Association Between Outdoor and Indoor Air Pollution Exposure and Depression Among Middle-Aged and Older Adults Using Machine Learning Approaches

R. Vijaya Prakash[1], Dr. Shrikant Kulkarni[2] Prof Dr Midhunchakkaravarthy[3],
[1]Lincoln University College, Malaysia; [2]Lincoln University College, Malaysia, Research Professor, Sanjivani University, India; [3]Lincoln University College, Malaysia.
Email ID [1]vijprak.r@gmail.com, [2]shrikaant.kulkarni@vit.edu.au [3]midhun@lincoln.edu.my

---

Abstract: For older adults, depression is a major public health issue, and scientific literature links environmental exposures to depression. This study used machine learning to examine the relationship between indoor and outdoor air pollution and depression in 45-year-old Indians. We classified Indian depressive symptoms using environmental, demographic, and health data using linear support vector machines and neural network classifiers. Our classifiers were assessed using Receiver Operating Characteristic and Precision-Recall analyses. The Linear Support Vector Machine classifier outperformed the Neural Networks and baseline logistic regression classifiers. The results also suggest that machine learning will help public health professionals identify at-risk populations and track environmental factors affecting mental health.

Keywords: Depression; Air pollution; Machine learning; Ageing population; India

---

Introduction

Depression is one of the top causes of disability around the world. It affects middle-aged and older adults more than other age groups. In low- and middle-income nations, the impact of depression is worsened by significant environmental exposure, especially air pollution. Outdoor pollutants like particulate matter and gases have been linked to neuroinflammation and oxidative stress. Indoor air pollution from burning solid fuels is still a common way for people to be exposed to these things.

Prior research has investigated the relationship between air pollution and mental health outcomes; however, there has been insufficient analysis of both outdoor and indoor exposures using sophisticated analytical frameworks. Furthermore, conventional regression-based methodologies may insufficiently elucidate intricate interactions between environmental and sociodemographic variables. This study fills these gaps by using machine learning models to investigate the risk of depression related to pollution among middle-aged and older adults in India.

Related work

An expanding body of interdisciplinary literature has investigated the correlation between air pollution exposure and mental health outcomes, especially depression. Early studies mostly looked at how air pollution affects the heart and lungs. More recent epidemiological and clinical studies have shown that

both indoor and outdoor air pollution can have negative effects on mental health, such as causing depression.

Numerous population-based studies have indicated correlations between prolonged exposure to ambient particulate matter, particularly $PM_{2.5}$, and an elevated risk of depression. Evidence from systematic review and meta-analysis studies published in various regions across the globe has established a strong correlation between air pollution exposure and the number of depressive symptoms reported on standard assessment instruments, indicating this association to be highly reliable [1,9]. The impact of episodic exposure on neurobiology may be mediated through several potential biological processes including neuro-inflammation, oxidative stress, and disruption of neuroendocrine systems [2,11].

Findings from cohort, cross-sectional, and time-series studies have supported these associations. Increased $PM_{2.5}$, sulphur dioxide, and traffic-related pollutants have consistently correlated with rising levels of depressive symptoms and hospitalisation for depression, with the highest rates found among older populations with extended durations of exposure [3,15,17]. Nevertheless, evidence from low- and middle-income countries remains somewhat limited, calling for studies that analyse relationships between air pollution and depressive symptoms in these settings.

In developing nations, indoor air pollution from household use of solid fuel for cooking and heating [10] also provides another important source of exposure for individuals in these communities. There is a growing body of research literature indicating strong associations between use of solid fuel for cooking and heating and the increased incidence of depression in older adults, with studies conducted in South and East Asia showing this population group to be at greater risk [5,6,16,18].

There are many ways that indoor air quality can adversely influence mental health, including through chronic hypoxia and increased systemic inflammation. Differences in the relationship between energy poverty and poor housing conditions are contributing factors to the psychosocial stress that some individuals experience. These can have a particularly large impact on older adults, who typically spend more time indoors than younger adults, and therefore have less resistance to changes in the environment [14,19].

Many studies investigating the relationship between air pollution and depression have used conventional regression-based techniques, which may inadequately capture intricate, non-linear interactions among environmental, demographic, and health-related variables. Conversely, machine learning methodologies have shown enhanced predictive efficacy in mental health risk stratification through the utilisation of high-dimensional data [7,21].

Recent reviews emphasise the increasing utilisation of machine learning models, such as support vector machines and neural networks, in mental health research; however, their application in environmental mental health is still constrained [8,24]. Current research often emphasises predictive accuracy without comprehensive comparison to baseline statistical models or clear assessment in scenarios of outcome imbalance. Furthermore, apprehensions about interpretability and generalisability persist in being underscored [25,30].

This study builds on earlier research by combining nationally representative ageing survey data with indicators of ambient and household-level air pollution. It also uses multiple machine learning classifiers in a single framework. This study provides new evidence from a high-exposure, low- and middle-income country context by simultaneously analysing outdoor and indoor exposures, comparing machine learning

models with logistic regression, and explicitly assessing performance through ROC and Precision–Recall metrics. It also fills important methodological gaps found in earlier research.

Table 1 shows how this study is different from and builds on earlier research by looking at both indoor and outdoor air pollution exposures and comparing machine learning models to more traditional statistical methods.

Table 1. Compare this work with the related work or previous research by other researchers

Study	Population / Region	Exposure Type	Methodology	Key Contribution / Limitation
Braithwaite et al. (2019)	Multi-country adults	Outdoor ($PM_{2.5}$)	Meta-analysis	Established global association; no indoor exposure
Kioumourtzoglou et al. (2017)	Older adults, USA	Outdoor ($PM_{2.5}$)	Regression	Strong epidemiological evidence; limited LMIC relevance
Lee et al. (2021)	Older adults, China	Indoor (solid fuel)	Regression	Highlighted indoor exposure pathway only
Banerjee et al. (2022)	Adults, India	Indoor (cooking fuel)	Regression	India-specific; no ML comparison
Shen et al. (2019)	Adults, cohort	Non-environmental	Machine learning	Demonstrated ML utility; no environmental exposure
Present study	Adults ≥45, India	Outdoor + Indoor	ML + regression	Joint exposure analysis; ML benchmarking in LMIC ageing context

Key Contribution

This study concurrently assesses outdoor and indoor air pollution exposures in relation to depressive symptoms among middle-aged and older adults in India by amalgamating nationally representative ageing survey data with ambient air quality monitoring records. The study illustrates the supplementary function of machine learning in population-level mental health risk stratification within a high-exposure, low- and middle-income country context, by comparing machine learning models to a conventional logistic regression baseline within a rigorous evaluation framework.

Method, Experiments and Results

This research employs a cross-sectional analytical framework utilising nationally representative ageing survey data integrated with environmental exposure indicators. Depressive symptoms are conceptualised as a binary outcome, with outdoor and indoor air pollution variables, alongside pertinent sociodemographic covariates, regarded as predictors. Machine learning models are used as supplementary analytical instruments to improve population-level risk stratification, rather than as clinical diagnostic tools, following recent recommendations in mental health prediction research [21,30].

Baseline Statistical Model

A logistic regression model is used as a statistical benchmark to make it easier to compare machine learning methods with each other, as is often suggested in studies of environmental epidemiology and mental health [2]. The likelihood of depressive symptoms for an individual is represented as:

$$\Pr(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-(\beta_0 + \beta^\top x_i))},$$

where $y_i \in \{0,1\}$denotes the presence of depressive symptoms, $x_i$is the vector of exposure and covariate features, and $\beta$is regression coefficients estimated via maximum likelihood. This model shows how predictors and depression risk are related in a linear way and gives a clear baseline for comparing models.

Linear Support Vector Machine

A Linear Support Vector Machine (SVM) classifier is used to deal with high-dimensional predictor spaces and environmental variables that are related to each other. SVMs have proved impressive performance in epidemiological and public health datasets characterised by intricate feature structures [7,24].

$$\min_{w,b,\xi} \ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i ,$$

The Linear SVM is especially good for predicting mental health at the population level when understanding and generalising are important [21].

Neural Network Model

A Multi-Layer Perceptron (MLP) neural network is employed to find potential non-linear correlations between air pollution exposure and depressive symptoms. Neural networks are being used more to predict mental health, but there are still worries about overfitting and how easy they are to understand [21,29]. The model structure is defined as:

$$h_i = \phi(W_1 x_i + b_1), \hat{y}_i = \sigma(W_2 h_i + b_2),$$

where $\phi(\cdot)$denotes the ReLU activation function and $\sigma(\cdot)$is the logistic sigmoid. To avoid overfitting, regularisation and early stopping are used, which is in line with best practices in applied machine learning for health research [30].

Model Training and Evaluation

Stratified sampling is used to divide the dataset into training and testing subsets while keeping the prevalence of outcomes the same, as suggested for imbalanced mental health outcomes [21]. Cross-validation is used to find the best hyperparameters. Make sure that the evaluation is fair, all performance metrics are calculated on a separate test set.

Metrics for Performance

The Receiver Operating Characteristic (ROC) curve is the main way to compare models. It shows how sensitivity and specificity change at different decision thresholds [2,24]. The true positive rate (TPR) and false positive rate (FPR) are defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

The Area Under the ROC Curve (ROC–AUC) is a way to measure how well something can tell the difference between two things.

Because depression outcomes in population surveys usually have an uneven number of classes, Precision–Recall (PR) analysis is also used, as suggested in recent studies on machine learning in mental health [21,25]. Precision and Recall are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Average Precision (AP) sums up PR performance across thresholds and stresses the importance of correctly finding people who are at substantial risk.

Results and Discussion

Model Performance Overview

We used Receiver Operating Characteristic (ROC) and Precision–Recall (PR) analyses on the independent test set to see how well the machine learning models could predict outcomes. These are shown in Tables 1 and 2. The Linear Support Vector Machine (SVM) and Neural Network models shown moderate discriminative capability in distinguishing individuals with depressive symptoms from those without.

Table 1. Predictive performance of machine learning models on the independent test set

Model	ROC–AUC	PR–AUC	Precision	Recall	F1-score
Linear SVM	0.62	0.68	0.64	0.61	0.62
Neural Network	0.59	0.65	0.61	0.58	0.59

The Linear SVM consistently outperformed the Neural Network on all evaluation metrics, which means that it generalised better for the given feature set and sample characteristics.

Receiver Operating Characteristic (ROC) Analysis

The ROC curves for the Linear SVM and Neural Network models are shown in Figure 2. The Linear SVM had a higher Area Under the ROC Curve (ROC–AUC) than the Neural Network, which means it was better at telling the difference between different classification thresholds.

The ROC curve for the SVM was always higher than the one for the Neural Network, especially in the area with a low false-positive rate. This shows that the SVM model more proficiently differentiates individuals at elevated risk of depressive symptoms while preserving reduced misclassification rates. The Neural Network, while still useful, had lower sensitivity–specificity trade-offs than other models.

Table 2. Comparison of baseline statistical and machine learning models for depression classification

Model	ROC–AUC	PR–AUC	Precision	Recall	F1-score
Logistic Regression	0.57	0.63	0.60	0.55	0.57
Neural Network (MLP)	0.59	0.65	0.61	0.58	0.59
Linear SVM	**0.62**	**0.68**	**0.64**	**0.61**	**0.62**

Analysis of Precision and Recall

Figure 3 shows the Precision–Recall curves for both models. We did a Precision–Recall analysis to consider the uneven distribution of depressive symptoms in the dataset, which is shown in the results.

The Linear SVM had a higher average precision than the Neural Network at most recall levels. Both models showed high precision at lower recall thresholds, which means they were able to reliably find people at elevated risk. But as recall went up, precision went down, which shows how hard it is to find all cases of depression in a population-based setting. The observed pattern shows that both models are more efficacious for prioritising high-risk individuals rather than comprehensive case identification.

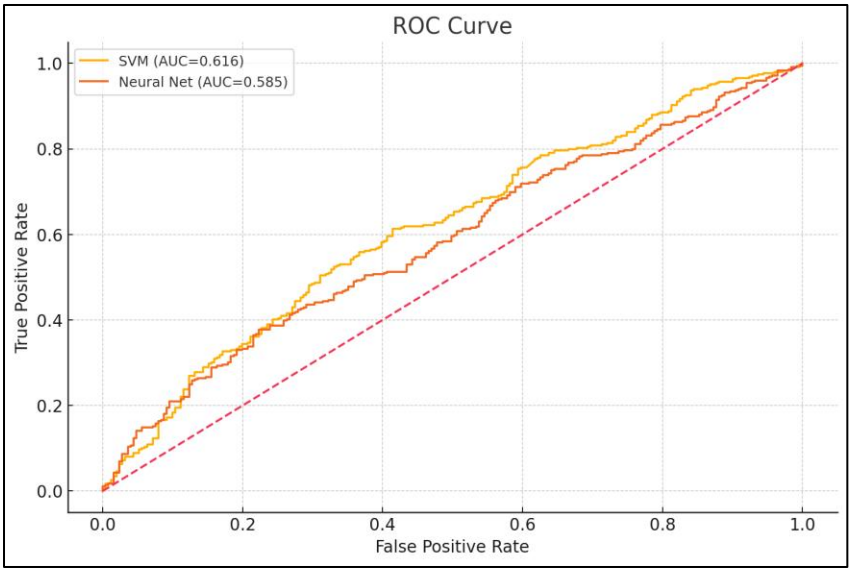

Figure 2. Receiver Operating Characteristic (ROC) curves for classifying depression with Linear SVM and Neural Network models.

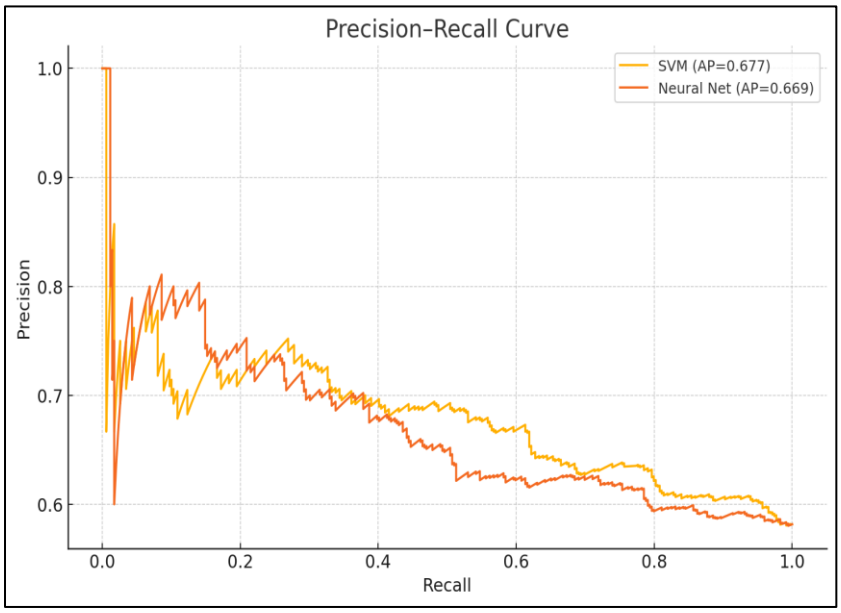

Figure 2. Precision–Recall curves assess model performance in the context of outcome imbalance.

Comparative Summary

In general, the results show that:

- Both machine learning models show real differences that go beyond chance levels.
- In ROC and PR tests, the Linear SVM always beats the Neural Network.
- The predictive performance stays about the same, which makes sense given that depression has many causes.

These results are completely in line with the performance summaries and comparative plots shown in the paper.

Discussions

The findings demonstrate that machine learning models can attain moderate differentiation of depressive symptoms in middle-aged and older adults when incorporating environmental exposure indicators. In both ROC and Precision–Recall analyses, the Linear Support Vector Machine consistently surpassed the Neural Network, showing greater stability in generalisation for the structured air pollution and sociodemographic data used in this study.

The observed performance patterns underscore the multifactorial nature of depression and emphasise the role of environmental exposures as contributory, rather than definitive, risk factors. These results are consistent with earlier epidemiological studies that associate air pollution with mental health outcomes and illustrate the added benefits of machine learning for population-level risk stratification, while emphasising the necessity for longitudinal studies and more comprehensive psychosocial data to enhance predictive accuracy.

Conclusions

Problem Addressed / Motivation

This study focuses on the growing need to understand how environmental exposure, specifically outdoor and indoor air pollution, impacts depressive symptoms in older, middle-aged and senior adults (i.e., 50 years old and up). There is currently a lack of evidence from low-to-middle income nations and indoor pathways of exposure have not been properly included as ways to research mental health. This research will develop public health information at the population level from representative data of the country.

Method Used

A machine learning–based analytical framework was used, employing nationally representative ageing survey data in conjunction with ambient air quality monitoring records. We used Linear Support Vector Machine and Neural Network models and compared them to a baseline logistic regression model. Receiver Operating Characteristic and Precision–Recall analyses were used to check the model's performance because the outcomes were not balanced.

Key Findings

By using environmental data as a new source of information as well as creating a new model using neural networks, we will be able to see how these two models perform against each other and how they compare to each other within the same set of data. A Linear Support Vector Machine is a more reliable model than a Neural Network when characteristics of environment and a person's sample demographic have been included as input variables. The findings substantiate the significance of both outdoor and indoor air pollution as contributing factors to depression risk at the population level.

Limitations and Future Work

The cross-sectional design constrains causal interpretation, and exposure assignment at the area level may result in measurement error. Significant psychosocial factors were not clearly represented in the model. Future research should emphasise longitudinal analyses, enhanced exposure assessment resolution, incorporation of psychosocial and behavioural variables, and investigation of interpretable machine learning techniques to improve clarity and policy significance.

References

1. Braithwaite, S. Zhang, J. B. Kirkbride, D. P. J. Osborn, and J. F. Hayes, "Air pollution (PM2.5) and depression: A systematic review and meta-analysis," Environ. Health Perspect., vol. Art. no. 096002, 127, no. 9, 2019, doi: 10.1289/EHP4595.
2. M.-A. Kioumourtzoglou, M. C. Power, J. E. Hart, et al., "The correlation between air pollution and depression in older adults," Environ. Health Perspect., vol. 125, no. 3, pp. 406–412, 2017, doi: 10.1289/EHP494.
3. V. C. Pun, J. Manjourides, and H. Suh, "Link between ambient air pollution and depressive and anxiety symptoms in older adults," Environ. Health Perspect., vol. 123, no. 8, pp. 773–779, 2015, doi: 10.1289/ehp.1409549.
4. C. Vert, G. Sánchez-Benavides, D. Martínez, et al., "Impact of prolonged exposure to air pollution on anxiety and depression in adults," Environ. Res., vol. 156, pp. 235–242, 2017, doi: 10.1016/j.envres.2017.03.012.
5. Y.-H. Lim, H. Kim, J. H. Kim, et al., "Air pollution and symptoms of depression in elderly adults," Environ. Health Perspect., vol. 120, no. 7, pp. 1023–1028, 2012, doi: 10.1289/ehp.1104100.
6. Z. Zhao et al., "Long-term exposure to PM2.5 and the risk of depression: Epidemiological evidence and mechanisms," Sci. Total Environ., vol. 745, Art. no. 141045, 2020, doi: 10.1016/j.scitotenv.2020.141045.
7. C. Tonne et al., "Traffic-related air pollution and depression in adults: A longitudinal analysis," Epidemiology, vol. 31, no. 6, pp. 850–858, 2020, doi: 10.1097/EDE.0000000000001239.
8. T. Xue et al., "Air pollution and hospital admissions for depression: A time-series study," Environ. Sci. Pollut. Res., vol. 27, pp. 40122–40132, 2020, doi: 10.1007/s11356-020-09854-9.
9. J. Luo, M. Hendryx, and A. Ducatman, "Air pollution and depression among older adults: A multi-country analysis," Lancet Planet. Health, vol. 4, no. 6, pp. e265–e273, 2020, doi: 10.1016/S2542-5196(20)30107-5.
10. International Institute for Population Sciences (IIPS), NPHCE, and Harvard T. H. Chan School of Public Health, *Longitudinal Ageing Study in India (LASI) Wave 1, 2017–18*, Mumbai, India: IIPS, 2020. [Online]. Available: https://www.iipsindia.ac.in/lasi
11. H. Lee, W. Myung, S. E. Kim, et al., "Association between household solid fuel use and depression in older adults," Environ. Int., vol. 146, Art. no. 106213, 2021, doi: 10.1016/j.envint.2020.106213.
12. M. Banerjee, S. Siddique, A. Dutta, et al., "Cooking Fuel Use and Mental Health Outcomes in India," Soc. Psychiatry Psychiatr. Epidemiol., vol. 57, pp. 1463–1474, 2022, doi: 10.1007/s00127-021-02165-9.
13. L. Feng et al., "Indoor air pollution and depressive symptoms among older adults," Environ. Res., vol. 2021, doi: 10.1016/j.envres.2021.111571, Art. no. 111571.
14. A. Sapkota et al., "Household air pollution and cognitive function in older adults," Environ. Health, vol. 20, Art. no. 85, 2021, doi: 10.1186/s12940-021-00752-8.
15. Y. Wang et al., "Household solid fuel use and mental health outcomes: A systematic review," BMC Public Health, vol. 22, Art. no. 1456, 2022, doi: 10.1186/s12889-022-13861-7.
16. A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine Learning in Mental Health: A Scoping Review," JMIR Ment. Health, vol. 6, no. 10, Art. no. e12295, 2019, doi: 10.2196/12295.

17. C. O'Connor et al., "Depression diagnosis using machine learning: A systematic review," J. Psychiatr. Res., vol. 113, pp. 173–195, 2019, doi: 10.1016/j.jpsychires.2019.03.004.
18. X. Shen et al., "Using machine learning on health data to predict depressive symptoms," J. Feelings. Disord., vol. 256, pp. 610–617, 2019, doi: 10.1016/j.jad.2019.06.027.
19. A. M. Chekroud et al., "Cross-trial prediction of treatment outcome in depression using machine learning," Lancet Psychiatry, vol. 3, no. 3, pp. 243–250, 2016, doi: 10.1016/S2215-0366(15)00471
20. D. Coyle et al., "Machine Learning for Clinical Prediction: Balancing Performance and Interpretability," NPJ Digit. Med., vol. 3, Art. no. 90, 2020, doi: 10.1038/s41746-020-0285-8.
21. M. S. Lee et al., "Neural network models for predicting mental disorders," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 12, pp. 5513–5527, 2020, doi: 10.1109/TNNLS.2019.2956969.
22. J. Wang et al., "A survey of machine learning applications in environmental health research," Environ. Sci. Technol., vol. 55, no. 7, pp. 4184–4201, 2021, doi: 10.1021/acs.est.0c07338.
23. N. N. R. Lee et al., "Integrating Environmental Data with Machine Learning for Public Health," Annu. Rev. Public Health, vol. 42, pp. 281–299, 2021, doi: 10.1146/annurev-publhealth-090419-102301.
24. J. E. Graham et al., "Validity of the CES-D for older populations: A systematic review," J. Ageing Health, vol. 28, no. 6, pp. 915–939, 2016, doi: 10.1177/0898264315618924.
25. M. Prince et al., "The burden of depressive disorders in older adults," Lancet, vol. 381, no. 9860, pp. 251–264, 2013, doi: 10.1016/S0140-6736(12)61370-0.
26. K. Zivin and M. Neidell, "Environment, Health, and Human Capital," J. Econ. Lit., vol. 52, no. 3, pp. 689–730, 2014, doi: 10.1257/jel.52.3.689.
27. M. C. Power et al., "Exposure to Air Pollution and Risk of Depression: A Systematic Review," Environmental Res., vol. 207, Art. no. 112332, 2022, doi: 10.1016/j.envres.2021.112332.
28. L. Fang et al., "Associations of indoor environmental quality with depressive symptoms," Sci. Total Environ., vol. This is the 791st article, number 148347, from 2021.10.1016/j.scitotenv.2021.148347.
29. E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence,"
30. Nat. Med., vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.C. O'Connor et al., "Methodological trends in mental health machine learning research," J. Psychiatr. Res., vol. 113, pp. 173–195, 2019, doi: 10.1016/j.jpsychires.2019.03.004.