PREDICTION AND KEY CHARACTERISTICS OF ALL-CAUSE MORTALITY IN MAINTENANCE HEMODIALYSIS PATIENTS

Mu Xiangwei[1], Zhu Mingjie[2], Liu Shuxin[2], Li Kequan[1],
You Lianlian[2] and Che Shuang[2]

[1]School of Maritime Economics and Management,
Dalian Maritime University, Dalian 116026, Liaoning, China
[2]Dalian Key Laboratory of Intelligent Blood Purification, Dalian Municipal
Central Hospital affiliated with Dalian Medical University,
Dalian 116033, Liaoning, China

ABSTRACT

Predict and analyze key features of all-cause death in maintenance hemodialysis patients to provide guidance for later diagnosis and treatment. Four machine learning methods were used to establish an all-cause death prediction model for maintenance hemodialysis patients and compare their performance. Analyze the key characteristics that have an important impact on all-cause death, and conduct user portraits for patients of different ages and genders. After comparison, the random forest algorithm works best, and an important factor affecting the all-cause death of patients is obtained. Among them, the all-cause death of all patients is related to factors such as albumin, blood potassium, blood magnesium, and urea; With age, the importance of factors such as blood sodium and phosphorus increases, and the importance of factors such as cardiac ultrasound ejection fraction decreases. Finally, there were also differences in the importance of analyzing patients of different ages and different sexes affecting their all-cause death. It is useful for residents to adjust their dialysis index timely.

KEYWORDS

Maintenance Hemodialysis, All-cause Mortality, Randomized Forest, Feature Importance, Prognosis.

1. INTRODUCTION

The growing number of patients with end-stage renal disease plagues global public health, and hemodialysis is one of the more important treatments, the rationalization of which contributes to the prognosis of patients. According to statistics, the number of patients with end-stage renal disease requiring hemodialysis treatment is estimated to be between 4.9 million and 7.08 million worldwide [1].

Although China has invested heavily in the treatment of end-stage renal disease and the reimbursement rate for dialysis treatment has received a large increase from 50% to 80%, the mortality rate of maintenance dialysis patients is still at a high level and their all-cause mortality rate is much higher than that of the normal population. Studies in Europe and the United States and other countries have shown that the mortality rate within 1 year of dialysis is about 20% [2]. Meanwhile, the mortality rate of maintenance hemodialysis patients in the United States has risen

David C. Wyld et al. (Eds): SIP, CRBL, WiMNeT, BIOM, EDUPT, NLDM, SOFTFM, COMIT - 2022
pp. 97-107, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121710

since the outbreak of the new crown epidemic in 2020, with a mortality rate of 37% for patients receiving dialysis treatment in 2020 [3]. There are many factors influencing the mortality rate of dialysis patients, and dialysis equipment cannot analyze the past of dialysis patients to adjust the dialysis dose according to past data, and machine learning can precisely address this deficiency. The aim of this study was to investigate all-cause mortality in maintenance hemodialysis patients and its associated influencing factors. The analysis yielded important influencing factors that could provide guidance for patient prognosis.

2. MATERIALS AND METHODS

2.1. Data Source and Processing

2.1.1. Data Source

Data for this study were obtained from a large hemodialysis center in northeastern China, and data were recorded for 5 years for 1052 patients, including basic patient information such as gender, age, primary disease, and age on dialysis, baseline data such as cardiac ultrasound ejection fraction, reduced diastolic function, and heart failure, and blood chemistries including creatinine, albumin, potassium, sodium, and calcium, which were recorded quarterly. Due to the presence of survival, death, and departure data in the dataset, where the departure data are the result of patients being transferred to a hospital or undergoing renal transplantation, etc., it is not instructive for studying all-cause mortality. Therefore, this part of the data was chosen to be excluded from this paper. The study was approved by the hospital ethics committee, and all patients gave informed consent and signed an extensive informed consent form.

2.1.2. Data Processing

In this paper, we introduce a padding method using missing values of adjacent data points - K-Nearest Neighbor algorithm [4-6], which identifies adjacent points by measuring the distance between points, estimates and fills missing values, and is robust to noise.

The complete samples in dataset D are denoted as D_i and the samples containing missing values are denoted as D_a. The complete samples are clustered and the Euclidean distance is used to calculate the clustering similarity between dataset D_i and D_a. The distance between samples $X_i(X_{i1}, X_{i2}, \cdots, X_{in})$ and $X_j(X_{j1}, X_{j2}, \cdots, X_{jn})$ is denoted as:

$$dist = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jK})^2} \qquad (1)$$

In the end, there were 9,992 processed data. Forty-six features were incorporated, as shown in Figure 1.

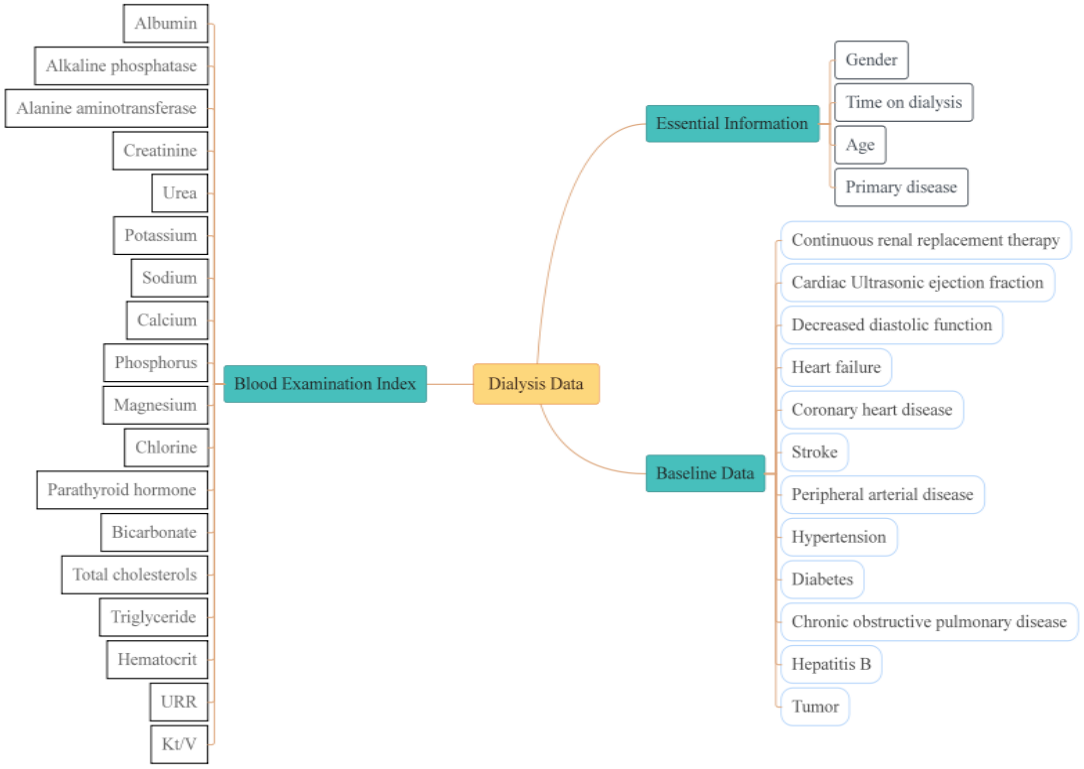

Figure 1. Dialysis data characteristics

Among them, the primary diseases include hypertensive nephropathy, diabetic nephropathy, chronic interstitial nephritis, hypertensive benign renal arteriosclerosis and polycystic kidney. Sodium, magnesium, and potassium plasma were included in the pre-dialysis and post-dialysis data.

2.2. Methods

In this paper, four machine learning algorithms-random forest algorithm, support vector machine algorithm, Adaboost algorithm, and logistic regression algorithm are used to construct a prediction model for all-cause mortality in maintenance hemodialysis patients. Due to the complexity of dialysis treatment, the algorithms with specific formulas such as K-NN and naïve bayes cannot describe them in detail, so the above algorithm was selected to establish an all-cause death model for maintenance hemodialysis patients, At the same time, the effect of different age groups and patient gender on mortality, and the user portrait.

2.2.1. Methods

(1) Random forest algorithm

Random forest is a supervised learning algorithm that consists of a series of independent decision trees trained by multiple Bagging integrated learning techniques for regression and classification. The variability between models is increased by constructing different training sets; the implementation of this algorithm obtains a sequence of classification models by training k rounds of initial data and uses this sequence of classification models to form a multi-classification system,

and finally, the final classification results of the system use simple majority voting method with the following classification decisions [7]:

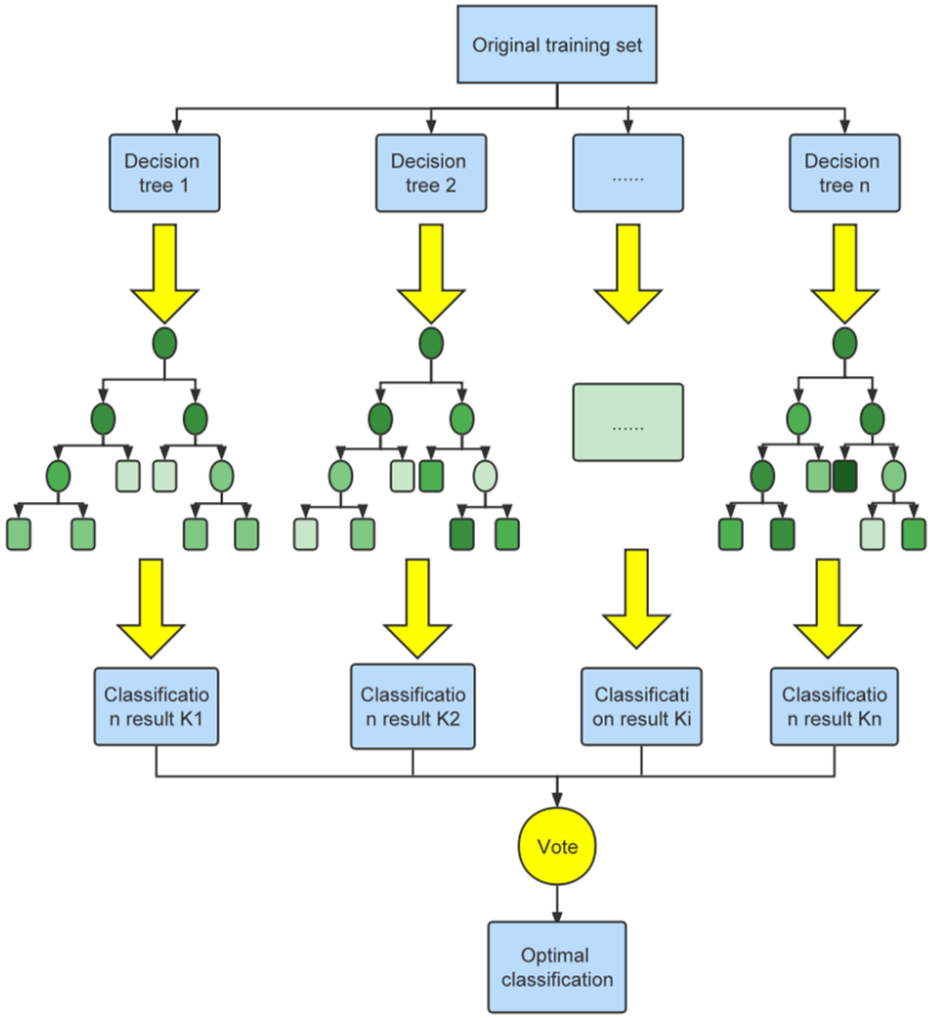

Figure 2. Schematic diagram of random forest

(2) Support vector machine algorithm

Support vector machine [8] is a classification model and the algorithm solves the binary classification problem by finding the optimal hyperplane. The optimal hyperplane is a multidimensional plane which finds the maximum interval of the binary classification problem and the solution of the algorithm can be expressed as the following problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2}$$

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, \cdots, N \tag{3}$$

(3) Adaboost algorithm

Similar to the random forest algorithm, the principle of the Adaboost algorithm [9] is to combine multiple if classifiers by certain methods to form a strong classifier. The algorithm is implemented by an iterative method. In the implementation of the algorithm, one weak classifier is trained at a time, and in the next training, the last trained weak classifier is used for iteration, and the implementation process of the algorithm is shown in Figure 3.

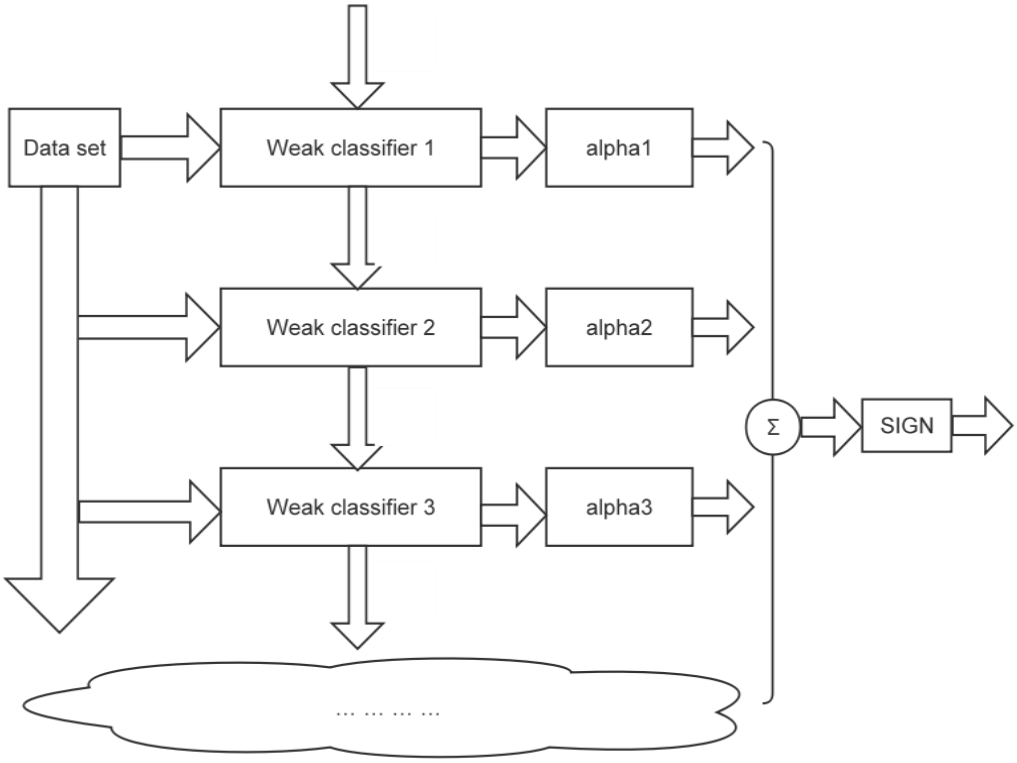

Figure 3. Schematic diagram of Adaboost algorithm

(4) Logistic regression algorithm

Similar to linear regression, the logistic regression algorithm [10] is a classical binary classification algorithm, and the algorithm is implemented by adding a Sigmoid function to the results of linear regression, transforming the linear regression results into results between 0 and 1.

2.2.2. Construction of Machine Learning Models

The 9,992 cases of data are divided into training set and test set according to the ratio of 7:3, and 70% of the data are used for training to build the best model, and the remaining 30% of the data are used as test set for testing for model evaluation and result output.
For model building, this paper introduces the 10-fold cross-validation method. The 10-fold cross-validation method is to divide the data set into ten parts, and select nine data in turn for training and one data for testing, and finally take the mean value of 10 test results to evaluate the performance of each model, and the specific results are shown in Table 2.

2.2.3. Evaluation of the Model

In this paper, the confusion matrix is used to evaluate the trained models, as shown in Table 1.

Table 1. Confusion matrix.

Confusion matrix		**True Value**	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

Based on the confusion matrix, the model evaluation metrics: accuracy, precision, recall, F1-score and ROC curve can be calculated.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

F1-score:

$$\text{F1} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

3. RESULTS

3.1. Comparison of Model Performance

The four models were trained to obtain the confusion matrix. As shown in Table 2, the random forest algorithm is greater than the remaining three algorithms in terms of accuracy, precision, recall, F1-score and AUC score, and all four metrics are above 90%. Collectively, the random forest algorithm has the best performance among the four algorithms.

Table 2. Results of model predictions.

model	**accuracy**	**precision**	**recall**	**F1-score**	**AUC**
Random forest algorithm	90%	90%	99%	93.50%	99%
Adaboost algorithm	84%	83%	84%	83%	87.9%
Support vector machine algorithm	79%	83%	79%	70%	75.8%
Logistic regression algorithm	79%	75%	79%	74%	77.3%

3.2. Feature Importance of Random Forest and Adaboost Models

In the training of Random Forest with Adaboost model, the importance of each influencing factor (feature) was obtained, and it was ranked to extract the features that have a greater impact on all-cause mortality, and the results are shown in Figure 4.

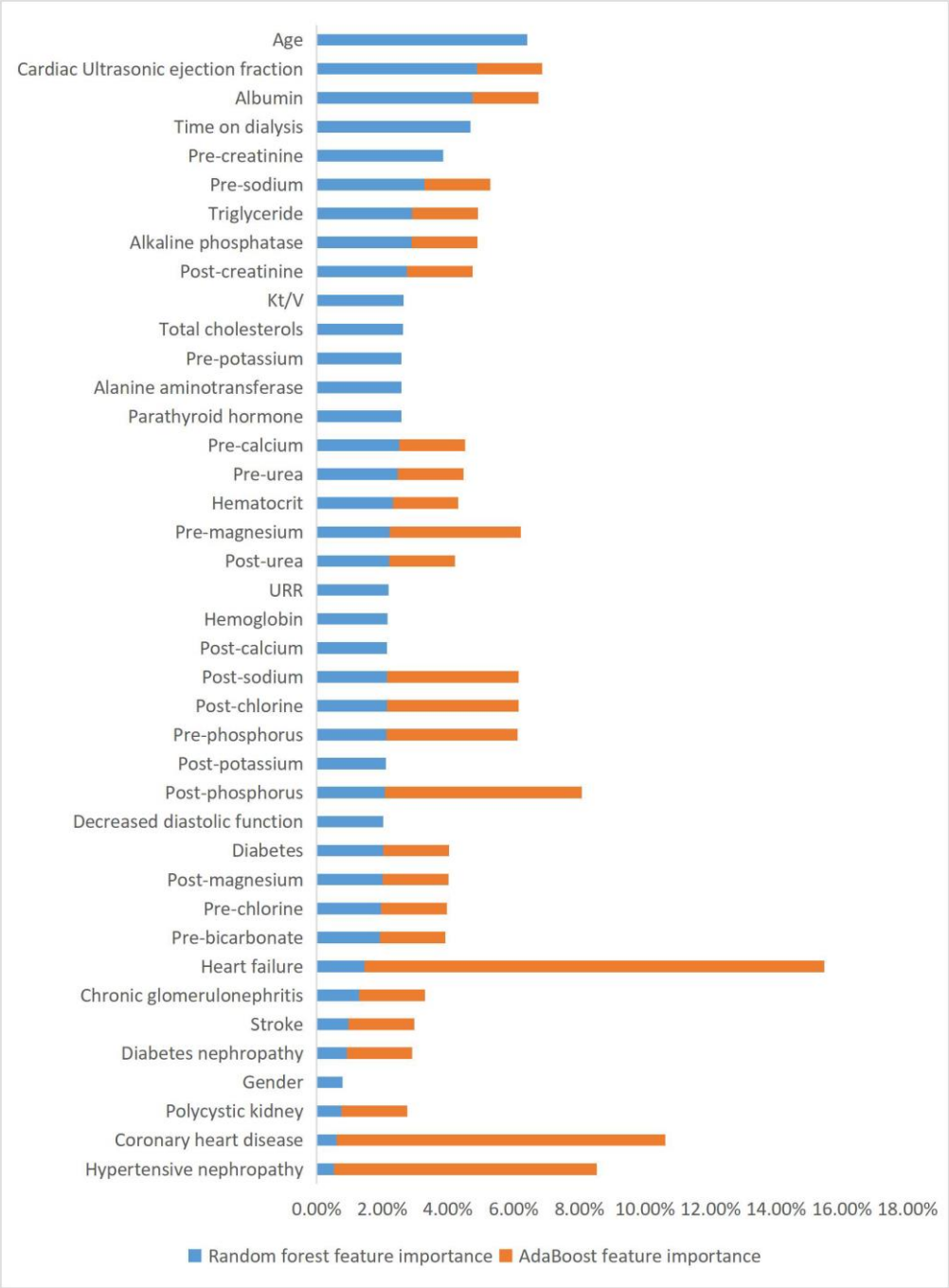

Figure 4. Ranking of feature importance

3.3. Further Analysis of the Random Forest Algorithm

The patient groups were divided according to age groups [11] into group A (<40), group B (40-59), group C (60-79), and group D (>=80). A random forest model was developed for the data of each group of patients separately, and a large difference in the importance of 16 characteristics, including dialysis age, cardiac ultrasound ejection fraction, and sodium, was found in the different age groups of patients, as shown in Figure 5.

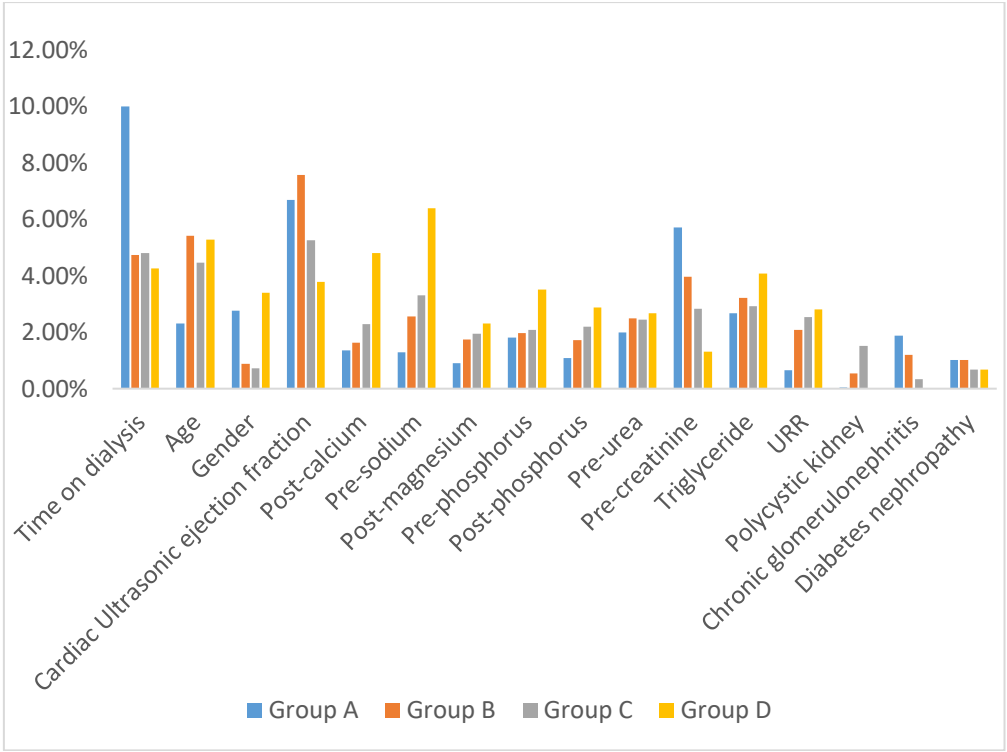

Figure 5. Fluctuations in the importance of age-segmented characteristics

For all patients, the weight of the effect of gender on all-cause mortality was relatively small, but after grouping by age group, the effect of gender characteristics on all-cause mortality was found to be higher in groups A and D. Therefore, the gender of patients in groups A and D was divided to explore the differences between factors influencing all-cause mortality in patients of different genders under this age group, and certain characteristics with greater variability were derived, as shown in Table 3.

Table 3. Gender classification assessment weights/% for groups A and D.

Features	**Group A**		**Group D**	
	Male	Female	Male	Female
Time on dialysis	16.61%	5.82%	17.00%	5.41%
Age	3.83%	2.76%	7.06%	4.74%
Cardiac Ultrasonic ejection fraction	2.31%	5.89%	0.64%	3.22%
Pre-calcium	1.23%	2.67%	1.69%	1.83%
Post-calcium	1.33%	1.58%	2.33%	2.48%
Pre-sodium	1.68%	1.90%	2.28%	3.52%
Post-sodium	1.31%	0.49%	1.33%	1.24%
Pre-magnesium	1.56%	0.51%	5.70%	1.03%
Total cholesterols	3.50%	1.46%	2.97%	2.19%
Parathyroid hormone	4.04%	0.94%	2.31%	1.31%
Alanine aminotransferase	1.95%	0.85%	5.19%	2.20%
Kt/v	1.49%	0.95%	1.25%	0.79%

4. CONCLUSIONS

Using four algorithms, all-cause mortality in maintenance hemodialysis patients was focused on, and the Random Forest algorithm was found to have the best performance among the four algorithms, while for all patients, the features with higher impact on all-cause mortality were derived using the Random Forest algorithm and Adaboost algorithm, as shown in Figure 5; finally, a gender- and age-specific user portrait of patients was conducted using the Random Forest algorithm for further analysis, with the following results:

Findings from a study of all patients: Random forest and Adaboost algorithms were used for the analysis of the importance of characteristics of all patients, and both reached consistent conclusions in the analysis of the importance of most characteristics, and the experimental results showed that in the patient population, the characteristics that had a greater impact on all-cause mortality were age on dialysis, age, albumin, potassium, cardiac ultrasound ejection fraction, calcium, sodium, potassium, urea, and alkaline phosphatase (alp). For the above findings, it has been found that age, diabetes, calcium, albumin, sodium, potassium, and alkaline phosphatase are associated with all-cause mortality in patients. According to a related study, malnutrition is one of the complications of maintenance hemodialysis patients and affects their prognosis, and serum albumin level is an important factor in determining the nutritional status of patients. Therefore, the study of serum albumin levels has significance for the prognosis of patients. It has been found,that alkaline phosphatase has a correlation with cardiac lesions and is its independent risk factor, which affects the survival rate of patients. The important features derived from the training can be focused on for further studies.

Conclusions from patients by age group: When the patient population was divided into four groups according to age, the importance of certain features fluctuated widely across age groups. Features that showed an overall increasing trend in importance with increasing age were: post-calcium, pre-sodium, post-magnesium, pre-phosphorus, post-phosphorus, and URR. The features that showed an overall decreasing trend in importance with decreasing age were: cardiac ultrasound ejection fraction, pre-creatinine, and chronic glomerulonephritis in primary disease.

Conclusions from grouping patients in groups A and D according to gender: Certain characteristics were analyzed in the two groups of patients with a large difference in their importance. Among them, in groups A and D, the importance of characteristics such as cardiac ultrasound ejection fraction, pre-calcium, pre-sodium, and so on were significantly higher in female patients than in male patients under this age group; while for characteristics such as dialysis age, age, pre-magnesium, total cholesterol, parathyroid hormone, alanine transferase, and Kt/v were higher in males than in females, as shown in Table 4. Therefore, in patients under this age group, dialysis should be performed with differentiation according to gender Treating, for different index values with different characteristics in different genders, dialysis should be reasonably arranged and timely interventions should be made to achieve the goal of reducing all-cause mortality in patients.

In this paper, the importance ranking and analysis of the characteristics related to all-cause mortality of all patients as well as maintenance hemodialysis patients of different ages and genders were performed, but the classification of patients was not detailed enough. Therefore, a more in-depth analysis will be conducted in the future to provide a more detailed user profile of patients, to identify patients accurately and efficiently, and to improve diagnostic efficiency.

ACKNOWLEDGEMENTS

This work was supported by Dalian Maritime University Undergraduate Education and Teaching Reform Research Project 2022-63, Humanities and Social Sciences Research Project of the Ministry of Education 18YJC630124, Liaoning Provincial Department of Education Science and Technology Research Project L2014203, Liaoning Provincial Social Science Planning Fund Project L14BGL012. We thank Dalian Municipal Central Hospital for the help with the clinical knowledge related to the extraction of our dataset and for providing clinical insights on the case study.

REFERENCES

[1] Thomas M. Cover, and Peter E. Hart."Nearest neighbor pattern classification.." IEEE Trans. Information Theory 13.1(1967).
[2] Tadashi Yamamoto, et al."Predialysis and Postdialysis pH and Bicarbonate and Risk of All-C ause and Cardiovascular Mortality in Long-term Hemodialysis Patients." American Journal of Kidney Diseases 66.3(2015).
[3] Lu Jiayue, et al."The relationship between survival rate and intradialytic blood pressure changes in maintenance hemodialysis patients.." Renal failure 39.1(2017).
[4] Viknesh Selvarajah, et al."Pre-dialysis systolic blood pressure-variability is independently associated with all-cause mortality in incident haemodialysis patients.." PLoS ONE 9.1(2017).
[5] Lv Ji-Cheng and Zhang Lu-Xia."Prevalence and Disease Burden of Chronic Kidney Disease.." Advances in experimental medicine and biology 1165.(2019).
[6] Mei Wang, et al."SKNN Algorithm for Filling Missing Oil Data Based on KNN." IOP Conference Series: Materials Science and Engineering 612.3(2019).
[7] Wenbo Wu, et al."A Cross-Sectional Machine Learning Approach for Hedge Fund Return Prediction and Selection." Management Science 67.7(2020). doi:10.1287/mnsc.2020.3696.
[8] "Cancer; Researchers from Department of Computer Sciences and Engineering Provide Details of New Studies and Findings in the Area of Cancer (Detecting biomarkers from microarray data using distributed correlation based gene selection)." Computers, Networks & Communications. (2020).
[9] Chen Y W, et al."[Early mortality and risk analysis in adult patients with maintenance hemodialysis].." Zhonghua nei ke za zhi 60.1(2021).
[10] Battisti Rodrigo, et al."Machine learning modeling and genetic algorithm-based optimization of a novel pilot-scale thermosyphon-assisted falling film distillation unit." Separation and Purification Technology 259.(2021). doi:10.1016/J.SEPPUR.2020.118122.
[11] Johansen Kirsten L., et al."US Renal Data System 2020 Annual Data Report: Epidemiology of Kidney Disease in the United States." American Journal of Kidney Diseases 77.4S1(2021).
[12] Theodorakopoulou Marieta, et al."SEX DIFFERENCES IN AMBULATORY BLOOD PRESSURE TRAJECTORIES AND BLOOD PRESSURE VARIABILITY IN HEMODIALYSIS PATIENTS." Journal of Hypertension 40.Suppl 1(2022).
[13] Bridle Tristen G, et al."Physiologically relevant hCys concentrations mobilize MeHg from rabbit serum albumin to form MeHg-hCys complexes.." Metallomics : integrated biometal science 14.3(2022).
[14] Patel Harsh, et al."Oropharyngeal cancer patient stratification using random forest based-learning over high-dimensional radiomic features.." Scientific reports 11.1(2021).
[15] Musbah Hmeda, Aly Hamed H.,and Little Timothy A.."Energy management of hybrid energy system sources based on machine learning classification algorithms." Electric Power Systems Research 199.(2021). doi:10.1016/J.EPSR.2021.107436.
[16] Ren Jiaolong, et al."Design optimization of cement grouting material based on adaptive boosting algorithm and simplicial homology global optimization." Journal of Building Engineering 49.(2022). [4]Wang Xingcheng, Meng Cheng,and Wang Yuanyu."Insight for the construction of R-T phase boundary in KNN piezoceramics from the view of energy band structure and electron density." Ceramics International 47.20(2021).
[17] Kathuria, Charu, Mehrotra, Deepti,and Misra, Navnit Kumar."A novel random forest approach to predict phase transition." International Journal of System Assurance Engineering and Management .prepublish(2021).

[18] Jiang Chengcheng, et al."Spatial modeling of gully head erosion on the Loess Plateau using a certainty factor and random forest model." Science of the Total Environment 783.(2021).

[19] Ali Ahmad Hassan, et al."A Model Incorporating Serum Alkaline Phosphatase for Prediction of Liver Fibrosis in Adults with Obesity and Nonalcoholic Fatty Liver Disease.." Journal of clinical medicine 10.15(2021).

AUTHORS

Mu Xiangwei, Associate Professor at Dalian Maritime University; His research focuses on machine learning and medical big data.