased predicted probability of hemodialysis. Additionally, lower GCS on admission and blood hemoglobin level were also associated with higher rates of hemodialysis.

Time to admission and chronic lithium exposure demonstrated smaller SHAP magnitudes. Shorter time to admission was associated with higher predicted risk, a pattern interpreted as reflecting confounding

by severity, whereby patients with more severe clinical presentations were more likely to present earlier for medical care.

Calibration and Recalibration

Initial calibration assessment of the RF model on the test set demonstrated modest miscalibration, with a Brier score of 0.126, a negative calibration intercept, and a calibration slope greater than 1, indicating systematic overestimation of risk and compressed probability spread (**Figure 4**).

Post-hoc probability recalibration using Platt scaling substantially improved calibration (**Figure 4**). Following recalibration, the Brier score decreased to 0.086, the calibration slope approached unity (1.06), and the calibration intercept moved closer to zero. Discriminative performance was preserved after recalibration, with a test-set AUROC of 0.87. Isotonic regression also improved calibration but offered no advantage over Platt scaling and demonstrated slightly lower discriminative performance (**Figure 4**).

Decision Curve Analysis

decision curve analysis (DCA) using Platt-calibrated probabilities demonstrated that the RF model provided a positive net clinical benefit across a broad range of clinically relevant threshold probabilities when compared with treat-all and treat-none strategies. The greatest net benefit was observed within intermediate threshold ranges, corresponding to scenarios in which clinical decision-making regarding hemodialysis is most uncertain (**Figure 4**).

These findings indicate that use of the calibrated prediction model may support individualized decision-making regarding hemodialysis in LP, rather than uniform application of treatment strategies.

Simple bedside score (nomogram)

Six variables were retained in the final score, consisting of first serum lithium level, severe neurological manifestations, GCS, age, hemoglobin, and alkaline phosphatase. Continuous variables were discretized using clinically relevant thresholds and our summary results (**Table 1**): first lithium concentration (<1, 1-2, 2-3, >3 mEq/L), GCS (<9 vs. ≥9), age (>30 years), hemoglobin (<12.5, 12.5-13.0, >13 g/dL), and alkaline phosphatase (>170 vs. ≤170 IU/L). The resulting LR coefficients were scaled into integer values, ranging from 0 to 6 points per variable (**Supplementary Tables**). The total score ranged from 0 to 17, with patients classified as Low risk (0-3 points), Moderate risk (4-7), and High risk (≥8), using the threshold of calibration plots.

The score achieved a test-set AUROC of 0.79 (95% CI: 0.24-1.00) and demonstrated clear stratification across risk categories. Confusion matrix is presented in **Table 3**. As shown in **Figure 5**, most patients were in the low- or moderate-risk categories, while a small subset were classified as high-risk. The score's structure and variable contribution are summarized in **Figure 5**, with complete logistic coefficients and scoring logic available in **Supplementary Tables**. A printable nomogram-style layout for

bedside use is provided in **Supplementary Tables**. To note, due to the zero-integer effect of alkaline phosphatase, we removed it from the nomogram and final bedside score system.

DISCUSSION

In this study, we developed an interpretable ML model to support hemodialysis decision-making in LP, achieving robust discrimination and calibration. Our constrained RF model demonstrated high accuracy in identifying patients who ultimately required hemodialysis, with a test AUROC of 0.83 and a favorable prevention of overfitting. Notably, post-hoc Platt scaling was applied to calibrate the model's probability outputs, ensuring that predicted risks correspond to actual outcome frequencies¹⁰. DCA further confirmed the model's clinical utility, showing that the RF would achieve a positive net benefit across a broad range of threshold probabilities for initiating dialysis, thereby improving upon the treat-all or treat-none extremes⁸. Finally, a bedside scoring system and nomogram for prediction of the risk of hemodialysis in LP maximized the clinical utility of our paper, bolding this papers methodological approach and results among similar studies in the literature. In practical terms, using the model to guide dialysis decisions would add clinical value by correctly identifying high-risk patients (and avoiding unnecessary dialysis in low-risk cases) within the threshold probability ranges that clinicians consider relevant. Taken together, these findings indicate that our ML model can accurately predict the need for extracorporeal removal in LP and could meaningfully assist decision-making in the ED or intensive care unit (ICU); however, external validation and further training on extra data is required for more confident interpretations.

The EXTRIP workgroup's consensus guidelines (2015) represent the current clinical standard for dialysis indications in LP¹¹. EXTRIP recommends hemodialysis in cases of severe LP, specifically if kidney function is impaired and serum lithium > 4.0 mEq/L, or if there is any decreased level of consciousness, seizures, or life-threatening dysrhythmias, irrespective of lithium level¹¹. Additionally, a serum lithium > 5 mEq/L or failure to reach a safe lithium level (< 1.0 mEq/L) following 36 hours of intoxication, regardless of clinical manifestations, are considered for hemodialysis¹¹. A retrospective analysis by Buckley et al. found that strict application of EXTRIP criteria would have recommended dialysis in 58% of chronic or acute-on-chronic LP cases, whereas in practice only 2.5% of patients were actually dialyzed³. This indicates that the EXTRIP thresholds, while safe, may not be specific; many patients might undergo unnecessary dialysis if guidelines were strictly followed. Our model was derived from real-world outcomes in 207 patients and inherently learned a more refined decision boundary. It identified the nuanced combinations of factors (lithium level plus clinical signs and patient factors) that truly associated with needing dialysis. As a result, the model could complement the EXTRIP guidance by providing patient-specific risk estimates. This more individualized approach aligns with recent suggestions to narrow the indications for lithium dialysis to those at highest risk of neurotoxicity^{3,12}. In essence, our data-driven model supports the spirit of the EXTRIP guidelines, i.e., prompt dialysis in severe cases, but could reduce over-triage by integrating multiple predictors into a single risk score rather than relying on any one threshold.

Beyond expert guidelines, few published studies have applied advanced predictive modeling to LP outcomes, making our paper an early contribution to this domain. One recent study by Mehrpour et al. analyzed the United States National Poison Data System with ML, using an RF model to predict severe outcomes in acute lithium overdoses¹³. They reported excellent performance (approximately 98–100% accuracy) for distinguishing cases with major vs. minor clinical effects in a very large dataset¹³. Notably, that study's SHAP analysis highlighted features like drowsiness, ataxia, age, and abdominal pain as key predictors¹³. We didn't include abdominal pain in our analysis due to uncertainty of a confident report in our retrospective setting; however, other predictors of Mehrpour et al. study was compatible with our SHAP analysis. Our model specifically targets the decision for hemodialysis, a more concrete and intervention-focused endpoint. In that respect, prior literature has mostly been limited to descriptive series. For example, a Taiwanese cohort of 36 LP patients found only 7 (19%) underwent hemodialysis, all of whom had significantly worse neurologic status and lower GCS (presumably due to decreased rate of lithium elimination in the nervous system¹⁴), and more complications than those managed conservatively¹². To our knowledge, no previous model has combined calibration and interpretability in this context. Thus, our work not only confirms known risk factors (and aligns with established criteria like EXTRIP³) but also pioneers a quantitative decision-support tool. It advances the field by moving from coarse heuristics and broad guidelines to a personalized risk prediction, which can be especially valuable in settings where expert toxicology or nephrology input is not immediately available.

As expected, the blood lithium concentration is a primary driver of toxicity severity. However, its interpretation must account for the kinetics of exposure (acute vs. chronic). In acute overdose, lithium may reach very high serum levels shortly after ingestion, but clinical toxicity can be delayed because lithium has not yet penetrated into vital tissues such as brain^{15,16}. Patients with acute LP can thus appear relatively well initially despite elevated serum levels, as the drug is largely confined to the intravascular space. In contrast, chronic LP (or acute-on-chronic) involves prolonged accumulation; where lithium equilibrates between serum and intracellular compartments over days, so a moderate serum level in a chronic patient may actually indicate a large total body burden and significant lithium in the nervous system^{15,16}. Mechanistically, a high lithium level suggests impending or ongoing end-organ effects (e.g., neuronal dysfunction, cardiovascular instability, renal impairment due to lithium-induced nephrogenic diabetes insipidus) and thus correlates with needing dialysis to prevent further tissue accumulation^{15,17}. These are compatible with our report of lithium concentration being the most important feature in the trained ML model.

The inclusion of alkaline phosphatase as a top predictor is initially surprising, but it can be explained by alkaline phosphatase's role as a general marker of organ stress and its links to lithium's systemic effects. Alkaline phosphatase is an enzyme found in liver, bone, and other tissues; elevated alkaline phosphatase in this context might signify cholestatic hepatic injury, bone turnover, or other metabolic stress^{18,19}. One hypothesis is that alkaline phosphatase reflects renal and hepatic stress due to LP¹⁸; however, it is important to note that an animal study stated reduced alkaline phosphatase concentration in animal models treated with lithium over time^{18,19}, making the true singular role of alkaline

phosphatase in LP a debating matter. Another possibility is that alkaline phosphatase elevation is related to lithium-induced hyperparathyroidism and bone metabolism changes during chronic therapy²⁰. Long-term lithium use is known to disrupt calcium regulation and can cause hyperparathyroidism in some patients. Chronic hyperparathyroidism leads to increased bone resorption and high bone alkaline phosphatase levels²⁰. This latter concept is probably a better conclusion, as our SHAP analysis showed a bidirectional effect of high alkaline phosphatase levels, and this feature's overall role is potentially affected by our chronic toxicity variable (Fig. 3).

Limitations and Future Directions

This study has some limitations. First, it was based on a relatively small, single-center cohort (n = 207, 13% requiring hemodialysis) from an Iranian referral hospital, which may limit generalizability. Local practice patterns and retrospective data collection may have introduced selection or documentation biases. While model development incorporated strict CV and calibration techniques, external validation on larger, multicenter datasets remains essential. Additionally, the retrospective design limits the ability to assess whether the model improves real-world clinical outcomes, and missing data in this design also limits the feature entry in ML model development. Additionally, while the bedside score offers interpretability and ease of use, it simplifies continuous variables into discrete bins, potentially losing granularity. Its development was based on a modest sample size with few hemodialysis events, and external validation is needed to confirm generalizability across different clinical settings. Future work should focus on external validation across diverse healthcare settings and prospective implementation trials to assess clinical impact. Integration into user-friendly tools, such as mobile apps or EHR-based decision support, may enable frontline providers, particularly in low-resource settings, to benefit from individualized risk prediction. Broadly, our approach illustrates the potential of interpretable AI to enhance toxicology decision-making and warrants further development across other high-risk poisoning scenarios.

METHODS

Study Design and Settings

This observational study used retrospective data collected from electronic health records and databases for patients admitted to Loghman Hakim Hospital, a referral poisoning center in Tehran, Iran, from January 2019 to March 2025. Confirmed LP Patients aged 15 to 75 years were included in the study if they didn't have a past medical history of end-stage disorder or cancers. LP was confirmed via an acute suicidal attempt of ingestion of lithium, or via clinical presentation of chronic LP confirmed by serum lithium concentration. Data gathering was performed by three investigators and supervised for any discrepancies by three clinical toxicology fellowships. Clinical aspects of choices for interpretations and feature selection were also validated by the consensus of four toxicology fellowships. The overall approach of the study design is depicted in Fig. 1.

Ethical and Protocol Approval

Ethical approval

was granted by the institutional review board at Shahid Beheshti University of Medical Sciences (ethical code: IR.SBMU.RETECH.REC.1404.049). The study protocol was also confirmed by the institutional review board Shahid Beheshti of Medical Sciences. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from either the participants or their families in cases where the participants were unable to provide consent themselves.

Data Preparation and Feature Selection

A set of demographic features, history and clinical examinations, vital signs on admission, first day laboratory findings, and a group of outcome features were collected for all subject, of which the hemodialysis is the primary outcome for this project. Continuous variables were summarized using appropriate descriptive statistics, and categorical variables were summarized as frequencies and percentages. Missing data were present in several continuous variables, with missingness below 30% for all included features.

Feature selection was primary conducted using a two-stage inferential approach in Stata 18. First, univariate LR analyses were performed to screen candidate predictors, using a liberal inclusion threshold ($p < 0.20$) to avoid premature exclusion of potentially relevant variables. Consequently, Firth penalized logistic regression was used to mitigate small-sample bias and separation²¹ (**Supplementary Tables 1 and 2**).

Given the limited discriminatory performance of purely inferential multivariable models and the study's predictive aim, final predictor selection was based on a combination of clinical relevance and stability across ML feature importance analyses. An RF regressor analysis on training set (to prevent data leakage into and test sets) was performed, in which features were ranked by importance, and candidates were reviewed for clinical plausibility before inclusion²².

Train–Test Split and Cross-Validation

The dataset was randomly split into training (80%) and test (20%) sets using stratified sampling to preserve the outcome distribution^{21,23}. The test set was held out and used exclusively for final model evaluation. Within the training set, five-fold stratified CV was employed for model tuning and internal validation⁴. All preprocessing steps, including imputation and scaling, were performed within model-specific pipelines and exclusively inside cross-validation loops, ensuring strict prevention of data leakage²⁴.

Model Development

Four supervised learning models were evaluated: elastic-net regularized LR, linear SVM, shallow ANN, and constrained RF.

For models requiring feature scaling, missing continuous variables were imputed using the median and standardized using z-score normalization within each cross-validation fold. For tree-based models, median imputation was applied without scaling. Class imbalance was addressed using class-weighted loss functions where applicable ²⁵.

Hyperparameter tuning was conducted using grid search within cross-validation, with the AUROC as the primary optimization metric. Model performance was evaluated on the independent test set using AUROC, accuracy, sensitivity, specificity, and F1-score ²⁶.

Model Selection and Explainability

The model with the most favorable balance of discrimination and classification performance on the test set and best prevention of overfitting was selected as the final model. Model explainability was assessed using SHAP with TreeExplainer. SHAP summary plots were generated to quantify the global contribution and directionality of each predictor to the model's output ⁷. SHAP results were interpreted as associative explanations of the model's behavior rather than causal effects, with particular attention to clinically plausible patterns and potential confounding by disease severity.

Calibration Assessment

Model calibration was evaluated on the test set using multiple complementary approaches ¹⁰. Probabilistic accuracy was quantified using the Brier score. Visual calibration was assessed using calibration curves with quantile-based binning. Calibration intercept and slope were estimated via LR of observed outcomes on the logit-transformed predicted probabilities. If miscalibration was present, post-hoc probability recalibration was performed using Platt scaling (sigmoid calibration), fitted on the training set via five-fold CV and applied to the test set. Calibration performance before and after recalibration was compared, while discrimination was reassessed to ensure preservation of ranking ability ¹⁰.

Decision Curve Analysis

Clinical utility was evaluated using DCA based on the final calibrated predicted probabilities. Net benefit was calculated across a range of clinically relevant threshold probabilities and compared with treat-all

and treat-none strategies. This analysis assessed whether use of the prediction model could provide incremental clinical value in guiding hemodialysis decisions⁸.

Bedside Risk Assessment Tool Development

To enhance clinical applicability, we sought to derive a simplified, point-based risk scoring system that could be used at the bedside to support hemodialysis decisions in lithium poisoning. This approach involved translating a multivariable model into an interpretable additive scale using clinically meaningful categorical thresholds.

Candidate predictors were selected based on their clinical plausibility, availability at the time of emergency department presentation, and predictive relevance in earlier modeling phases. Each continuous predictor was binned into clinically relevant categories, and all variables were encoded as binary or ordinal factors. We then fit an LR model on the training set using these categorized variables. Regression coefficients were scaled and rounded to derive integer-based point values. The resulting total scores were used to stratify patients into risk groups for model interpretability and potential bedside deployment.

Software and Analysis

The statistical analyses of summary data and between group (no hemodialysis vs. hemodialysis) differences were performed using Stata 18 (StataCorp. 2023). Kolmogorov-Smirnov test for normality was performed on continuous variables, and they were compared between groups using either t-test or Mann-Whitney U test. Binary variables were compared using either Chi-square or Fisher exact test. Finally, ordinal variables (GCS on admission and type of poisoning) were compared using trend test. A p-value < 0.05 was considered significant. Also, the model development and regressor-based steps, along with explainability, calibration, and DCA were performed using Python 3.

Declarations

ACKNOWLEDGMENT

The present work was supported by the Toxicological Research Center, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences.

AUTHOR CONTRIBUTION

Kamran Rezaei, Shahin Shadnia, and Peyman Erfan Talab Evini conceptualized and administered the project. Methodology was confirmed by **Babak Mostafazadeh, Shahin Shadnia, Mitra Rahimi, Sayed Masoud Hosseini, and Joshua King**. Formal analysis was performed by **Sayed Masoud Hosseini and Kamran Rezaei**. Software, visualizations, machine learning model development, and assessments were

performed by **Kamran Rezaei** and **Fatemeh Saber**. Validation and supervision were done by **Sayed Masoud Hosseini** and **Peyman Erfan Talab Evini**. Data curation process was done by **Kamran Rezaei**, **Sarina Sadat Abouei Mehrizi**, and **Pooya Eini**. Writing of the first draft was performed by **Kamran Rezaei**, and Review-editing of the manuscript was performed by **Babak Mostafazadeh**, **Shahin Shadnia**, **Mitra Rahimi**, **Joshua King**, and **Peyman Erfan Talab Evini**.

DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are not publicly available due to ethical restrictions elaborated by the Loghman Hakim Hospital data center, but are available from the corresponding author upon reasonable request.

DISCLOSURE OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

No funding source was used for this project.

References

1. Hoffman, R. S. Evidence-based recommendations for haemodialysis in lithium-poisoned patients: Getting from where we are to where we want to be. *British Journal of Clinical Pharmacology* **86**, 528 (2020).
2. Ivkovic, A. & Stern, T. A. Lithium-induced neurotoxicity: clinical presentations, pathophysiology, and treatment. *Psychosomatics* **55**, 296–302 (2014).
3. Buckley, N. A. *et al.* Haemodialysis for lithium poisoning: translating EXTRIP recommendations into practical guidelines. *British journal of clinical pharmacology* **86**, 999–1006 (2020).
4. Kaliappan, J. *et al.* Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise. *Diagnostics (Basel)* **13** (2023). <https://doi.org/10.3390/diagnostics13101692>
5. Mohr, F. & van Rijn, J. N. Learning curves for decision making in supervised machine learning: a survey. *Machine Learning* **113**, 8371–8425 (2024).
6. Mehrpour, O., Abdollahi, M. & Sharifi, M. D. Oxidative stress and hyperglycemia in aluminum phosphide poisoning. *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences* **19**, 196 (2014).
7. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J Comput Aided Mol Des* **34**, 1013–1026 (2020). <https://doi.org/10.1007/s10822-020-00314-0>

8. Van Calster, B. *et al.* Reporting and interpreting decision curve analysis: a guide for investigators. *European urology* **74**, 796–804 (2018).
9. Reimann, F. & Whyte, I. Chronic lithium toxicity. *Aust Prescr* **45**, 93–94 (2022).
<https://doi.org/10.18773/austprescr.2022.024>
10. Ojeda, F. M. *et al.* Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Stat Med* **42**, 5451–5478 (2023). <https://doi.org/10.1002/sim.9921>
11. Decker, B. S. *et al.* Extracorporeal treatment for lithium poisoning: systematic review and recommendations from the EXTRIP workgroup. *Clinical Journal of the American Society of Nephrology* **10**, 875–887 (2015).
12. Liu, Y.-H. *et al.* Hemodialysis treatment for patients with lithium poisoning. *International journal of environmental research and public health* **19**, 10044 (2022).
13. Mehrpour, O., Vohra, V., Nakhaee, S., Mohtarami, S. A. & Shirazi, F. M. Machine learning for predicting medical outcomes associated with acute lithium poisoning. *Scientific Reports* **15**, 14468 (2025).
14. Konieczny, K., Detraux, J. & Bouckaert, F. The Syndrome of Irreversible Lithium-Effectuated Neurotoxicity: A Scoping Review. *Alpha Psychiatry* **25**, 190 (2024).
15. Baird-Gunning, J., Lea-Henry, T., Hoegberg, L. C., Gosselin, S. & Roberts, D. M. Lithium poisoning. *Journal of intensive care medicine* **32**, 249–263 (2017).
16. Ehrlich, B. E., Clausen, C., Gosenfeld, L. F. & Diamond, J. M. Lithium concentration in the muscle compartment of manic-depressive patients during lithium therapy. *Journal of psychiatric research* **18**, 139–148 (1984).
17. Munshi, K. R. & Thampy, A. The syndrome of irreversible lithium-effectuated neurotoxicity. *Clinical neuropharmacology* **28**, 38–49 (2005).
18. Ahmad, M., Elnakady, Y., Farooq, M. & Wadaan, M. Lithium induced toxicity in rats: blood serum chemistry, antioxidative enzymes in red blood cells and histopathological studies. *Biological and Pharmaceutical Bulletin* **34**, 272–277 (2011).
19. Broulik, P., Štěpán, J., Souček, K. & Pacovský, V. Alterations in human serum alkaline phosphatase and its bone isoenzyme by chronic administration of lithium. *Clinica chimica acta* **140**, 151–155 (1984).
20. Zamani, A., Omrani, G. R. & Nasab, M. M. Lithium's effect on bone mineral density. *Bone* **44**, 331–334 (2009).
21. Park, S. Y. *et al.* Artificial neural network approach for acute poisoning mortality prediction in emergency departments. *Clinical and experimental emergency medicine* **8**, 229 (2021).
22. Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications* **134**, 93–101 (2019).
23. El-Banna, A., Ragab, H., Rohiem, R. & Mohamed Saleh, A. Implementation of machine learning models bridges the prognostic gap in Aluminum Phosphide poisoning. *Egyptian Journal of Forensic*

Sciences **15**, 1–15 (2025).

24. DOMNIK, J. & HOLLAND, A. On data leakage prevention and machine learning. *35th Bled eConference Digital Restructuring and Human (Re) action*, 695 (2022).
25. Le, D. N., Le, H. X., Ngo, L. T. & Ngo, H. T. Transfer learning with class-weighted and focal loss function for automatic skin cancer classification. *arXiv preprint arXiv:2009.05977* (2020).
26. Naidu, G., Zuva, T. & Sibanda, E. M. in *Computer science on-line conference*. 15–25 (Springer).

Tables

Table 1. Descriptive summary table of the included patients.

Variable	No hemodialysis	Hemodialysis	Total	p-value	Missing (%)
N	180 (87.0%)	27 (13.0%)	207 (100.0%)	-	-
Age	28.78 (12.99)	40.92 (15.42)	30.37 (13.91)	<0.001	0 (0)
Sex	-	-	-	0.11	0 (0)
Males	36 (20.0%)	9 (33.3%)	45 (21.7%)	-	-
Females	144 (80.0%)	18 (66.7%)	162 (78.3%)	-	-
Last amount of lithium ingestion in milligrams (acute/acute on chronic)*	3,000	6,000	3,000	0.04	60 (28.99)
Poisoning type	-	-	-	<0.001	0 (0)
Acute**	43 (23.9%)	0 (0.0%)	43 (20.8%)	-	-
Acute on chronic	132 (73.3%)	22 (81.5%)	154 (74.4%)	-	-
Chronic **	5 (2.8%)	5 (18.5%)	10 (4.8%)	-	-
Coingestion	161 (89.4%)	18 (66.7%)	179 (86.5%)	<0.001	-
Approximate time elapsed to ED admission (minutes)*	180	150	150	0.46	28 (13.53)
Gastric lavage before ED admission**	12 (6.7%)	1 (3.7%)	13 (6.3%)	1.00	0 (0)
Signs and symptoms on ED admission					
Nausea	47 (26.1%)	6 (22.2%)	53 (25.6%)	0.71	0 (0)
Vomiting	29 (16.1%)	6 (22.2%)	35 (16.9%)	0.14	0 (0)
Diarrhea**	5 (2.8%)	0 (0.0%)	5 (2.4%)	1.00	0 (0)
Neurological manifestations (abnormal motor, myoclonus, seizure)	31 (17.2%)	15 (55.6%)	46 (22.2%)	<0.001	0 (0)

Past medical history					
History of lithium poisoning**	19 (10.6%)	5 (18.5%)	24 (11.6%)	0.25	0 (0)
Allergy **	6 (3.3%)	0 (0.0%)	6 (2.9%)	1.00	0 (0)
Ischemic heart disease**	5 (2.8%)	2 (7.4%)	7 (3.4%)	1.00	0 (0)
Hypertension**	6 (3.3%)	2 (7.4%)	8 (3.9%)	0.21	0 (0)
Hyperthyroidism**	2 (1.1%)	0 (0.0%)	2 (1.0%)	1.00	0 (0)
Hypothyroidism**	5 (2.8%)	2 (7.4%)	7 (3.4%)	0.51	0 (0)
Diabetes mellitus**	3 (1.7%)	3 (11.1%)	6 (2.9%)	0.09	0 (0)
Hepatic disorders**	0 (0.0%)	1 (3.7%)	1 (0.5%)	-	0 (0)
Renal disorders**	0 (0.0%)	2 (7.4%)	2 (1.0%)	-	0 (0)
Neurological disorders**	14 (7.8%)	4 (14.8%)	18 (8.7%)	0.39	0 (0)
Psychiatric disorders	165 (91.7%)	27 (100.0%)	192 (92.8%)	0.60	0 (0)
Vital signs on admission					
GCS	-	-	-	<0.001	0 (0)
3	0 (0.0%)	1 (3.7%)	1 (0.5%)	-	-
4	0 (0.0%)	1 (3.7%)	1 (0.5%)	-	-
5	3 (1.7%)	2 (7.4%)	5 (2.4%)	-	-
6	1 (0.6%)	4 (14.8%)	5 (2.4%)	-	-
7	2 (1.1%)	0 (0.0%)	2 (1.0%)	-	-
8	2 (1.1%)	0 (0.0%)	2 (1.0%)	-	-
9	1 (0.6%)	0 (0.0%)	1 (0.5%)	-	-
10	4 (2.2%)	1 (3.7%)	5 (2.4%)	-	-
11	15 (8.3%)	2 (7.4%)	17 (8.2%)	-	-
12	11 (6.1%)	2 (7.4%)	13 (6.3%)	-	-
13	141 (78.3%)	14 (51.9%)	155 (74.9%)	-	-
14	0 (0.0%)	1 (3.7%)	1 (0.5%)	-	-
15	0 (0.0%)	1 (3.7%)	1 (0.5%)	-	-

Pulse oximetry*	96.24 (2.43)	95.96 (4.22)	96.20 (2.73)	0.39	5 (2.42)
Heart rate	84.07 (14.04)	85.96 (16.99)	84.31 (14.42)	0.53	6 (2.9)
Respiratory rate	15.98 (2.80)	16.00 (3.20)	15.98 (2.84)	0.97	10 (4.83)
Mean arterial pressure*	88.33	88.33	88.33	0.58	8 (3.86)
Temperature (Celsius)*	37.01	37.01	37.01	0.32	70 (33.82%)
First laboratory findings					
Hemoglobin	12.99 (1.62)	12.519 (2.55)	12.93 (1.77)	0.19	2 (0.97)
White blood cell count*	8.55	9.30	8.70	0.04	2 (0.97)
Platelet count*	244	243	243	0.77	2 (0.97)
Hematocrit	38.96 (4.19)	37.97 (6.13)	38.83 (4.49)	0.28	2 (0.97)
Aspartate transaminase*	24	25	24	0.60	2 (0.97)
Alanine transaminase*	18	17	18	0.87	2 (0.97)
Alkaline phosphatase*	169	176	170	0.72	2 (0.97)
Lactate dehydrogenase*	375	418	384	0.14	24 (11.59)
Creatine phosphokinase*	85	104	86	0.28	18 (8.7)
Serum urea*	25	30	26	0.02	0 (0)
Creatinine*	1.0	1.1	1.0	<0.001	0 (0)
Blood sugar	95	108	96	0.16	46 (22.22)
Serum sodium	136.42 (2.99)	135.77 (4.12)	136.34 (3.16)	0.32	0 (0)
Serum potassium*	3.8	4.0	3.9	0.03	0 (0)
Venous Blood gas: pH	7.38 (0.05)	7.37 (0.07)	7.38 (0.05)	0.46	8 (3.86)
Venous Blood gas: HCO3*	24.10	25.20	24.20	0.50	8 (3.86)
Venous Blood gas: pCO2*	39.90	40.80	40.00	0.90	8 (3.86)
Venous Blood gas: pO2*	42.00	44.50	44.30	0.84	8 (3.86)

First lithium concentration	1.26 (0.65)	2.75 (1.58)	1.46 (0.96)	<0.001	0 (0)
Outcome variables					
Worst lithium concentration*	1.27	2.30	1.40	<0.001	0 (0)
Confusion	9 (5.0%)	19 (70.4%)	28 (13.5%)	<0.001	0 (0)
Seizure **	0 (0.0%)	2 (7.4%)	2 (1.0%)	-	0 (0)
Refractory shock or dysrhythmia	0 (0.0%)	8 (29.6%)	8 (3.9%)	<0.001	0 (0)
Lithium concentration over 4 mmol/L	0 (0.0%)	8 (29.6%)	8 (3.9%)	<0.001	0 (0)
Lithium concentration, not decreased to 1 mmol/L in 36 hours	9 (5.0%)	15 (55.6%)	24 (11.6%)	<0.001	0 (0)
Length of hospital stay (hours)*	20	68	22	<0.001	1 (0.48)
In hospital mortality	0 (0.0%)	5 (18.5%)	5 (2.4%)	-	0 (0)

Continuous variables are summarized as mean and standard deviation (for normally distributed variables) or median and interquartile range (for non-normally distributed variables), and binary/ordinal/categorical variables are summarized as frequency (percentage).

*Non-normal distribution; differences analyzed using Mann-Whitney U test sum test.

**Differences assessed using Fisher's exact test, rather than chi-square test.

Table 2. Results of evaluation metrics in each trained model.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score	AUROC on test set (95% CI)	Mean CV AUROC (95% CI)
Elastic-Net LR*	79	80	78	0.47	0.79 (0.43-1.00)	0.91 (0.79-1.02)
Linear SVM**	88	60	91	0.54	0.79 (0.48-0.97)	0.90 (0.78-1.02)
Shallow ANN***	73	60	75	0.35	0.60 (0.48-0.97)	0.90 (0.77-1.03)
Constrained RF****	83	80	83	0.53	0.83 (0.65-0.97)	0.90 (0.80-0.99)

*Best params: {'clf__C': 0.1, 'clf__l1_ratio': 0.25}.

**Best params: {'clf__C': 0.01}.

***Best configuration: {'hidden_units': 16.0, 'l2': 0.1}

****Best params: {'clf__max_depth': 4, 'clf__max_features': 'sqrt', 'clf__min_samples_leaf': 10, 'clf__n_estimators': 300}

Table 3. Confusion matrices for each trained model.

Model		Observed positive	Observed negative
Elastic-Net LR	<i>Predicted positive</i>	4	8
	<i>Predicted negative</i>	1	29
Linear SVM	<i>Predicted positive</i>	3	3
	<i>Predicted negative</i>	2	34
Shallow ANN	<i>Predicted positive</i>	3	9
	<i>Predicted negative</i>	2	28
Constrained RF	<i>Predicted positive</i>	4	6
	<i>Predicted negative</i>	1	31

Figures

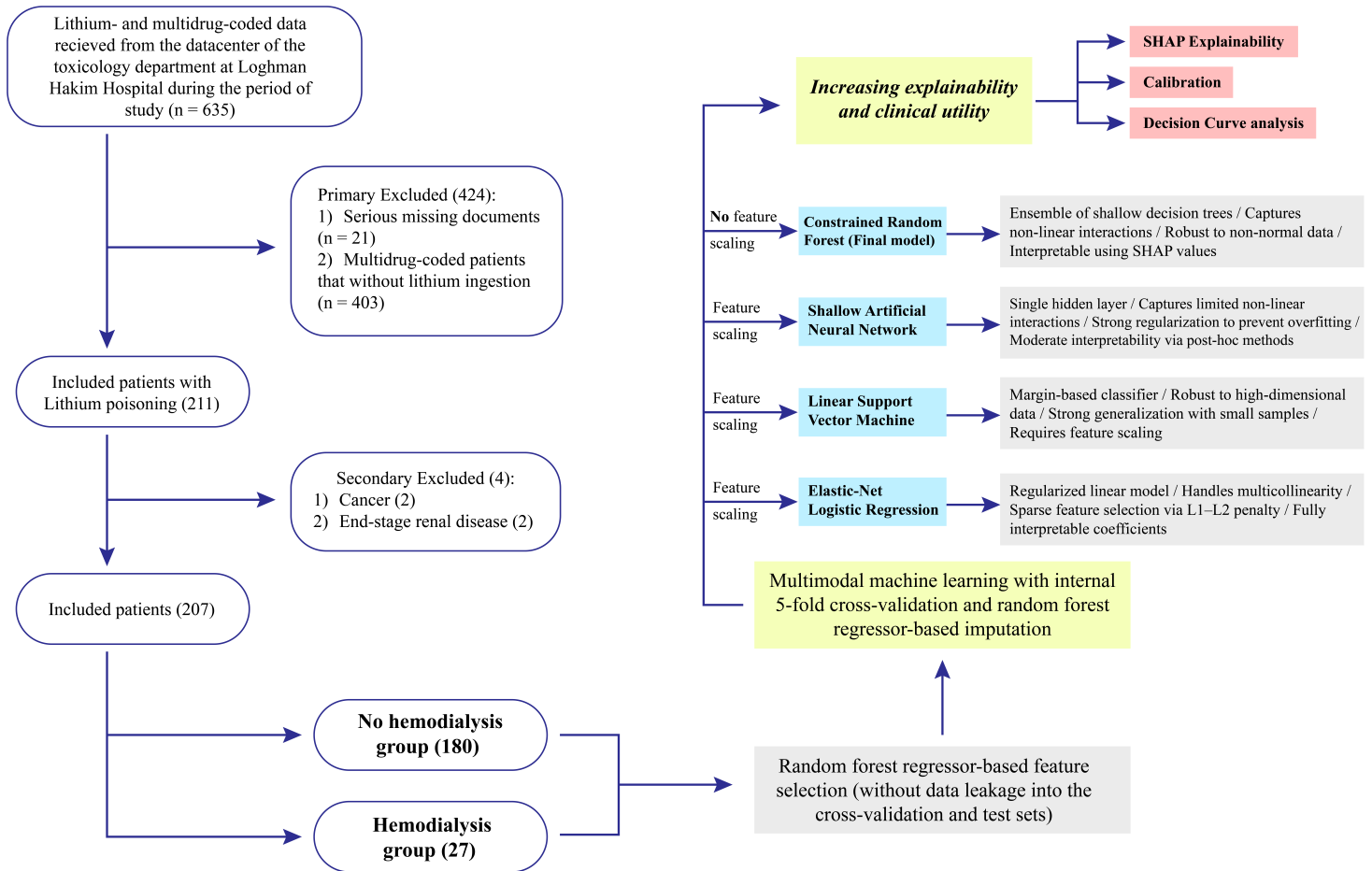


Figure 1

Patient selection flowchart, and a summary of roadmap from patient selection to model development.

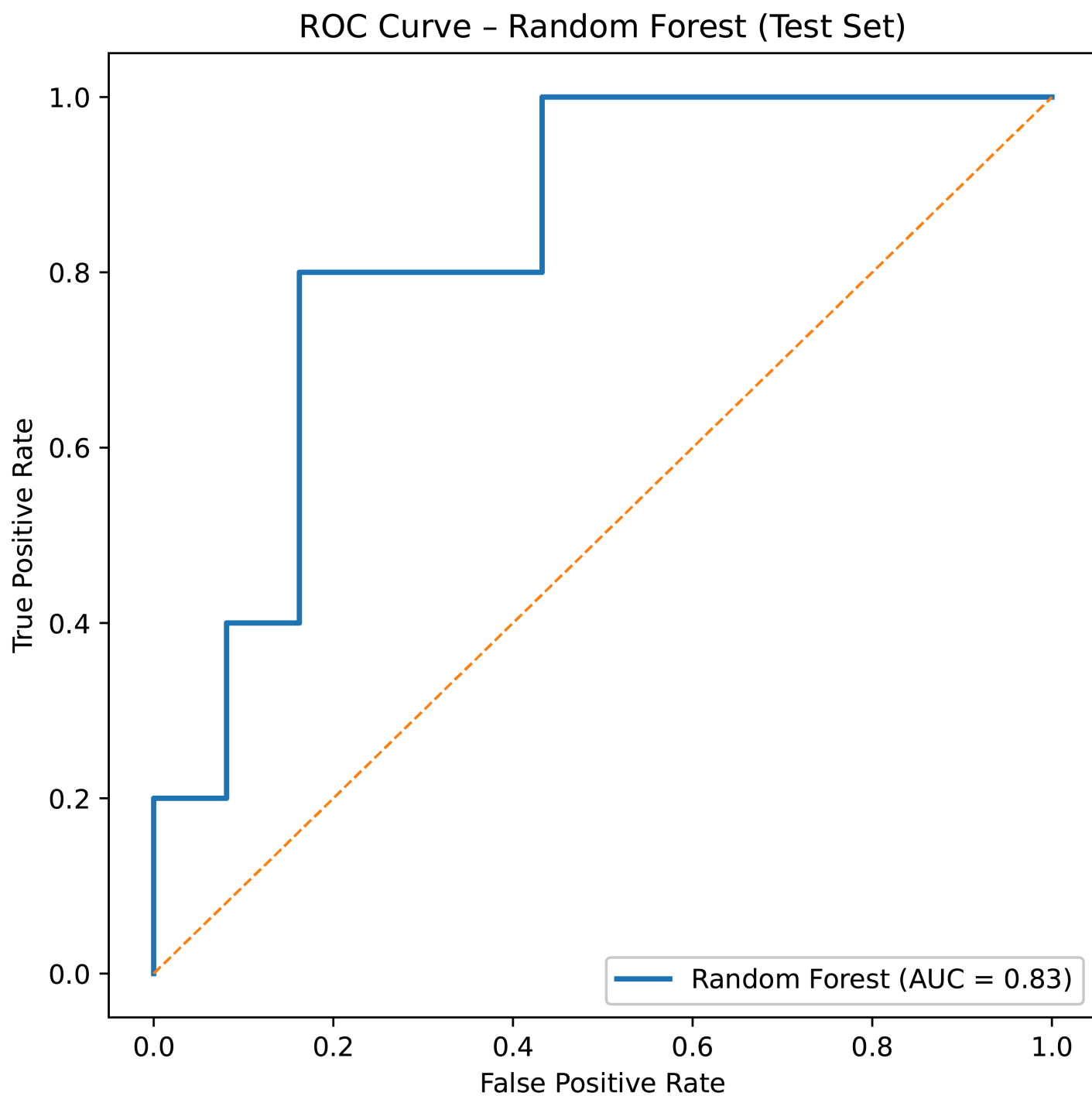


Figure 2

Receiver Operating Characteristic (ROC) curve for the selected RF model.

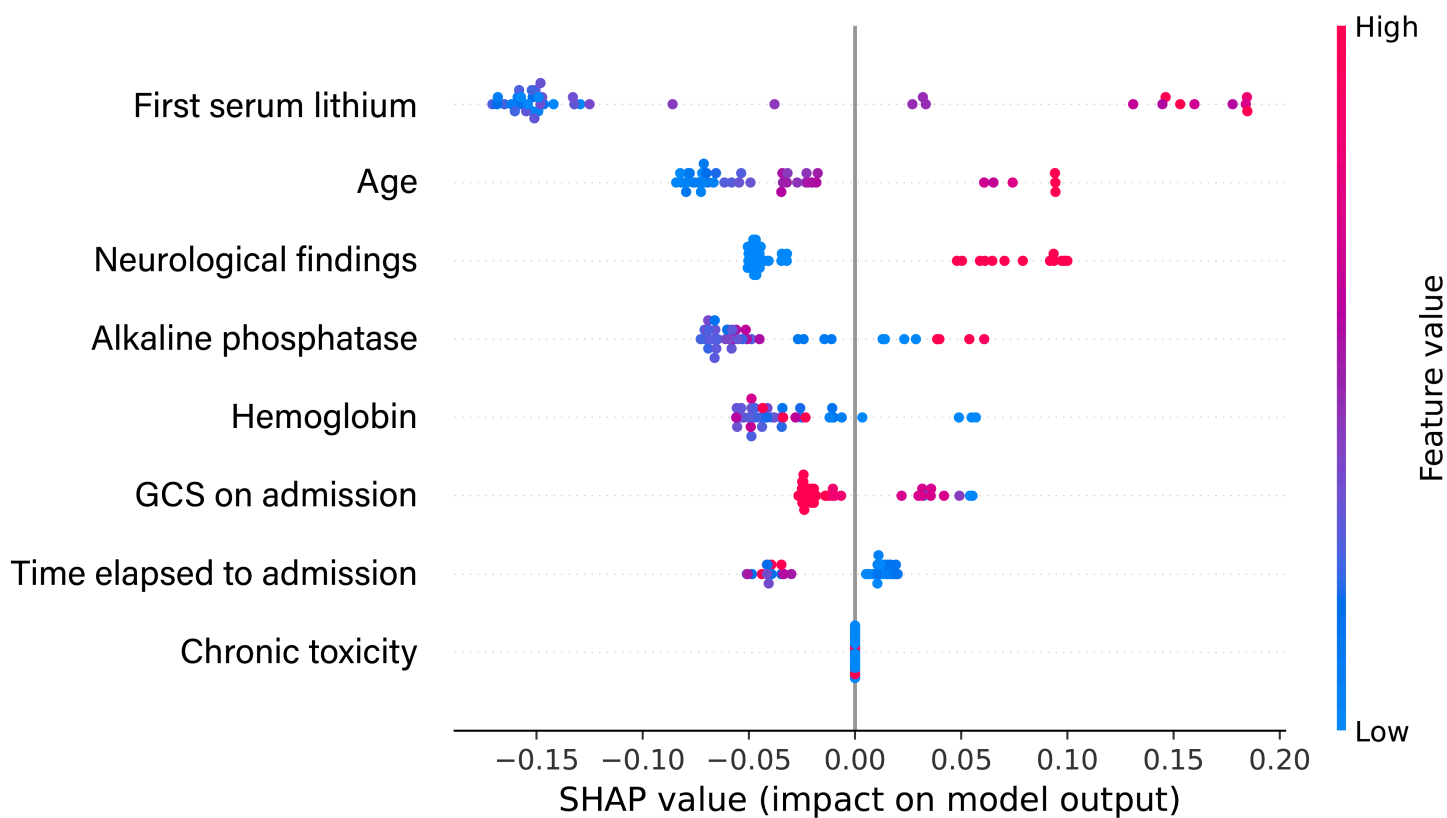


Figure 3

SHAP summary plot depicting the most important features and their contribution to the risk of hemodialysis in lithium poisoning.

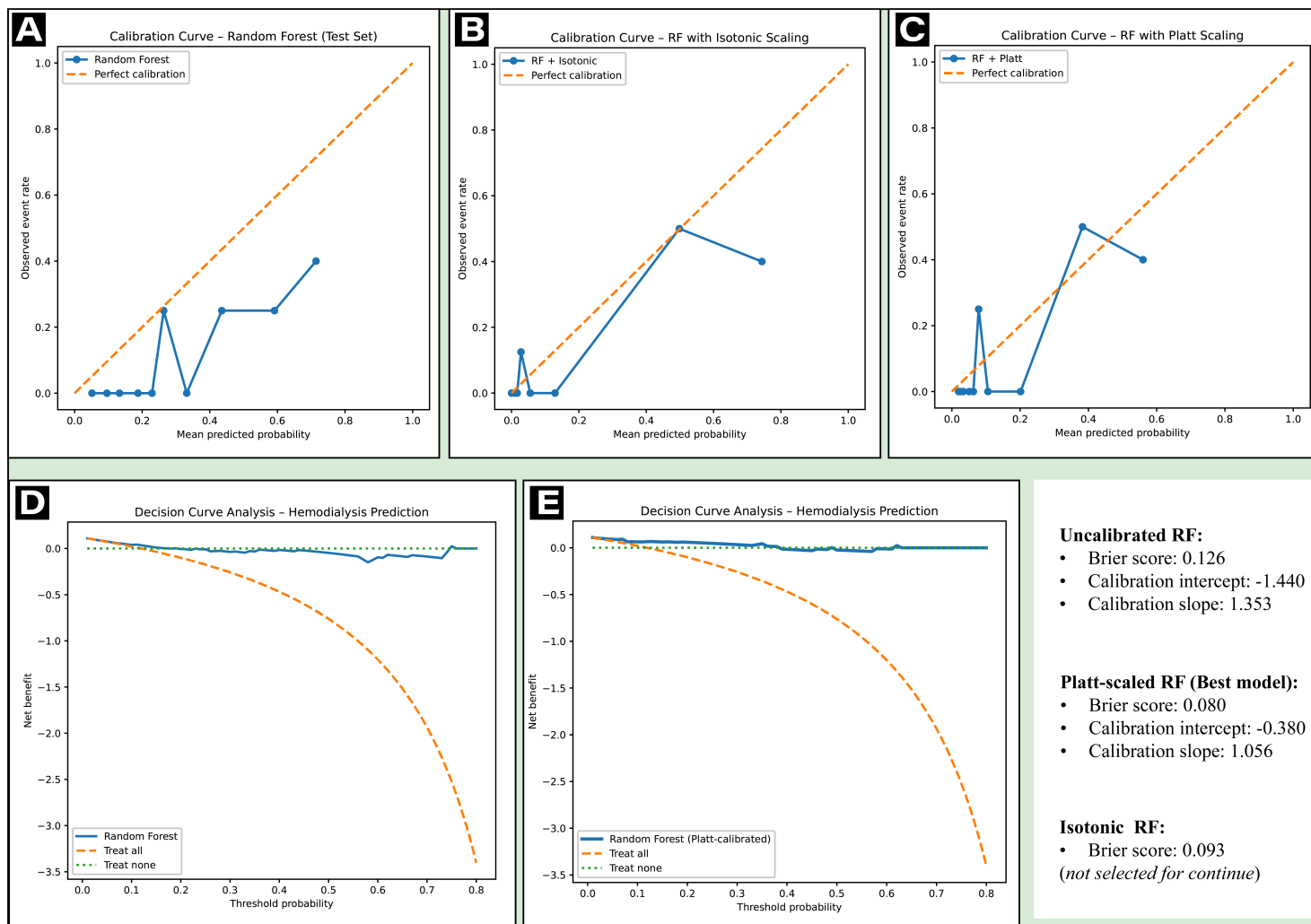


Figure 4

Calibration curves for (A) non-calibrated RF model, (B) recalibrated RF model with isotonic scaling, and (C) recalibrated RF model with Platt scaling. Also, decision curve plots were depicted for (D) non-calibrated RF model and (E) recalibrated RF model with Platt scaling.

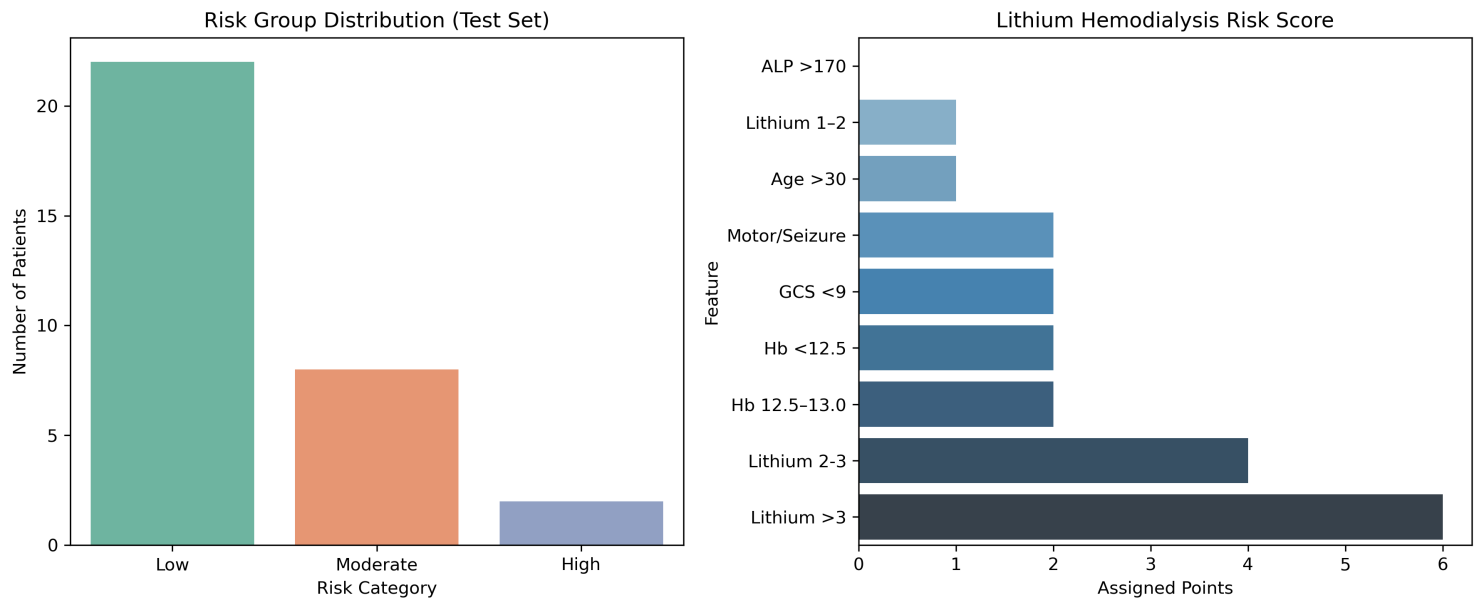


Figure 5

Risk group distribution assessed on the test set (left), and summary of the scoring structure across each variable in the bedside score system. ALP: alkaline phosphatase, Lithium: first lithium concentration, Hb: hemoglobin, Motor/Seizure: assigned variable for severe neurological manifestations.

Supplementary Files

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

- [MLModelsPythonCodes.rar](#)
- [SupplementarymaterialsNestedcvRFtopfeatures.xls](#)
- [SupplementaryTables.docx](#)
- [SupplementaryMaterialNomogram.pdf](#)