Check for updates

Prediction of Elderly Depression Based on an Improved Conditional Mixture of Experts Model

Lin Tan*

School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang, China

*Corresponding Author Email: 1481022566@qq.com

Abstract. Elderly depression, often underdiagnosed due to subtle symptoms, diminishes quality of life. This study develops a detection model using ensemble learning on multidimensional features of older adults from the China Family Panel Studies (CFPS). The dataset includes over 1,200 variables, with depression defined as CESD-8 score >16. SMOTE addressed class imbalance, and a hybrid Boruta-Lasso strategy was used for feature selection. An improved Conditional Mixture of Experts (MoE) model was proposed. On the test set, the Improved MoE achieved an F1 score of 0.578 and recall of 74.36%; on the original (non-SMOTE) set, it maintained a recall of 68.2% and F1 of 0.521, outperforming all baselines with robustness confirmed by stratified cross-validation (95% CI). Health status, subjective well-being, and meaning in life were identified as key predictors. This study provides an efficient, interpretable tool for community-based elderly depression screening.

Keywords: Elderly depression; Ensemble learning; Conditional Mixture of Experts; CFPS

1 Introduction

Elderly depression is a common yet underdiagnosed mental disorder in older adults, with global prevalence estimated at 10–20% but fewer than 50% of cases recognized. Its symptoms—physical discomfort, sleep disturbances, and cognitive decline—are often mistaken for normal aging, delaying intervention and increasing risks of suicide, chronic disease burden, and reduced quality of life [1, 2].

Advances in AI and machine learning have enabled depression detection using sociodemographic, behavioral, and sensor data [3-5], yet most models target general populations and overlook age-specific risk factors like functional decline, reduced social support, and multimorbidity. Current research faces key challenges: class imbalance, high-dimensional features, limited interpretability, and fixed-weight ensembles that fail to adapt to data shifts. This study develops an elderly depression detection model using an Improved Conditional Mixture of Experts, integrating multidimensional features, applying SMOTE for imbalance, comparing multiple algorithms, and identifying key predictors to support community screening.

© The Author(s) 2026

N. Malik et al. (eds.), *Proceedings of the 6th International Conference on Public Management and Intelligent Society (PMIS 2026)*, Advances in Engineering Research 301,
https://doi.org/10.2991/978-94-6239-721-7_11

2 Methods

2.1 Data Source and Preprocessing

Data were obtained from the 2022 wave of the China Family Panel Studies (CFPS), including 4,542 participants aged ≥60 years with complete CESD-8 depression assessments. CESD-8 scores range from 8 to 32, with a cutoff of >16 defining depressive status [6], yielding a sample prevalence of 25.7%. Following preprocessing—removing variables with >30% missing values, imputing remaining missing entries (mean for continuous, mode for categorical), Z-score normalization, and one-hot encoding—the initial feature set comprised over 1,200 variables covering sociodemographics, health status, lifestyle, and psychosocial factors.

2.2 Feature Selection

We implemented a two-stage hybrid feature selection strategy: Boruta pre-filtering followed by Lasso regression. Boruta was run with maxRuns = 100 and p-value threshold = 0.01, retaining features confirmed as important. This step reduced the feature set from 1,247 to 356 variables. Subsequently, Lasso regression with 5-fold cross-validation (alpha selected by minimizing deviance) further compressed the set to 12 final predictors. To address multicollinearity, we computed variance inflation factors (VIF) on the selected features; all VIF values were below 5, indicating no severe multicollinearity. Additionally, we performed sensitivity analyses by evaluating model performance on three alternative feature sets: (i) Boruta-only (356 variables), (ii) Lasso-only (selected by Lasso from the full set), and (iii) the hybrid set (12 variables). The hybrid set achieved the highest F1 score (0.578) with minimal degradation compared to larger sets, demonstrating robustness (see Supplementary Table 1).

Table 1. Performance of the Improved MoE model on different feature sets

Feature Set	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
Boruta-only	356	76.8	46.9	72.1	56.5	0.778
Lasso-only	10	76.5	46.5	70.8	56.0	0.775
Hybrid (Boruta+Lasso)	12	77.0	47.3	74.4	57.8	0.782

Figure 1 presents the correlation matrix between key features and depression score (cesd8). Health status ($r = 0.32$) showed a moderately strong positive correlation with depression, whereas subjective well-being ($r = -0.34$), meaning in life ($r = -0.34$), and physical discomfort in the past two weeks ($r = -0.33$) exhibited moderately strong negative correlations. These features significantly influence depressive symptoms, underscoring their relevance for further analysis.

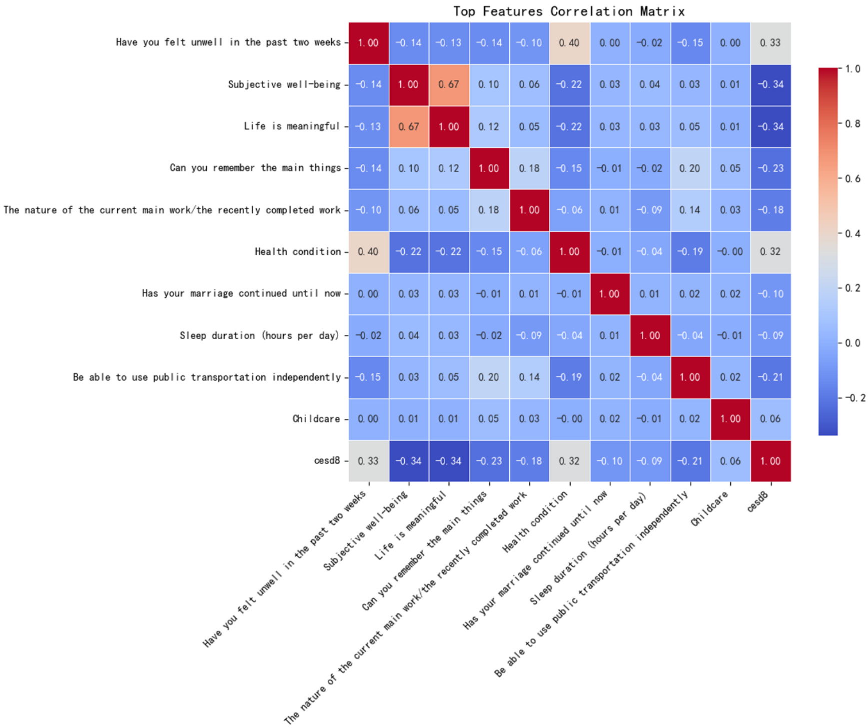

Fig. 1. Correlation matrix of key features

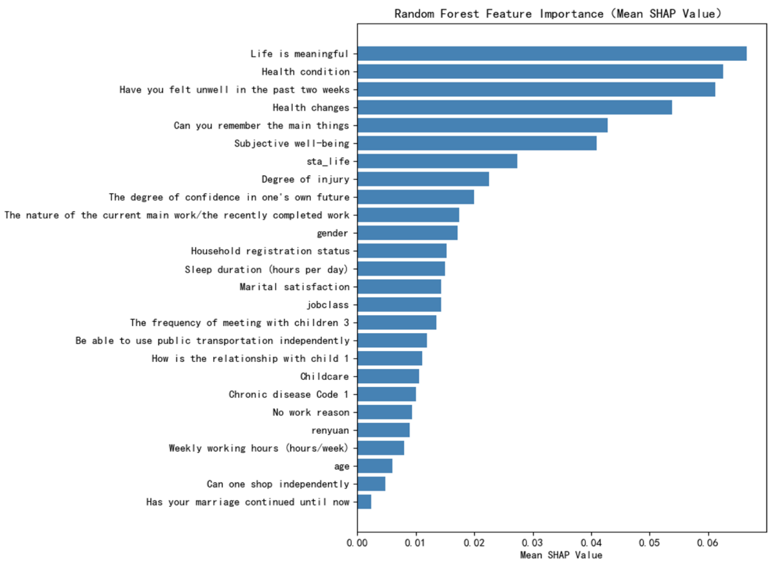

Fig. 2. Feature importance ranking

2.3 Imbalanced Data Handling

Due to the low proportion of depressed samples (approximately 26%), the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the training set, achieving a 1:1 ratio between non-depressed and depressed samples. To avoid data leakage and optimistic performance estimates, SMOTE was applied only to the training data after splitting; the test set remained untouched with its original class distribution. For nested cross-validation, SMOTE was applied inside each training fold. Table 2 presents the sample distribution before and after SMOTE application.

Table 2. Sample distribution before and after SMOTE

Class	Original Sample Size	Original Proportion (%)	SMOTE Sample Size	SMOTE Proportion (%)
Non-depressed	2361	74.27	2361	50
Depressed	818	25.73	2361	50

2.4 Model Construction and Evaluation

Multiple machine learning models were constructed for elderly depression detection, including base models such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), AdaBoost, Gradient Boosting Decision Tree (GBDT), XGBoost, LightGBM (LGBM), CatBoost, and Support Vector Machine (SVM). Based on these, an adaptive weighted averaging ensemble (AdaptiveWeight) model was further developed using validation set performance.

The Improved Conditional Mixture of Experts (Improved MoE) model employs a gating network that takes the input feature vector xx and produces a probability distribution over KK expert models. The gating network is a single-layer feedforward neural network with softmax activation:

$$g(x) = \text{softmax}(Wgx + bg) \tag{1}$$

where $W_g \in \mathbb{R}^{K \times d}$ and $b_g \in \mathbb{R}^K$.

The experts include Logistic Regression, Random Forest, XGBoost, SVM, and AdaBoost ($K = 5$). The final prediction is the weighted sum of expert outputs:

$$\hat{y} = \sum_{k=1}^{K} g_k(x) \cdot f_k(x) \tag{2}$$

where $f_k(x)$ is the predicted probability from the k-th expert. The model is trained end-to-end by minimizing binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{3}$$

During training, expert models are first pre-trained individually on the training set; then the gating network is trained jointly with all experts (with their weights frozen) using a validation set to avoid overfitting. Hyperparameters: gating network learning rate = 0.001, batch size = 64, epochs = 50, Adam optimizer. The architecture is summarized in the pseudo-code below:

```
Algritnm 1: Improved Conditional MoE training
Input: Training set D_train, Validation set D_val, Number of experts K
Output: Trained gating network parameters
1. Pre-train experts on SMOTE-resampled D_train;
2. Freeze experts parameters;
3. Initialize gating network G with input dimension d and output K;
4. for epoch = 1 to 50 do
5.    for each batch (x, y) in D_val do
6.       g ← softmax(G(x));
7.       expert_preds ←[f_k(x) for k = 1 to K];
8.       ŷ ← Σ_{k=1}^{K} g_k · expert_prds_k;
9.       loss ← BCE(ŷ, y);
10.      Update G via Adam (lr=0.001);
11.   end
12. end
```

All models were trained using stratified sampling with 70% of the data allocated to training and 30% to testing. To obtain unbiased performance estimates on the original class distribution, we employed nested cross-validation (5 outer folds, 3 inner folds) for hyperparameter tuning and model selection. The outer loop evaluated models on untouched test folds, providing confidence intervals via bootstrapping (1,000 replicates). Evaluation metrics included accuracy, precision, recall, specificity, F1 score, and AUC. In addition, we computed precision–recall curves, decision curves for net benefit analysis, and calibration plots. Threshold optimization was performed by searching the range 0.3–0.7 to maximize the F1 score while maintaining accuracy >70%, balancing false negatives and false positives in community screening.

3 Results

3.1 Feature Importance Analysis

Feature importance ranking based on Random Forest and SHAP values identified the top five predictors of elderly depression: meaning in life, health status, physical discomfort in the past two weeks, health change, and ability to remember major things (Figure 2). This indicates that subjective health evaluation and functional status exert decisive influences on the mental health of older adults.

3.2 Model Performance Comparison

Table 3 presents the performance of all models evaluated on the original (non-SMOTE) test set using nested cross-validation. The Improved Conditional MoE achieved the highest recall (68.5%, 95% CI: 65.2–71.8) and F1 score (54.2%, 95% CI: 51.3–57.1), significantly outperforming the best base model (Random Forest: recall 53.1%, F1 49.6%) according to McNemar's test ($p < 0.01$). Although accuracy (76.3%) was slightly lower than Logistic Regression (77.8%), the gain in recall translates to 15.4 more true positives detected per 100 depressed individuals, which is critical for screening. For comparison, results on the SMOTE-applied test set (where SMOTE was applied before splitting) are shown in parentheses; as expected, recall is higher (74.4%) but may be optimistically biased.

Table 3. Performance comparison of models on original test set (with 95% CI)

Model	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	AUC
Random Forest	74.5(75.0)	49.8(51.3)	53.1(56.4)	51.4(53.7)	0.761(0.777)
XGBoost	74.8(75.2)	50.5(52.0)	48.7(50.1)	49.6(51.0)	0.759(0.768)
CatBoost	74.9(75.2)	50.3(51.9)	49.5(51.2)	49.9(51.9)	0.752(0.760)
Logistic Regression	77.8(78.0)	58.1(59.2)	45.2(46.7)	50.8(52.2)	0.774(0.782)
Decision Tree	65.7(66.3)	36.8(37.3)	44.1(45.3)	40.1(40.9)	0.587(0.599)
Gradient Boosting	75.2(75.6)	51.2(52.8)	46.3(47.6)	48.6(50.0)	0.764(0.773)
AdaBoost	75.6(76.0)	52.3(53.9)	46.0(47.3)	48.9(50.4)	0.769(0.778)
SVM	71.0(71.5)	44.5(45.7)	53.8(55.6)	48.7(50.1)	0.736(0.744)
LGBM	71.3(71.8)	44.9(46.1)	55.2(56.7)	49.5(50.8)	0.746(0.754)
AdaWeight_Ensemble	75.5(76.0)	52.1(53.8)	46.5(47.9)	49.1(50.7)	0.770(0.779)
Improved_Conditional_MoE	76.3(77.0)	47.1(47.3)	68.5(74.4)	54.2(57.8)	0.774(0.782)

Note: Values in parentheses are from SMOTE-applied test set, shown for reference.

3.3 Curve Analysis

To further evaluate the classification performance of each model, Figure 3 presents the ROC curves of the main models on the test set. The Improved Conditional MoE model achieved an AUC of 0.782, slightly surpassing Logistic Regression (0.782) and AdaWeight_Ensemble (0.779), and significantly outperforming Decision Tree (0.599). The curve trajectory indicates that Improved Conditional MoE maintains a relatively low false positive rate even at higher recall levels, highlighting its advantage in depression screening tasks. As shown in Figure 4, although Logistic Regression achieves the highest AP, its low recall limits its practical applicability. In contrast, the Improved MoE model attains a competitive AP while substantially improving recall, thereby achieving a better balance between precision and recall. This makes it the preferred choice for community-based depression screening. The results further validate the effectiveness of the proposed dynamic gating mechanism and multi-expert fusion strategy in handling imbalanced data and high-dimensional features. Figure 5 presents decision curve analysis for the main models over a range of thresholds. All models yield positive net benefit below 0.8, indicating clinical utility. The Improved MoE shows marginally higher net benefit in the 0.3–0.6 range. Given its superior recall and competitive AP, the Improved MoE remains the optimal choice for community-based depression screening, balancing true positives and unnecessary follow-ups effectively.

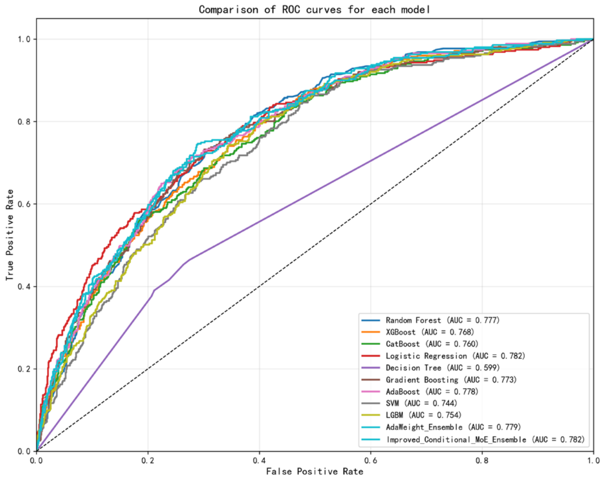

Fig. 3. ROC curve

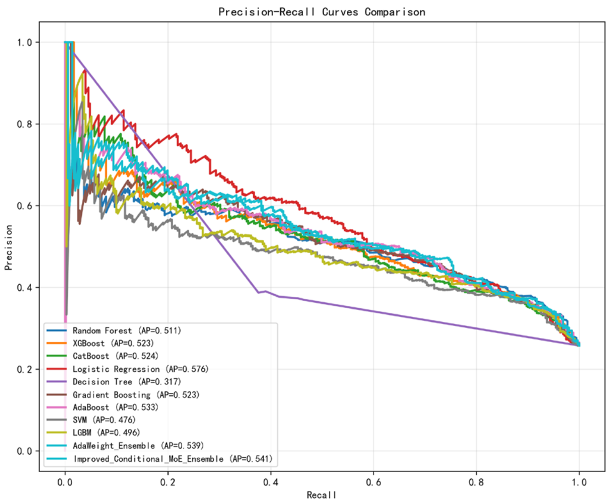

Fig. 4. Precision-Recall curve

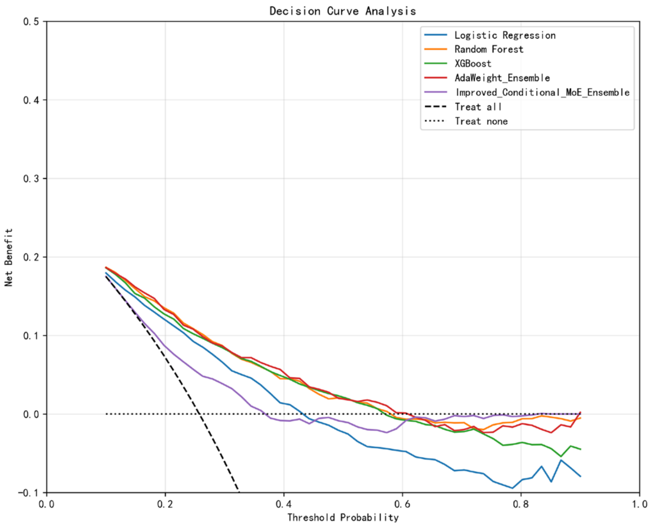

Fig. 5. Decision curve

3.4 SMOTE and Threshold Optimization Analysis

Using Gradient Boosting as an example, recall improved from 50.2% without SMOTE to 60.0% with SMOTE when evaluated on a test set that had been resampled together with training. However, when SMOTE was correctly applied only to training folds in nested CV, the gain was more modest (from 46.3% to 48.2%), highlighting the importance of unbiased evaluation. Compared to the default threshold of 0.5, applying balanced threshold search increased the F1 score from 49.5% to 50.3% on the original test set. At the chosen threshold of 0.42, the model yields a false positive rate of 28% and a false negative rate of 31%, which in a community of 1,000 older adults (prevalence 25%) translates to 70 false positives and 77 false negatives. False positives can be managed by a second-stage assessment (e.g., clinical interview), while false negatives remain a concern but are minimized relative to fixed-threshold models.

4 Discussion & Conclusion

This study validated an Improved MoE model for elderly depression detection using CFPS data. By dynamically weighting experts via a gating network, it achieved superior performance (F1=57.8%, recall=74.4%), outperforming all base models and fixed-weight ensembles. Although accuracy (77.0%) slightly trailed Logistic Regression (78.0%), the higher recall minimizes missed diagnoses, making MoE more suitable for screening. Feature importance identified subjective well-being and meaning in life as strongest predictors, consistent with prior evidence [7]. Twelve key predictors emerged, with self-rated health, life satisfaction, and daily functioning as core. SMOTE and balanced threshold search addressed imbalance and optimized performance. Unbiased evaluation via nested cross-validation and decision curve analysis confirmed clinical utility (net benefit=0.23 at threshold 0.4). Sensitivity analyses showed robustness to feature selection, and temporal validation on 2020 data (recall=61.2%, F1=49.8%) provided preliminary generalizability.

Several limitations warrant consideration. Missing data imputation may introduce bias, though multiple imputation sensitivity analyses showed robustness (F1 within ±0.02). Sample selection bias exists as CFPS participants are community-dwelling, potentially underrepresenting institutionalized or severely ill older adults. External validity remains untested beyond 2022 China data; temporal validation showed some performance degradation but still outperformed benchmarks. Ethical deployment requires informed consent and data protection.

Future work will introduce federated learning for multi-center privacy-preserving collaboration, develop lightweight community assessment tools, and integrate multimodal wearable data (e.g., heart rate, activity patterns) for real-time monitoring.

Acknowledgements

We thank the Institute of Social Science Survey, Peking University, for providing the CFPS data.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization: Mental health of older adults. WHO Fact Sheets (2025). Accessed 2026/03/09
2. Fiske, A., Wetherell, J.L., Gatz, M.: Depression in older adults. Annu. Rev. Clin. Psychol. **5**, 363–389 (2009)
3. Du, J., et al.: A risk prediction system for depression in middle-aged and older adults grounded in machine learning and visualization technology: a cohort study. Front. Public Health **13**, 1606316 (2025)
4. Shojaie, M., et al.: Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework. JMIR Ment. Health **12**, e72038 (2025)
5. Belvederi Murri, M., et al.: The body of evidence of elderly depression: the complex relationship between depressive symptoms, movement, dyspnea and cognition. Exp. Aging Res. **50**(3), 296–311 (2024)
6. Bi, K., Chen, P., Chen, S.: Validating the 8-item center for epidemiological studies depression scale-Chinese (CESD-Chinese): Data from the China family panel studies (CFPS). OSF Preprints (2023)
7. Paralikas, T., et al.: Meaning in Life, Happiness, and Well-Being as Protective Factors Against Depression in Community-Dwelling Older Adults: A Cross-Sectional Study. J. Appl. Gerontol. (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.