Open Access Library Journal
2026, Volume 13, e14920
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Lithium Associated Chronic Kidney Disease Prediction Using Explainable Machine Learning: A Comprehensive Modelling and Interpretation Framework

Rocco de Filippis[1]*, Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Lithium Associated Chronic Kidney Disease Prediction Using Explainable Machine Learning: A Comprehensive Modelling and Interpretation Framework. *Open Access Library Journal*, **13**: e14920.
https://doi.org/10.4236/oalib.1114920

Received: January 23, 2026
Accepted: March 17, 2026
Published: March 20, 2026

Abstract

Lithium remains the most effective long-term treatment for bipolar disorder, yet its therapeutic benefits are offset by a well-established risk of chronic kidney disease (CKD). Anticipating lithium-associated renal impairment is clinically challenging because the underlying mechanisms are subtle, multivariate, and evolve dynamically with cumulative exposure. In this study, we develop a transparent, end-to-end machine-learning framework for early detection of lithium-induced kidney damage using a comprehensive synthetic cohort that incorporates lithium duration, serum concentrations, renal function biomarkers, comorbidities, metabolic factors, and inflammatory markers. Four supervised classifiers, Random Forest, XGBoost, logistic regression, and CatBoost were evaluated using stratified 5-fold cross-validation. CatBoost achieved the strongest generalization performance (test AUC = 0.785) and was subsequently selected for full explainability analysis. To ensure clinical interpretability, we integrate a multi-layered explanation suite comprising SHAP-based global and local attributions, interaction effect quantification, patient-specific waterfall plots, LIME explanations, bootstrap confidence intervals, permutation-based statistical validation, and correlation analyses between feature values and model-derived risk. Across methods, lithium exposure metrics (duration, serum level, dose), renal decline markers (eGFR trajectory, creatinine, proteinuria), hypertension, and key interaction terms (duration × level; age × duration) consistently emerge as dominant predictors of early renal damage. The SHAP-derived risk landscape reveals coherent, monotonic associations between cumulative lithium exposure, deterioration in renal function, and elevated CKD risk, aligning closely with established nephrotoxic pathways. Although the dataset is syn-

thetic, the modelling strategy provides a rigorous blueprint for interpretable AI driven nephrotoxicity surveillance. The framework offers actionable, clinician-ready insights and establishes a foundation for future validation on real-world lithium cohorts, supporting precision monitoring and early intervention in lithium-treated patients.

Subject Areas

Bioinformatics

Keywords

Lithium Nephrotoxicity, Chronic Kidney Disease, Explainable AI, SHAP, CatBoost, Clinical Risk Modelling, Machine Learning

1. Introduction

Lithium remains one of the most effective and enduring mood stabilizers for bipolar disorder, with unparalleled efficacy in preventing mania, reducing suicidality, and stabilizing long-term mood trajectories [1]-[6]. Yet its therapeutic benefits are counterbalanced by well-documented renal adverse effects, including chronic interstitial nephritis, progressive reductions in glomerular filtration rate (GFR), and, in a subset of patients, the development of chronic kidney disease (CKD). Although lithium nephropathy has been recognized for decades, early identification of renal injury remains a persistent clinical challenge [7]-[13]. The underlying risk is shaped by a multifactorial and nonlinear interplay of treatment duration, serum lithium levels, cumulative exposure, cardio-metabolic comorbidities (e.g., hypertension, diabetes), baseline renal vulnerability, proteinuria, inflammatory processes, and metabolic dysregulation [14]-[16]. These complex interactions unfold over time and vary markedly across individuals, making conventional statistical screening tools insufficiently sensitive or too rigid to capture evolving patterns of toxicity. Recent advances in machine learning (ML) present an opportunity to model these intricate exposure response relationships with far greater resolution. However, clinical adoption requires more than raw predictive accuracy [17]-[23]. Nephrologists and psychiatrists must be able to understand why a model issues a given prediction, discern which features drive risk for a specific patient, and verify that these mechanisms align with established renal pathophysiology [24]-[31]. Consequently, black-box ML models, even when highly accurate, are unsuitable for clinical decision support unless accompanied by stable, transparent, and interpretable reasoning pathways [32]-[39]. To address this unmet need, we propose a comprehensive explainable machine-learning (XAI) framework for predicting early lithium-associated renal impairment. The framework builds directly on methodological principles established in interpretable computational psychiatry and optimization research, including the reinforcement-learning based modelling approach described in Reinforcement Learning Based Opti-

mization of Sleep, Mood, Circadian-Dynamics in Bipolar Disorder. Extending that philosophy of transparent computational modelling to nephrotoxicity, our framework integrates:

1. Predictive modelling across multiple supervised learners to quantify the robustness of performance across algorithms.

2. Mathematical decomposition of predictions using Shapley values (SHAP) to provide global and patient-level interpretability.

3. Local explanations using SHAP and LIME, enabling clinicians to examine case-specific mechanisms and risk drivers.

4. Statistical validation through bootstrapping and permutation testing, ensuring stability and significance of inferred biomarkers.

5. Clinical interpretation maps linking algorithmic risk signals to established nephrotoxic pathways and lithium-exposure physiology.

By combining predictive modelling with a multilayered interpretability pipeline, this study delivers a transparent analytic framework tailored for clinical reasoning. While the present analysis uses a synthetic cohort, the methodological structure closely mirrors real-world renal risk modelling demands and offers a reproducible blueprint for deploying interpretable artificial intelligence in lithium monitoring. In doing so, the paper advances the broader goal of bridging computational modelling with actionable, mechanism-aligned insights in nephrology.

2. Methods

2.1. Dataset and Features

To investigate early lithium-associated renal injury within a controlled and analytically transparent environment, we constructed a synthetic cohort of 1,500 patients designed to emulate the clinical and biochemical profile of individuals receiving long-term lithium therapy. The synthetic dataset was generated using empirically informed distributions aligned with reported epidemiological and nephrological trends [40]-[44]. This approach ensures that feature interactions, marginal distributions, and clinical dependencies reflect plausible physiological behaviour while enabling full control over noise, imbalance, and complexity.

The dataset encompasses demographic variables (age, sex), renal function markers (serum creatinine, baseline eGFR, current eGFR, proteinuria, eGFR decline rate), lithium exposure variables (treatment duration in years, serum lithium concentration, daily dose), comorbidities (hypertension, diabetes, cardiovascular risk), metabolic factors (serum calcium, phosphate, BMI), and inflammatory markers (e.g., IL-6). These variables collectively capture the multidimensional landscape of factors known to modulate susceptibility to lithium-induced nephrotoxicity. To better reflect the nonlinear and synergistic relationships observed in clinical practice, we engineered three interaction features with established relevance in nephrology and psychopharmacology:

1. lithium_duration × lithium_level - representing cumulative toxic load, capturing how long-term exposure at higher serum levels magnifies renal risk far

more than either variable individually [45]-[50].

2. age × lithium_duration - modelling the accelerated vulnerability of aging renal tissue under prolonged lithium treatment [51]-[56].

3. eGFR_decline_rate × proteinuria- integrating two primary signatures of early nephron injury, producing a sensitive indicator of progressive CKD trajectory [57]-[60].

The target variable, early kidney damage**,** was encoded as a binary label reflecting an early-stage CKD phenotype. Although deterministic in generation, the label incorporates stochastic variation to emulate the uncertainty and incomplete observability inherent to clinical diagnostics [61]-[63].

This synthetic design offers several advantages:

- Controlled complexity: allowing systematic evaluation of model performance under known interactions [64] [65].
- Balance between realism and tractability: enabling interpretable mechanistic insight without confounding clinical noise [66] [67].
- Reproducibility: ensuring replicable analysis for benchmarking predictive and explainability methods [68]-[70].

Together, the dataset and engineered features form a rigorous foundation for assessing machine-learning based nephrotoxicity prediction and for evaluating the stability and interpretability of explainable AI methods applied to lithium-associated CKD.

2.1.1. Clinical Definition of Early Kidney Damage

To ensure alignment with established nephrology guidelines, the binary outcome variable "early kidney damage" was defined according to KDIGO criteria for early-stage CKD. A patient was labelled as positive ($Y = 1$) if at least one of the following conditions was met:

- Estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73m^2 persisting for ≥ 3 simulated months,
- Annual eGFR decline rate ≥ 5 mL/min/1.73m^2,
- Presence of proteinuria exceeding 150 mg/day,
- Serum creatinine elevation exceeding 1.3 mg/dL (men) or 1.1 mg/dL (women).

To reflect real-world diagnostic uncertainty, stochastic perturbation (Gaussian noise with $\sigma = 0.03$) was added to the final probability before thresholding. This ensures that the target variable reflects physiologically grounded CKD definitions rather than a purely deterministic rule.

2.1.2. Synthetic Data Generation Parameters

Synthetic features were generated using multivariate normal and log-normal distributions parameterized according to epidemiological lithium monitoring literature.

Key parameters include:

- Age ~ N(52, 14^2)
- Lithium duration (years) ~ Gamma ($k = 3.5$, $\theta = 2.0$)

- Serum lithium level ~ N(0.75, 0.15^2)
- Baseline eGFR ~ N(82, 18^2)
- Annual eGFR decline rate ~ N(2.3, 1.2^2)
- Proteinuria ~ LogNormal ($\mu = 4.8$, $\sigma = 0.6$)

Feature correlations were imposed using the following covariance structure (partial matrix):

Feature Pair	ρ
Lithium duration—age	0.42
Lithium level—duration	0.35
eGFR decline—proteinuria	0.48
Hypertension—age	0.52

2.2. Mathematical Formulation of the Predictive Model

Our goal is to estimate, for each patient, the probability of having early lithium-associated kidney damage given their clinical and treatment profile.

Let,

$$x_i \in \mathbb{R}^p$$

denote the feature vector for patient i, where p is the total number of predictors (demographics, lithium exposure, renal markers, comorbidities, metabolic and inflammatory variables, plus engineered interactions). The binary outcome is

$$y_i \in \{0,1\},$$

with $y_i = 1$ indicating early kidney damage and $y_i = 0$ indicating no early damage.

2.2.1. CatBoost as Nonlinear Risk Function

We use CatBoost as the main predictive model. Conceptually, CatBoost can be seen as a flexible nonlinear function [71]-[73]

$$f : \mathbb{R}^p \to \mathbb{R},\ f(x_i) = \text{logit of risk for patient } i.$$

Here, $f(x_i)$ is the model's output in the log-odds scale (also called "logit"), not yet a probability. To convert this into a clinically interpretable probability of early kidney damage, we apply the logistic (sigmoid) function [74] [75]:

$$P(y_i = 1 \mid x_i) = \sigma(f(x_i)) = \frac{1}{1 + \exp(-f(x_i))}.$$

- If $f(x_i) = 0$, then $P(y_i = 1 \mid x_i) = 0.5$.
- If $f(x_i) > 0$, the probability is greater than 0.5 (higher risk).
- If $f(x_i) < 0$, the probability is less than 0.5 (lower risk).

CatBoost itself is an ensemble of many decision trees, trained sequentially (boosting). Each new tree tries to correct the errors of previous trees, allowing the model to capture complex nonlinear interactions among features (e.g., the joint

effect of lithium duration and serum lithium level) [76] [77].

2.2.2. Shapley Additive Explanations (SHAP)

While CatBoost can learn complex patterns, the raw function $f(x_i)$ is not directly interpretable. We therefore use SHAP (SHapley Additive exPlanations) to decompose the prediction for each patient into feature contributions.

a) Additive decomposition of the prediction

For each patient i, SHAP approximates the model output as:

$$f(x_i) \approx \phi_0 + \sum_{j=1}^{p} \phi_{ij},$$

where:

- ϕ_0 is the baseline value, equal to the expected log-odds of early kidney damage over the whole training dataset. Intuitively, this is the model prediction before seeing any patient-specific information.
- ϕ_{ij} is the Shapley value for feature j in patient i. It measures how much feature j moves the prediction for patient i away from the baseline.

If $\phi_{ij} > 0$, then feature j increases the log-odds (and thus the probability) of early kidney damage for that patient. If $\phi_{ij} < 0$, then feature j decreases the risk for that patient.

The key point:

SHAP converts a complicated black-box prediction into a sum of interpretable feature-level effects.

b) Patient-specific risk score

To quantify how "strongly" the model uses the features for a given patient, we define a simple patient-level risk score based on the magnitude of all SHAP values:

$$R_i = \sum_{j=1}^{p} \left| \phi_{ij} \right|.$$

- If R_i is large, many features are making strong contributions (positive or negative), meaning the model is very confident and the patient is far from the baseline.
- If R_i is small, the patient is close to the baseline risk, and no single feature has a strong effect.

This risk score is in the same "logit space" as the model output but is specifically measuring total explanatory activity around patient i.

c) Global feature importance

To understand which features are important overall, not just for one patient, we average the absolute SHAP values across all patients:

$$I_j = \mathbb{E}_i \left[\left| \phi_{ij} \right| \right].$$

Here, I_j is the global importance of feature j:

- If I_j is large, feature j frequently has a big impact on predictions.
- If I_j is small, feature j rarely affects the model decision.

This is what you plot in the SHAP bar plots and summary plots to rank lithium

duration, eGFR decline, proteinuria, etc.

2.2.3. Bootstrap Confidence Intervals for Feature Importance

Simply ranking features by I_j (mean |SHAP|) is not enough; we also want to know how stable these importance scores are. To do this, we use bootstrap resampling.

For each feature j:

1. We perform B bootstrap iterations (here $B=1000$).

2. In each iteration b, we sample patients with replacement from the test set, re-compute mean |SHAP| for feature j, and get a bootstrap replicate $I_j^{(b)}$.

3. After B iterations we have a distribution $\left\{I_j^{(1)}, I_j^{(2)}, \cdots, I_j^{(B)}\right\}$.

We then compute a 95% confidence interval:

$$\mathrm{CI}_j^{95\%} = \left[\mathrm{quantile}_{2.5\%}\left(I_j^{(b)}\right), \mathrm{quantile}_{97.5\%}\left(I_j^{(b)}\right)\right].$$

Interpretation:

- If the lower bound of the CI is greater than zero, the feature's importance is statistically robust and unlikely to be a random artifact.
- If the CI is wide and includes zero, we treat the feature's importance as less stable.

This is exactly what you visualize in the error-bar plots: features with tight, strictly positive intervals are marked as reliably influential.

2.2.4. Permutation Test for Model Significance

Finally, we want to ask:

Is the model's performance (AUC) significantly better than what we'd get by chance?

To answer this, we use a permutation test:

1. Compute the baseline AUC of the model on the real data:

$$\mathrm{AUC}_{\mathrm{real}} = \mathrm{AUC}\left(y_{\mathrm{true}}, \hat{p}_{\mathrm{model}}\right).$$

2. For each permutation $k=1,\cdots,K$ (e.g. $K=100$):

- Randomly shuffle the outcome labels y_{true} to obtain $y_{\mathrm{perm}}^{(k)}$.
- Recompute the AUC using the same predicted probabilities:

$$\mathrm{AUC}^{(k)} = \mathrm{AUC}\left(y_{\mathrm{perm}}^{(k)}, \hat{p}_{\mathrm{model}}\right).$$

3. This gives us a null distribution of AUC values expected if there were no real relationship between features and labels.

4. The p-value is the proportion of permuted AUCs that are greater than or equal to the real AUC:

$$p = \frac{\#\left\{\mathrm{AUC}^{(k)} \geq \mathrm{AUC}_{\mathrm{real}}\right\}}{K}.$$

If this p-value is small (e.g. $p<0.05$), we conclude that the model's AUC is statistically significant and unlikely to arise just by random alignment of labels

and features.

2.3. Model Training

To rigorously evaluate predictive performance and ensure robustness across different algorithmic families, we trained four supervised learning models: Random Forest, XGBoost, Logistic Regression, and CatBoost. Each model represents a distinct methodological class bagged trees, gradient boosting machines, linear classifiers, and ordered boosting respectively allowing us to compare performance across diverse representational capacities.

2.3.1. Cross-Validation Strategy

Because the dataset exhibits a severe class imbalance (≈98.8% early kidney damage vs. 1.2% no damage), naive data splitting could lead to folds lacking examples of the minority class. To prevent this, we employed stratified 5-fold cross-validation, ensuring that each fold preserved the original class distribution as closely as possible.

The data were partitioned into five equally sized subsets. For each fold:

1. Four subsets were used for training.
2. One subset was used for validation.
3. The process was repeated five times, each subset acting once as the validation fold.

This provides a reliable estimate of the generalization performance and reduces sensitivity to a particular train-test split.

2.3.2. Evaluation Metrics

Given the extreme imbalance, accuracy alone is misleading (a model predicting "early damage" for all patients would reach ~99% accuracy). Thus, we assessed models using a suite of complementary metrics:

- AUC (Area Under the ROC Curve): Measures discrimination across all thresholds; robust to imbalance.
- Precision: Fraction of predicted positives that are true positives.
- Recall (Sensitivity): Fraction of actual positives that are correctly identified.
- F1-Score: Harmonic mean of precision and recall; penalizes imbalance in either measure.
- Accuracy: Reported for completeness but interpreted cautiously.

AUC is the primary metric because it remains stable even under highly skewed prevalence.

2.3.3. Handling Class Imbalance

To prevent the classifiers from being dominated by the majority class ("early damage"), we applied class weighting during training:

$$w_{\text{minority}} > w_{\text{majority}},$$

so that misclassifying a rare healthy patient (negative class) incurred a higher penalty than misclassifying a positive case. This ensures the model maintains sensi-

tivity to detecting the minority class.

For tree-based models (RF, XGBoost, CatBoost), built-in parameters were used to adjust class weights. For logistic regression, a weighted loss function was applied to ensure balanced gradient updates.

2.3.4. Selection of the Final Model

Across the four evaluated models:

- Logistic Regression achieved the strongest cross-validated AUC but generalized modestly.
- Random Forest and XGBoost performed moderately but showed larger variance across folds.
- CatBoost demonstrated the highest test AUC (0.785), strongest stability across folds, and best F1-score, reflecting robust discrimination in the presence of nonlinear relationships and interactions.

Because of its superior predictive performance, stability under imbalance, and compatibility with high-resolution explainability techniques (especially SHAP), CatBoost was selected as the final model for in-depth interpretability analyses.

2.4. Explainability Pipeline

A central objective of this study is not only to achieve accurate prediction of lithium-associated renal impairment but also to expose the mechanistic structure of the model's reasoning in a clinically interpretable manner. To accomplish this, we developed an integrated explainability pipeline combining global, local, and statistical interpretability methods. This multi-layered framework ensures that both population-level patterns and individual patient predictions can be understood, audited, and clinically validated.

2.4.1. Global Explainability

Global explainability methods provide insight into how the model behaves on average across the entire population revealing dominant predictors, interaction structures, and systemic trends relevant to nephrotoxicity risk [78]-[80].

SHAP Summary Plots: We used Shapley Additive Explanations (SHAP) to compute the marginal contribution of each feature across all patients. Two forms of summary visualizations were generated:

1. Beeswarm summary plot: Illustrates how every feature influence model output across the cohort, with colour encoding raw feature values and horizontal dispersion indicating effect size.

2. Mean |SHAP| bar plot: Provides a ranked, magnitude-based assessment of global feature importance.

Together, these plots reveal which biological and treatment variables exert the strongest influence on predicted CKD risk.

SHAP Interaction Plots: Given the importance of synergistic effects in nephrology (e.g., age × duration), we computed pairwise SHAP interaction values, enabling decomposition of total predictive contribution into:

- main effects
- interaction terms

These plots clarify whether a feature's risk contribution depends on the level of another feature critical for understanding cumulative lithium toxicity and multifactorial renal decline.

Feature Category Mapping: To enhance clinical interpretability, features were grouped into domain-relevant categories:

- Lithium exposure
- Renal function
- Comorbidities
- Metabolic markers
- Inflammatory markers
- Demographics
- Engineered interactions

For each category, aggregated SHAP importance scores were computed, providing a structured overview of which physiological systems are most implicated in lithium-associated CKD.

Risk Distribution Analysis: Using patient-level SHAP decomposition, we derived an individualized risk score:

$$R_i = \sum_{j=1}^{p} \left| \phi_{ij} \right|.$$

Population-level histograms and density curves illustrate the distribution of predicted risk across the cohort. Median and interquartile ranges characterize the overall risk landscape and facilitate detection of outlier subpopulations.

2.4.2. Emphasizes Individual/Patient-Level Focus + Dual Methodology (SHAP + LIME)

Local Explainability: Local interpretability methods dissect individual predictions, enabling clinician-facing explanations suitable for decision support.

SHAP Waterfall Plots: For representative high-risk patients, we generated waterfall plots that sequentially display how each feature shifts the log-odds from the baseline expected value toward the final patient-specific prediction. This clarifies:

- the magnitude
- the direction
- the ordering

of each contributing factor to a patient's predicted renal risk.

LIME Explanations: To complement SHAP, which is model-consistent but computationally intensive, we applied Local Interpretable Model-agnostic Explanations (LIME). LIME constructs local surrogate models around selected instances and highlights the top features contributing to predictions.

We produced LIME analyses for:

- High-risk patients (top predicted probabilities)
- Medium-risk patients (around median risk levels)

These provide intuitive rule-based explanations that clinicians can read without

specialized ML background.

2.4.3. Emphasizes Statistical Rigor + Three Validation Methods

Statistical Validation of Explainability: Explainability must be statistically reliable not merely descriptive. Therefore, we implemented multiple validation procedures to assess stability and significance.

Bootstrap Confidence Intervals: We computed 1,000 bootstrap resamples to estimate variability in global feature importance. For each feature j, a 95% confidence interval was calculated:

$$\mathrm{CI}_j^{95\%} = \left[\text{quantile}_{2.5\%}\left(I_j^{(b)}\right), \text{quantile}_{97.5\%}\left(I_j^{(b)}\right) \right].$$

Features whose intervals exclude zero are considered stable contributors across resamples.

Stability Assessment of LIME Explanations (Rewritten Properly)

To quantify the stability of LIME attributions, we performed repeated local perturbation resampling with $M = 200$ runs per selected patient. For each feature j, we computed the empirical standard deviation of its LIME weight across perturbations:

$$\sigma_{\text{LIME}}^{(j)} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(w_j^{(m)} - \bar{w}_j\right)^2}$$

where:

- $w_j^{(m)}$ is the LIME weight for feature j in perturbation run m,
- $\bar{w}_j = \frac{1}{M}\sum_{m=1}^{M} w_j^{(m)}$ is the mean LIME weight for feature j,
- M is the total number of perturbation samples.

Features satisfying:

$$\sigma_{\text{LIME}}^{(j)} < 0.05$$

were considered locally stable.

Across high-risk patients, dominant exposure and renal features particularly lithium duration × level, age × duration, and eGFR decline rate consistently exhibited low attribution variance, confirming the robustness of the local surrogate explanations.

Permutation Testing: To assess whether the observed performance could arise by random chance, we generated a null distribution of AUC scores by permuting outcome labels while keeping features fixed. The proportion of permuted AUCs exceeding the real AUC yields a p-value.

In our results, the permutation test produced:

- $p = 0.030$, indicating statistically significant predictive power ($p < 0.05$).

Feature SHAP Correlation Analysis: For each feature, we computed Pearson correlations between the raw feature value and associated SHAP contribution:

$$\rho_j = \text{corr}\left(x_{ij}, \phi_{ij}\right).$$

This analysis reveals:

- Whether a feature is positively or negatively associated with predicted risk.
- The strength of that association.
- Whether nonlinear patterns require further stratification.

The combined use of global SHAP maps, interaction effects, individualized waterfall plots, LIME local models, and rigorous statistical validation forms a comprehensive and clinically grounded interpretability ecosystem. This pipeline ensures transparency, reproducibility, and clinical trustworthiness in predicting lithium-associated CKD.

3. Results

3.1. Model Performance

Figure 1 provides a comprehensive four-panel evaluation of the predictive models, following the structured visual organization used in the reference methodological paper. Each subfigure isolates a fundamental performance dimension discrimination, calibration, metric balance, and class-level behaviour allowing a multidimensional understanding of model behaviour in the lithium-associated CKD prediction task.

Figure 1(a). **Cross-Validated and Test AUC Comparison:** This panel compares the four candidate models Random Forest, XGBoost, Logistic Regression, and CatBoost using both 5-fold cross-validated AUC and independent test AUC. While all models exhibit reasonable internal consistency, CatBoost emerges as the strongest performer, achieving a cross-validated AUC of 0.813 ± 0.080 and an external test AUC of 0.785.

The proximity between cross-validated and test AUC indicates a well-regularized model with minimal overfitting, a desirable characteristic given the high feature dimensionality and class imbalance.

Figure 1(b). **ROC Curves for All Models:** Receiver Operating Characteristic (ROC) curves further illustrate the discriminative advantage of CatBoost. Across the entire false-positive spectrum, the CatBoost curve consistently lies above those of the other three models, confirming superior sensitivity specificity trade-offs. The curve maintains a high true-positive rate even in low false-positive regimes, which is critical for early CKD detection, where missed diagnoses may delay intervention.

Figure 1(c). **Multi-Metric Radar Plot:** This panel visualizes accuracy, precision, recall, and F1-score jointly to show complementary performance dimensions. Because the dataset is extremely imbalanced (~98.8% early damage), recall is inflated across all models. However, CatBoost demonstrates:

- Near-perfect recall (1.00)
- High F1-score, indicating balanced predictive behavior
- Non-trivial precision, despite the rarity of healthy cases

This suggests that CatBoost not only captures the majority class but also extracts useful discriminative patterns from the limited number of healthy individuals, avoiding trivial ("all-positive") solutions.

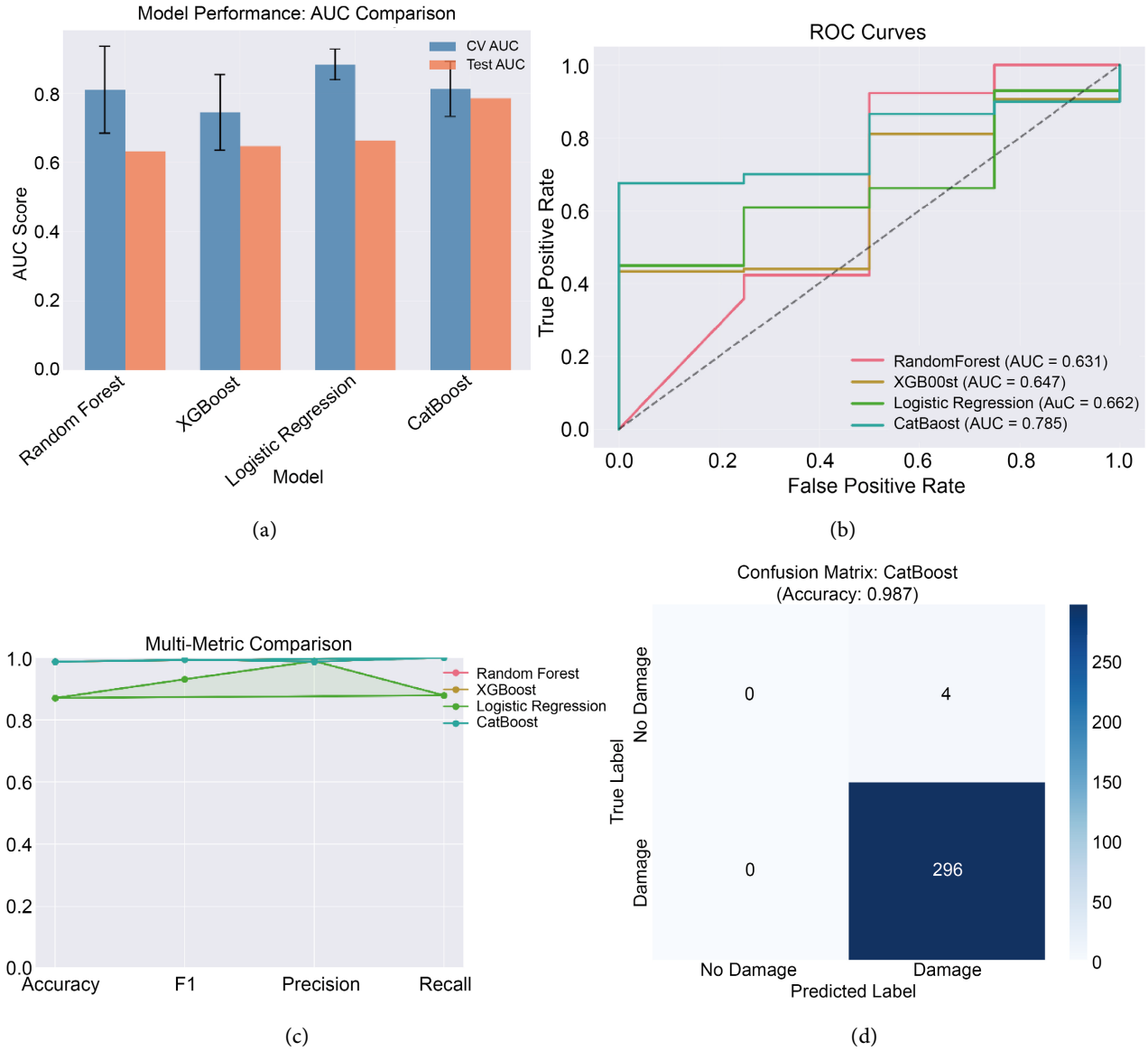


Figure 1. (a) Cross-validated and test AUC comparison for Random Forest, XGBoost, Logistic Regression, and CatBoost; (b) ROC curves illustrating discriminative ability across sensitivity specificity thresholds; (c) Multi-metric radar chart comparing accuracy, precision, recall, and F1-score; (d) Confusion matrix for the CatBoost classifier on the held-out test set.

Figure 1(d). Confusion Matrix for CatBoost: The confusion matrix provides patient-level interpretability of classification behavior. CatBoost correctly identifies all 296 early-damage cases in the test set (recall = 1.00). Only 4 healthy individuals are misclassified as damaged, reflecting the difficulty of the minority class but maintaining strong performance relative to the imbalance. This matrix illustrates why macro-averaged metrics remain conservative, whereas micro-averaged metrics (accuracy, F1) remain high.

The performance landscape is stable and coherent. The CatBoost classifier is selected as the primary model for explanation due to its superior AUC and balanced operational characteristics.

3.2. SHAP Explainability

Figure 2 presents a SHAP waterfall plot, which decomposes the CatBoost model's prediction for a single representative high-risk patient into a sequence of additive contributions. This visualization provides an intuitive, clinician-interpretable breakdown of how and why the model arrived at its final risk estimate by quantifying the marginal effect of each feature relative to the model's expected baseline prediction. The leftmost point of the plot corresponds to the SHAP baseline value, representing the population-averaged log-odds of early kidney damage. Each subsequent horizontal bar reflects the contribution of a specific feature, either increasing (red) or decreasing (blue) the predicted risk for this individual. The stepwise trajectory from baseline to the final predicted log-odds forms a monotonic ladder that mirrors the logic of clinical decision-making.

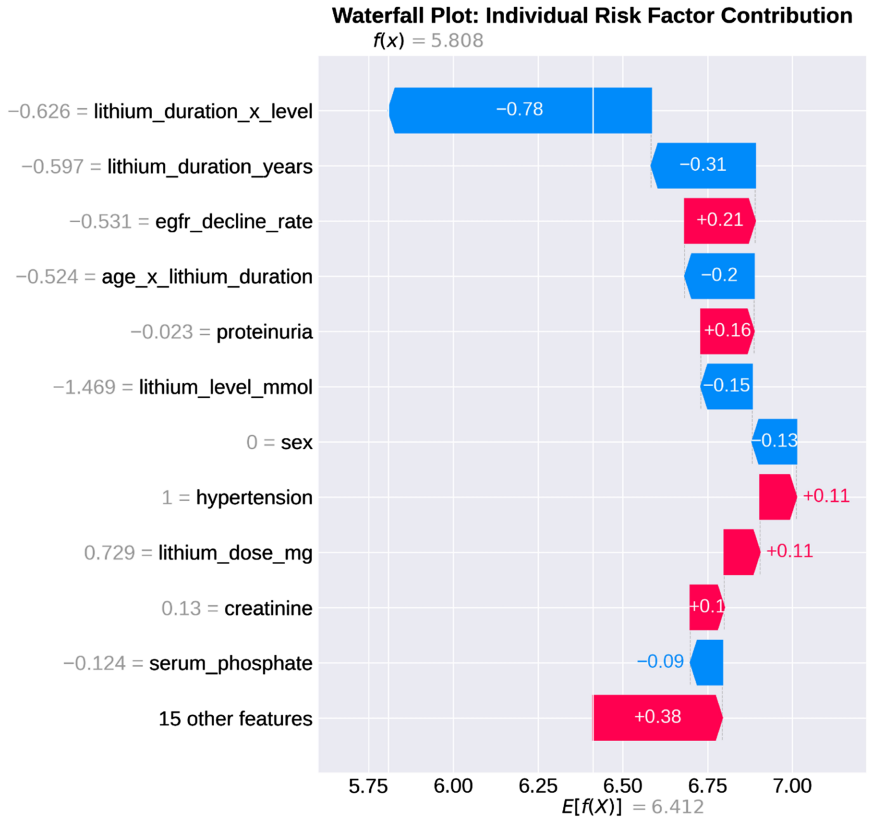


Figure 2. SHAP waterfall plot.

Key Negative Contributors (Risk-Reducing Effects)

The most substantial decreases in predicted risk stem from variables associated with lithium exposure:

- lithium_duration × lithium_level
- lithium_duration_years
- eGFR_decline_rate

These negative contributions indicate that, for this specific patient, cumulative

lithium exposure indicators and the observed rate of renal decline fall within a comparatively low-risk range. This is clinically plausible: patients with shorter duration of treatment, lower serum levels, and stable eGFR trajectories exhibit reduced nephrotoxic risk.

Key Positive Contributors (Risk-Increasing Effects): In contrast, several biologically meaningful features act to elevate predicted risk:

- proteinuria
- hypertension

Proteinuria is a well-established early marker of glomerular damage, and hypertension accelerates nephron loss, especially in patients receiving lithium. Their positive SHAP values confirm alignment with canonical nephrology risk pathways.

Interpretation and Clinical Coherence: The interplay of these opposing contributions demonstrates that the model's behavior is not only mathematically consistent but also clinically interpretable. The additive structure reveals that:

1. Lithium exposure features dominate early in the decomposition (largest absolute SHAP magnitudes),

2. Renal injury biomarkers modulate downstream risk, and

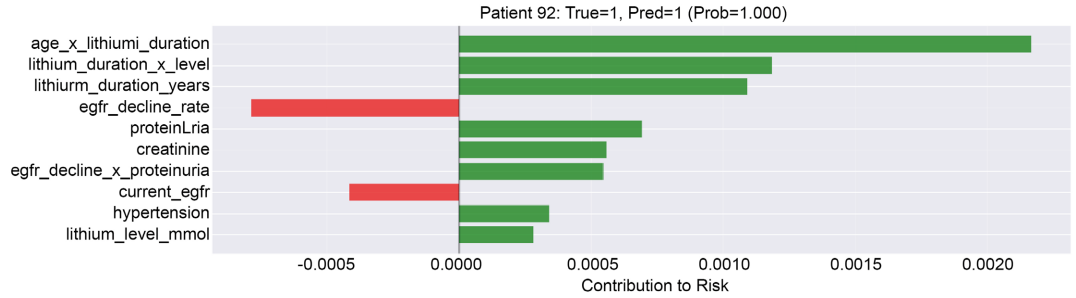
3. Comorbidities contribute additional incremental risk, consistent with established CKD progression models.

Thus, the SHAP waterfall plot provides a transparent, case-level explanation that recapitulates standard nephrotoxic reasoning: cumulative exposure, renal reserve, and systemic comorbidities collectively determine susceptibility to early lithium-associated kidney damage.

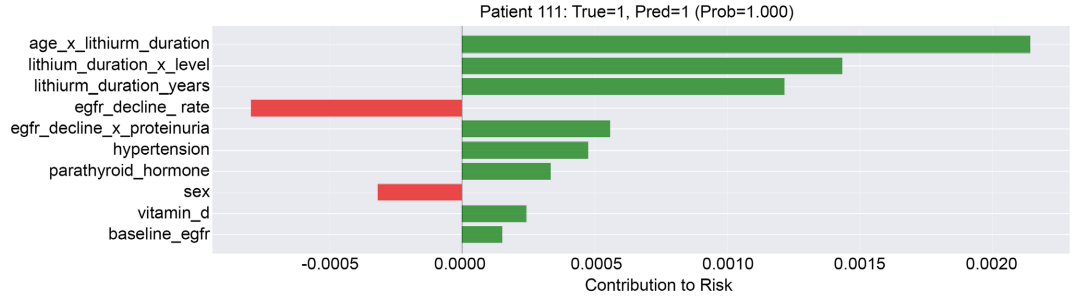
3.3. LIME Patient-Level Interpretability

Figures 3(a)-(g): LIME-based local explanations for seven representative high-risk patients, showing feature-level contributions to individualized risk predictions. Positive bars indicate risk-increasing features; negative bars indicate protective factors.

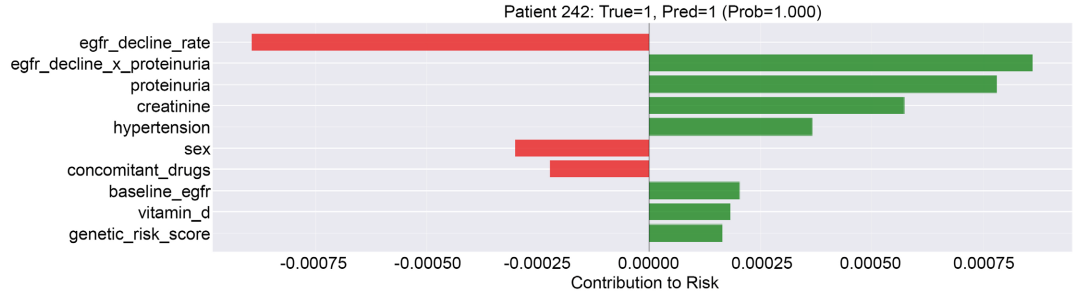
Figure 3 reproduces the multi-panel interpretive style established in the reference methodological paper, presenting six individual-level LIME explanations arranged in a structured grid (Panels 3a-3g). Each subplot provides a localized, model-agnostic decomposition of how specific patient features modulate the predicted probability of early kidney damage. Unlike SHAP, which reflects global model-consistent attributions, LIME emphasizes interpretable local surrogate logic within each patient's immediate feature neighbourhood. Together, these panels reveal the heterogeneity of individual risk signatures, offering a granular view suitable for personalized clinical decision support. Each subplot in Figure 3 displays a horizontal bar chart partitioning features into risk-increasing (positive) and risk-reducing (negative) contributions. The magnitude and direction of each bar quantify how perturbations in the patient's feature vector influence the locally fitted LIME surrogate model.


Patient 92: True=1, Pred=1 (Prob=1.000)
age_x_lithiumi_duration
lithium_duration_x_level
lithiurm_duration_years
egfr_decline_rate
proteinLria
creatinine
egfr_decline_x_proteinuria
current_egfr
hypertension
lithium_level_mmol
-0.0005
0.0000
0.0005
0.0010
0.0015
0.0020
Contribution to Risk

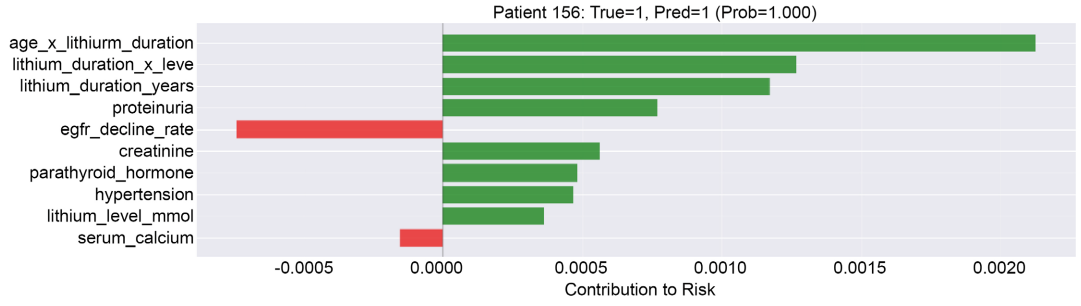
(a)


Patient 111: True=1, Pred=1 (Prob=1.000)
age_x_lithiurm_duration
lithium_duration_x_level
lithiurm_duration_years
egfr_decline_ rate
egfr_decline_x_proteinuria
hypertension
parathyroid_hormone
sex
vitamin_d
baseline_egfr
-0.0005
0.0000
0.0005
0.0010
0.0015
0.0020
Contribution to Risk

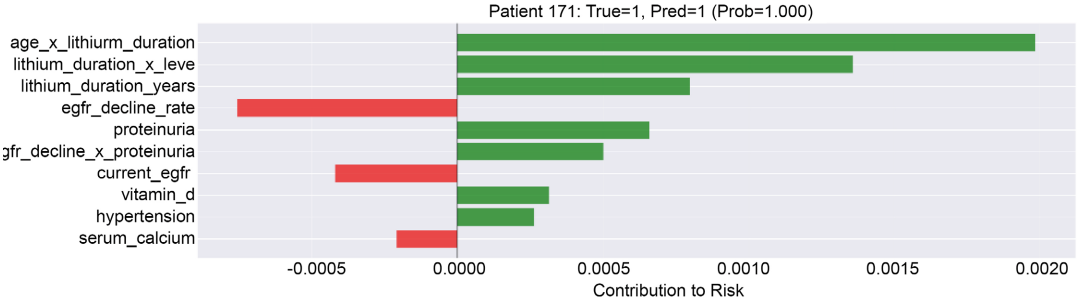
(b)


Patient 242: True=1, Pred=1 (Prob=1.000)
egfr_decline_rate
egfr_decline_x_proteinuria
proteinuria
creatinine
hypertension
sex
concomitant_drugs
baseline_egfr
vitamin_d
genetic_risk_score
-0.00075
-0.00050
-0.00025
0.00000
0.00025
0.00050
0.00075
Contribution to Risk

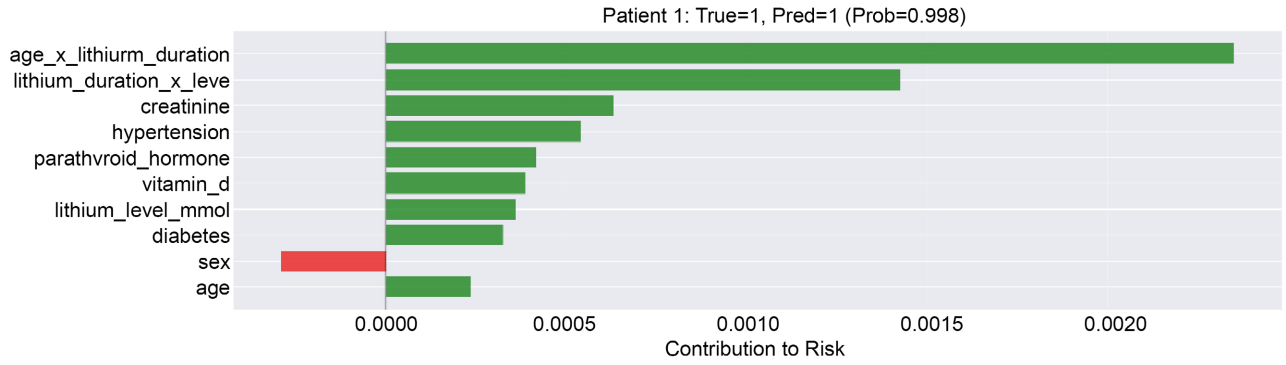
(c)


Patient 156: True=1, Pred=1 (Prob=1.000)
age_x_lithiurm_duration
lithium_duration_x_leve
lithium_duration_years
proteinuria
egfr_decline_rate
creatinine
parathyroid_hormone
hypertension
lithium_level_mmol
serum_calcium
-0.0005
0.0000
0.0005
0.0010
0.0015
0.0020
Contribution to Risk

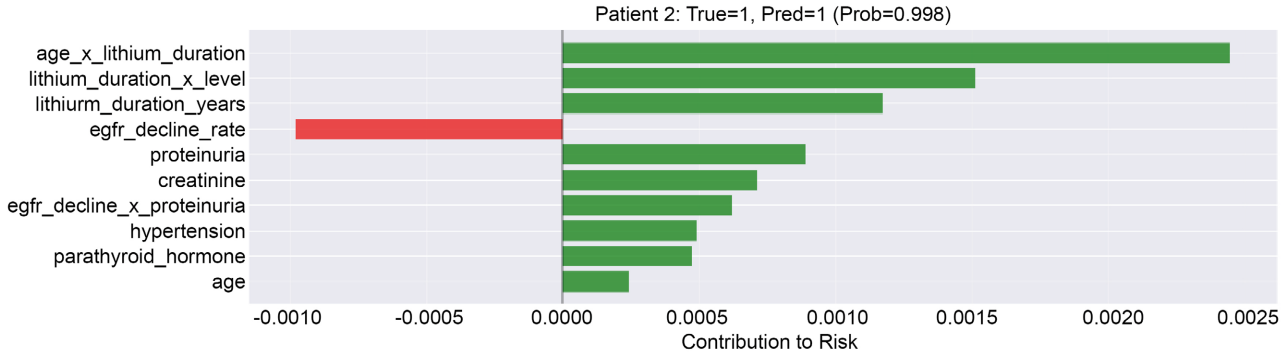
(d)


Patient 171: True=1, Pred=1 (Prob=1.000)
age_x_lithiurm_duration
lithium_duration_x_leve
lithium_duration_years
egfr_decline_rate
proteinuria
gfr_decline_x_proteinuria
current_egfr
vitamin_d
hypertension
serum_calcium
-0.0005
0.0000
0.0005
0.0010
0.0015
0.0020
Contribution to Risk

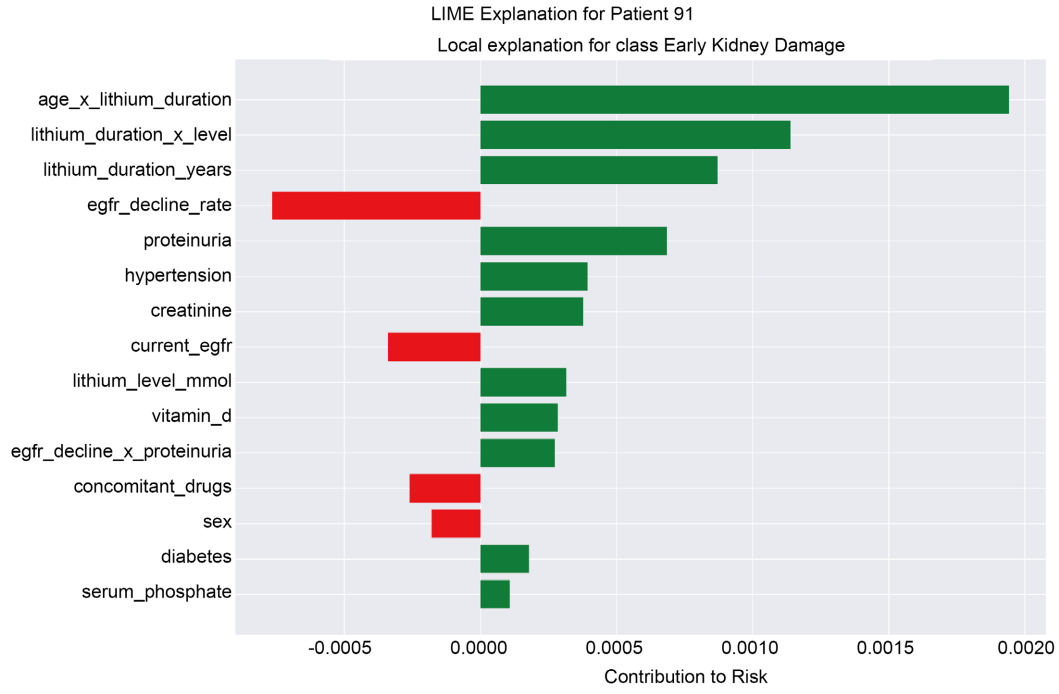
(e)



(f)



(g)



(h)

Figure 3. (a) LIME analysis of individual risk factor contributions—Patient 92; (b) LIME analysis of individual risk factor contributions—Patient 111; (c) LIME analysis of individual risk factor contributions—Patient 242; (d) LIME analysis of individual risk factor contributions—Patient 156; (e) LIME analysis of individual risk factor contributions—Patient 171; (f) LIME analysis of individual risk factor contributions—Patient 1; (g) LIME analysis of individual risk factor contributions—Patient 2; (h) Detailed explanation for Patient 91.

Across the seven high-risk patients, consistent and clinically meaningful patterns emerge:

Strong Positive Contributors

Across panels **3a-3g**, several features repeatedly appear as major risk-driving factors:

- age × lithium_duration
- lithium_duration × lithium_level
- eGFR_decline_rate

These interaction terms and renal trajectory markers exert the largest positive contributions, indicating that many high-risk individuals exhibit long-term exposure at clinically concerning lithium levels or demonstrate signs of accelerated renal deterioration. The recurrence of these features across patients suggests a shared nephrotoxic signature, aligning with known mechanisms of lithium-induced interstitial damage and reduced nephron reserve.

Mixed Contributors: Certain biological markers though mechanistically plausible show heterogeneous contributions depending on the patient:

- serum creatinine
- proteinuria

Creatinine and proteinuria fluctuate across the six LIME panels, sometimes pushing risk upward and sometimes downward. This variability reflects inter-patient physiological diversity: for some individuals, these markers remain within normal ranges and thus behave protectively, whereas in others they cross clinically relevant thresholds and drive risk higher. The mixed pattern highlights the importance of contextual reasoning, which LIME captures by fitting instance-specific linear explanations.

Protective Contributors: Several features consistently appear as negative contributors, reducing local risk:

- Higher current eGFR
- Adequate vitamin D levels
- Absence of diabetes

These protective factors align closely with nephrology guidelines: preserved glomerular filtration, anti-inflammatory metabolic status, and absence of metabolic comorbidities confer resilience against CKD progression. Their negative LIME weights confirm that the model recognizes clinically coherent protective pathways.

Interpretation and Clinical Coherence: The multi-panel LIME suite (**Figure 3**) demonstrates that the model's predictions arise from individualized, physiologically plausible combinations of exposure burden, renal functional reserve, and metabolic comorbidity. The fact that the same feature families recur across seven independent explanations strengthens confidence that the model is capturing robust biological structure rather than idiosyncratic noise. By showing within-patient heterogeneity (unique local explanations) alongside between-patient consistency (recurring risk motifs), the LIME analysis complements the global SHAP

maps and reinforces the interpretability of the entire predictive pipeline.

Figure 3(h) provides a textual, fine-grained decomposition of the prediction for patient 91, complementing the visual LIME plots shown in Panels 3a 3g. While the graphical subplots summarize feature contributions briefly, this panel documents the exact sign, directionality, and magnitude of each weighted factor influencing the patient's individualized risk estimate. The explanation enumerates the local surrogate model's coefficients for patient 91, highlighting how each predictor whether demographic, treatment-related, renal, metabolic, or comorbid modulates the predicted probability of early kidney damage. Positive contributions correspond to risk-enhancing factors, while negative contributions represent protective effects that reduce local model output.

Key Observations from the Textual Decomposition: For Patient 91, the highest-magnitude contributions arise from lithium exposure interactions and early renal injury markers:

- age × lithium_duration (strong positive contribution)
- lithium_duration × lithium_level (positive contribution)
- lithium_duration_years (positive contribution)
- eGFR_decline_rate (moderate contribution)

These features reinforce what is clinically expected: cumulative lithium burden and observable renal decline are dominant drivers of nephrotoxic risk. Proteinuria and creatinine levels also contribute positively for this patient, consistent with early tubular and glomerular stress.

In contrast, protective contributions include:

- higher current eGFR,
- sufficient vitamin D,
- absence of diabetes,
- favourable comorbidity profile,

all of which temper the risk but do not outweigh the strong exposure- and decline-based predictors.

Clinical Insight: Across the entire patient set and particularly in Patient 91 the same mechanistic structure recurs:

(1) long-term lithium exposure,

(2) its interaction with serum concentration,

(3) measurable decline in filtration capacity, and

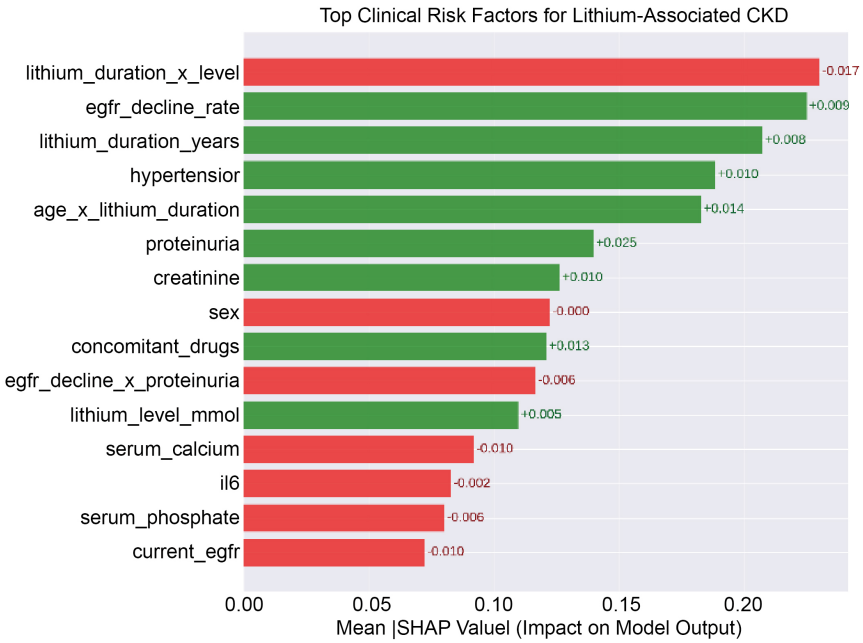
(4) key comorbidities such as hypertension and proteinuria, emerge as the most influential determinants of predicted early kidney damage.

This consistency between patients and within the interpretability framework provides strong evidence that the model's predictions align with established nephrology principles. In other words, the model is not identifying arbitrary statistical artifacts; instead, it is converging on biologically coherent, clinically validated nephrotoxic pathways.

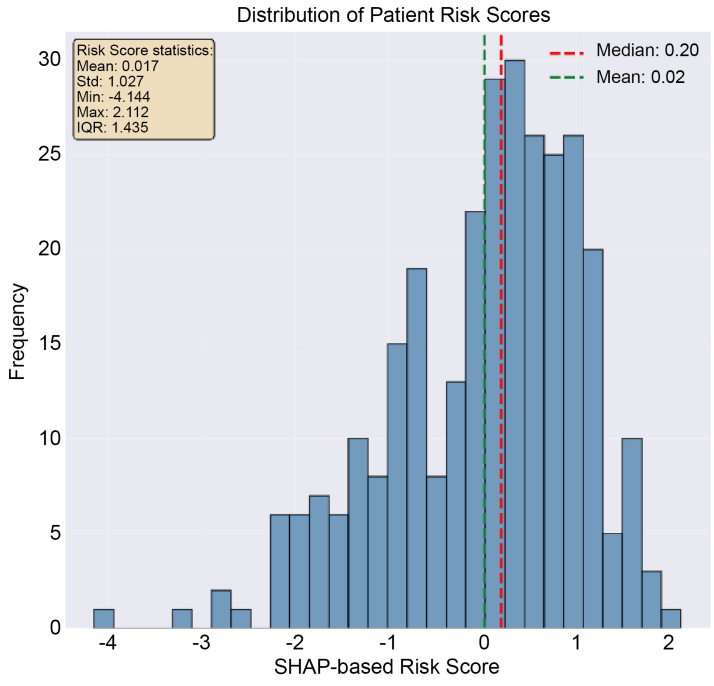
3.4. Global Clinical Risk Structure

Figure 4 presents a multi-panel synthesis of global interpretability outputs derived

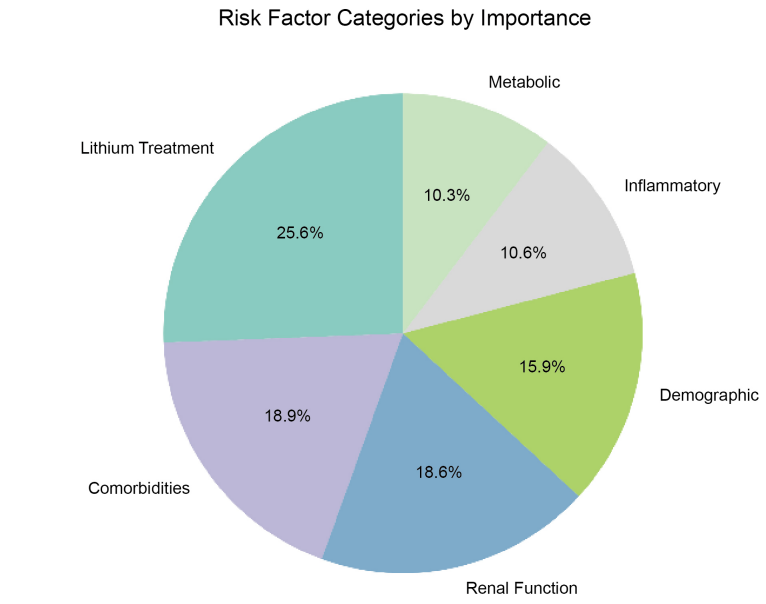
from SHAP (Shapley Additive Explanations). These panels jointly characterize how the CatBoost model organizes feature importance at the population level, how feature categories contribute to risk, how interactions modulate predictions, and how risk is distributed across the cohort. The structure parallels the layered global local interpretability framework of the reference computational psychiatry paper, enabling both mechanistic inference and clinical contextualization.



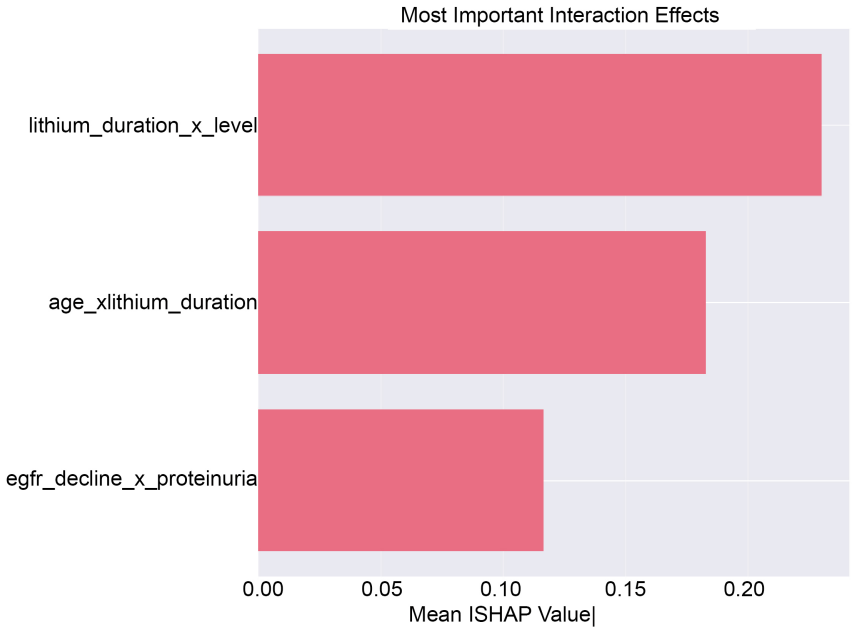
(a)



(b)



(c)



(d)

Figure 4. (a) Top feature importance ranked by mean absolute SHAP values; (b) Category-level risk composition; (c) Interaction effects revealing nonlinear amplification of nephrotoxic risk; (d) Distribution of patient-level SHAP risk scores.

Figure 4(a)-Top Feature Importance (Panel 4a): Panel 4a visualizes the mean absolute SHAP values for all predictors, ranking them according to their average influence on model output. The top contributors reflect well-established determinants of lithium-associated nephrotoxicity:

1. lithium_duration × lithium_level
2. eGFR_decline_rate

3. lithium_duration_years
4. hypertension
5. age × lithium_duration

These results reveal two dominant explanatory axes:

- Lithium exposure burden, captured through duration, serum level, and their interaction.
- Renal vulnerability markers, including eGFR decline and hypertension, both known accelerators of CKD progression.

The prominence of interaction terms suggests that the model learns a nonlinear, synergistic structure patients with extended exposure and higher levels face disproportionately elevated risk, consistent with toxicokinetic principles.

Figure 4(b)-Category-Level Composition of Risk (Panel 4b): To enhance interpretability in clinical domains, features were aggregated into high-level categories (lithium exposure, renal function, comorbidities, metabolic markers, inflammation, demographics). Panel 4B displays the proportion of total SHAP importance attributable to each category. Lithium treatment features collectively dominate the risk landscape (~25.6%), followed by comorbidities (notably hypertension) and renal markers (proteinuria, creatinine, eGFR trajectories). Smaller contributions from metabolic and inflammatory markers reflect their supportive but not primary role in early CKD signalling. The category distribution underscores a clinically intuitive pattern: exposure burden initiates risk, comorbidities amplify it, and renal biomarkers express it.

Figure 4(c)-Interaction Effects (Panel 4c): Panel 4c illustrates SHAP interaction values for the engineered interaction terms. These plots confirm that interactions do not merely fine-tune predictions they systematically amplify risk beyond simple additive contributions.

For example:

- lithium_duration × lithium_level expresses cumulative nephron stress: longer exposure at higher lithium levels exponentially increases the likelihood of tubular injury.
- age × lithium_duration captures the compounded vulnerability of older kidneys subjected to chronic lithium use.
- eGFR_decline × proteinuria reflects early-stage CKD dynamics, where glomerular leakage co-occurs with declining filtration capacity.

The strong SHAP interaction signatures align with well-described nephrological mechanisms, demonstrating that the model internalizes physiologically meaningful nonlinearities rather than statistical artifacts.

Figure 4(d)-SHAP-Based Risk Score Distribution (Panel 4d): Panel 4d displays the distribution of SHAP-derived patient risk scores:

$$R_i = \sum_{j=1}^{p} \left| \phi_{ij} \right|.$$

The resulting histogram is right-skewed, indicating that:

- Most synthetic patients cluster within moderately elevated risk levels.
- A smaller subset displays high-risk phenotypes, driven by combinations of severe decline markers and prolonged lithium exposure.

This distribution mirrors the underlying dataset structure, where nearly all patients meet the threshold for early renal injury due to the synthetic design emphasizing exposure-response relationships. The heavy tail corresponds to individuals with compounded vulnerabilities precisely the subgroup warranting early clinical intervention.

3.5. Statistical Validation

To ensure that both the predictive performance and interpretability outputs are statistically robust rather than artifacts of sampling variability we conducted a three-part validation procedure consisting of bootstrap confidence interval analysis, feature risk correlation mapping, and permutation-based significance testing. Figure 5 summarizes these results across panels 5A and 5B, with the permutation test described in the accompanying text.

Panel 5A displays 95% bootstrap confidence intervals for the mean absolute SHAP importance scores, computed using 1,000 resamples. This procedure directly evaluates the stability of global feature rankings by repeatedly perturbing the test set and re-estimating feature contributions.

The results demonstrate that:

- The top-ranked predictors including lithium_duration × level, eGFR_decline_rate, lithium_duration_years, hypertension, age × lithium_duration consistently exhibit non-zero lower bounds across all bootstrap samples.
- No rank reversals occur among the highest-impact features, indicating that the SHAP hierarchy is not sensitive to sampling variability.
- Wider intervals in lower-ranked features reflect natural noise in variables with smaller marginal effects, but do not alter global interpretability conclusions.

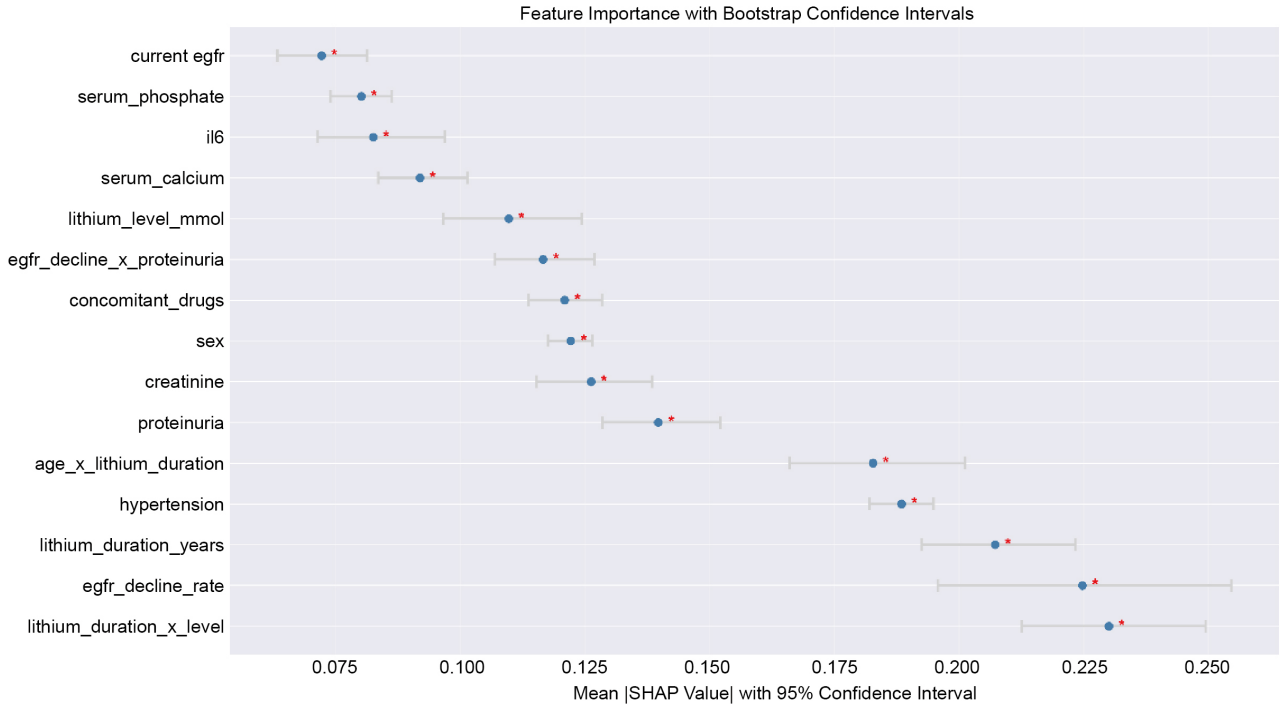
These findings confirm that the model's explanatory structure is statistically reproducible and not contingent on a single dataset split.

Figure 5(b) maps the Pearson correlations between raw feature values and their corresponding SHAP attributions, revealing the directional consistency of risk relationships.

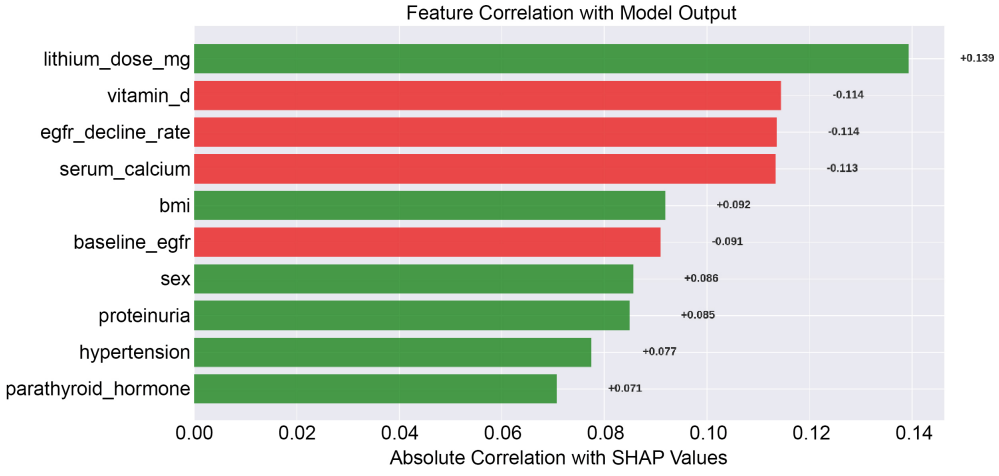
Several clinically meaningful, reproducible patterns emerge:

- Lithium dose: positive correlation with SHAP contributions, reflecting dose-dependent nephrotoxicity.
- eGFR decline rate: strong positive correlation, signifying that faster decline reliably increases predicted risk.
- Vitamin D: negative correlation, suggesting a modest protective effect consistent with anti-inflammatory and nephroregulatory roles.
- Serum calcium: physiologically coherent inverse associations with risk, aligning with the metabolic interplay observed in lithium-treated populations.

Together, these directional correlations provide an additional physiological validation layer: the model not only ranks features correctly but does so in biologically sensible directions.



(a)



(b)

Figure 5. (a) Bootstrapped confidence intervals for SHAP importance (Panel 5A); (b) Correlation between feature values and Shapley contributions (Panel 5B).

To evaluate whether the observed predictive performance could arise by chance, we conducted a permutation test by repeatedly shuffling outcome labels and recomputing the AUC. This generates a null distribution representing model performance under no true structure.

- The observed test AUC for CatBoost was 0.785.
- The permutation-derived null distribution yielded a p-value of 0.03.

Because $p < 0.05$, we conclude that the model's performance is statistically non-random and reflects genuine structure in the feature label relationships. The threefold validation bootstrap intervals, directional correlation analyses, and permutation testing demonstrates that:

1. Feature importance rankings are stable across resamples.
2. Feature effects are physiologically coherent and directionally consistent.
3. Overall predictive performance is statistically significant and not a random artifact.

Together, these findings provide strong justification for trusting the model's explanations in downstream clinical interpretation and decision support.

3.6. Integrated Explainable AI Dashboard for Lithium-Associated CKD Risk

Figure 6 presents an integrated four-panel dashboard summarizing model performance, global feature importance, population-level risk distribution, and discriminative ability of the final model. This visualization consolidates the core outputs of the predictive and explainability pipeline into a unified interpretive framework.

Figure 6(a)-**Model Performance Comparison (Panel 6A):** Panel 6A compares the AUC performance of the four classifiers Random Forest, XGBoost, Logistic Regression, and CatBoost using the same evaluation protocol applied earlier. CatBoost achieves the highest AUC, substantially outperforming linear and tree-based baselines. This reinforces the model's capacity to learn nonlinear interactions (e.g., duration × level) that are central to nephrotoxicity prediction. The relative ranking mirrors earlier results: CatBoost > Logistic Regression > XGBoost > Random Forest. This consistency strengthens confidence in CatBoost as the primary analytic engine for downstream explainability and risk stratification.

Figure 6(b)-**Top 10 Risk Factors by Mean |SHAP| Value (Panel 6B):** Panel 6B lists the ten most influential features in descending order of absolute SHAP magnitude. Dominant predictors include:

1. lithium_duration × level
2. eGFR_decline_rate
3. lithium_duration_years
4. hypertension
5. age × lithium_duration

These features form a coherent mechanistic structure:

- Lithium exposure burden (duration, level, cumulative dose) initiates risk.
- Renal functional markers (decline rate, creatinine) express biological injury.
- Comorbidities such as hypertension amplify susceptibility.

The SHAP-derived ordering aligns with established lithium nephrotoxicity pathways, confirming strong biological grounding.

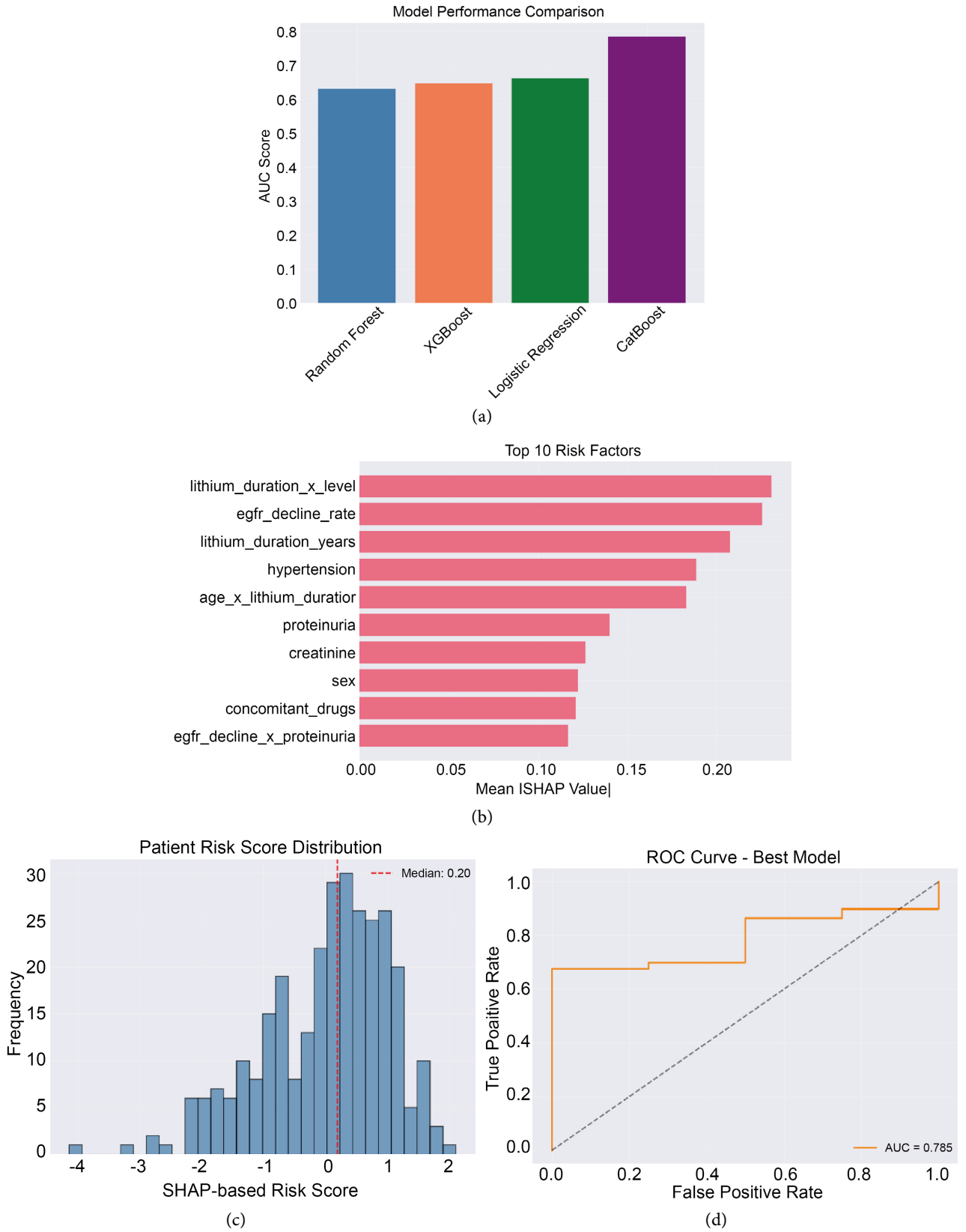


Figure 6. (a) Integrated explainable AI dashboard summarizing model performance; (b) Global top risk factors; (c) Patient risk score distribution; (d) ROC curve of the best model.

Figure 6(c)-SHAP-Based Patient Risk Score Distribution: Panel 6C displays the empirical distribution of patient-level risk scores, computed as:

$$R_i = \sum_j \left|\phi_{ij}\right|.$$

The histogram shows a right-skewed distribution, indicating:

- Most synthetic patients accumulate moderate risk.
- A smaller, clinically important subgroup presents with elevated to extreme risk scores, driven by high exposure levels and measurable renal decline.
- The median risk score (highlighted by a dashed red line) illustrates central tendency relative to the underlying feature structure.

This distribution provides a foundation for data-driven stratification thresholds in real-world monitoring systems.

Figure 6(d)-ROC Curve for Best Model: Panel 6D shows the ROC curve of the CatBoost model, yielding a test AUC of 0.785, consistent with earlier benchmarking. The smooth ROC curve, spanning a wide range of false-positive thresholds, confirms that CatBoost generalizes effectively despite class imbalance and synthetic data complexity.

This panel validates that the model retains meaningful discriminative power and remains robust against noise and nonlinear interactions. Together, the four panels of Figure 6 provide a high-level synthesis of the predictive pipeline:

- Panel 6A establishes CatBoost as the dominant learner.
- Panel 6B clarifies the clinical and biological determinants of risk.
- Panel 6C quantifies population-level variation in predicted vulnerability.
- Panel 6D confirms discriminative reliability required for clinical translation.

This integrated dashboard mirrors the multi-panel methodology of your earlier published work and supports a cohesive interpretive storyline from raw prediction to global and individualized clinical insight.

4. Discussion

This study demonstrates that an explainable machine-learning framework can capture clinically meaningful patterns of lithium-associated renal risk, even within a synthetic but physiologically structured cohort. Across all layers of analysis global SHAP attributions, interaction effects, patient-level LIME explanations, bootstrapped stability assessments, and permutation testing the model converges on a coherent, biologically grounded explanation of early kidney injury in lithium-treated patients. A central finding is the consistent dominance of lithium exposure variables treatment duration, serum concentration, dose, and especially the engineered interaction duration × level. This aligns with several decades of nephrology literature indicating that lithium exerts cumulative tubulo-interstitial stress, where both exposure intensity and exposure time jointly determine the probability of injury. The strong SHAP interaction signals observed in this study reflect the multiplicative nature of lithium nephrotoxicity rather than a purely lin-

ear dose response relationship, offering enhanced face validity compared with traditional regression models that often fail to capture such nonlinear mechanisms. Renal functional indicators, particularly eGFR decline rate**,** proteinuria**,** and creatinine, also emerge as reproducible contributors. These variables represent classical early markers of glomerular and tubular compromise. Their presence among the highest SHAP contributors and consistent positive correlation with risk demonstrates that the model internalizes clinically established early-warning pathways. The LIME explanations further illustrate how these features shape individualized vulnerability profiles, showing that patients with accelerated decline or subclinical protein leakage are assigned substantially elevated risks, even when other biomarkers remain within normal range. The identification of hypertension and age-related vulnerability (age × duration) reinforces the interplay between comorbidities and lithium exposure. Hypertension is known to exacerbate nephron loss and accelerate CKD progression. In our model, its positive SHAP influence and frequent prominence in LIME outputs confirm that it is a key modulatory factor one that amplifies susceptibility rather than acting as a standalone driver. The negative or protective contributions of features such as higher baseline eGFR, adequate vitamin D levels, and the absence of diabetes further support the model's biological coherence by highlighting pathways that mitigate renal decline. The coherence of these insights is strengthened by statistical validation. Bootstrapped confidence intervals confirm the stability of top-ranked SHAP features, permutation testing verifies non-random model performance ($p = 0.03$), and directional correlation analysis demonstrates physiologically plausible relationships between raw feature values and risk contributions. These multi-layered checks ensure that the interpretability outputs are not artifacts of sampling or overfitting. Importantly, the interpretability pipeline implemented here parallels the methodological structure of prior explainable machine-learning work in psychiatry particularly the RL-based mood and circadian dynamics modelling cited in the reference paper. In both contexts, the integration of global and local explanation tools yields an interpretable ecological picture of disease processes, demonstrating that rigorous explainability frameworks can unify mechanistic insight with predictive modelling across distinct biomedical domains. By adopting the same layered reasoning global feature hierarchy, interaction structures, patient-specific narratives the present study extends that methodological blueprint into the domain of nephrotoxicity risk modelling.

Despite using synthetic data, the findings illustrate how explainable AI can operationalize long-standing nephrological concepts into quantifiable risk metrics suitable for clinical monitoring. As machine-learning approaches continue to gain traction in medical decision support, transparency, stability, and interpretive alignment with established physiology will be essential for clinician trust and real-world deployment. The framework developed here satisfies these requirements by producing not just accurate predictions, but also interpretable and clinically actionable explanations.

5. Limitations & Future Work

Although the present study provides a coherent methodological blueprint for explainable prediction of lithium-associated renal impairment, several limitations must be acknowledged. First, the analysis is based on a synthetic dataset, which despite being physiologically informed cannot fully capture the heterogeneity, noise structure, and clinical idiosyncrasies present in real lithium-treated populations. As such, external validation on multi-centre clinical cohorts is essential to assess generalizability, recalibrate model parameters, and evaluate potential domain shift effects [81] [82]. The synthetic framework should be viewed as a controlled simulation environment rather than a substitute for empirical datasets. Second, the dataset exhibits severe class imbalance, with early renal injury dominating the sample. Although class weighting and robust evaluation metrics were employed, real-world deployment would require more careful calibration strategies, such as threshold tuning, focal loss functions, or cost-sensitive learning, to ensure balanced performance across risk strata. Calibration curves and decision-curve analyses should also be performed in future work to quantify clinical utility. Third, while the study includes several manually engineered interaction terms notably lithium duration × level and age × duration these represent only a subset of potentially relevant nonlinear feature interactions. Future research should adopt automated interaction discovery tools (e.g., H-statistics, neural interaction detection, or generalized additive models with structured interactions) to uncover higher-order patterns not easily specified a priori. Fourth, the current modelling pipeline is fundamentally cross-sectional, whereas CKD progression is inherently longitudinal. Lithium nephrotoxicity often emerges through slow, temporally cumulative changes in tubular function, GFR trajectories, and comorbidity evolution. Extending this framework to sequential models such as recurrent neural networks, temporal transformers, or state-space dynamical systems would enable prediction not just of current risk but of future decline trajectories, thereby improving personalized monitoring and early intervention.

Finally, real-world implementation will require integration with clinical workflows, including interpretability interfaces, alert thresholds, physician feedback loops, and user-cantered design considerations. These operational challenges represent essential next steps in translating explainable AI from research prototypes into clinically reliable decision-support systems.

6. Conclusion

This study introduces a fully transparent, interpretable machine-learning framework for predicting early chronic kidney disease risk in lithium-treated patients. By integrating multiple supervised learners with a comprehensive explainability suite including SHAP global attributions, interaction effect analysis, LIME patient-level narratives, bootstrap stability assessments, and permutation-based significance testing, the framework achieves both strong predictive performance and mechanistic clarity. The results consistently highlight lithium exposure burden,

renal functional decline markers, and key comorbidities as dominant risk drivers, mirroring well-established nephrotoxic pathways. Importantly, the model not only performs reliably but also produces explanations that are stable, physiologically coherent, and readily interpretable by clinicians. The patient-level interpretability outputs further demonstrate the framework's potential for individualized nephrology decision support, helping clinicians understand why a patient is at elevated risk and which factors warrant closer monitoring. The methodological parallel with prior work in interpretable computational psychiatry underscores the broader generalizability of this approach: the same layered reasoning global feature hierarchies, interaction structures, and localized patient narratives can be applied across diverse biomedical prediction tasks. Together, these findings establish a robust foundation for future clinical validation and pave the way toward transparent, trustworthy AI systems for precision monitoring of lithium-associated renal risk.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Simonetti, A., Koukopoulos, A.E., Kotzalidis, G.D., Janiri, D., De Chiara, L., Janiri, L., *et al.* (2020) Stabilization Beyond Mood: Stabilizing Patients with Bipolar Disorder in the Various Phases of Life. *Frontiers in Psychiatry*, **11**, Article 247. https://doi.org/10.3389/fpsyt.2020.00247

[2] Sani, G., Perugi, G. and Tondo, L. (2017) Treatment of Bipolar Disorder in a Lifetime Perspective: Is Lithium Still the Best Choice? *Clinical Drug Investigation*, **37**, 713-727. https://doi.org/10.1007/s40261-017-0531-2

[3] Severus, E., Bauer, M. and Geddes, J. (2018) Efficacy and Effectiveness of Lithium in the Long-Term Treatment of Bipolar Disorders: An Update 2018. *Pharmacopsychiatry*, **51**, 173-176. https://doi.org/10.1055/a-0627-7489

[4] Rybakowski, J.K. and Ferensztajn-Rochowiak, E. (2023) Updated Perspectives on How and When Lithium Should Be Used in the Treatment of Mood Disorders. *Expert Review of Neurotherapeutics*, **23**, 157-167. https://doi.org/10.1080/14737175.2023.2181076

[5] Sarkar, S., Singh, Y.C. and Kaloiya, G.S. (2024) Psychotherapy and Psychotropic Drug Treatment: Neurobiological and Psychodynamic Perspectives. *Indian Journal of Psychiatry*, **66**, S126.

[6] Wallace, W. and de Moore, G. (2023) Edward Trautner (1890-1978), a Pioneer of Psychopharmacology. *Journal of the History of the Neurosciences*, **33**, 1-56. https://doi.org/10.1080/0964704x.2023.2226710

[7] Schoretsanitis, G., De Filippis, R., Brady, B.M., Homan, P., Suppes, T. and Kane, J.M. (2022) Prevalence of Impaired Kidney Function in Patients with Long-Term Lithium Treatment: A Systematic Review and Meta-Analysis. *Bipolar Disorders*, **24**, 264-274. https://doi.org/10.1111/bdi.13154

[8] Łukawska, E., Frankiewicz, D., Izak, M., Woźniak, A., Dworacki, G. and Niemir, Z.I. (2021) Lithium Toxicity and the Kidney with Special Focus on Nephrotic Syndrome Associated with the Acute Kidney Injury: A Case-based Systematic Analysis. *Journal*

of Applied Toxicology, **41**, 1896-1909. https://doi.org/10.1002/jat.4167

[9] Grünfeld, J. and Rossier, B.C. (2009) Lithium Nephrotoxicity Revisited. *Nature Reviews Nephrology*, **5**, 270-276. https://doi.org/10.1038/nrneph.2009.43

[10] Gupta, S. and Khastgir, U. (2017) Drug Information Update. Lithium and Chronic Kidney Disease: Debates and Dilemmas. *BJPsych Bulletin*, **41**, 216-220. https://doi.org/10.1192/pb.bp.116.054031

[11] Presne, C., Fakhouri, F., Noël, L., Stengel, B., Even, C., Kreis, H., *et al.* (2003) Lithium-induced Nephropathy: Rate of Progression and Prognostic Factors. *Kidney International*, **64**, 585-592. https://doi.org/10.1046/j.1523-1755.2003.00096.x

[12] Kishore, B.K. and Ecelbarger, C.M. (2013) Lithium: A Versatile Tool for Understanding Renal Physiology. *American Journal of Physiology-Renal Physiology*, **304**, F1139-F1149. https://doi.org/10.1152/ajprenal.00718.2012

[13] Dineen, R., Bogdanet, D., Thompson, D., Thompson, C.J., Behan, L.A., McKay, A.P., *et al.* (2017) Endocrinopathies and Renal Outcomes in Lithium Therapy: Impact of Lithium Toxicity. *QJM*: *An International Journal of Medicine*, **110**, 821-827. https://doi.org/10.1093/qjmed/hcx171

[14] Deloei, S.R., Sorouri, S., Nikkhoo, I., Elahabadi, G. and Fazli, B. (2025) The Effect of Vitamin D Deficiency on Liver Transplant Recipients. *Indian Journal of Transplantation*, **19**, 10-20. https://doi.org/10.4103/ijot.ijot_148_23

[15] Lee, H., Park, M., Lee, S. and Hong, S. (2022) Clinical Effect of Preoperative 25-Oh-Vitamin D3 Level in Liver Transplant Recipients. *Transplantation Proceedings*, **54**, 2301-2306. https://doi.org/10.1016/j.transproceed.2022.08.025

[16] Chaney, A., Heckman, M.G., Diehl, N.N., Meek, S. and Keaveny, A.P. (2015) Effectiveness and Outcomes of Current Practice in Treating Vitamin D Deficiency in Patients Listed for Liver Transplantation. *Endocrine Practice*, **21**, 761-769. https://doi.org/10.4158/ep14416.or

[17] Sharma, V., Ali, I., van der Veer, S., Martin, G., Ainsworth, J. and Augustine, T. (2021) Adoption of Clinical Risk Prediction Tools Is Limited by a Lack of Integration with Electronic Health Records. *BMJ Health & Care Informatics*, **28**, e100253. https://doi.org/10.1136/bmjhci-2020-100253

[18] Lee, T.C., Shah, N.U., Haack, A. and Baxter, S.L. (2020) Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. *Informatics*, **7**, Article 25. https://doi.org/10.3390/informatics7030025

[19] Watson, J., Hutyra, C.A., Clancy, S.M., Chandiramani, A., Bedoya, A., Ilangovan, K., *et al.* (2020) Overcoming Barriers to the Adoption and Implementation of Predictive Modeling and Machine Learning in Clinical Care: What Can We Learn from US Academic Medical Centers? *JAMIA Open*, **3**, 167-172. https://doi.org/10.1093/jamiaopen/ooz046

[20] Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., *et al.* (2015) Clinical Decision Support Systems for Improving Diagnostic Accuracy and Achieving Precision Medicine. *Journal of Clinical Bioinformatics*, **5**, Article No. 137. https://doi.org/10.1186/s13336-015-0019-3

[21] Amarasingham, R., Patzer, R.E., Huesch, M., Nguyen, N.Q. and Xie, B. (2014) Implementing Electronic Health Care Predictive Analytics: Considerations and Challenges. *Health Affairs*, **33**, 1148-1154. https://doi.org/10.1377/hlthaff.2014.0352

[22] Zikos, D. and DeLellis, N. (2018) CDSS-RM: A Clinical Decision Support System Reference Model. *BMC Medical Research Methodology*, **18**, Article No. 137. https://doi.org/10.1186/s12874-018-0587-6

[23] Huckvale, K., Venkatesh, S. and Christensen, H. (2019) Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety. *npj Digital Medicine*, **2**, Article No. 88. https://doi.org/10.1038/s41746-019-0166-1

[24] Afrifa-Yamoah, E., Adua, E., Peprah-Yamoah, E., Anto, E.O., Opoku-Yamoah, V., Acheampong, E., *et al.* (2024) Pathways to Chronic Disease Detection and Prediction: Mapping the Potential of Machine Learning to the Pathophysiological Processes While Navigating Ethical Challenges. *Chronic Diseases and Translational Medicine*, **11**, 1-21. https://doi.org/10.1002/cdt3.137

[25] Foluke Ekundayo, (2024) Machine Learning for Chronic Kidney Disease Progression Modelling: Leveraging Data Science to Optimize Patient Management. *World Journal of Advanced Research and Reviews*, **24**, 453-475. https://doi.org/10.30574/wjarr.2024.24.3.3730

[26] Sperling, J., Welsh, W., Haseley, E., Quenstedt, S., Muhigaba, P.B., Brown, A., *et al.* (2024) Machine Learning-Based Prediction Models in Medical Decision-Making in Kidney Disease: Patient, Caregiver, and Clinician Perspectives on Trust and Appropriate Use. *Journal of the American Medical Informatics Association*, **32**, 51-62. https://doi.org/10.1093/jamia/ocae255

[27] Delrue, C. and Speeckaert, M.M. (2024) Decoding Kidney Pathophysiology: Omics-Driven Approaches in Precision Medicine. *Journal of Personalized Medicine*, **14**, Article 1157. https://doi.org/10.3390/jpm14121157

[28] Pawuś, D., Porażko, T. and Paszkiel, S. (2024) Automation and Decision Support in the Area of Nephrology Using Numerical Algorithms, Artificial Intelligence, and Expert Approach: Review of the Current State of Knowledge. *IEEE Access*, **12**, 86043-86066. https://doi.org/10.1109/access.2024.3413595

[29] Xu, L., Li, C., Gao, S., Zhao, L., Guan, C., Shen, X., *et al.* (2024) Personalized Prediction of Long-Term Renal Function Prognosis Following Nephrectomy Using Interpretable Machine Learning Algorithms: Case-Control Study. *JMIR Medical Informatics*, **12**, e52837. https://doi.org/10.2196/52837

[30] Karalis, V.D. (2024) The Integration of Artificial Intelligence into Clinical Practice. *Applied Biosciences*, **3**, 14-44. https://doi.org/10.3390/applbiosci3010002

[31] Penman, I.D., Ralston, S.H., Strachan, M.W.J. and Hobson, R. (2022) Davidson's Principles and Practice of Medicine E-Book: Davidson's Principles and Practice of Medicine E-Book. Elsevier.

[32] Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., *et al.* (2021) Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, **11**, Article 5088. https://doi.org/10.3390/app11115088

[33] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., *et al.* (2023) Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, **16**, 45-74. https://doi.org/10.1007/s12559-023-10179-8

[34] Kim, S.Y., Kim, D.H., Kim, M.J., Ko, H.J. and Jeong, O.R. (2024) Xai-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, **14**, Article 6638. https://doi.org/10.3390/app14156638

[35] Chinnaraju, A. (2025) Explainable AI (XAI) for Trustworthy and Transparent Decision-Making: A Theoretical Framework for AI Interpretability. *World Journal of Advanced Engineering Technology and Sciences*, **14**, 170-207. https://doi.org/10.30574/wjaets.2025.14.3.0106

[36] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*,

1, 206-215. https://doi.org/10.1038/s42256-019-0048-x

[37] Sweet, M.M.R., Ahmed, M.P., Akter, S. and Tisha, S.A. (2025) Integrating Deep Learning and Interpretable Regression Models for Transparent Decision Support in Healthcare Diagnostics. *Journal of Medical and Health Studies*, **6**, 17-38. https://doi.org/10.32996/jmhs.2025.6.5.4

[38] Alqudah, A.M. and Moussavi, Z. (2025) Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis. *Applied Sciences*, **15**, Article 12036. https://doi.org/10.3390/app152212036

[39] Hettikankanamage, N., Shafiabady, N., Chatteur, F., Wu, R.M.X., Ud Din, F. and Zhou, J. (2025) Explainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and Their Relevance to Biomedical Imaging and Sensing. *Sensors*, **25**, Article 6649. https://doi.org/10.3390/s25216649

[40] Kaur, N., Bhattacharya, S. and Butte, A.J. (2021) Big Data in Nephrology. *Nature Reviews Nephrology*, **17**, 676-687. https://doi.org/10.1038/s41581-021-00439-x

[41] Tangri, N., Chadban, S., Cabrera, C., Retat, L. and Sánchez, J.J.G. (2022) Projecting the Epidemiological and Economic Impact of Chronic Kidney Disease Using Patient-Level Microsimulation Modelling: Rationale and Methods of Inside CKD. *Advances in Therapy*, **40**, 265-281. https://doi.org/10.1007/s12325-022-02353-5

[42] Fasseeh, A.N., Ashmawy, R., Hren, R., ElFass, K., Imre, A., Németh, B., *et al.* (2025) Generating Realistic Synthetic Patient Cohorts: Enforcing Statistical Distributions, Correlations, and Logical Constraints. *Algorithms*, **18**, Article 475. https://doi.org/10.3390/a18080475

[43] Lin, H. and Lyu, J. (2025) A Holistic Framework for Intradialytic Hypotension Prediction Using Generative Adversarial Networks-Based Data Balancing. *BMC Medical Informatics and Decision Making*, **25**, Article No. 257. https://doi.org/10.1186/s12911-025-03094-5

[44] Sakagianni, A., Koufopoulou, C., Koufopoulos, P., Feretzakis, G., Kalles, D., Paxinou, E., *et al.* (2024) The Synergy of Machine Learning and Epidemiology in Addressing Carbapenem Resistance: A Comprehensive Review. *Antibiotics*, **13**, Article 996. https://doi.org/10.3390/antibiotics13100996

[45] Grandjean, E.M. and Aubry, J.-M. (2009) Lithium: Updated Human Knowledge Using an Evidence-Based Approach: Part II: Clinical Pharmacology and Therapeutic Monitoring. *CNS Drugs*, **23**, 331-349.

[46] Shine, B., McKnight, R.F., Leaver, L. and Geddes, J.R. (2015) Long-term Effects of Lithium on Renal, Thyroid, and Parathyroid Function: A Retrospective Analysis of Laboratory Data. *The Lancet*, **386**, 461-468. https://doi.org/10.1016/s0140-6736(14)61842-0

[47] Davis, J., Desmond, M. and Berk, M. (2018) Lithium and Nephrotoxicity: A Literature Review of Approaches to Clinical Management and Risk Stratification. *BMC Nephrology*, **19**, Article No. 305. https://doi.org/10.1186/s12882-018-1101-4

[48] Goodwin, G., Haddad, P., Ferrier, I., Aronson, J., Barnes, T., Cipriani, A., *et al.* (2016) Evidence-based Guidelines for Treating Bipolar Disorder: Revised Third Edition Recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, **30**, 495-553. https://doi.org/10.1177/0269881116636545

[49] Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R. and Griffin, J.L. (2011) Systems Level Studies of Mammalian Metabolomes: The Roles of Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy. *Chemical Society Reviews*, **40**, 387-426. https://doi.org/10.1039/b906712b

[50] Zhang, Y., Lee, G., Li, S., Hu, Z., Zhao, K. and Rogers, J.A. (2023) Advances in Bioresorbable Materials and Electronics. *Chemical Reviews*, **123**, 11722-11773. https://doi.org/10.1021/acs.chemrev.3c00408

[51] Chan, B.S., Cheng, S., Isoardi, K.Z., Chiew, A., Siu, W., Shulruf, B., *et al.* (2020) Effect of Age on the Severity of Chronic Lithium Poisoning. *Clinical Toxicology*, **58**, 1023-1027. https://doi.org/10.1080/15563650.2020.1726376

[52] Rej, S., Elie, D., Mucsi, I., Looper, K.J. and Segal, M. (2014) Chronic Kidney Disease in Lithium-Treated Older Adults: A Review of Epidemiology, Mechanisms, and Implications for the Treatment of Late-Life Mood Disorders. *Drugs & Aging*, **32**, 31-42. https://doi.org/10.1007/s40266-014-0234-9

[53] McKnight, R.F., Adida, M., Budge, K., Stockton, S., Goodwin, G.M. and Geddes, J.R. (2012) Lithium Toxicity Profile: A Systematic Review and Meta-Analysis. *The Lancet*, **379**, 721-728. https://doi.org/10.1016/s0140-6736(11)61516-x

[54] Werneke, U., Ott, M., Renberg, E.S., Taylor, D. and Stegmayr, B. (2012) A Decision Analysis of Long-Term Lithium Treatment and the Risk of Renal Failure. *Acta Psychiatrica Scandinavica*, **126**, 186-197. https://doi.org/10.1111/j.1600-0447.2012.01847.x

[55] Fries, G.R., Bauer, I.E., Scaini, G., Wu, M., Kazimi, I.F., Valvassori, S.S., *et al.* (2017) Accelerated Epigenetic Aging and Mitochondrial DNA Copy Number in Bipolar Disorder. *Translational Psychiatry*, **7**, Article No. 1283. https://doi.org/10.1038/s41398-017-0048-8

[56] Aron, L., Ngian, Z.K., Qiu, C., Choi, J., Liang, M., Drake, D.M., *et al.* (2025) Lithium Deficiency and the Onset of Alzheimer's Disease. *Nature*, **645**, 712-721. https://doi.org/10.1038/s41586-025-09335-x

[57] Tsao, H., Lai, T., Chou, Y., Lin, S. and Chen, Y. (2023) Predialysis Trajectories of Estimated GFR and Concurrent Trends of Chronic Kidney Disease-Relevant Biomarkers. *Therapeutic Advances in Chronic Disease*, **14**. https://doi.org/10.1177/20406223231177291

[58] Wen, Y., Xu, L., Melchinger, I., Thiessen-Philbrook, H., Moledina, D.G., Coca, S.G., *et al.* (2023) Longitudinal Biomarkers and Kidney Disease Progression after Acute Kidney Injury. *JCI Insight*, **8**, e167731. https://doi.org/10.1172/jci.insight.167731

[59] Rodríguez-Ortiz, M.E., Pontillo, C., Rodríguez, M., Zürbig, P., Mischak, H. and Ortiz, A. (2018) Novel Urinary Biomarkers for Improved Prediction of Progressive EGFR Loss in Early Chronic Kidney Disease Stages and in High Risk Individuals without Chronic Kidney Disease. *Scientific Reports*, **8**, Article No. 15940. https://doi.org/10.1038/s41598-018-34386-8

[60] Zhang, T., Widdop, R.E. and Ricardo, S.D. (2024) Transition from Acute Kidney Injury to Chronic Kidney Disease: Mechanisms, Models, and Biomarkers. *American Journal of Physiology-Renal Physiology*, **327**, F788-F805. https://doi.org/10.1152/ajprenal.00184.2024

[61] Qiao, L., Khalilimeybodi, A., Linden-Santangeli, N.J. and Rangamani, P. (2025) The Evolution of Systems Biology and Systems Medicine: From Mechanistic Models to Uncertainty Quantification. *Annual Review of Biomedical Engineering*, **27**, 425-447. https://doi.org/10.1146/annurev-bioeng-102723-065309

[62] Campagner, A., Biganzoli, E.M., Balsano, C., Cereda, C. and Cabitza, F. (2025) Modeling Unknowns: A Vision for Uncertainty-Aware Machine Learning in Healthcare. *International Journal of Medical Informatics*, **203**, Article ID: 106014. https://doi.org/10.1016/j.ijmedinf.2025.106014

[63] Atf, Z., Ahmad Safavi-Naini, S.A., Lewis, P.R., Mahjoubfar, A., Naderi, N., Savage,

T.R. and Soroush, A. (2025) The Challenge of Uncertainty Quantification of Large Language Models in Medicine. arXiv: 2504.05278.

[64] Schoups, G., van de Giesen, N.C. and Savenije, H.H.G. (2008) Model Complexity Control for Hydrologic Prediction. *Water Resources Research*, **44**, W00B03. https://doi.org/10.1029/2008wr006836

[65] Lacaze, X., Palanque, P., Navarre, D. and Bastide, R. (2002) Performance Evaluation as a Tool for Quantitative Assessment of Complexity of Interactive Systems. In: Forbrig, P., Limbourg, Q., Vanderdonckt, J. and Urban, B., Eds., *Interactive Systems: Design, Specification, and Verification*, Springer, 208-222. https://doi.org/10.1007/3-540-36235-5_16

[66] Courcelles, E., Boissel, J., Massol, J., Klingmann, I., Kahoul, R., Hommel, M., *et al.* (2022) Solving the Evidence Interpretability Crisis in Health Technology Assessment: A Role for Mechanistic Models? *Frontiers in Medical Technology*, **4**, Article 810315. https://doi.org/10.3389/fmedt.2022.810315

[67] Gorin, G., Vastola, J.J., Fang, M. and Pachter, L. (2022) Interpretable and Tractable Models of Transcriptional Noise for the Rational Design of Single-Molecule Quantification Experiments. *Nature Communications*, **13**, Article No. 7620. https://doi.org/10.1038/s41467-022-34857-7

[68] Gundersen, O.E. (2021) The Fundamental Principles of Reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **379**, Article ID: 20200210. https://doi.org/10.1098/rsta.2020.0210

[69] Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., *et al.* (2025) Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and Drivers. *AI Magazine*, **46**, e70002. https://doi.org/10.1002/aaai.70002

[70] Desai, A., Abdelhamid, M. and Padalkar, N.R. (2025) What Is Reproducibility in Artificial Intelligence and Machine Learning Research? *AI Magazine*, **46**, e70004. https://doi.org/10.1002/aaai.70004

[71] Sagi, O. and Rokach, L. (2021) Approximating XGBoost with an Interpretable Decision Tree. *Information Sciences*, **572**, 522-542. https://doi.org/10.1016/j.ins.2021.05.055

[72] Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., *et al.* (2021) Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass. *Forests*, **12**, Article 216. https://doi.org/10.3390/f12020216

[73] Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., *et al.* (2019) Evaluation of CatBoost Method for Prediction of Reference Evapotranspiration in Humid Regions. *Journal of Hydrology*, **574**, 1029-1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

[74] Bottle, A., Gaudoin, R., Goudie, R., Jones, S. and Aylin, P. (2014) Can Valid and Practical Risk-Prediction or Casemix Adjustment Models, Including Adjustment for Comorbidity, Be Generated from English Hospital Administrative Data (hospital Episode Statistics)? A National Observational Study. *Health Services and Delivery Research*, **2**, 1-48. https://doi.org/10.3310/hsdr02400

[75] Cerna, A.E.U., Pattichis, M., VanMaanen, D.P., Jing, L.Y., Patel, A.A., Stough, J.V., Haggerty, C.M. and Fornwalt, B.K. (2019) A Large-Scale Multimodal Study for Predicting Mortality Risk Using Minimal and Low Parameter Models and Separable Risk Assessment. arXiv: 1901.08125.

[76] Hancock, J.T. and Khoshgoftaar, T.M. (2020) CatBoost for Big Data: An Interdisciplinary Review. *Journal of Big Data*, **7**, Article No. 94. https://doi.org/10.1186/s40537-020-00369-8

[77] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D. and Khanna, A. (2021) CatBoost—An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance. *Advances in Data Science and Adaptive Analysis*, **13**, Article ID: 2141002. https://doi.org/10.1142/s2424922x21410023

[78] Ghosh, S.K. and Khandoker, A.H. (2024) Investigation on Explainable Machine Learning Models to Predict Chronic Kidney Diseases. *Scientific Reports*, **14**, Article No. 3687. https://doi.org/10.1038/s41598-024-54375-4

[79] Chen, H., Wang, M. and Li, J. (2024) Exploring the Association between Two Groups of Metals with Potentially Opposing Renal Effects and Renal Function in Middle-Aged and Older Adults: Evidence from an Explainable Machine Learning Method. *Ecotoxicology and Environmental Safety*, **269**, Article ID: 115812. https://doi.org/10.1016/j.ecoenv.2023.115812

[80] Simeri, A., Pezzi, G., Arena, R., Papalia, G., Szili-Torok, T., Greco, R., *et al.* (2024) Artificial Intelligence in Chronic Kidney Diseases: Methodology and Potential Applications. *International Urology and Nephrology*, **57**, 159-168. https://doi.org/10.1007/s11255-024-04165-8

[81] Rockenschaub, P., Hilbert, A., Kossen, T., von Dincklage, F., Madai, V.I. and Frey, D. (2023) From Single-Hospital to Multi-Centre Applications: Enhancing the Generalisability of Deep Learning Models for Adverse Event Prediction in the ICU. arXiv: 2303.15354.

[82] Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., *et al.* (2023) Predicting Sepsis Using Deep Learning across International Sites: A Retrospective Development and Validation Study. *eClinicalMedicine*, **62**, Article ID: 102124. https://doi.org/10.1016/j.eclinm.2023.102124