Heliyon 10 (2024) e36878

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Predicting all-cause mortality and premature death using interpretable machine learning among a middle-aged and elderly Chinese population

Qi Yu, Lingzhi Zhang, Qian Ma, Lijuan Da, Jiahui Li, Wenyuan Li *

Center of Clinical Big Data and Analytics of The Second Affiliated Hospital and Department of Big Data in Health Science School of Public Health, Zhejiang University School of Medicine, Hangzhou, 310058, Zhejiang, China

ARTICLE INFO

Keywords:
Machine learning
Epidemiology
All-cause mortality
Cohort study
Prediction model

ABSTRACT

Objective: To develop machine learning-based prediction models for all-cause and premature mortality among the middle-aged and elderly population in China.
Method: Adults aged 45 years or older at baseline of 2011 from the China Health and Retirement Longitudinal Study (CHARLS) were included. The stacked ensemble model was built utilizing five selected machine learning algorithms. These models underwent training and testing using the CHARLS 2011–2015 cohort (derivation cohort) and subsequently underwent external validation using the CHARLS 2015–2018 cohort (validation cohort). SHapley Additive exPlanations (SHAP) was introduced to quantify the importance of risk factors and explain machine learning algorithms.
Result: In derivation cohort, a total of 10,677 subjects were included, 478 died during the follow-up. The stacked ensemble model demonstrated the highest efficacy in terms of its discrimination capability for predicting all-cause mortality and premature death, with an AUC[95 % CI] of 0.826 [0.792–0.859] and 0.773[0.725–0.821], respectively. In validation cohort, the corresponding AUC[95 % CI] were 0.803[0.743–0.864] and 0.791[0.719–0.863], respectively. Risk factors including age, sex, self-reported health, activities of daily living, cognitive function, ever smoker, levels of systolic blood pressure, Cystatin C and low density lipoprotein were strong predictors for both all-cause mortality and premature death.
Conclusion: Stacked ensemble models performed well in predicting all-cause and premature death in this Chinese cohort. Interpretable techniques can aid in identifying significant risk factors and non-linear relationships between predictors and mortality.

1. Introduction

Life expectancy and mortality are key indicators of public health. The seventh national census in 2020 revealed that the percentage of individuals aged 60 and above in China is approximately 18.70 %. The confluence of an aging population and evolving lifestyles has yielded a perceptible rise in the incidence of chronic illnesses, subsequently impacting mortality rates [1,2]. Given the changing circumstances, it became paramount to precisely discern risk stratification and implement timely interventions [3,4].

* Corresponding author. Department of Big Data in Health Science School of Public Health, and Center of Clinical Big Data and Analytics of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China.

E-mail address: wenyuanli@zju.edu.cn (W. Li).

https://doi.org/10.1016/j.heliyon.2024.e36878
Received 4 February 2024; Received in revised form 15 August 2024; Accepted 23 August 2024

Extensive prior studies have endeavored to explore the intricate relationships between specific predictors and death. These factors encompassed a range of aspects, including physical activity, alcohol consumption, and self-reported health status [5–7]. Furthermore, some studies have aimed to amalgamate risk factors from diverse domains to enhance predictive accuracy. For instance, Ganna and Ingelsson [8] constructed Cox regression models, utilizing 13 risk factors for men and 11 for women. Their approach yielded predictions of 5-year all-cause mortality, with area under the curve (AUC) of 0.8 (95 % CI 0.77–0.83) and 0.79 (95 % CI 0.76–0.83), respectively. Recently, the utilization of machine learning techniques has been increasing in constructing prediction models [9–13].

In contrast to traditional regression models, machine learning models offer an alternative approach to building prediction models, and may fit the non-linear interactions between variables by minimizing errors between predicted and observed labels. Previous research have used machine learning models to predict death in different databases [14–19]. Notably, numerous studies have harnessed diverse machine learning algorithms to forecast mortality using various databases. For instance, a prospective cohort study [18] conducted in the UK employed ways such as random forest (RF), deep learning and Cox regression to predict premature death. Remarkably, RF (AUC 0.783, 95 % CI 0.776–0.791) and deep learning (AUC 0.790, 95 % CI 0.783–0.797) exhibited superior predictive performance in comparison to standard regression methods (AUC 0.751, 95 % CI 0.748–0.767). In another study [14] using data from 15,933 patients who newly visited the cardiovascular institute, five machine learning methods were constructed to predict death, and support vector machine (SVM) algorithm had the best AUC (0.900) among all machine learning algorithms.

However, most of the studies predicting mortality were based on a single machine learning classifier. This framework may underestimate model uncertainty and lead to overconfidence in prediction model performance. To mitigate this limitation, the approach of ensemble learning was the method to enhance model performance by integrating multiple classifiers to solve the same problem [20–22]. The ensemble technique usually reduced the variance in the model, and leads to the most possible stable models with the best performance. Ensemble learning has exhibited its effectiveness across a spectrum of applications, including cognitive impairment prediction [9], cancer prediction [23], heart disease detection [24] and so on. Furthermore, in addition to building a prediction model using ensemble machine learning, we also used the SHapley Additive exPlanations (SHAP) method, which provides various interpretations according to the contribution of each predictor in the algorithm. The SHAP method has been used in many research to interpret machine learning models [25,26].

The China Health and Retirement Longitudinal Study (CHARLS) cohort [27] was an nationally representative cohort. Our study used these data to build risk prediction models with stacking ensemble technique, and identified top predictors by explaining complex machine learning algorithms with Shapley values. Our study further explored non-linear relationships of biomarkers or physical examination factors with the risk of mortality.

2. Method

2.1. Study sample

The CHARLS study started in 2011, with subsequent follow-up visits conducted in 2013, 2015, and 2018. The probability proportional to scale sampling method was performed to select a representative sample of Chinese individuals. Medical history and physical examination were collected in each wave, and biomarkers were collected in 2011 and 2015 waves. Details information about study design and cohort profile have been published elsewhere [27].

In our study, derivation cohort were derived from the 2011–2015 CHARLS cycles (n = 17,708). After excluding participants at baseline who did not participate in the later follow-up survey (n = 1417), those who did not participate in blood measure at baseline (n

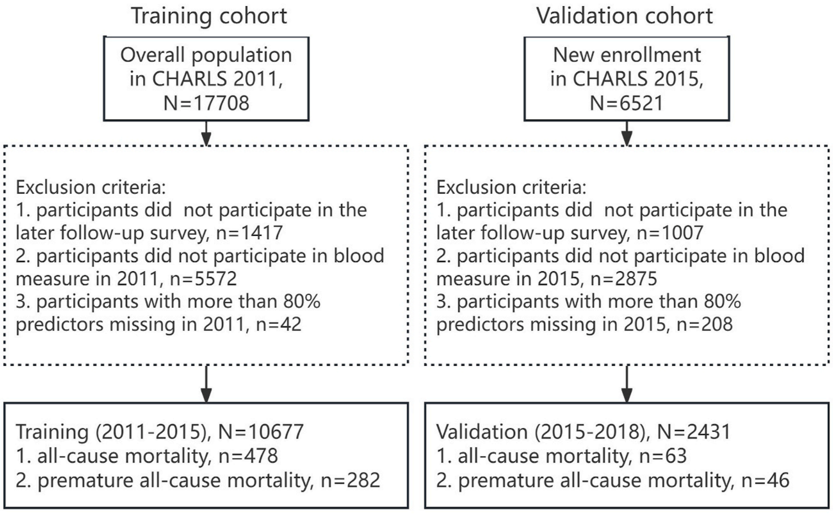
Training cohort
Validation cohort
Overall population in CHARLS 2011, N=17708
New enrollment in CHARLS 2015, N=6521
Exclusion criteria:
1. participants did not participate in the later follow-up survey, n=1417
2. participants did not participate in blood measure in 2011, n=5572
3. participants with more than 80% predictors missing in 2011, n=42
Exclusion criteria:
1. participants did not participate in the later follow-up survey, n=1007
2. participants did not participate in blood measure in 2015, n=2875
3. participants with more than 80% predictors missing in 2015, n=208
Training (2011-2015), N=10677
1. all-cause mortality, n=478
2. premature all-cause mortality, n=282
Validation (2015-2018), N=2431
1. all-cause mortality, n=63
2. premature all-cause mortality, n=46

Fig. 1. Subject selection flowchart.

= 5572), and individuals with more than 80 % predictors missing at baseline (n = 42). Finally, 10,677 subjects were available for analysis (Fig. 1).

2.2. Data collection

The CHARLS questionnaire included demographic information, lifestyle factors, medical conditions, physical examination, health indicators and biochemical indicators. Systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse were measured thrice, and the average of three measurements was calculated. Health indicators included self-reported health, activities of daily living (ADL), cognitive function and CES-D score. Self-reported health was answered in five categories: very good, good, fair, poor, and very poor. The ADL included six indexes: bathing, dressing, eating, getting in/out of bed, using the toilet, and controlling urination. Individuals who reported some difficulty or could not do the activity were scored 1 point, and those without difficulty were scored 0 point (total ADL score ranged from 0 to 6). In CHARLS 2011, the 10-item version of the CES-D [28] was applied to investigate condition of depression (total score ranged from 0 to 30), and lower sum scores indicate better mental status. Cognitive performance was calculated using three cognitive tests including Telephone Interview of Cognitive Status (TICS) [29,30] (total score ranged from 0 to 10), figure drawing test (total score ranged from 0 to 1), and word recall (total score ranged from 0 to 20) [31,32]. Cognitive performance was the sum score of TICS, figure drawing test and word recall test (ranging from 0 to 31), and higher sum scores indicate better cognitive function. Finally, after excluding variables with more than 30 % missing, 49 variables were included as potential predictors in this study (Table S1 in Supplementary material).

2.3. Outcome ascertainment

The date of death was documented as occurring between the date of the last survey and the date of the survey in which the death was reported (the midpoint between the 2 surveys). We then calculated the survival time as the interval between the first survey and the time of death. In our study, premature mortality was defined based on the average life expectancy in China in 2011, with men dying before 72.7 years and women dying before 76.9 years [33].

2.4. Statistical analysis

Continuous predictors of derivation cohort were depicted as mean ± standard deviation, while categorical predictors of derivation cohort were depicted as number (percentage). Between group differences were performed using ANOVA for continuous predictors and Chi-square test for categorical predictors.

- Data splitting

The CHARLS 2011–2015 cohort was utilized as the derivation cohort, while the CHARLS 2015–2018 cohort was utilized as the validation cohort. We randomly split the derivation cohort into train and test datasets at a 7:3 ratio.

- Data processing

The RF imputation algorithm was developed using only the training set [34]. Imputation was conducted separately for the train and test datasets, utilizing the imputation method derived from the training dataset. This approach ensures that the testing set remains unaffected during the imputation process. Categorical predictors were transformed into binary parameters using one-hot encoding, and numerical predictors were transformed using Z-Score standardization. Subjects who died or survived during the follow-up were considered as positive or negative labels, respectively.

- Feature selection

We used LASSO regression to reduce the number of predictors using the L1 norm to penalize the regression coefficients, some of which shrunk to zero [35]. Variable selection was based solely on the training set without any information from the testing set. This prevents any knowledge about the testing data from influencing the variable selection process, thus maintaining the independence of the testing set. After that, predictors with non-zero regression coefficients (N = 40) were used in both the train and test datasets during the modelling process.

- Model construction and evaluation

A 5-fold cross-validation was conducted on the training set to acquire optimal hyper-parameters values based on the AUC [36]. Seven commonly used classifiers were initially applied, and five classifiers (RF, SVM, logistic regression (LR), light gradient boosting machine (LightGBM), and extra trees (ET)) were selected based on the performance of AUC. Models were trained and evaluated on training set and testing set (derivation cohort), and externally evaluated on evaluation cohort. Metrics including accuracy, sensitivity, specificity, AUC and Brier score were utilized to evaluate the model performance. Delong's test was conducted to assess the difference in AUCs between fusion model and base algorithms [37].

The stacked ensemble modelling strategy was used to improve the model performance (Fig. S1). The stacking technique mainly consists of two layers: base models and meta models. The predictions of base-level models were selected as the input of meta-level models [21]. In our study, five base models (LR, RF, SVM, LightGBM and ET) were constructed to obtain a new training set for the meta-level classifier. Among this training set, each base model provides a prediction probability over the outcome values. The LR classifier was used to reduce the complexity of the model at the meta-level. We also assessed the stacked ensemble model for all-cause mortality across age and sex.

In our derivation cohort, only 478 participants (4.48 %) died during follow-up.

Given the imbalanced nature of our data, we applied the "class_weight" hyperparameter [38]. The "class_weight" was set to "balanced," which means that the model automatically adjusts the weights of the classes inversely proportional to their frequencies. This method assigns higher weights to the minority class and lower weights to the majority class, enabling the model to prioritize the minority class during the training process.

SHAP was utilized to evaluate the importance of the various predictors [39]. SHAP was based on cooperative game theory and computes Shapley values to attribute the model's output to different features. By using SHAP values, we can understand how each predictor contributes to the model's output, thereby gaining valuable insights into the model's decision-making process. SHAP summary plot were generated to observe the importance of each predictor. We also conducted stratified analyses using SHAP summary plot to explore whether the meaningful risk factors of mortality varied by age of 70 years old or sex. To explore the non-linear relationships of biomarkers or physical examination factors with the risk of mortality, we reconstructed the RF model to fit the raw data (without data scaling), and generated a SHAP dependency plot. Tree-based models, such as decision trees and random forests, typically do not require scaling of numerical features. The SHAP dependency plot can facilitate the exploration of the association between the raw variables and outcomes.

In our study, two-sided P-values <0.05 were deemed statistically significant. Construction and interpretation of the machine learning models were conducted using *scikit-learn* and *shap* package in Python (3.6.5). We utilized R software (4.0.2) for additional statistical analyses.

Table 1
Selected baseline characteristics of participants in our study sample.

Characteristics	Survival (n = 10199)	Died (n = 478)	P-value
Age(year)	58.68 ± 9.09	68.40 ± 10.52	<0.001
Sex			
Women	5435(53.29 %)	195(40.79 %)	<0.001
Men	4764(46.71 %)	283(59.21 %)	
Education level			
Primary school and below	9169(89.9 %)	453(94.77 %)	0.002
Middle school degree	903(8.85 %)	20(4.18 %)	
College degree and above	127(1.25 %)	5(1.05 %)	
Marital status			
married	10118(99.21 %)	471(98.54 %)	0.113
single	81(0.79 %)	7(1.46 %)	
Site			
urban	3609(35.39 %)	142(29.71 %)	0.011
rural	6590(64.61 %)	336(70.29 %)	
Ever drinker			
no	6222(61.01 %)	263(55.02 %)	0.008
yes	3966(38.89 %)	215(44.98 %)	
Current drinker			
no	6813(66.8 %)	329(68.83 %)	0.363
yes	3382(33.16 %)	149(31.17 %)	
Ever smoker			
no	6225(61.04 %)	223(46.65 %)	<0.001
yes	3972(38.94 %)	255(53.35 %)	
Current smoker			
no	7048(69.1 %)	305(63.81 %)	0.191
yes	2970(29.12 %)	147(30.75 %)	
BMI(kg/m^2)	24.11 ± 28.72	22.10 ± 3.72	<0.001
SBP(mmHg)	129.23 ± 21.26	138.60 ± 24.12	<0.001
DBP(mmHg)	75.29 ± 12.15	76.67 ± 13.55	0.065
Waist(cm)	84.38 ± 12.57	83.36 ± 12.36	0.064
Weight(kg)	58.88 ± 11.63	54.90 ± 11.64	<0.001
Self-reported health			
very good	567(5.56 %)	12(2.51 %)	<0.001
good	1657(16.25 %)	49(10.25 %)	
fair	5107(50.07 %)	184(38.49 %)	
poor	2400(23.53 %)	173(36.19 %)	
very poor	463(4.54 %)	60(12.55 %)	
Cognitive function	13.00 ± 5.81	9.77 ± 6.12	<0.001
CESD score	8.54 ± 6.35	10.17 ± 6.98	<0.001

3. Result

3.1. Baseline characteristics

In derivation cohort, a total of 10,677 subjects aged 45 or older were participated in the baseline wave of 2011, and 478 subjects died, including 282 premature death. Participants who died during the follow-up tended to be older, male, single, with low education level, with poor cognitive function and self-reported health status, and more likely to smoke or drink alcohol (Table 1).

3.2. Model performance

The outcomes of the all-cause mortality prediction are presented in Table 2 and Fig. S2 in supplementary material. In the derivation cohort, the stacked ensemble algorithm demonstrates the highest AUC[95 % CI] of 0.826[0.792–0.859], compared to LR (0.818 [0.783–0.854, $P < 0.05$), RF (0.81[0.777–0.844], $P \geq 0.05$), SVM (0.815[0.777–0.852], $P \geq 0.05$), lightGBM (0.798[0.761–0.834], $P < 0.01$) and ET (0.811[0.775–0.848], $P \geq 0.05$). In the validation cohort, the stacked ensemble model also exhibits the highest discriminatory capability (AUC[95 % CI]: 0.803[0.743–0.864]). Similar results are seen in the stacked ensemble model employing limited features using LASSO feature selection, with an AUC[95 % CI] of 0.824[0.788–0.86] and 0.807[0.749–0.866] in derivation and validation cohort, respectively. To predict premature death (Table 3 and Fig. S2 in supplementary material), the stacked ensemble algorithm demonstrates the highest AUC[95 % CI] of 0.773[0.725–0.821] in derivation cohort, compared to LR (0.772[0.724–0.819], $P \geq 0.05$), RF (0.758[0.704–0.813], $P \geq 0.05$), SVM (0.749[0.692–0.806], $P \geq 0.05$), lightGBM (0.747[0.695–0.798], $P < 0.05$) and ET (0.698[0.643–0.754], $P < 0.001$). In the validation cohort, the stacked ensemble model also exhibits the highest discriminatory capability (AUC[95 % CI]: 0.791[0.719–0.863]). Similar results are seen in the stacked ensemble model employing limited features using LASSO feature selection, with an AUC[95 % CI] of 0.771[0.724–0.819] and 0.781[0.706–0.856] in derivation and validation cohort, respectively. Furthermore, among the various models, the stacked ensemble, SVM and LightGBM models demonstrated a good calibration ability for predicting all-cause mortality (Fig. S3 in supplementary material).

3.3. Model interpretation

SHAP values of the top 15 features for predicting all-cause mortality and premature death using stacked ensemble model are presented in Fig. 2. In SHAP summary plot, a negative SHAP value represents a reduced risk of death, while a positive SHAP value represents an increased risk of death. As expected, age emerged as the top-ranked predictor for all-cause mortality, followed by Cystatin C, cognitive function, pulse, mean corpuscular volume and self-reported health (Fig. 2A). In terms of premature mortality, important predictors included age, Cystatin C, self-reported health, pulse, ever smoker and cognitive function (Fig. 2B).

The SHAP dependency plot was applied to observe the SHAP value among different quantitative biomarkers and physical examination indicators using RF model estimation. Interaction effects were also represented by vertical dispersion of the data points (Fig. 3 and Fig. S5 in Supplementary material). In the case of SBP and DBP, the plots show a J-shaped relationship between SBP or DBP level and the SHAP value. The relationship between cholesterol levels (LDL, HDL, TC and TG) and the SHAP value was U-shaped, with low and high concentrations indicating a higher risk of death. We observed a higher SHAP value in participants with low level of weight, waist, or BMI, and higher level of pulse, especially for the elderly population. Cystatin C has the highest SHAP value among all biomarkers. It can be confirmed that the SHAP value of Cystatin C remained stable in the beginning, and shapely increased at a Cystatin C of 1 mg/L, especially for elderly adults.

4. Discussion

In this study, we used stacked ensemble models to predict death in 10,677 Chinese individuals. The stacked ensemble model reached notably high discrimination ability. In addition, we found age, sex, self-reported health, activities of daily living, cognitive function, ever smoker, levels of SBP, Cystatin C and LDL-C were strong predictors of either all-cause mortality or premature mortality.

Previous mortality prediction models were mostly developed based on traditional Cox regression models. A study [8] included 498, 103 UK Biobank participants explored the sex-specific associations of 655 variables (demographics, disease history, health indicators and lifestyle) with the risk of 5-year all-cause mortality, 13 variables for men and 11 variables for women were finally selected to construct the Cox regression model that predicted 5-year all-cause mortality, with an AUC of 0.80 (95 % CI 0.77–0.83) and 0.79 (95 % CI 0.76–0.83), respectively. Recently, machine learning has proven to be an invaluable tool for developing predictive models [40–43]. Weng et al. [18] constructed traditional Cox models, random forest and deep learning models to predict premature mortality using 60 baseline variables among nearly 0.5 million participants, and the models reached an AUC of 0.751 (95 % CI 0.748–0.767), 0.783 (95 % CI 0.776–0.791) and 0.790 (95 % CI 0.783–0.797) respectively. Machine learning models theoretically have higher discrimination ability compared with traditional regression models because they could reflect complex relationships between different variables. In another retrospective cohort study [15] of 6520 patients, EuroSCORE II, traditional regression models, and machine learning classifiers were used to predict the in-hospital mortality after elective cardiac surgery: the machine learning model reached an AUC of 0.795 (95 % CI 0.755–0.834), which was significantly higher than the EuroSCORE II (AUC = 0.737, 95 % CI 0.691–0.783) or traditional regression model (AUC = 0.742, 95 % CI 0.698–0.785). However, using a single machine learning algorithm may suffer from limitations such as uncertainty. In the current study, instead of using single machine learning models, we applied stacked ensemble models, which fused the results of all five machine learning algorithms (base learner), and improved the model performance

Table 2
Performance of different models for predict all-cause mortality.

Model	All features					LASSO features				
	accuracy	sensitivity	specificity	Brier score	AUC[95 % CI]	accuracy	sensitivity	specificity	Brier score	AUC[95 % CI]
Derivation cohort										
LR	0.774	0.678	0.779	0.171	0.818[0.783–0.854]*	0.769	0.72	0.772	0.164	0.814[0.776–0.852]**
RF	0.828	0.587	0.839	0.179	0.81[0.777–0.844]	0.796	0.629	0.803	0.195	0.807[0.773–0.842]
SVM	0.765	0.727	0.766	0.039	0.815[0.777–0.852]	0.765	0.72	0.767	0.039	0.817[0.779–0.854]
lightGBM	0.807	0.601	0.816	0.148	0.798[0.761–0.834]**	0.782	0.657	0.788	0.168	0.816[0.783–0.849]
ET	0.772	0.713	0.774	0.208	0.811[0.775–0.848]	0.775	0.72	0.778	0.204	0.817[0.782–0.851]
Fusion model	0.769	0.699	0.773	0.17	0.826[0.792–0.859]	0.764	0.713	0.767	0.17	0.824[0.788–0.86]
Validation cohort										
LR	0.784	0.667	0.787	0.171	0.794[0.734–0.854]	0.782	0.698	0.785	0.164	0.801[0.74–0.861]
RF	0.83	0.619	0.835	0.178	0.782[0.72–0.844]	0.803	0.651	0.807	0.192	0.784[0.724–0.844]
SVM	0.771	0.698	0.773	0.026	0.798[0.738–0.858]	0.773	0.714	0.775	0.026	0.809[0.752–0.867]
lightGBM	0.823	0.667	0.827	0.135	0.778[0.71–0.845]	0.77	0.73	0.771	0.176	0.782[0.717–0.847]*
ET	0.791	0.651	0.794	0.206	0.781[0.717–0.846]	0.778	0.651	0.781	0.204	0.782[0.723–0.842]
Fusion model	0.788	0.746	0.789	0.168	0.803[0.743–0.864]	0.778	0.714	0.78	0.169	0.807[0.749–0.866]

RF: Random forest, LR: Logistic regression, SVM: Support vector machine, LightGBM: Light gradient boosting machine, ET: Extra trees.
*P < 0.05,**P < 0.01,***P < 0.001 for difference of AUCs between fusion model and base models.

Table 3
Performance of different models for predict premature all-cause mortality.

Model	All features					LASSO features				
	accuracy	sensitivity	specificity	Brier score	AUC[95 % CI]	accuracy	sensitivity	specificity	Brier score	AUC[95 % CI]
Derivation cohort										
LR	0.756	0.647	0.759	0.187	0.772[0.724–0.819]	0.755	0.612	0.759	0.19	0.77[0.722–0.818]
RF	0.766	0.624	0.771	0.229	0.758[0.704–0.813]	0.75	0.624	0.754	0.204	0.749[0.698–0.801]
SVM	0.732	0.647	0.735	0.027	0.749[0.692–0.806]	0.734	0.6	0.738	0.027	0.758[0.707–0.809]
lightGBM	0.77	0.576	0.776	0.183	0.747[0.695–0.798]*	0.77	0.576	0.776	0.184	0.746[0.695–0.798]*
ET	0.728	0.482	0.736	0.21	0.698[0.643–0.754]***	0.758	0.553	0.764	0.238	0.728[0.673–0.783]**
Fusion model	0.72	0.682	0.721	0.196	0.773[0.725–0.821]	0.734	0.647	0.737	0.193	0.771[0.724–0.819]
Validation cohort										
LR	0.765	0.717	0.766	0.187	0.786[0.713–0.859]	0.765	0.739	0.766	0.189	0.778[0.703–0.854]
RF	0.762	0.587	0.765	0.23	0.764[0.694–0.834]	0.787	0.565	0.792	0.197	0.755[0.679–0.83]
SVM	0.711	0.717	0.711	0.019	0.788[0.719–0.857]	0.753	0.717	0.753	0.019	0.773[0.694–0.853]
lightGBM	0.627	0.783	0.624	0.233	0.75[0.671–0.829]*	0.688	0.696	0.688	0.212	0.739[0.656–0.821]*
ET	0.7	0.674	0.701	0.213	0.756[0.685–0.828]	0.764	0.565	0.768	0.238	0.738[0.659–0.816]
Fusion model	0.719	0.783	0.718	0.2	0.791[0.719–0.863]	0.744	0.761	0.743	0.191	0.781[0.706–0.856]

RF: Random forest, LR: Logistic regression, SVM: Support vector machine, LightGBM: Light gradient boosting machine, ET: Extra trees.
*P < 0.05,**P < 0.01,***P < 0.001 for difference of AUCs between fusion model and base models.

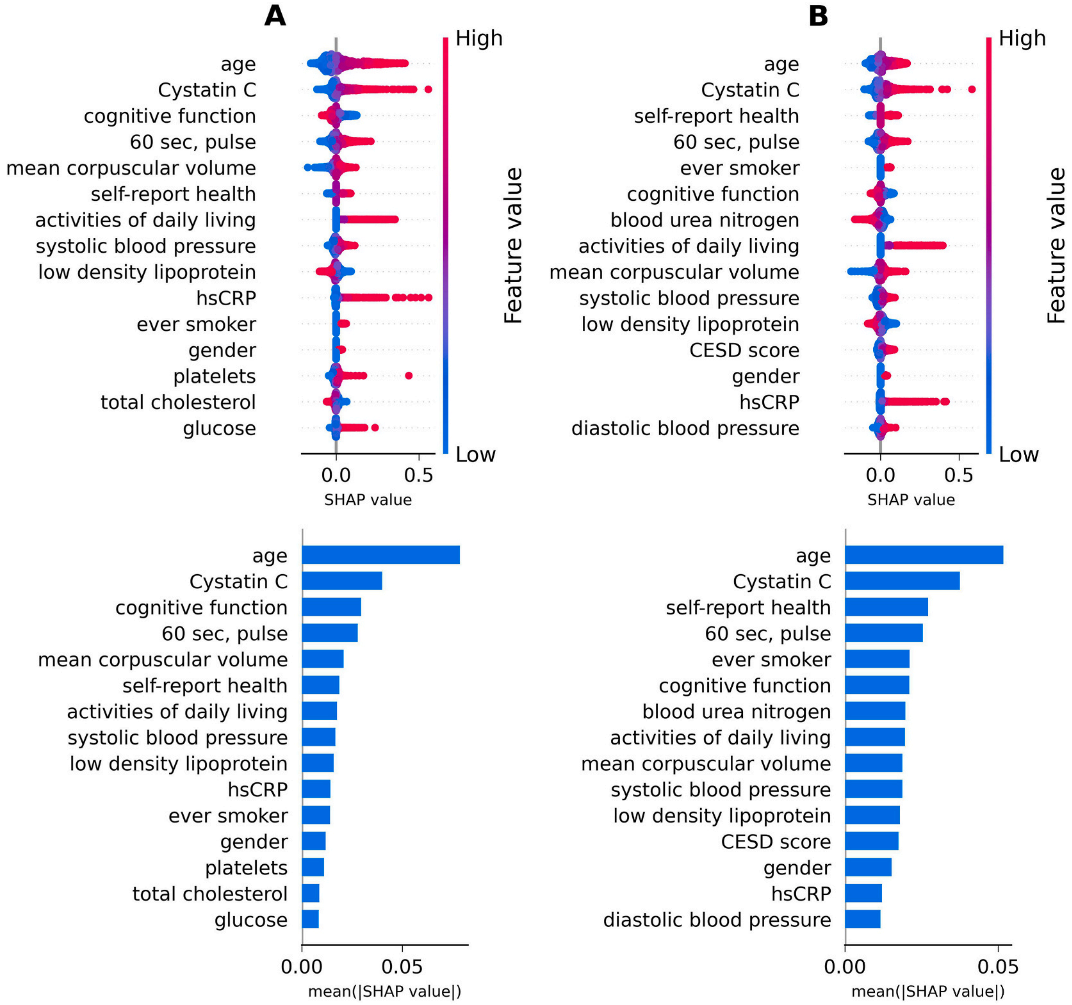

Fig. 2. SHAP plot of the stacked ensemble model for predicting all-cause mortality (A) and premature all-cause mortality (B).
In dot plot, the dot colors indicates values of predictor: red represents a larger value, and blue represents a smaller value. A negative SHAP value represents a reduced risk of death, while a positive SHAP value represents a increased risk of death. In bar plot, each bar signifies the importance of the respective variable.

even higher.

In a study included 13,611 adults aged 52 or older from the National Health and Retirement Study [16], Cox regression and machine learning models were performed to identify the most important risk factors of mortality over a 6-year follow-up period. Smoking, alcohol and lack of physical activity were found to be the strongest predictors of mortality among older American adults. What's more, in the study [8] using UK Biobank data, the authors explored the sex-specific associations of 655 variables (demographics, disease history, health indicators and lifestyle) with the risk of 5-year all-cause mortality, and found that self-reported health status was most closely associated with all-cause mortality in men and a former cancer diagnosis was most closely associated with all-cause mortality in women. Moreover, in a study [44] using data from CHARLS, plasma Cystatin C and hsCRP were found to be independent indicators for all-cause mortality. We found that age, sex, self-reported health, activities of daily living, cognitive function, ever smoker, levels of SBP, Cystatin C and LDL-C were associated with the risk of mortality, which is to some extent consistent with previous findings.

The SHAP method can also identify the nonlinear relationship between exposure and outcome variables. We observed a U-shaped association between cholesterol level (LDL-C, HDL-C, TG and TC) and all-cause mortality. A recent study [45] with 108,243 Danish general population using Cox regression model also suggested that low and high levels of LDL-C were associated with higher risk of death. Another systematic review [46] including 30 cohorts with 68,094 elderly adults showed that lower LDL-C was associated with all-cause mortality, particularly for individuals aged 60 years or older. Consistent with previous research [47,48], our study observed a J-shaped association of SBP and DBP with mortality risk. What's more, we have obtained a stronger association between mortality and low weight, waist, and BMI, especially for the elderly population. Previous observational research have also reported death risk of weight loss for elderly adults [49,50]. In the current study, Cystatin C contributed the most to mortality prediction among all

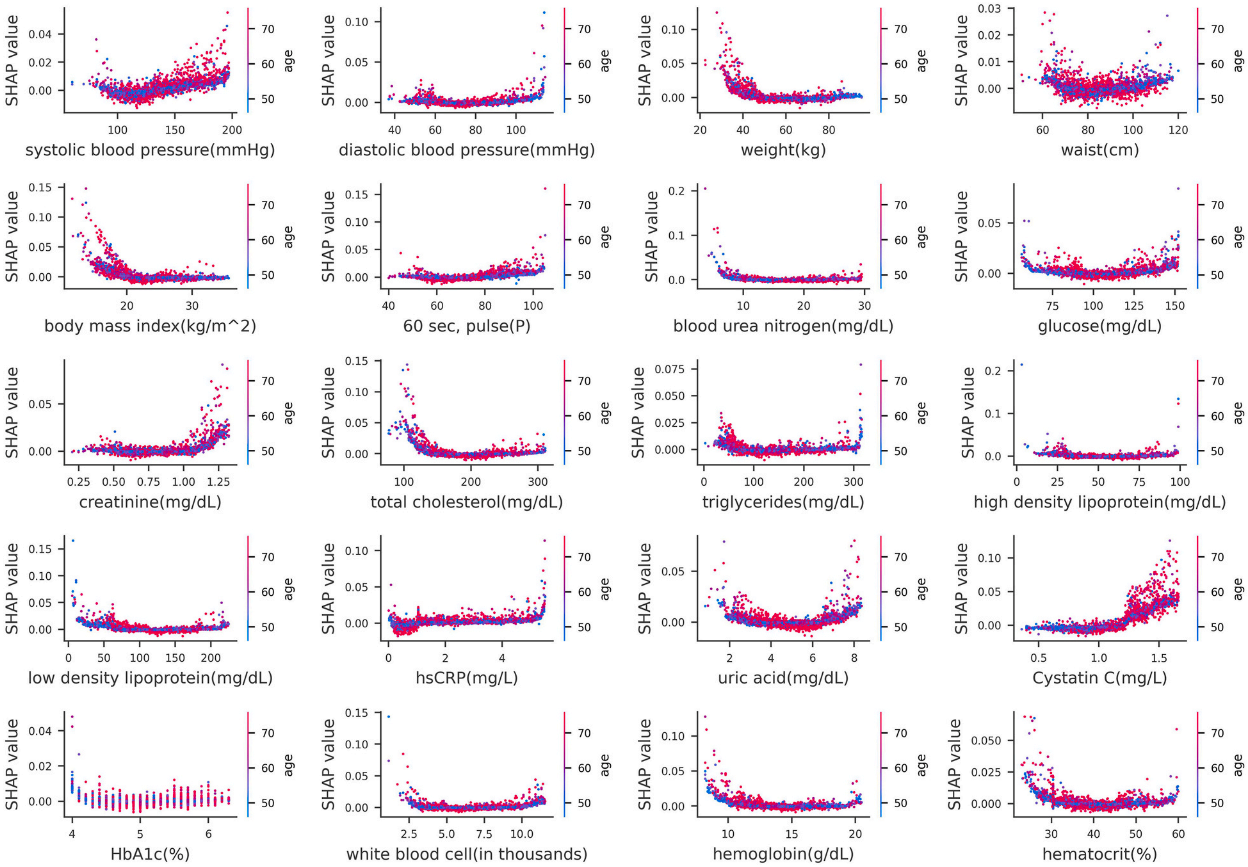

Fig. 3. SHAP dependence plot of biomarker and physical examination indicators of the RF model for predicting all-cause mortality.
The dot colors indicates values of predictor: red represents a larger value, and blue represents a smaller value. A negative SHAP value represents a reduced risk of death, while a positive SHAP value represents a increased risk of death.

biomarkers, particularly for elderly individuals. This association has been observed in another study using HRS cohort [51].

By incorporating diverse base models that capture different aspects of mortality risk, the ensemble method can better account for the complexity and variability present in data [52]. In our findings, although the stacking ensemble model exhibited a modest enhancement in discriminative ability compared to the base model, a slight enhancement in the predictive model's discriminative ability within a large-scale population can yield substantial advantages in risk stratification. In contrast, a simple algorithm for mortality prediction was well-suited for areas with limited clinical data and low-resource settings due to its ease of use and understanding. Furthermore, explainable techniques, such as SHAP method, can provide clinicians with transparent insights by identifying significant variables related to death and complex relationships between variables.

The current study has a few limitations that should be noted. First, some risk factors including disease history or general health indicators were based on self-reported. In addition, some potential predictors such as dietary patterns or physical activity level were not included. Second, participants who did not provide fasting blood samples at baseline were excluded. However, it is unlikely that these missingness were differential. Third, due to the absence of specific data on the time of death within the cohort, we were unable to accurately estimate the participants' survival time or further construct time-to-event prediction models. Fourth, modeling and validation were conducted using CHARLS data, without external database validation. However, we utilized participants newly enrolled in 2015 for external validation, which differ from the baseline population in 2011. Future studies regarding AI application tool can be developed to promote clinical utility.

However, there are some strengths as well. One important advance of our study was that the data were from a well-conducted cohort with a relatively large sample size, providing a representative sample of general Chinese populations. Second, the current study developed stacked ensemble models for predicting all-cause and premature death. We observed a relatively high AUC and showed that the ensemble model performed better than single machine learning models. Third, unlike previous prediction models that mainly focused on explaining the importance of predictors, this study performed a comprehensive interpretation of the machine learning models based on complex feature contributions.

5. Conclusion

Based on a nationwide cohort of Chinese adults, we successfully constructed machine learning models to predict all-cause and premature mortality using basic demographics, physical examinations, lifestyle, comorbidity and biomarkers. We analyzed and identified important risk factors of mortality using the SHAP method. The current study showed the feasibility, practicability, and the advantages of applying ensemble models in constructing prediction models. Future studies that utilize trajectories of important mortality predictors may further improve the model performance.

Funding

This study was supported by National Natural Science Foundation for Young Scientists of China (82203984), Zhejiang Key Laboratory of Intelligent Preventive Medicine (2020E10004), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01007), Key Research and Development Program of Zhejiang Province (2020C03002) and Healthy Zhejiang One Million people Cohort (K20230085).

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations. This research uses data from the China Health and Retirement Longitudinal Study (CHARLS). The CHARLS was approved by the Biomedical Ethics Review Committee of Peking University and all participants provided written informed consent. The study was approved by the Biomedical Ethics Review Committee of Peking University.

Consent for publication

Not Applicable.

Availability of data and materials

This research uses data from the China Health and Retirement Longitudinal Study (CHARLS), which can be downloaded at https://www.cpc.unc.edu/projects/china/data/datasets. The questionnaire has been published elsewhere, which can be downloaded at https://www.cpc.unc.edu/projects/china/data/questionnaires.

CRediT authorship contribution statement

Qi Yu: Writing – review & editing, Writing – original draft, Methodology, Data curation. **Lingzhi Zhang:** Writing – review & editing, Visualization, Validation, Data curation. **Qian Ma:** Writing – review & editing, Validation. **Lijuan Da:** Writing – review & editing, Visualization. **Jiahui Li:** Writing – review & editing, Visualization, Validation. **Wenyuan Li:** Writing – review & editing, Visualization, Validation, Conceptualization, Methodology.

Declaration of competing interest

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

Acknowledgment

This study uses data from the China Health and Retirement Longitudinal Study (CHARLS). We thank the China Center for Economic Research at Peking University for providing us with the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36878.

References

[1] M. Zhou, H. Wang, X. Zeng, et al., Mortality, morbidity, and risk factors in China and its provinces, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017, Lancet 394 (10204) (2019) 1145–1158, https://doi.org/10.1016/S0140-6736(19)30427-1.

[2] L.M. Wang, Z.H. Chen, M. Zhang, et al., Zhonghua Liuxingbingxue Zazhi 40 (3) (2019) 277–283, https://doi.org/10.3760/cma.j.issn.0254-6450.2019.03.005.

[3] S.P. Bell, A. Saraf, Risk stratification in very old adults: how to best gauge risk as the basis of management choices for patients aged over 80, Prog. Cardiovasc. Dis. 57 (2) (2014) 197–203, https://doi.org/10.1016/j.pcad.2014.08.001.

[4] L.C. Yourman, S.J. Lee, M.A. Schonberg, E.W. Widera, A.K. Smith, Prognostic indices for older adults: a systematic review, JAMA 307 (2) (2012) 182–192, https://doi.org/10.1001/jama.2011.1966.

[5] C.P. Wen, J.P. Wai, M.K. Tsai, et al., Minimum amount of physical activity for reduced mortality and extended life expectancy: a prospective cohort study, Lancet 378 (9798) (2011) 1244–1253, https://doi.org/10.1016/S0140-6736(11)60749-6.

[6] Z. Hongli, X. Bi, N. Zheng, C. Li, K. Yan, Joint effect of alcohol drinking and tobacco smoking on all-cause mortality and premature death in China: a cohort study, PLoS One 16 (1) (2021) e0245670, https://doi.org/10.1371/journal.pone.0245670.

[7] Y. Fan, D. He, Self-rated health, socioeconomic status and all-cause mortality in Chinese middle-aged and elderly adults, Sci. Rep. 12 (1) (2022) 9309, https://doi.org/10.1038/s41598-022-13502-9.

[8] A. Ganna, E. Ingelsson, 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study, Lancet 386 (9993) (2015) 533–540, https://doi.org/10.1016/S0140-6736(15)60175-1.

[9] S. Wang, W. Wang, X. Li, et al., Using machine learning algorithms for predicting cognitive impairment and identifying modifiable factors among Chinese elderly people, Front. Aging Neurosci. 14 (2022) 977034, https://doi.org/10.3389/fnagi.2022.977034.

[10] J.L. Speiser, K.E. Callahan, D.K. Houston, et al., Machine learning in aging: an example of developing prediction models for serious fall injury in older adults, J Gerontol A Biol Sci Med Sci 76 (4) (2021) 647–654, https://doi.org/10.1093/gerona/glaa138.

[11] C. Ye, J. Li, S. Hao, et al., Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm, Int J Med Inform 137 (2020) 104105, https://doi.org/10.1016/j.ijmedinf.2020.104105.

[12] X. Shi, Y. Cui, S. Wang, Y. Pan, B. Wang, M. Lei, Development and validation of a web-based artificial intelligence prediction model to assess massive intraoperative blood loss for metastatic spinal disease using machine learning techniques, Spine J. 24 (1) (2024) 146–160, https://doi.org/10.1016/j.spinee.2023.09.001.

[13] A.A. Abujaber, I. Albalkhi, Y. Imam, A. Nashwan, N. Akhtar, I.M. Alkhawaldeh, Machine learning-based prognostication of mortality in stroke patients, Heliyon 10 (7) (2024) e28869, https://doi.org/10.1016/j.heliyon.2024.e28869. Published 2024 Apr 3.

[14] S. Sakr, R. Elshawi, A.M. Ahmed, et al., Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project, BMC Med Inform Decis Mak 17 (1) (2017) 174, https://doi.org/10.1186/s12911-017-0566-6.

[15] J. Allyn, N. Allou, P. Augustin, et al., A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis, PLoS One 12 (1) (2017) e0169772, https://doi.org/10.1371/journal.pone.0169772.

[16] E. Puterman, J. Weiss, B.A. Hives, et al., Predicting mortality from 57 economic, behavioral, social, and psychological factors, Proc Natl Acad Sci U S A 117 (28) (2020) 16273–16282, https://doi.org/10.1073/pnas.1918455117.

[17] S. Tedesco, M. Andrulli, M.Å. Larsson, et al., Comparison of machine learning techniques for mortality prediction in a prospective cohort of older adults, Int J Environ Res Public Health 18 (23) (2021) 12806, https://doi.org/10.3390/ijerph182312806.

[18] S.F. Weng, L. Vaz, N. Qureshi, J. Kai, Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches, PLoS One 14 (3) (2019) e0214365, https://doi.org/10.1371/journal.pone.0214365.

[19] W. Qiu, H. Chen, A.B. Dincer, S. Lundberg, M. Kaeberlein, S.I. Lee, Interpretable machine learning prediction of all-cause mortality, Commun. Med. 2 (2022) 125, https://doi.org/10.1038/s43856-022-00180-x. Published 2022 Oct 3.

[20] Clarke Bertrand, Comparing bayes model averaging and stacking when model approximation error cannot be ignored, J. Mach. Learn. Res. 4 (2003) 683–712.

[21] Saso Džeroski, Bernard Ženko, Is combining classifiers with stacking better than selecting the best one? Mach. Learn. 54 (3) (2004) 255–273.

[22] V. Genre, G. Kenny, A. Meyler, et al., Combining expert forecasts: can anything beat the simple average? Int. J. Forecast. 29 (1) (2013) 108–121.

[23] H. Kwon, J. Park, Y. Lee, Stacking ensemble technique for classifying breast cancer, Healthc Inform Res 25 (4) (2019) 283–288, https://doi.org/10.4258/hir.2019.25.4.283.

[24] J. Zhang, H. Zhu, Y. Chen, et al., Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors, BMC Med Inform Decis Mak 21 (1) (2021) 187, https://doi.org/10.1186/s12911-021-01535-5.

[25] R.C. Kessler, M.S. Bauer, T.M. Bishop, et al., Evaluation of a model to target high-risk psychiatric inpatients for an intensive postdischarge suicide prevention, Intervention JAMA Psychiatry (2023) e224634, https://doi.org/10.1001/jamapsychiatry.2022.4634.

[26] M. Liu, J. Zhou, Q. Xi, et al., A computational framework of routine test data for the cost-effective chronic disease prediction, Brief Bioinform (2023) bbad054, https://doi.org/10.1093/bib/bbad054.

[27] Y. Zhao, Y. Hu, J.P. Smith, J. Strauss, G. Yang, Cohort profile: the China health and retirement longitudinal study (CHARLS), Int. J. Epidemiol. 43 (1) (2014) 61–68, https://doi.org/10.1093/ije/dys203.

[28] H. Chen, A.C. Mui, Factorial validity of the center for epidemiologic studies depression scale short form in older population in China, Int. Psychogeriatr. 26 (1) (2014) 49–57, https://doi.org/10.1017/S1041610213001701.

[29] S.E. Cook, M. Marsiske, K.J. McCoy, The use of the Modified Telephone Interview for Cognitive Status (TICS-M) in the detection of amnestic mild cognitive impairment, J Geriatr Psychiatry Neurol 22 (2) (2009) 103–109, https://doi.org/10.1177/0891988708328214.

[30] M. Zuo, C. Gan, T. Liu, J. Tang, J. Dai, X. Hu, Physical predictors of cognitive function in individuals with hypertension: evidence from the CHARLS basline survey, West. J. Nurs. Res. 41 (4) (2019) 592–614, https://doi.org/10.1177/0193945918770794.

[31] A.C. Bender, A.M. Austin, F. Grodstein, J.P.W. Bynum, Executive function, episodic memory, and Medicare expenditures, Alzheimers Dement 13 (7) (2017) 792–800, https://doi.org/10.1016/j.jalz.2016.12.013.

[32] X. Pan, Y. Luo, D. Zhao, L. Zhang, Associations among drinking water quality, dyslipidemia, and cognitive function for older adults in China: evidence from CHARLS, BMC Geriatr. 22 (1) (2022) 683, https://doi.org/10.1186/s12877-022-03375-y. Published 2022 Aug 18.

[33] Xing Zhang, Chen Xu, Ran Guangming, Ma Yuanxiao, Adult children's support and self-esteem as mediators in the relationship between attachment and subjective well-being in older adults, Pers. Indiv. Differ. 97 (2016) 229–233.

[34] A.D. Shah, J.W. Bartlett, J. Carpenter, O. Nicholas, H. Hemingway, Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study, Am. J. Epidemiol. 179 (6) (2014) 764–774, https://doi.org/10.1093/aje/kwt312.

[35] F. Mu, M. Wang, X. Zeng, F. Wang, Predicting risk of subsequent pregnancy loss among women with recurrent pregnancy loss: an immunological factor-based multivariable model, Am. J. Reprod. Immunol. 91 (3) (2024) e13837, https://doi.org/10.1111/aji.13837.

[36] Bergstra James, Bengio Yoshua, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

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

[38] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, in: The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 1–8.

[39] S.M. Lundberg, B. Nair, M.S. Vavilala, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749–760, https://doi.org/10.1038/s41551-018-0304-0.

[40] Y. Cui, X. Shi, Y. Qin, et al., Establishment and validation of an interactive artificial intelligence platform to predict postoperative ambulatory status for patients with metastatic spinal disease: a multicenter analysis, Int. J. Surg. 110 (5) (2024) 2738–2756, https://doi.org/10.1097/JS9.0000000000001169. Published 2024 May 1.

[41] M. Lei, B. Wu, Z. Zhang, et al., A web-based calculator to predict early death among patients with bone metastasis using machine learning techniques: development and validation study, J. Med. Internet Res. 25 (2023) e47590, https://doi.org/10.2196/47590. Published 2023 Oct 23.

[42] S.K. Biswas, A. Nath Boruah, R. Saha, R.S. Raj, M. Chakraborty, M. Bordoloi, Early detection of Parkinson disease using stacking ensemble method, Comput Methods Biomech Biomed Engin 26 (5) (2023) 527–539, https://doi.org/10.1080/10255842.2022.2072683.

[43] A. Gialluisi, A. Di Castelnuovo, M.B. Donati, G. de Gaetano, L. Iacoviello, Moli-sani Study Investigators, Machine learning approaches for the estimation of biological aging: the road ahead for population studies, Front. Med. 6 (2019) 146, https://doi.org/10.3389/fmed.2019.00146.

[44] Y. Shen, Y. Zhang, S. Xiong, X. Zhu, C. Ke, High-sensitivity C-reactive protein and cystatin C independently and jointly predict all-cause mortality among the middle-aged and elderly Chinese population, Clin. Biochem. 65 (2019) 7–14, https://doi.org/10.1016/j.clinbiochem.2018.12.012.

[45] C.D.L. Johannesen, A. Langsted, M.B. Mortensen, B.G. Nordestgaard, Association between low density lipoprotein and all cause and cause specific mortality in Denmark: prospective cohort study [published correction appears in BMJ. 2021 Feb 12;372:n422], BMJ 371 (2020) m4266, https://doi.org/10.1136/bmj.m4266. Published 2020 Dec 8.

[46] U. Ravnskov, D.M. Diamond, R. Hama, et al., Lack of an association or an inverse association between low-density-lipoprotein cholesterol and mortality in the elderly: a systematic review, BMJ Open 6 (6) (2016) e010401, https://doi.org/10.1136/bmjopen-2015-010401. Published 2016 Jun 12.

[47] H. Gao, K. Wang, F. Ahmadizar, et al., Changes in late-life systolic blood pressure and all-cause mortality among oldest-old people in China: the Chinese longitudinal healthy longevity survey, BMC Geriatr. 21 (1) (2021) 562, https://doi.org/10.1186/s12877-021-02492-4. Published 2021 Oct 18.

[48] M. Li, Z. Su, H. Su, et al., Effect of blood pressure on the mortality of the elderly population with (pre)frailty: results from NHANES 1999-2004, Front Cardiovasc Med 9 (2022) 919956, https://doi.org/10.3389/fcvm.2022.919956. Published 2022 Aug 1.

[49] F.D.C. De Stefani, P.S. Pietraroia, M.M. Fernandes-Silva, J. Faria-Neto, C.P. Baena, Observational evidence for unintentional weight loss in all-cause mortality and major cardiovascular events: a systematic review and meta-analysis, Sci. Rep. 8 (1) (2018) 15447, https://doi.org/10.1038/s41598-018-33563-z. Published 2018 Oct 18.

[50] A.A. Javed, R. Aljied, D.J. Allison, L.N. Anderson, J. Ma, P. Raina, Body mass index and all-cause mortality in older adults: a scoping review of observational studies, Obes. Rev. 21 (8) (2020) e13035, https://doi.org/10.1111/obr.13035.

[51] J. Wu, Y. Liang, R. Chen, et al., Association of plasma cystatin C with all-cause and cause-specific mortality among middle-aged and elderly individuals: a prospective community-based cohort study, Sci. Rep. 12 (1) (2022) 22265, https://doi.org/10.1038/s41598-022-24722-4. Published 2022 Dec 23.

[52] R. Malhotra, M. Khanna. Particle swarm optimization-based ensemble learning for software change prediction, Inf. Software Technol. 102 (2018) 65–84, https://doi.org/10.1016/j.infsof.2018.05.007.