

RESEARCH ARTICLE

Development and external validation of an interpretable machine learning model for depressive symptom screening in Chinese adults aged 50 years or older with diabetes

Supplementary File

Table S1. Missingness of candidate variables in the CHARLS cohort before multiple imputation

Variable	Overall CHARLS cohort (<i>N</i> = 2,187)	Training set (<i>n</i> = 1,530)	Internal validation set (<i>n</i> = 657)
Age	0 (0.00%)	0 (0.00%)	0 (0.00%)
Gender	0 (0.00%)	0 (0.00%)	0 (0.00%)
Educational attainment	0 (0.00%)	0 (0.00%)	0 (0.00%)
Current employment status	1 (0.05%)	0 (0.00%)	1 (0.15%)
Health insurance	0 (0.00%)	0 (0.00%)	0 (0.00%)
Internet use	0 (0.00%)	0 (0.00%)	0 (0.00%)
Sleep duration	0 (0.00%)	0 (0.00%)	0 (0.00%)
Household registration status	0 (0.00%)	0 (0.00%)	0 (0.00%)
Physical activity	0 (0.00%)	0 (0.00%)	0 (0.00%)
Social isolation	0 (0.00%)	0 (0.00%)	0 (0.00%)
Marital status	0 (0.00%)	0 (0.00%)	0 (0.00%)
Living with children	15 (0.69%)	12 (0.78%)	3 (0.46%)
Number of household members	0 (0.00%)	0 (0.00%)	0 (0.00%)
Smoking status	0 (0.00%)	0 (0.00%)	0 (0.00%)
Alcohol consumption	0 (0.00%)	0 (0.00%)	0 (0.00%)
Life satisfaction	0 (0.00%)	0 (0.00%)	0 (0.00%)
Stroke	94 (4.30%)	71 (4.64%)	23 (3.50%)
Monthly social participation	0 (0.00%)	0 (0.00%)	0 (0.00%)
Satisfaction with relationships with children	13 (0.59%)	11 (0.72%)	2 (0.30%)
Liver disease	0 (0.00%)	0 (0.00%)	0 (0.00%)
Kidney disease	0 (0.00%)	0 (0.00%)	0 (0.00%)
Digestive disease	0 (0.00%)	0 (0.00%)	0 (0.00%)
Hypertension	0 (0.00%)	0 (0.00%)	0 (0.00%)
Dyslipidemia	0 (0.00%)	0 (0.00%)	0 (0.00%)
Heart disease	0 (0.00%)	0 (0.00%)	0 (0.00%)
Diabetes medication	149 (6.81%)	107 (6.99%)	42 (6.39%)
Cancer	0 (0.00%)	0 (0.00%)	0 (0.00%)
Arthritis	0 (0.00%)	0 (0.00%)	0 (0.00%)
Chronic lung disease	0 (0.00%)	0 (0.00%)	0 (0.00%)
History of falls	0 (0.00%)	0 (0.00%)	0 (0.00%)
Vision	4 (0.18%)	3 (0.20%)	1 (0.15%)
Hearing	2 (0.09%)	2 (0.13%)	0 (0.00%)
Self-perceived health status	2 (0.09%)	2 (0.13%)	0 (0.00%)
History of hip fracture	0 (0.00%)	0 (0.00%)	0 (0.00%)
IADL	1 (0.05%)	0 (0.00%)	1 (0.15%)
ADL	1 (0.05%)	0 (0.00%)	1 (0.15%)
Cognitive function	211 (9.65%)	149 (9.74%)	62 (9.44%)

Notes: Values are presented as *n* (%). Missingness was assessed after eligibility and participant-level exclusions and before multiple imputation. Variables with more than 20% missingness in the training set were excluded before model development.

Abbreviation: CHARLS: China Health and Retirement Longitudinal Study.

Table S2. Definitions of covariates

Variable	Definition
Age	Continuous variable in the model analysis.
Sleep time	Continuous variable in the model analysis.
Social isolation score	Continuous variable ranging from 0 to 4, with higher scores indicating greater social isolation.
Number of family members	Continuous variable in the model analysis.
IADL score	Continuous variable ranging from 0 to 5, calculated as the sum of five instrumental activities of daily living (doing household chores, cooking, shopping, managing finances, and taking medications). Each item was coded as 0/1 according to whether difficulty was present, with higher scores indicating poorer functional status.
ADL score	Continuous variable ranging from 0 to 6, calculated as the sum of six activities of daily living (dressing, bathing, eating, getting in/out of bed, toileting, and continence control). Each item was coded as 0/1 according to whether difficulty was present, with higher scores indicating poorer functional status.
Cognitive score	Continuous variable ranging from 0 to 21, calculated as the sum of the TICS-10 score (0–10; orientation and attention), word recall score (0–10; average of immediate and delayed recall, assessing episodic memory), and figure drawing score (0–1; visuospatial ability). Higher scores indicate better cognitive function.
Gender	0 = Female, 1 = Male
Education	0 = Below high school, 1 = High school, 2 = College or above
Marital status	0 = Others (separated, divorced, widowed, or never married), 1 = Married or partnered (first married, remarried, or civil partner)
Household registration	0 = Rural, 1 = Urban
Employment status	0 = No, 1 = Yes
Medical insurance	0 = No, 1 = Yes
Internet use	0 = No, 1 = Yes
Physical activity	Number of days per week of vigorous activity?
	Number of moderate physical activity days per week?
	0 = No, 1 = Yes
	0 = No, 1 = Yes
	0 = No (Less than twice a week), 1 = Yes (at least once a week)
Living with children	0 = No, 1 = Yes
Smoking status	0 = No (Not a current smoker), 1 = Yes (Current smoker)
Alcohol consumption	0 = No (Not drinking now)
	1 = Yes (Drinking now)

(cont'd...)

Table S2. (Continued)

Variable	Definition
Life satisfaction	Self-reported life satisfaction. 1 = Not satisfied; 2 = Average; 3 = Satisfied
Stroke	0 = No, 1 = Yes
Sociability	Do you socialize monthly? 0 = No, 1 = Yes
Child relationship satisfaction	Self-reported satisfaction with relationship with children. 1 = Not satisfied; 2 = Average; 3 = Satisfied
Liver disease	Self-reported physician-diagnosed liver disease 0 = No, 1 = Yes
Kidney disease	Self-reported diagnosed kidney disease by doctors 0 = No, 1 = Yes
Digestive disease	Self-reported diagnosed digestive disease by doctors 0 = No, 1 = Yes
Hypertension	Self-reported diagnosed hypertension by doctors 0 = No, 1 = Yes
Dyslipidemia	Self-reported diagnosed dyslipidemia by doctors 0 = No, 1 = Yes
Heart disease	Self-reported diagnosed heart disease by doctors 0 = No, 1 = Yes
Diabetes medication use	0 = No, 1 = Yes
Cancer	Self-reported diagnosed cancer by doctors 0 = No, 1 = Yes
Arthritis	Self-reported diagnosed arthritis by doctors 0 = No, 1 = Yes
Chronic lung disease	Self-reported diagnosed chronic lung disease by doctors 0 = No, 1 = Yes
Falls	0 = No, 1 = Yes
Vision	1 = Poor; 2 = Average; 3 = Good
Hearing	1 = Poor; 2 = Average; 3 = Good
Self-perceived health status	1 = Poor; 2 = Average; 3 = Good
Hip fracture	0 = No, 1 = Yes

Abbreviations: ADL: Activities of daily living; IADL: Instrumental activities of daily living; TICS-10: Telephone Interview of Cognitive Status-10.

Table S3. Key methodological and reproducibility details

Topic	Clarification
Internal validation set	The CHARLS holdout sample was used for internal validation/model selection; the Xijing Hospital cohort was reserved for external validation.
Multiple imputation	MICE: m = 5, maxit = 15, seed = 2026. The outcome was neither imputed nor used as an imputation predictor; validation rows were ignored during imputation-model fitting.
Feature selection	LASSO was repeated in each imputed training dataset using lambda.1se; variables selected in at least three of five imputations formed the 16-variable set.
Model tuning and selection	Hyperparameter tuning used stratified 5-fold cross-validation with AUC as the tuning metric within each imputed training dataset. XGBoost was fixed before external validation using discrimination, Brier score, calibration, decision-curve analysis, and deployability.
Hyperparameter search spaces	GLMNet: penalty 10^{-4} to $10^{-0.5}$, mixture 0.2–1.0; SVM: cost 10^{-2} to 10, sigma 10^{-3} to 1; neural network: hidden units 1 to data-dependent upper ≤ 8 , penalty 10^{-5} to 10^{-1} , epochs 50–200; XGBoost/LightGBM: 20 Latin-hypercube points, trees 150–500, depth 2–5, sample 0.6–0.9, mtry data-dependent, learn_rate 10^{-3} to $10^{-0.7}$, min_n 2–30 (XGBoost) or 5–50 (LightGBM), and loss_reduction 10^{-10} to 31.6 for XGBoost; RF/ExtraTrees: 800 trees, mtry 2 to p predictors, min_n 20–80, depth 3–8.
Imputation handling	Models were fitted separately in each imputed dataset, and performance was summarized across imputations. The calculator used the first-imputation XGBoost workflow after predictor selection.
Software and seeds	R 4.5.2; key packages: mice 3.19.0, glmnet 4.1.10, tidymodels 1.4.1, xgboost 3.2.0.1, pROC 1.19.0.1, shapviz 0.10.3, and shiny 1.14.0. Seeds: 123, 2026, 2127–2531, and 2027–2031.
Calculator model	Deployed XGBoost: mtry = 3, trees = 472, min_n = 9, tree_depth = 4, learn_rate = 0.006724, loss_reduction = 0.000051, sample_size = 0.7585.
Calibration and thresholds	Slopes >1 and the negative external intercept indicate non-ideal calibration; local recalibration should be considered. Metrics used 0.50; risk categories were < 0.20, 0.20 to < 0.50, and ≥ 0.50 .

Abbreviations: AUC: Area under the curve; GLMNet: Generalized linear model with elastic net regularization; LASSO: Least absolute shrinkage and selection operator; LightGBM: Light Gradient Boosting Machine; MICE: Multiple imputation by chained equations; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme Gradient Boosting.

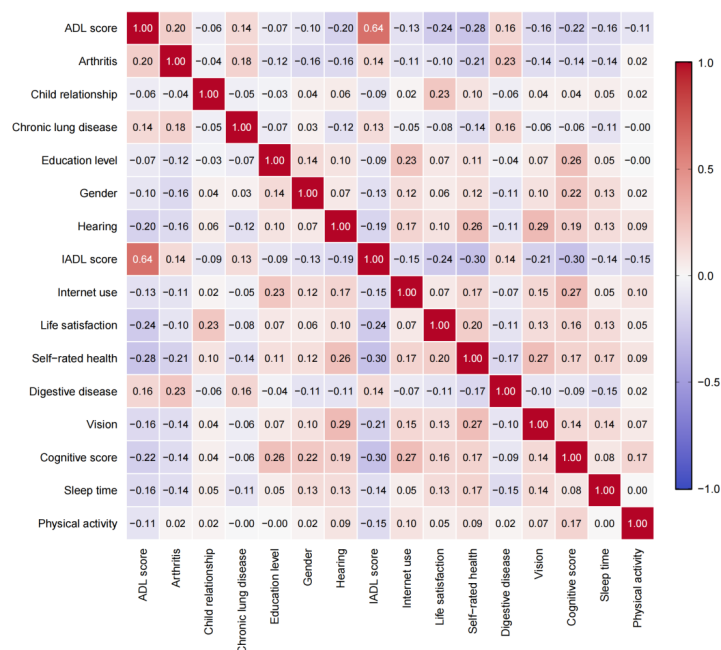


Figure S1. Pearson correlation matrix of the predictors included in the final XGBoost model for identifying depressive symptoms. The heatmap illustrates the pairwise Pearson correlation coefficients among the predictors included in the final model. The correlation coefficient is shown in each cell, and the color intensity indicates the strength and direction of the association.

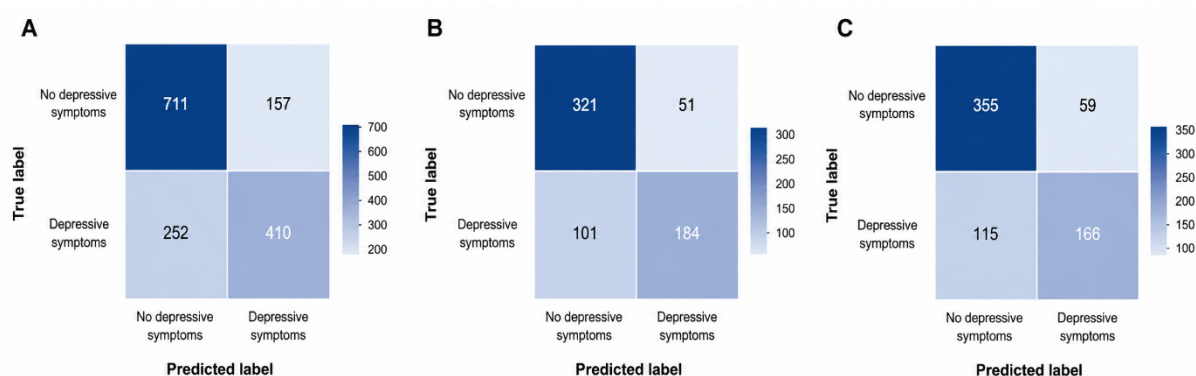


Figure S2. Confusion matrices of the final XGBoost model in the training, internal validation, and external validation sets. The matrices show the classification results of the final model at the default threshold of 0.5, including correctly and incorrectly classified cases with and without depressive symptoms. (A) Training set. (B) Internal test set. (C) External validation set.

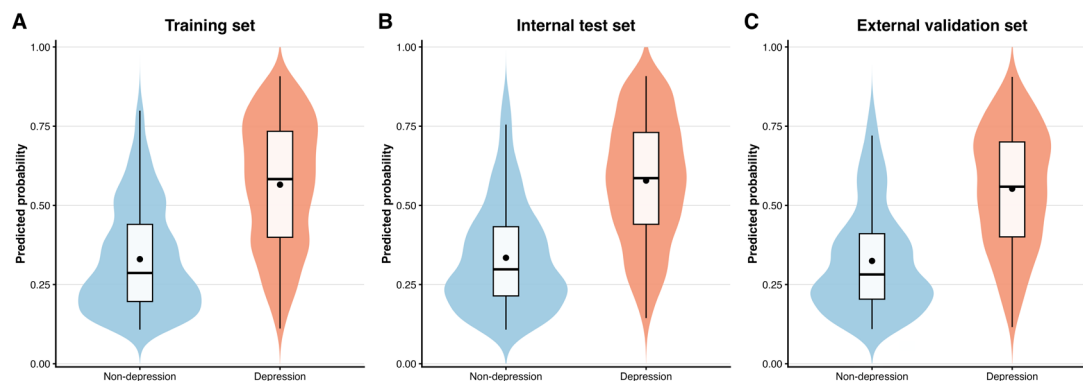


Figure S3. Distribution of predicted probabilities generated by the final XGBoost model according to depressive symptom status in the (A) training, (B) internal validation, and (C) external validation sets. Violin plots with embedded boxplots illustrate the distribution of predicted risk probabilities in participants with and without depressive symptoms across the three datasets, showing clear separation between the two groups.

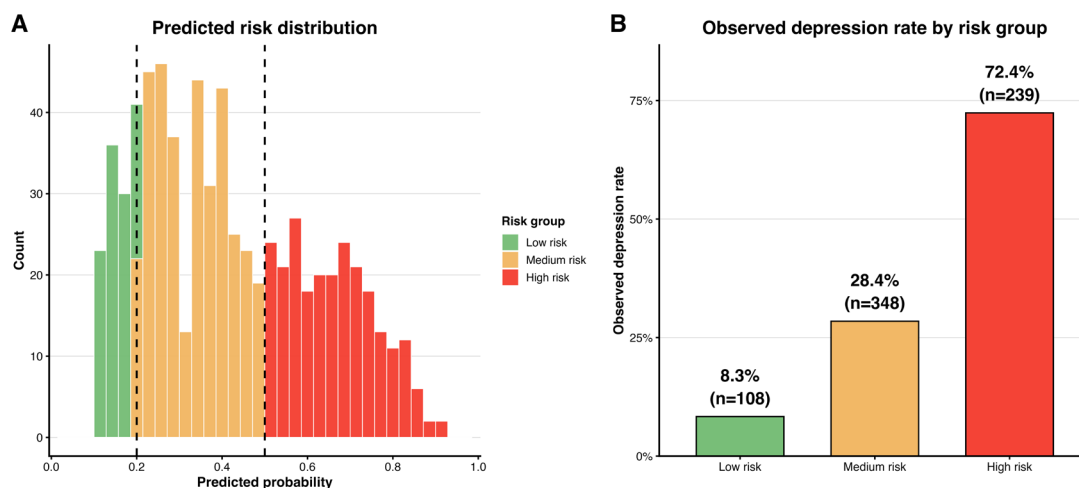


Figure S4. Risk stratification performance of the final XGBoost model in the external validation set. (A) The distribution of predicted probabilities across the predefined low-, medium-, and high-risk groups. (B) The observed rates of depressive symptoms within each risk group together with the corresponding sample sizes.