In this study, we developed and externally validated an explainable ML model for predicting depression risk among middle-aged and older adults in East Asia, using two large nationally representative cohorts (CHARLS and KLOSA). Among eight algorithms, CatBoost consistently achieved the best discrimination and calibration. SHAP analysis enabled transparent interpretation, highlighting self-rated health, bodily pain, health problems limit work, gender, and child/grandchild able to help respondents with future adl needs as the most influential predictors. These predictors align with known risk factors of late-life depression, lending credibility to the model. To enhance usability, we implemented a traffic-light stratification system, categorizing individuals into high (red), intermediate (yellow), and low (green) risk tiers. Finally, the model was deployed as a multilingual interactive web-based platform, supporting Chinese, Korean, and English. Collectively, these findings demonstrate that explainable ML models can provide scalable, interpretable, and clinically relevant tools for community-based depression screening.

Most existing prediction models for depression risk have been regression-based, relying on assumptions of linearity and often yielding modest discrimination and limited generalizability[1] . For instance, population-based regression models using self-reported health and sociodemographic factors have achieved AUROCs of only 0.65–0.70 [2]. Recent ML studies have reported improved performance, but often lacked external validation or interpretability. A multicenter study of vascular depression applied LightGBM with SHAP explanations and demonstrated strong performance across cohorts, but focused narrowly on cardiovascular comorbidities [3]. In Korea, ML algorithms applied to KLoSA data predicted late-onset depression, yet without explainable frameworks or web-based tools [4]. Other ML studies have primarily used Western cohorts or single-center data, limiting cross-cultural generalizability [5, 6]. Our study extends this literature by systematically benchmarking eight algorithms, integrating SHAP explanations for interpretability, validating across two East Asian cohorts, and deploying a multilingual interactive tool with a traffic-light risk stratification system. These combined features provide added value over previous models, bridging methodological innovation with translational utility.

The predictors identified by SHAP were clinically plausible and consistent with epidemiological evidence. Poor self-rated health was the strongest predictor, reinforcing its established role as a comprehensive indicator of morbidity and mortality, and its strong association with depression risk in older adults [7]. Health problems limit work further contributed to risk, in line with studies linking disability and restricted daily activities to higher depressive symptom burden and poorer prognosis [8]. Bodily pain emerged as another critical determinant, echoing evidence that chronic pain both exacerbates depressive symptoms and complicates treatment outcomes [2]. Gender differences were evident, with women at significantly higher risk of depression, consistent with longitudinal studies across multiple cohorts [9]. Conversely, social support served as a protective factor, aligning with evidence from population-based cohorts that connectedness and perceived support buffer against depression in older adults [10]. These findings support the face validity of our model and underscore the importance of multidimensional psychosocial and functional determinants in late-life depression.

Beyond predictive accuracy, this study highlights several translational innovations that may enhance real-world applicability. The traffic-light stratification system simplifies complex ML outputs into intuitive categories: high-risk (“red”) individuals can be prioritized for immediate referral or closer monitoring, intermediate-risk (“yellow”) individuals for preventive interventions, and low-risk (“blue”) individuals for routine follow-up. This approach parallels ML applications in oncology and dementia, where risk-tiering facilitated adoption by clinicians and caregivers [11]. The multilingual web-based platform further expands accessibility, allowing clinicians, public health practitioners, and community workers across China, Korea, and other settings to engage with the tool in their native languages. This supports global mental health equity initiatives and aligns with national strategies such as *Healthy China 2030* and international calls for scalable community-based interventions. Importantly, by integrating SHAP-based explanations, our tool provides transparency into individual risk profiles, potentially enhancing clinician trust and patient engagement. Together, these features demonstrate how interpretable ML can support equitable, scalable, and community-based mental health screening, moving beyond black-box predictions toward actionable clinical decision support.

This study has several limitations. First, the cross-sectional and observational design restricts causal inference and the ability to capture temporal dynamics of depression risk. Second, although validated across two East Asian cohorts, broader evaluation in diverse ethnic and cultural populations is needed to establish global generalizability. Third, while the CESD-10 is widely used and validated, it may not fully capture the spectrum of depressive disorders, and future studies should incorporate structured diagnostic assessments. Fourth, the absence of biological markers such as genetic, neuroimaging, or inflammatory data constrained the comprehensiveness of our model, though integrating multimodal data has shown promise in improving predictive accuracy and biological plausibility [12].

In conclusion, we present one of the first externally validated, interpretable ML models for depression risk prediction in East Asia, integrating a traffic-light stratification system and multilingual web-based deployment. The model demonstrated strong discrimination and calibration, and SHAP analysis provided clinically plausible explanations. Future research should validate the model across diverse populations, incorporate longitudinal and multimodal data, and extend prediction frameworks to comorbid outcomes such as anxiety and cognitive impairment. Implementation research is also warranted to evaluate real-world effectiveness, cost-effectiveness, and clinician uptake. Embedding the traffic-light system into electronic health records and community health workflows could further facilitate large-scale, equitable depression screening worldwide.

[1] CHEKROUD A M, ZOTTI R J, SHEHZAD Z. Cross-trial prediction of treatment outcome in depression: a machine learning approach [J]. Lancet Psychiatry, 2016, 3(3): 243-50.

[2] BAIR M J, DAVIS G E, KROENKE K. Depression and pain comorbidity: update [J]. BMJ, 2023, 381: e072890.

[3] ZHANG R, LI Y, WANG Y. Identification and validation of an explainable machine learning model for vascular depression: a multicentre cohort study [J]. BMC Medicine, 2025, 23: 448.

[4] KIM J, LEE H, PARK S. Prediction of late-onset depression in elderly Koreans using machine learning algorithms [J]. Scientific Reports, 2025, 15: 85157.

[5] GARRIGA R. A machine learning model to predict mental health crisis using electronic health records [J]. Nature Medicine, 2022, 28: 1240-8.

[6] SHATTE A B R, HUTCHINSON D M, TEAGUE S J. Machine learning in mental health: a systematic review of the literature [J]. Psychological Medicine, 2019, 49(9): 1426-48.

[7] LIVINGSTON G, HUNTLEY J, SOMMERLAD A. Dementia prevention, intervention, and care: 2024 report of the Lancet Commission [J]. Lancet, 2024, 403(10424): 1299-346.

[8] STEFFENS D C. Late-life depression, antidepressant treatment, and cognitive decline [J]. American Journal of Psychiatry, 2024, 181(6): 457-68.

[9] SINGH-MANOUX A, DUGRAVOT A, FOURNIER A. Trajectories of depressive symptoms before diagnosis of dementia: a 28-year follow-up study [J]. JAMA Psychiatry, 2017, 74(7): 712-8.

[10] DORESTE-MENDEZ R, LEE J, ALMEIDA O P. Perception of social support and cognitive performance in older adults with depression [J]. JAMA Network Open, 2023, 6(3): e233978.

[11] HOU F, LI C, WU Y. Development and validation of an interpretable machine learning model for predicting distant metastasis in papillary thyroid cancer: a multicentre study [J]. EClinicalMedicine, 2024, 67: 102913.

[12] PIGONI A. Machine learning for suicide risk prediction: methodological and clinical utility [J]. Molecular Psychiatry, 2024, 29: 161-75.