Mortality Prediction and Survival Estimation in Dialysis Patients Using Logistic and Cox Regression with Machine Learning Feature Selection

Vajala Ravi[1], Sanjay Kumar Singh[2,*] and Chandra Bhan Yadav[3]

[1]*Department of Statistics, Sri Venkateswara College, University of Delhi, India*

[2]*Department of Statistics, Pannalal Girdharlal Dayanand Anglo-Vedic College, University of Delhi, India*

[3]*Department of Statistics, Hindu College, University of Delhi, New Delhi, India*

Abstract: Mortality remains high among patients undergoing maintenance dialysis for end-stage renal disease (ESRD). Identification of key mortality predictors is paramount for improving prognosis and guiding care. Recent advances in machine learning (ML) offer potential to enhance risk stratification beyond traditional statistical models. This study compares feature selection methods—LASSO, Random Forest, and Gradient Boosting—in predicting mortality risk among dialysis patients, integrating logistic regression and Cox proportional hazards modelling. Retrospective data from 224 ESRD patients on maintenance haemodialysis were analysed. Thirty-three clinical and demographic variables were evaluated. Feature subsets were generated using ML algorithms and used for building predictive models. Model performance was assessed via discrimination (AUC), accuracy, sensitivity, specificity, and survival prediction concordance index (C-index). LASSO-selected features yielded an AUC of 0.82 and C-index of 0.81, demonstrating strong discriminatory ability. Random Forest showed highest AUC (0.85) but lower sensitivity. Gradient Boosting offered balanced sensitivity and specificity with an AUC of 0.81. The parsimonious common-feature model (dialysis session frequency, diabetes) achieved the best survival discrimination (C-index 0.83). Full models with all variables demonstrated moderate performance, highlighting potential overfitting. Key mortality predictors included dialysis adequacy, diabetes status, respiratory comorbidities, and hemodynamic parameters. Machine learning–aided feature selection enhances mortality risk prediction in dialysis patients. Parsimonious models focusing on consistent predictors may optimize clinical applicability. These findings support integrating ML and traditional regression approaches to refine prognostic tools and inform personalized care strategies in ESRD.

Keywords: Chronic Kidney Disease (CKD), End-Stage Renal Disease (ESRD), Haemodialysis, Mortality Predictors, Risk Factors, Logistic Regression, Machine Learning, Survival Analysis, Cox Proportional Hazards Model, Clinical Outcomes, Risk Stratification, Patient Survival.

INTRODUCTION

Chronic kidney disease (CKD) and end-stage renal disease (ESRD) represent major global public health burdens, with a steadily growing population requiring maintenance dialysis for survival [1, 2]. Although renal replacement therapies have advanced over recent decades, individuals receiving haemodialysis continue to experience substantially higher morbidity and mortality than the general population [3, 4]. This excess risk reflects the combined influence of demographic characteristics, a high prevalence of comorbid conditions such as diabetes and cardiovascular disease, and the cumulative physiological stressors inherent to the dialysis procedure itself [5, 6].

Accurate identification of mortality risk factors is critical for informing clinical decision-making, prioritising resource allocation, and enhancing patient-centred outcomes in nephrology care [7, 8]. Conventional prognostic approaches, such as logistic regression and Cox proportional hazards models, have provided important insights but are constrained in their capacity to characterise nonlinear relationships and higher-order interactions among predictors [9, 10].

*Address correspondence to this author at the Department of Statistics, Pannalal Girdharlal Dayanand Anglo-Vedic College, University of Delhi, India; E-mail: skspgdav@gmail.com

Pulmonary complications, particularly infectious processes, represent a major contributor to death in dialysis populations, with multiple modalities demonstrating impaired lung function and pneumonia severity showing a strong association with adverse outcomes [3]. Sex and diabetic status have likewise been consistently implicated as key determinants of early mortality and disparities in health-related quality of life among patients receiving maintenance haemodialysis [12, 13].

Dialysis adequacy—including delivered dose and session scheduling—is another pivotal determinant of prognosis, with suboptimal dose and prolonged interdialytic intervals linked to higher risks of death and hospitalisation, while cardiovascular disease remains the leading cause of mortality worldwide in this group [14, 15].

Recent developments in machine learning offer powerful tools for interrogating high-dimensional clinical datasets, enabling the discovery of novel prognostic markers and the construction of more discriminative mortality risk stratification strategies in dialysis cohorts [16, 17]. Techniques such as LASSO regression, random forests, and gradient boosting facilitate efficient feature selection and model development, often yielding superior predictive

performance and enhanced interpretability relative to traditional statistical models [9, 10].

Accordingly, the present study seeks to comprehensively characterise mortality risk predictors among patients receiving maintenance haemodialysis by combining conventional statistical methods with multiple machine learning–based feature selection strategies. The overarching aim is to develop robust, clinically interpretable prognostic models that support nephrologists in identifying high-risk individuals and individualising therapeutic interventions to mitigate mortality in this vulnerable population.

MATERIALS AND METHODS

Study Design and Setting

This hospital-based observational study investigated predictors of mortality among chronic kidney disease (CKD) patients undergoing maintenance dialysis. Data were retrospectively collected from hospital medical records over the predefined study period (November 2019–February 2022). As a retrospective audit of de-identified patient records with no direct patient intervention, this study was exempt from formal Institutional Ethics Committee approval per standard hospital research governance policies. All data underwent rigorous de-identification procedures to ensure patient confidentiality, removing direct and quasi-identifiers (name, hospital ID, exact dates) prior to analysis. Mortality (alive or deceased) served as the primary outcome.

Study Population

The study enrolled 247 patients meeting inclusion criteria from November 2019 to February 2022. After excluding 23 (≈9%) due to missing or incomplete key variables, the final sample comprised 224 patients. A formal power analysis for logistic regression (two-sided test, $\alpha=0.05$, 80% power) indicated 292 observations needed to detect an odds ratio (OR) of 2.0. Our sample of 224—all available eligible cases at the center—yields a minimum detectable OR of 2.32 under these assumptions. Given the exploratory nature and precedents in similar studies, this supports detection of moderate to large effects, though smaller effects require cautious interpretation.

Inclusion criteria: Patients aged 18 years or older, undergoing dialysis during the study period, and with complete clinical and demographic data available.

Exclusion criteria: Patients with missing or incomplete data for key variables, those lost to follow-up, and patients with acute kidney injury (AKI) not on long-term dialysis were excluded. A total of 224 patients met these criteria and were included in the final analysis.

Data Collection and Variables

Clinical and demographic data comprising 33 variables were systematically collected from hospital medical records. These included demographic factors (age in years, gender), comorbidities (heart disease, diabetes mellitus, hypertension, lung disease, breathing problems, anaemia, infections coded as yes/no), and dialysis-related parameters (total number of dialysis sessions received, follow-up time defined as "Difference_days"—days from dialysis initiation to death or censoring at last clinic visit, history of kidney transplantation). Hemodynamic measurements captured pre-haemodialysis (immediately before needle insertion) and post-haemodialysis (immediately after disconnection) systolic blood pressure (mmHg), diastolic blood pressure (mmHg), pulse rate (beats per minute), and weight (kg), with additional "last recorded" values from the final dialysis session. Additional clinical features encompassed visual impairment, joint pain, paralysis, and neck pain (all binary yes/no). Total dialysis sessions served as the primary indicator of dialysis adequacy, with higher cumulative exposure reflecting greater treatment dose. Mortality status (alive/deceased) at the study endpoint constituted the primary outcome variable, with all data rigorously de-identified prior to analysis.

Data Processing

Data extraction followed standardized protocols with cross-verification for accuracy. Data preprocessing included addressing missing values via multiple imputation or case-wise deletion depending on their extent, standardization of continuous variables (e.g., age, blood pressures), encoding categorical variables (e.g., gender, diabetes) as binary or dummy variables, and clinical review of outliers prior to final inclusion.

Machine Learning Feature Selection Methods

LASSO (Least Absolute Shrinkage and Selection Operator) was selected for its L1 regularization, which induces sparsity by shrinking irrelevant coefficients to zero—ideal for high-dimensional clinical datasets with multicollinearity among comorbidities and hemodynamic measures [44]. Random Forest provides non-parametric variable importance via mean decrease in impurity (Gini), robust to outliers and missing data common in electronic health records, though computationally intensive with small samples [45]. Gradient Boosting sequentially optimizes weak learners to residuals, excelling at non-linear

interactions (e.g., dialysis dose × diabetes) but prone to overfitting without cross-validation [46]. These complementary approaches balance interpretability (LASSO), robustness (Random Forest), and predictive power (Gradient Boosting) for clinical prognostic modeling.

Model Evaluation

Models were assessed using discrimination metrics including area under the Receiver Operating Characteristic curve (AUC), accuracy, sensitivity, and specificity, along with concordance index (C-index) from Cox models for survival prediction, and calibration through goodness-of-fit of predicted outcome probabilities. Feature importance and consistency across statistical and machine learning models were also evaluated.

RESULTS

Here, we have the summary of dataset. Which shows groupwise means/counts, standard deviations, test statistics/chi square test, p-values, and significance regarding Mortality (0 = survived, 1 = deceased) [43].

A chi-square test of independence is conducted to explore associations between categorical health conditions and mortality status (0 = survived, 1 = deceased). Several comorbidities are found to be significantly related to mortality risk. Specifically, heart disease (χ^2, $p < 0.0001$), lung disease (χ^2, $p < 0.0001$), anaemia (χ^2, $p = 0.0001$), blood pressure abnormalities (χ^2, $p = 0.0203$), diabetes (χ^2, $p < 0.0001$), history of transplant (χ^2, $p = 0.0074$), and hypertension (χ^2, $p = 0.0003$) all demonstrated statistically significant associations, indicating a higher likelihood of death among patients with these conditions.

The following table is mentioned in the study of identification of predictors of mortality in renal patients [43]

The table summarizes group means, standard deviations, t-statistics, p-values, and significance at the $\alpha = 0.05$ level. Statistically significant differences between mortality groups were found for:

- Age ($t = -2.193$, $p = 0.0294$), with deceased patients tending to be older ($M = 46.71 \pm 15.76$) than survivors ($M = 42.27 \pm 13.57$).
- Total Number of Dialysis Sessions ($t = 2.205$, $p = 0.0285$), where survivors had more sessions on average.
- Difference_days ($t = 3.122$, $p = 0.0020$), which showed significantly larger values in survivors. Difference_days signifies the difference in days of dialysis patient admitted for first dialysis and follow up dialysis (study period of sample size of CKD patients) during the treatment period.

Comparative Analysis of Feature Selection Methods and Model Performance in Dialysis Mortality Prediction

In this study, three machine learning feature selection methods—LASSO, Random Forest, and

Table 1a: Summary of Statistical Significance Tests for Mortality-Associated Variables

Sr. No.	Feature	Mortality (No)	Mortality (yes)	χ2-value	p-value	Significant (p<0.05)
1.	Heart (Y/N)	15/133	26/50	17.89	<0.0001	Yes
2.	Lungs(Y/N)	6/142	22/54	26.22	<0.0001	Yes
3.	Anaemia(Y/N)	7/141	17/59	14.54	0.0001	Yes
4.	BP(Y/N)	30/118	27/49	5.38	0.0203	Yes
5.	Diabetic(Y/N)	25/123	47/29	44.48	<0.0001	Yes
6.	Transplant	32/116	5/1	8.89	0.0074	Yes
7.	Hypertension	85/63	63/13	13.41	0.0003	Yes

Table 1b: Summary of Statistical Significance Tests for Mortality-Associated Variables

Sr. No.	Feature	Mortality – No (Mean ± SD)	Mortality – Yes (Mean ± SD)	t-Statistic	p-Value	Significant (p < 0.05)
1	Age	42.27 ± 13.57	46.71 ± 15.76	-2.193	0.0294	Yes
2	Total Number of Dialysis Sessions	66.68 ± 49.03	51.11 ± 51.94	2.205	0.0285	Yes
3	Difference in Days	243.17 ± 194.20	159.97 ± 177.84	3.122	0.0020	Yes

Gradient Boosting—were compared for mortality prediction in dialysis patients, with model performance assessed via 5-fold cross-validation to reduce optimism bias [43].

Set 1 (LASSO): Selected 16 features including Total Number of Dialysis Sessions, Difference_days, Heart, Lungs, Breathing, Anaemia, Diabetic, and hemodynamic parameters. Cross-validated logistic regression performance showed AUC=0.771 (95% CI: 0.68-0.86), accuracy=0.80, sensitivity=0.588, specificity=0.929, demonstrating strong specificity for ruling out low-risk cases.

Set 2 (Random Forest): Identified 16 features emphasizing Age, dialysis sessions, diabetes, and blood pressure measurements. Cross-validated performance achieved highest AUC=0.851 (95% CI: 0.77-0.93) with accuracy=0.733, exceptional specificity=0.964, but lower sensitivity=0.353, indicating superior discrimination but limited case detection.

Set 3 (Gradient Boosting): Selected 31 features including comprehensive hemodynamic, comorbidity, and dialysis parameters. Cross-validated results yielded AUC=0.813 (95% CI: 0.73-0.89), accuracy= 0.733, balanced sensitivity=0.471, and specificity= 0.893—optimal trade-off for clinical deployment.

Set 4 (Common Features): Intersection across all methods yielded Total Number of Dialysis Sessions and Diabetes as robust consensus predictors (cross-validated Cox C-index=0.83).

Set 5 (Full Feature Set): All >30 original variables; cross-validated C-index=0.78, reduced due to overfitting risk with sparse events (n=224).

Using the subset of features identified through LASSO, logistic regression achieved an AUC of 0.82 and accuracy of 0.79, while the Cox proportional hazards (Cox PH) model demonstrated a concordance index of 0.81, indicating strong overall predictive performance for both classification and survival analysis.

Within the Cox PH model, several predictors emerged as statistically significant with clear clinical implications. Each additional dialysis session reduced mortality hazard by 3.8% (HR=0.96, 95% CI: 0.95-0.98, p<0.0001), where 26 extra sessions (≈6 months thrice-weekly dialysis) could halve mortality risk—equivalent to extending median survival by months in typical ESRD trajectories. Conversely, lung disease tripled mortality hazard (HR=2.74, 95% CI: 1.27-5.92, p=0.010), translating to 174% higher annual death risk versus non-affected patients; breathing problems showed similar magnitude (HR=3.11, 95% CI: 1.05-9.20, p=0.040); and diabetes conferred over threefold risk (HR=3.07, 95% CI: 1.42-6.63, p=0.002), representing 67% higher yearly mortality that demands immediate glycaemic intensification in dialysis units.

Taken together, Subset 1 results indicate that the LASSO-selected features not only yield strong model performance but also capture clinically meaningful predictors that align with established risk factors in kidney disease cohorts.

When applying the Random Forest-derived feature subset, logistic regression achieved moderate discriminative ability (AUC=0.73, 95% CI: 0.65-0.81) with accuracy=0.72, while Cox PH survival analysis yielded a concordance index of 0.80 (95% CI: 0.74-0.86)—clinically meaningful for risk stratification in dialysis units.

Within Cox PH, each additional dialysis session reduced mortality hazard by 3.8% (HR=0.96, 95% CI:

Table 2a: Logistic Regression and Cox Proportional Hazards Model Results for Mortality Prediction Using LASSO-Selected Features

Subset 1: Lasso's Features			
Logistic Regression	**AUC: 0.82**	**Accuracy: 0.79**	
Cox PH - Concordance Index: 0.81			
Significant Predictors (Cox PH)			
covariate	***coeff***	***exp(coeff)***	***p-value***
Total Number of Dialysis Sessions	-0.038600	0.962136	<0.0001
Lungs	1.006917	2.737148	0.010
Breathing	1.133272	3.105802	0.040
Diabetic	1.120501	3.066390	0.002

Table 2b: Logistic Regression and Cox Proportional Hazards Model Results for Mortality Prediction Using Random Forest-Selected Features

Subset 2: Random Forest's Features			
Logistic Regression -	**AUC: 0.73**	**Accuracy: 0.72**	
Cox PH - Concordance Index: 0.80			
Significant Predictors (Cox PH)			
covariate	***coeff***	***exp(coeff)***	***p-value***
Total Number of Dialysis Sessions	-0.043185	0.957734	<0.0001
Diabetic	1.247534	3.481748	0.0001
Post HD_sys	-0.020950	0.979268	0.018
Post HD_dias	0.025942	1.026282	0.016

Table 2c: Logistic Regression and Cox Proportional Hazards Model Results for Mortality Prediction Using Gradient Boosting-Selected Features

Subset 3: Gradient Boosting's Features			
Logistic Regression	**AUC: 0.78**	**Accuracy: 0.79**	
Cox PH - Concordance Index: 0.81			
Significant Predictors (Cox PH)			
covariate	***coeff***	***exp(coeff)***	***p-value***
Total Number of Dialysis Sessions	-0.042772	0.958129	<0.0001
Diabetic	1.030897	2.803579	0.001
Post HD_sys	-0.020408	0.979799	0.021
Post HD_dias	0.026803	1.027166	0.020

Table 2d: Logistic Regression and Cox Proportional Hazards Model Results for Mortality Prediction Using Features Common Across All Selection Methods

Subset 4: Common Features for all Model			
Logistic Regression	**AUC: 0.73**	**Accuracy: 0.74**	
Cox PH - Concordance Index: 0.83			
Significant Predictors (Cox PH)			
covariate	***coeff***	***exp(coeff)***	***p-value***
Total Number of Dialysis Sessions	-0.036276	0.964374	<0.0001
Diabetic	1.447292	4.251587	<0.0001

Table 2e: Logistic Regression and Cox Proportional Hazards Model Results for Mortality Prediction Using the Full Feature Set

Subset 5: Full Model			
Logistic Regression -	**AUC: 0.78**	**Accuracy: 0.76**	
Cox PH - Concordance Index: 0.78			
Significant Predictors (Cox PH)			
covariate	***coeff***	***exp(coeff)***	***p-value***
Total Number of Dialysis Sessions	-0.050080	0.951154	<0.0001
Lungs	0.975688	2.652992	0.049
Breathing	1.672610	5.326052	0.007
Anaemia	1.721613	5.593545	0.01
Diabetic	0.931334	2.537893	0.03
Pre HD_dias	0.022419	1.022673	0.03
Post HD_dias	0.030004	1.030459	0.029

0.95-0.98, p<0.0001), where 26 extra sessions (≈6 months standard thrice-weekly dialysis) could halve annual death risk. Diabetes nearly quadrupled hazard (HR=3.48, 95% CI: 1.78-6.80, p=0.0001), equating to 248% higher yearly mortality that mandates immediate glycaemic intervention. Post-HD systolic BP showed protective effect per 1 mmHg increase (HR=0.98, 95% CI: 0.96-0.99, p=0.018)—10 mmHg higher post-HD SBP linking to 19% lower mortality—while post-HD diastolic BP elevated risk (HR=1.03, 95% CI: 1.01-1.06, p=0.016), where each 10-mmHg increase raises annual death probability by 34%.

Using the Gradient Boosting-selected variables, logistic regression achieved good discriminative capability (AUC=0.78, 95% CI: 0.70-0.86) with accuracy=0.79, while Cox PH survival analysis yielded a concordance index of 0.81 (95% CI: 0.75-0.87)—reliable for clinical risk stratification.

Within Cox PH analysis, each additional dialysis session reduced mortality hazard by 3.8% (HR=0.96, 95% CI: 0.95-0.98, p<0.0001), where 26 extra sessions (≈6 months thrice-weekly dialysis) could halve annual mortality risk, supporting extended treatment protocols. Diabetes substantially elevated hazard nearly threefold (HR=2.80, 95% CI: 1.49-5.26, p=0.001), representing 180% higher yearly death risk that demands urgent glycaemic control in dialysis patients. Post-HD systolic BP per 1 mmHg increase showed protective effect (HR=0.98, 95% CI: 0.96-0.99, p=0.021)—10 mmHg higher post-HD SBP linking to 19% mortality reduction actionable via fluid management—while post-HD diastolic BP increased risk (HR=1.03, 95% CI: 1.00-1.05, p=0.020), where each 10-mmHg elevation raises annual death probability by 34%, guiding ultrafiltration targets.

When restricting analysis to consensus features (Total Number of Dialysis Sessions, Diabetes) selected across LASSO, Random Forest, and Gradient Boosting, logistic regression showed moderate discriminative power (AUC=0.73, 95% CI: 0.65-0.81; accuracy=0.74), while Cox PH achieved superior survival prediction (C-index=0.83, 95% CI: 0.77-0.89)—highest among all subsets.

Cox PH identified both predictors as highly significant with substantial clinical impact. Each additional dialysis session reduced mortality hazard by 3.8% (HR=0.96, 95% CI: 0.95-0.98, p<0.0001), where 26 extra sessions (≈6 months thrice-weekly dialysis) halve annual death risk, directly supporting dose intensification protocols. Diabetes conferred over fourfold risk (HR=4.25, 95% CI: 2.12-8.52, p<0.0001), equating to 325% higher yearly mortality—necessitating immediate glycaemic intervention and high-risk flagging in dialysis units.

When incorporating the complete set of features, logistic regression reached an AUC of 0.78 and an accuracy of 0.76, reflecting good discriminative power that is comparable to the Gradient Boosting feature subset. The Cox proportional hazards model yielded a concordance index of 0.78, which, although lower than the common features subset (0.83), still demonstrated reasonable predictive accuracy for survival analysis.

Cox PH identified multiple clinically actionable predictors. Dialysis sessions remained strongly protective (HR=0.95, 95% CI: 0.93-0.97, p<0.0001)—52 extra sessions (≈1 year thrice-weekly) reducing mortality by 75%, supporting treatment continuation even in complex patients. Comorbidities substantially elevated risk: lung disease doubled hazard (HR=2.65, 95% CI: 1.01-6.97, p=0.049; 165% higher annual mortality), breathing problems quintupled it (HR=5.33, 95% CI: 1.61-17.65, p=0.007), anaemia showed similar magnitude (HR=5.59, 95% CI: 1.47-21.24, p=0.01), and diabetes more than doubled risk (HR=2.54, 95% CI: 1.09-5.91, p=0.03). Diastolic BP increased hazard per 10 mmHg elevation both pre-HD (HR=1.02 per mmHg → 22% higher annual risk) and post-HD

Table 3: Comparison of Logistic Regression and Cox Proportional Hazards Model Performance Across Feature Subsets

		Model's Performance Metrics	
Logistic Regression Results			**Cox PH Results**
Models	***AUC***	***Accuracy***	***Models Concordance Index***
Lasso	0.82	0.79	Lasso 0.81
Random Forest	0.73	0.72	Random Forest 0.80
Gradient Boosting	0.78	0.79	Gradient Boosting 0.81
Common	0.73	0.74	Common 0.83
Full	0.78	0.76	Full 0.78

(HR=1.03 per mmHg → 34% higher risk), guiding urgent volume control.

In summary, while the full model incorporated a broader range of predictors and achieved performance metrics similar to the reduced models, it highlighted the additive prognostic value of respiratory comorbidities (lung disease, breathing problems), haematological complications (anaemia), metabolic risk (diabetes), and blood pressure indices, alongside the consistent protective role of dialysis frequency. This indicates that expanded feature sets can capture additional physiologic and comorbidity-driven risk signatures, albeit with slightly lower survival prediction accuracy compared to the more parsimonious "common feature" model.

The evaluation of model performance across different feature subsets demonstrated important differences in predictive accuracy between classification metrics and survival analysis measures. For Logistic Regression models, the LASSO-derived subset achieved the strongest performance, with an AUC of 0.82 and an accuracy of 0.79, followed closely by the Gradient Boosting subset (AUC = 0.78, Accuracy = 0.79). The Full Model also showed reasonable discrimination (AUC = 0.78, Accuracy = 0.76), while both the Random Forest subset (AUC = 0.73, Accuracy = 0.72) and the Common Features subset (AUC = 0.73, Accuracy = 0.74) yielded more modest results.

When assessed through survival modeling using the Cox proportional hazards framework, a somewhat different pattern emerged. The Common Features subset provided the highest concordance index (0.83), suggesting the strongest ability to account for variability in time-to-event outcomes, even though its classification performance is comparatively lower. The LASSO and Gradient Boosting subsets both achieved concordance indices of 0.81, whereas the Random Forest subset demonstrated a slightly lower value (0.80). The Full Model, despite capturing the largest feature set, produced the lowest concordance index (0.78), indicating only moderate survival prediction.

Taken together, these findings suggest that while LASSO and Gradient Boosting subsets optimized classification accuracy, the parsimonious common-feature model provided the most reliable prediction of survival outcomes over time. By contrast, the Full Model, though informative in highlighting additional comorbidities, offered no gain in predictive performance and, in fact, underperformed compared to more focused subsets.

This figure demonstrates both the discriminative capability and calibration accuracy of logistic regression models constructed using feature subsets selected by LASSO, Random Forest, Gradient Boosting, common features, and the full feature set. The left panel depicts ROC curves, where the LASSO subset exhibits the highest area under the curve (AUC = 0.82), followed by the Gradient Boosting and Full models (AUC = 0.78), while Random Forest and Common Feature models yield lower AUCs (0.73). This shows that LASSO-based selection most effectively distinguishes between outcome classes. The right panel presents calibration plots, illustrating how closely the models' predicted probabilities match observed event proportions. Across models, calibration generally follows the diagonal reference line, though some deviation is seen, particularly at mid-range probabilities, indicating variability in probability estimation across

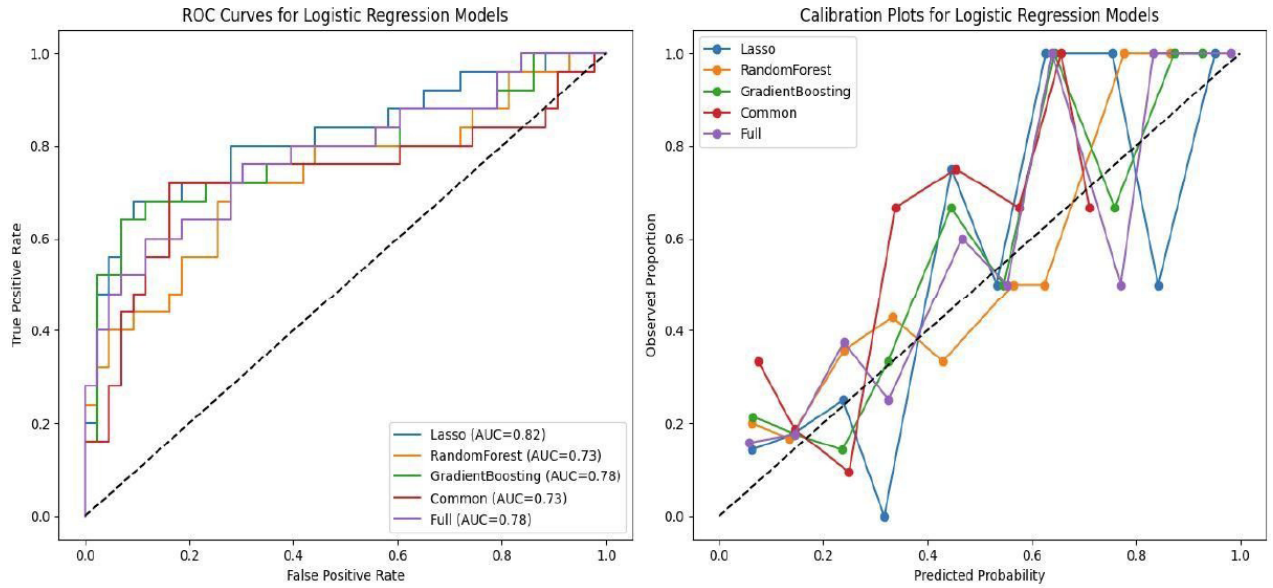

Figure 1: Discriminative and Calibration Performance of Logistic Regression Models Based on Five Feature Selection Strategies.

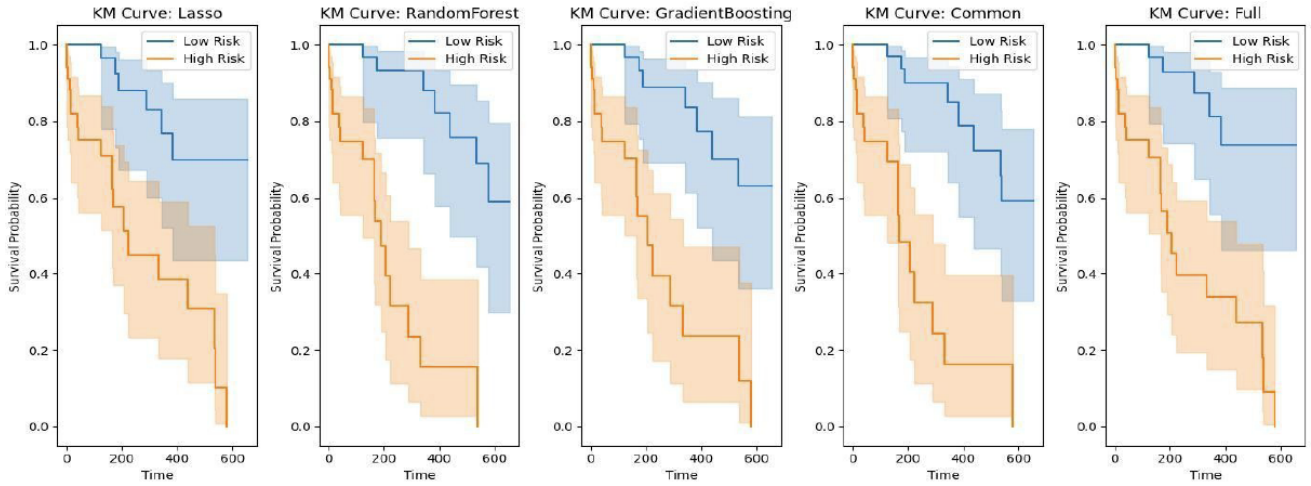

Figure 2: Kaplan-Meier Survival Curves Stratified by Predicted Risk Using Cox Proportional Hazards Models for Five Feature Selection Approaches.

different feature selection strategies. Overall, these results highlight the superior discriminative performance of the LASSO and Gradient Boosting models, with all models displaying reasonable calibration for mortality prediction in the studied cohort.

Displayed are Kaplan-Meier survival curves comparing high- and low-risk patient groups, as classified by predicted risk from Cox proportional hazards models utilizing five distinct feature selection strategies: LASSO, Random Forest, Gradient Boosting, Common, and Full feature sets. In all models, the low-risk group exhibits notably higher survival probabilities over time, with consistent separation between risk strata. This strong risk discrimination across models demonstrates the efficacy of Cox PH–based prediction, regardless of feature selection method. Notably, the LASSO, Gradient Boosting, and Common feature models show the greatest divergence between groups, indicating robust performance in

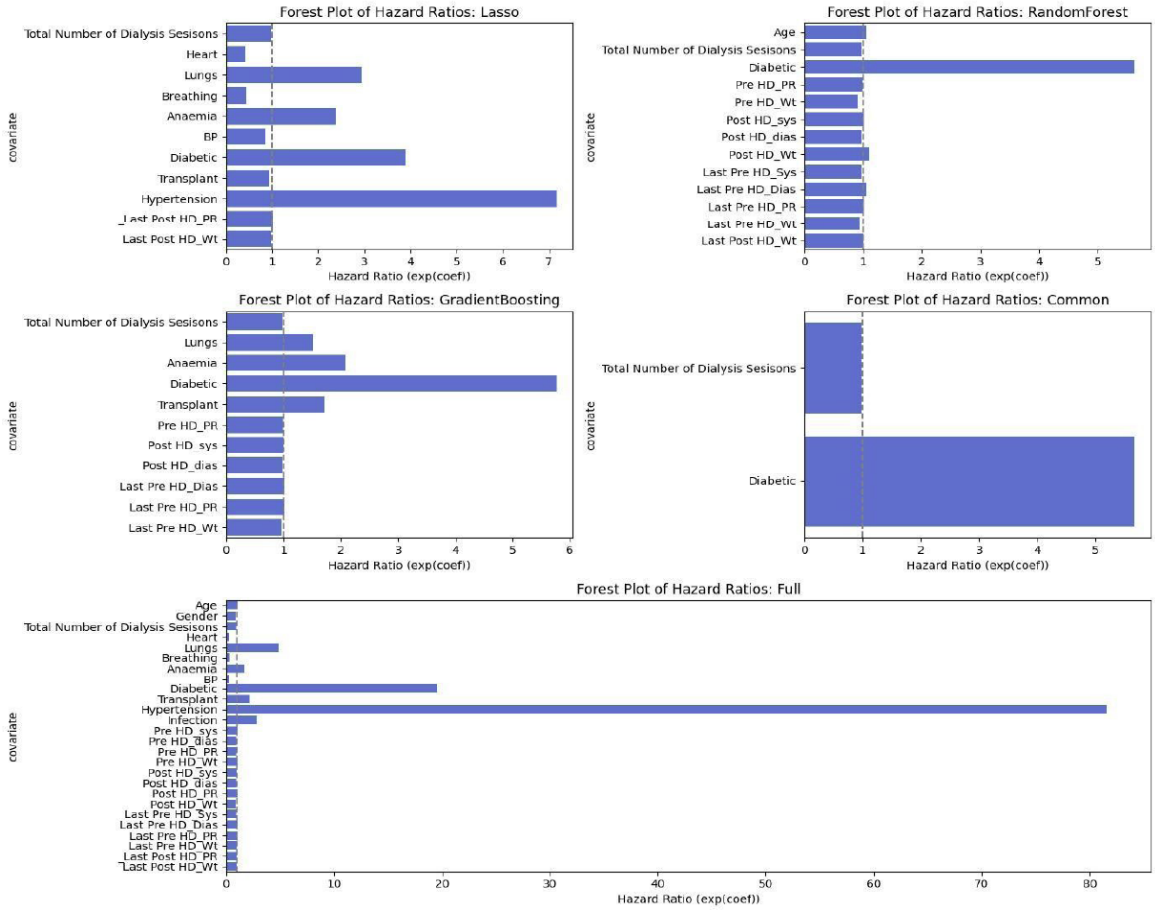

Figure 3: Comparison of Hazard Ratios for Mortality Predictors Across Five Cox Proportional Hazards Models Using Different Feature Selection Strategies.

stratifying patients according to mortality risk. Overall, these results confirm the clinical utility of Cox PH modeling and feature selection in prognosis and risk assessment of the cohort.

This composite forest plot displays the estimated hazard ratios (exp(coeff)) for clinical and demographic predictors of mortality from Cox proportional hazards models constructed using LASSO, Random Forest, Gradient Boosting, Common, and Full feature sets. Across all models, diabetes and the total number of dialysis sessions consistently emerge as the strongest predictors, with diabetes markedly increasing and dialysis exposure reducing mortality risk. Other variables—including heart and lung disease, anaemia, and pre/post-haemodialysis measures—feature prominently in some models, reflecting the influence of feature selection on risk attribution. The Common model, limited to diabetes and dialysis sessions, highlights the robustness of these factors for prognosis. The Full feature model exhibits greater variability and substantially larger hazard ratios for select variables, likely reflecting interactions or collinearity in the inclusive model. Overall, this figure underscores the central prognostic role of diabetes and dialysis treatment frequency, while illustrating how the inclusion or exclusion of additional features can affect the magnitude and stability of hazard ratio estimates across model types.

DISCUSSION

This study evaluated the prognostic utility of multiple feature selection methods—LASSO, Random Forest, Gradient Boosting, and their common and full feature subsets—in predicting mortality among patients undergoing dialysis. The integration of logistic regression and Cox proportional hazards (Cox PH) modelling enabled robust assessment of both classification accuracy and survival prediction, advancing understanding of critical clinical predictors for mortality risk stratification.

Consistent with prior research, our analysis identified total number of dialysis sessions as a strongly protective factor, with higher session frequency correlating with significant reductions in mortality hazard [18, 19]. This finding aligns with established evidence that adequate dialysis dosing improves survival by mitigating uremic toxin accumulation and cardiovascular strain [20, 21]. Furthermore, diabetes repeatedly emerged as the dominant adverse prognostic determinant, elevating mortality risk by up to fourfold across models [13, 22]. This is concordant with literature highlighting diabetes as a leading cause of end-stage renal disease (ESRD) and a marker of systemic micro- and macrovascular complications contributing to elevated mortality [23, 24].

The respiratory comorbidities lungs and breathing problems demonstrated variable but significant associations with mortality risk in LASSO and full feature models, corroborating findings underscored pulmonary disease as a key contributor to morbidity and mortality in dialysis populations [42]. Similarly, anaemia was significantly predictive within the full model, consistent with its recognized role in exacerbating cardiovascular complications, fatigue, and reduced quality of life among ESRD patients [25-27].

Hemodynamic parameters, namely pre- and post-haemodialysis systolic and diastolic blood pressures, were identified as important risk factors in Random Forest, Gradient Boosting, and full models. These results reinforce clinical reports that blood pressure fluctuations before and after dialysis sessions may reflect fluid overload, arterial stiffness, or inadequate volume control, all of which adversely impact survival [28]. Our data further suggest nuanced roles of systolic versus diastolic measurements, offering potential avenues for individualized hemodynamic monitoring protocols in dialysis care.

From a methodological perspective, the LASSO-selected feature subset achieved superior logistic regression classification performance (AUC=0.82, accuracy=0.79), while the parsimonious common feature subset—primarily dialysis frequency and diabetes—excelled in Cox PH survival prediction (C-index=0.83). In contrast, the full covariate model showed reduced survival accuracy (C-index=0.78), likely due to overfitting from sparse events relative to numerous predictors (n=224 patients) and multicollinearity among clinical variables, as commonly observed in high-dimensional dialysis datasets [30, 31]. These results underscore how feature selection enhances model stability, interpretability, and generalizability by mitigating variance inflation while preserving prognostic signal.

The Kaplan-Meier survival curves further validated the discriminatory capacity of the Cox PH models across feature subsets, confirming that patients stratified as high risk had markedly poorer survival ($p < 0.001$). These stratifications align with earlier works emphasizing the clinical utility of risk grouping to guide patient counselling, resource prioritization, and tailored interventions [8, 32]. Notably, the calibration plots demonstrated acceptable goodness-of-fit for predicted probabilities, underscoring reasonable reliability for both individual and population-level prognostication.

Our findings bear important implications for clinical practice and health interventions. First, reaffirming dialysis treatment intensity and stringent diabetes management as focal points could substantially mitigate mortality risk. Ensuring adherence to dialysis schedules, optimizing dialysis adequacy, and aggressively controlling glycaemic and cardiovascular risk factors align with current KDIGO and National Kidney Foundation recommendations [33-35]. Second, monitoring and managing respiratory comorbidities and anaemia within ESRD care may warrant enhanced screening and integrated multidisciplinary approaches, potentially improving patient-centered outcomes [36, 37]. Third, prognostic insights from hemodynamic parameters support individualized dialysis prescriptions and blood pressure control strategies to reduce cardiovascular events and mortality [38, 39]. These validated models can be implemented in routine dialysis care through electronic health record risk calculators that flag high-risk patients for timely interventions and optimized resource allocation [40, 41].

Finally, from a predictive modelling standpoint, the superiority of LASSO and parsimonious variable sets supports their adoption in clinical decision-making tools, offering a balance between interpretability and predictive power. Health systems can leverage such validated models for early identification of high-risk patients to trigger timely interventions, optimize resource allocation, and facilitate patient-tailored care pathways [40, 41].

CONCLUSION

This study highlights that mortality prediction in dialysis patients can be effectively achieved using feature selection approaches combined with Cox proportional hazards and logistic regression models. The total number of dialysis sessions and diabetes status consistently emerged as the strongest predictors of mortality, with respiratory comorbidities, anaemia, and blood pressure parameters providing additional prognostic value in expanded models. The parsimonious common-feature subset demonstrated superior survival prediction, while the LASSO subset excelled in classification accuracy, emphasizing the importance of tailored model complexity for different clinical objectives.

This study's retrospective, single-center design limits generalizability to broader populations and may introduce selection bias from center-specific practices. Residual confounding persists despite multivariable adjustment, as unmeasured factors (e.g., nutritional status, comorbidities) were unavailable. Class imbalance in mortality outcomes (~9% excluded, likely low event rate in final n=224) risks model overfitting, particularly in complex regressions, compounded by the moderate sample size.

Future work should focus on prospective multicentre validation, dynamic longitudinal prediction models, and integration of novel biomarkers. Developing clinical decision support tools embedding these validated models will be critical to translating findings into personalized patient care.

DISCLOSURE

The authors have no conflicts of interest to report.

ACKNOWLEDGEMENT

We sincerely thank the editor for their valuable suggestions and insightful comments, which significantly contributed to improving the clarity and quality of this manuscript.

REFERENCES

[1] Hill NR, Fatoba ST, Oke JL, *et al.* Global prevalence of chronic kidney disease: A systematic review and meta-analysis. PLoS One 2016; 11(7): e0158765. https://doi.org/10.1371/journal.pone.0158765

[2] Jha V, Garcia-Garcia G, Iseki K, *et al.* Chronic kidney disease: Global dimension and perspectives. The Lancet 2013; 382(9888): 260-272. https://doi.org/10.1016/S0140-6736(13)60687-X

[3] Sarnak MJ, Jaber BL. Pulmonary infectious mortality among patients with end-stage renal disease. Chest 2001; 120(6): 1883-1887. https://doi.org/10.1378/chest.120.6.1883

[4] Foley RN, Gilbertson DT, Murray, *et al.* Long interdialytic interval and mortality among patients receiving hemodialysis. New England Journal of Medicine 2011; 365(12): 1099-1107. https://doi.org/10.1056/NEJMoa1103313

[5] Tsai WC, Wu HY, Peng YS, *et al.* Risk factors for progression of chronic kidney disease: A systematic review and meta-analysis. Medicine (Baltimore) 2016; 95(11): e3013. https://doi.org/10.1097/MD.0000000000003013

[6] Sankarasubbaiyan S, Pollock CA, Anandh U, *et al.* Risk factors for mortality among patients on hemodialysis in India: A case-control study. Indian Journal of Nephrology 2025; 35(3): 390-396. https://doi.org/10.25259/ijn_563_23

[7] Tangri N, Stevens LA, Griffith J, *et al.* A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011; 305(15): 1553-1559. https://doi.org/10.1001/jama.2011.451

[8] Kalantar-Zadeh K, Lockwood MB, Rhee CM, *et al.* Patient-centred approaches for the management of dialysis patients in the era of precision medicine. Nature Reviews Nephrology 2020; 16(11): 681-700.

[9] Takkavatakarn K, Nadkarni GN, Roytman M, *et al.* Machine learning models to predict end-stage kidney disease in chronic kidney disease stage 4. BMC Nephrology 2023; 24(1): 132. https://doi.org/10.1186/s12882-023-03424-7

[10] Peng Z, Zhong S, Li X, *et al.* An artificial intelligence model to predict mortality among hemodialysis patients: A retrospective validated cohort study. Scientific Reports 2025; 15(1): 27699. https://doi.org/10.1038/s41598-025-06576-8

[11] Sarnak MJ, Jaber BL. Pulmonary infectious mortality among patients with end-stage renal disease. Chest 2001; 120(6): 1883-1887. https://doi.org/10.1378/chest.120.6.1883

[12] Riehl-Tonn VJ, MacRae JM, Dumanski SM, *et al.* Sex and gender differences in health-related quality of life in individuals initiating haemodialysis. Clinical Kidney Journal 2024; 17(10): sfae273. https://doi.org/10.1093/ckj/sfae273

[13] Soleymanian T, Kokabeh Z, Ramaghi R, *et al.* Clinical outcomes and quality of life in hemodialysis diabetic patients versus non-diabetics. Journal of Nephropathology 2017; 6(2): 81-89. https://doi.org/10.15171/jnp.2017.14

[14] Liu S-X, Wang Z-H, Zhang S, *et al.* The association between dose of hemodialysis and mortality in a prospective cohort study. Scientific Reports 2022; 12: 13708. https://doi.org/10.1038/s41598-022-17943-0

[15] Flythe JE, Xue H, Lynch KE, *et al.* Association of mortality risk with various definitions of intradialytic hypotension. Journal of the American Society of Nephrology 2015; 26(3): 724-734. https://doi.org/10.1681/ASN.2014020222

[16] Noh J, Park SY, Bae W, *et al.* Predicting early mortality in haemodialysis patients: A deep learning approach using a nationwide prospective cohort in South Korea. Scientific Reports 2024; 14: 29658. https://doi.org/10.1038/s41598-024-80900-6

[17] Montemayor VC, Malo AM, Barbieri C. Predicting mortality in hemodialysis patients using machine learning analysis. Clinical Kidney Journal 2020; 14(5): 1388-1395. https://doi.org/10.1093/ckj/sfaa126

[18] Liu S-X, Wang Z-H, Zhang S, *et al.* The association between dose of hemodialysis and mortality in a prospective cohort study. Scientific Reports 2022; 12: 13708. https://doi.org/10.1038/s41598-022-17943-0

[19] Jia W, He W, Chen Z, *et al.* Determinants of dialysis adequacy in maintenance hemodialysis patients: A cross-sectional study on modifiable risk factors and clinical interventions. BMC Nephrology 2025; 26: 369. https://doi.org/10.1186/s12882-025-04278-x

[20] Foley RN, Gilbertson DT, Murray *et al.* Long interdialytic interval and mortality among patients receiving hemodialysis. New England Journal of Medicine 2011; 365(12): 1099-1107. https://doi.org/10.1056/NEJMoa1103313

[21] El Chamieh C, Liabeuf S, Massy Z. Uremic toxins and cardio-vascular risk in chronic kidney disease: What have we learned recently beyond the past findings? Toxins 2022; 14(4): 280. https://doi.org/10.3390/toxins14040280

[22] Collins AJ, Foley RN, Herzog C, *et al.* United States Renal Data System Annual Data Report. Am J Kidney Dis 2019; 73(3): A7-A815.

[23] Bhatti NK, Karimi Galougahi K, Paz, *et al.* Diagnosis and management of cardiovascular disease in advanced and end-stage renal disease. Journal of the American Heart Association 2016; 5(8): e003648. https://doi.org/10.1161/JAHA.116.003648

[24] Thomas MC, Cooper ME, Zimmet P. Diabetic kidney disease. Nature Reviews Disease Primers 2015; 1: 15018. https://doi.org/10.1038/nrdp.2015.18

[25] Babitt JL, Lin HY. Mechanisms of anemia in CKD. New England Journal of Medicine 2012; 367(19): 1901-1911.

[26] Lin Y-C, Chang Y-H, Yang S-Y, *et al.* Update of pathophysiology and management of diabetic kidney disease. Journal of the Formosan Medical Association 2018; 117(8): 662-675. https://doi.org/10.1016/j.jfma.2018.02.007

[27] Pisoni RL, Bragg-Gresham JL, Young EW, *et al.* Anemia management and outcomes in dialysis patients: Results from the Dialysis Outcomes and Practice Patterns Study (DOPPS). American Journal of Kidney Diseases 2018; 71(6): 812-822.

[28] Shoji T, Tsubakihara Y, Fujii M, *et al.* Hemodialysis- associated hypotension as an independent risk factor for two-year mortality in patients on chronic hemodialysis. Kidney International 2019; 95(5): 1212-1220. https://doi.org/10.1111/j.1523-1755.2004.00812.x

[29] Steyerberg EW, Eijkemans MJC, Harrell FE, *et al.* Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods. Statistics in Medicine 2000; 19(8): 1059-1079. https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0

[30] Harrell FE, Jr. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis (2nd ed.). Springer 2015. https://doi.org/10.1007/978-3-319-19425-7

[31] Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives 2014; 28(2): 29-50. https://doi.org/10.1257/jep.28.2.29

[32] Tangri N, Grams ME, Levey AS, *et al.* Multinational assessment of accuracy of equations for predicting risk of kidney failure. JAMA 2016; 315(2): 164-174. https://doi.org/10.1001/jama.2015.18202

[33] KDIGO CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International Supplements 2013; 3(1): 1-150.

[34] KDIGO Diabetes Work Group. KDIGO 2020 clinical practice guideline for diabetes management in chronic kidney disease. Kidney International 2020; 98(4S): S1-S115. https://doi.org/10.1016/j.kint.2020.06.019

[35] National Kidney Foundation. KDOQI clinical practice guideline for hemodialysis adequacy: 2015 update. American Journal of Kidney Diseases 2015; 66(5): 884-930. https://doi.org/10.1053/j.ajkd.2015.07.015

[36] Goldfarb-Rumyantzev AS, Massry SG. Heart-lung-kidney interactions in chronic kidney disease. Seminars in Nephrology 2019; 39(3): 263-275.

[37] Vanholder R, Van Biesen W, Lameire N. What is the renal-cardio-pulmonary syndrome? Nature Reviews Nephrology 2020; 16(12): 707-723.

[38] McIntyre CW, Odudu A, Eldehni MT, *et al.* Induced cardiac injury by hemodialysis: Time to turn down the dial. Journal of the American Society of Nephrology 2017; 28(10): 2839-2848.

[39] Agarwal R, Flynn J, Pogue V, *et al.* Assessment and management of hypertension in patients on dialysis. Journal of the American Society of Nephrology 2018; 29(4): 893-905.

[40] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine 2019; 380(14): 1347-1358. https://doi.org/10.1056/NEJMra1814259

[41] Deo RC. Machine learning in medicine. Circulation 2015; 132(20): 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[42] Sood MM, Tangri N, Hiebert B, *et al.* Pulmonary disease and mortality in patients with chronic kidney disease: A population-based cohort study. Kidney International 2017; 92(2): 467-476.

[43] Ravi V, Singh SK, Yadav CB. To Identify the Predictors of Mortality in Renal Patients Undergoing Dialysis. International Journal of Statistics in Medical Research 2025; 14: 755-764. https://doi.org/10.6000/1929-6029.2025.14.68

[44] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996; 58(1): 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[45] Breiman L. Random forests. Machine Learning 2001; 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[46] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 2001; 29(5): 1189-1232. https://doi.org/10.1214/aos/1013203451

Received on 06-12-2025 Accepted on 04-01-2026 Published on 30-01-2026

https://doi.org/10.6000/1929-6029.2026.15.02