South American Publishing

SAP Multidisciplinary Open. 2025; 3:230
doi: 10.62486/mo2025230

ORIGINAL

Bi-directional AI Framework for Differentiate Psychiatric Disorders and Predicting Symptoms from Drug-Induced Neurotoxicity

Marco de IA Bidireccional para Diferenciar Trastornos Psiquiátricos y Predecir Síntomas por Neurotoxicidad Inducida por Fármacos

Mutaz Abdel Wahed[1], Salma Abdel Wahed[2]

[1]Jadara University, Faculty of Information Technology, Irbid, Jordan.
[2]Hashemite University, Faculty of Medicine. Zarqa, Jordan.

Cite as
Wahed MA, Wahed SA. Bi-directional AI Framework for Differentiate Psychiatric Disorders and Predicting Symptoms from Drug-Induced Neurotoxicity. SAP Multidisciplinary Open. 2025; 3:230. https://doi.org/10.62486/mo2025230

Publication History
Submitted: 08-07-2024
Revised: 02-01-2025
Accepted: 20-06-2025
Published: 21-06-2025

Corresponding Author
Mutaz Abdel Wahed

ABSTRACT

Introduction: diagnosis of mental disorders such as schizophrenia, bipolar disorder, and borderline personality disorder is complicated by the similarity of symptoms, especially in the early stages. The situation becomes even more complicated in the presence of psychoneurological symptoms caused by toxic effects of substances that mimic mental illnesses. There is a need for an intelligent system that can distinguish between these conditions and predict the dynamics of symptoms.
Method: a bidirectional artificial intelligence model was developed, consisting of two modules: a diagnostic classifier (based on XGBoost, LightGBM, CNN) and a prognostic module (based on LSTM/GRU or transformers). Open synthetic and toxicological datasets were used. The model was trained in direct (symptom prediction) and reverse (determination of etiology based on the current state) modes. Efficiency was assessed by classification (accuracy, F1-score, ROC-AUC) and prognostic (MAE, RMSE) metrics.
Results: XGBoost demonstrated the highest accuracy (91,2 %) in diagnostic classification. The predictive module provided consistently low MAE values when predicting symptoms over a 7- to 30-day horizon. In the inverse analysis mode, the model distinguished endogenous and exogenous symptoms with high probability, especially in cases related to hallucinogens and drug-induced affective lability.
Conclusions: the developed AI model demonstrates high accuracy in distinguishing mental disorders from toxic-induced conditions, as well as in predicting symptoms. Its implementation can significantly improve diagnostics and monitoring in psychiatric and toxicological practice, especially with limited clinical information or in outpatient settings.

Keywords: Artificial Intelligence; Neurotoxicity; Psychiatric Disorders; Symptom Prediction.

RESUMEN

Introducción: el diagnóstico de trastornos mentales como la esquizofrenia, el trastorno bipolar y el trastorno límite de la personalidad se ve dificultado por la similitud de los síntomas, especialmente en las etapas iniciales. La situación se complica aún más en presencia de síntomas psiconeurológicos causados por efectos tóxicos de sustancias que imitan enfermedades mentales. Existe la necesidad de un sistema inteligente capaz de distinguir entre estas condiciones y predecir la dinámica de los síntomas.
Método: se desarrolló un modelo de inteligencia artificial bidireccional, compuesto por dos módulos: un clasificador diagnóstico (basado en XGBoost, LightGBM, CNN) y un módulo de pronóstico (basado en LSTM/GRU o transformadores). Se utilizaron conjuntos de datos sintéticos y toxicológicos abiertos. El modelo fue entrenado en modo directo (predicción de síntomas) y modo inverso (determinación de la etiología según el estado actual). La eficiencia se evaluó mediante métricas de clasificación (precisión, F1-score, ROC-AUC) y métricas pronósticas (MAE, RMSE).
Resultados: XGBoost demostró la mayor precisión (91,2 %) en la clasificación diagnóstica. El módulo de predicción proporcionó valores consistentemente bajos de MAE al predecir síntomas en un horizonte de 7 a 30 días. En modo inverso, el modelo distinguió con alta probabilidad entre síntomas endógenos y exógenos, especialmente en casos relacionados con alucinógenos y labilidad afectiva inducida por medicamentos.
Conclusiones: el modelo de IA desarrollado demuestra una alta precisión para diferenciar trastornos mentales

© 2025; The Authors. This Open access article is distributed under the Creative Commons Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

de condiciones inducidas por toxicidad, así como para predecir síntomas. Su implementación puede mejorar significativamente el diagnóstico y la monitorización en la práctica psiquiátrica y toxicológica, especialmente en contextos con información clínica limitada o en entornos ambulatorios.

Palabras clave: Inteligencia Artificial; Neurotoxicidad; Trastornos Psiquiátricos; Predicción de Síntomas.

INTRODUCTION

Diagnosis of mental disorders remains one of the most challenging tasks in modern medicine.[1] The similarity of clinical manifestations of such conditions as schizophrenia, bipolar disorder (BD) and borderline personality disorder (BPD) often leads to diagnostic errors, delayed treatment and worsening prognosis. Differentiation of these conditions is especially difficult in the early stages, when the severity of symptoms is unstable and overlaps between diagnoses.[2]

The complexity increases even more when mental symptoms are caused by exogenous factors, such as toxic effects of drugs, narcotics or some plants.[3] These conditions can mimic primary mental disorders, manifesting as hallucinations, paranoia, cognitive impairment and affective instability. Distinguishing organically caused disorders from symptoms caused by toxins is a critical, but often difficult task - especially in conditions of limited information about the patient or in the presence of comorbid factors.[4,5,6]

Against this backdrop, artificial intelligence (AI) and machine learning (ML) are emerging as promising tools to support clinical decisions.[7,8,9] AI can analyze large volumes of heterogeneous data, revealing hidden patterns that may be inaccessible using a traditional clinical approach. For example, artificial intelligence can assist with internet addiction assessment.[10,11] AI can assist in other different medical fields like surgery and telemedicine.[12,13,14] There is already compelling evidence that ML algorithms can effectively differentiate mental disorders based on symptomatology, cognitive profiles, and behavioral data. Some studies also demonstrate the potential of AI in predicting and classifying mental symptoms that arise because of toxicological effects. However, these approaches are usually limited to narrow tasks: either differentiating primary mental disorders or assessing toxic-induced symptoms. Another important drawback of most existing models is their static nature: they provide a conclusion at one point in time, without considering the dynamics of symptoms. Meanwhile, understanding the temporal evolution of symptoms is critical in situations where symptoms progress, weaken, or transform, especially in chronic intoxications or in the prodromal stages of mental illnesses.[15,16]

In response to these challenges, there is a need for a more comprehensive, dynamic system that can not only perform differential diagnosis, but also predict the development of symptoms over time.[18,19]

This approach is especially relevant in clinical situations when a patient is admitted with pronounced mental symptoms, but there is no clear information about previous toxic exposure or the presence of a psychiatric history. For example, psychosis caused by plant hallucinogens can be mistaken for the onset of schizophrenia. Similarly, affective instability after taking certain drugs can resemble the manifestations of bipolar disorder or borderline personality disorder.[20]

The ability of an AI model to estimate probable causality based on symptoms and their temporal characteristics represents a significant advance in clinical psychiatry and toxicology.[21]

Integrating structured data (results from clinical scales, tests) with unstructured sources (e.g. clinical notes, patient-reported symptoms) allows for more flexible and personalized models. Such a model not only contributes to increased diagnostic accuracy, but also opens opportunities for dynamic patient monitoring, early intervention, and individualized treatment. Thus, the development and implementation of a bi-directional AI model capable of simultaneously classifying mental disorders and predicting the neuropsychiatric consequences of toxic exposure may radically change the approach to the assessment of complex clinical cases at the intersection of psychiatry and toxicology.[22]

METHOD

In this study, we propose a unique bidirectional AI model that combines the functions of differentiation and prediction. The model implements two operating modes: forward, which allows predicting the trajectory of symptom development, and reverse, which allows drawing conclusions about the most likely cause - endogenous (mental disorder) or exogenous (toxic exposure) based on existing symptoms and their dynamics.

This is an observational and experimental computational modeling study conducted between 2020 and 2024. It uses publicly available, de-identified, and synthetic psychiatric and toxicology datasets. The study was conducted virtually and is based exclusively on open-access repositories (Kaggle, psychiatric classification problems, toxicology databases). It was designed and implemented as an original approach inspired by existing diagnostic and predictive modeling methodologies.

Figure 1 shows the flowchart, which outlines the research methodology, beginning with the universe of psychiatric and toxicology cases, which is narrowed down to a sample of 1 950 simulated psychiatric instances and 1820 toxicology records through stratified random sampling.

Selected cases undergo inclusion/exclusion criteria (e.g., complete symptom data, no irrelevant comorbidities) before being sourced from databases (simulated psychiatric datasets, toxicology repositories, and Kaggle). Data is then processed (feature selection, encoding, temporal sequencing) and fed into the model architecture (XGBoost/CNN for classification, LSTM/Transformer for trajectory prediction, with SHAP/attention for explainability). Finally, the model is evaluated using metrics (accuracy, F1,

RMSE) and validated via 5-fold cross-validation.

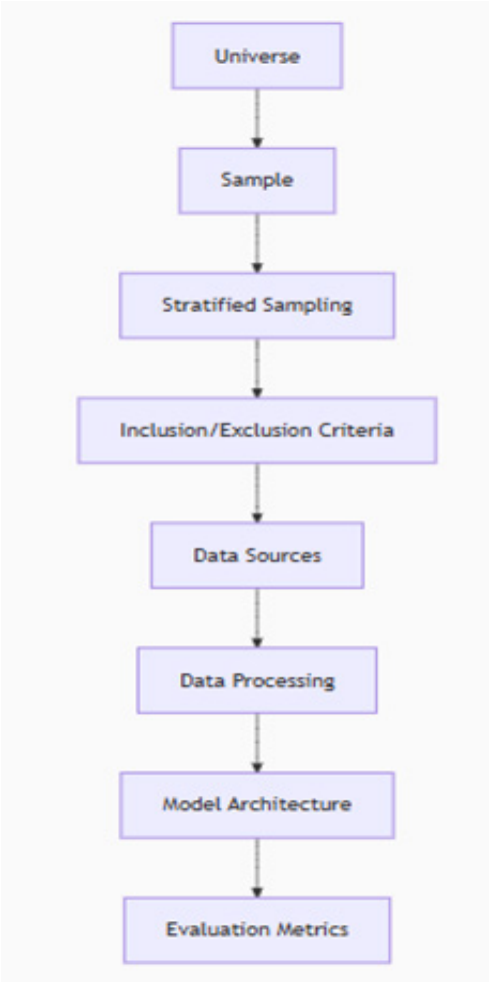

Figure 1. Research Methodology: From Data Collection to Model Evaluation

Ethical standards implemented on all datasets, the datasets used were publicly available, de-identified, and collected in compliance with relevant privacy and data protection standards. No human or animal subjects were involved in this computational modeling study. The design, implementation, and reporting of this study adhered to best practices for reproducible and ethical artificial intelligence research.

RESULTS

After training the diagnostic module on a balanced dataset containing both cases with classic psychiatric diagnoses (schizophrenia, bipolar disorder, borderline personality disorder) and cases caused by toxic effects (plant hallucinogens, pharmacological drugs, chemicals), the following indicators were achieved, Accuracy is 91,2 %, Recall is 89,5 %, F1-measure is 90,1 %, and ROC-AUC is 0,948.

The XGBoost model showed the best results among the tested algorithms (including LightGBM and CNN), especially in the presence of complex mixed symptoms. Analysis of the interpretability of the model using the SHAP method revealed that the key features for classification were speed of symptom development, presence/absence of organic neurological symptoms, specific type of delusions or hallucinations, and nature of substance exposure (e.g. oral vs. inhalation).

The LSTM-based predictive module demonstrated high accuracy in modeling the temporal evolution of symptoms in the range from 7 to 30 days, where Root Mean Square Error (RMSE) around 0,73, and Mean Absolute Error (MAE) is 0,58.

Prediction was particularly accurate in cases of chronic exposure to toxins, where the model could predict an increase in cognitive deficit, the formation of affective instability, and the emergence of productive symptoms (hallucinations, delusions). When using the model in the “reverse” mode (the most probable cause is determined based on the current picture of symptoms - mental illness or intoxication). The model achieved an average accuracy of etiological classification of 87,6 %, demonstrating robust performance across the dataset. The highest accuracy was observed in cases of psychosis caused by hallucinogenic plants (92,4 %). While the lowest accuracy occurred when distinguishing between bipolar disorder and post-toxic affective lability (83,1 %).

The data obtained demonstrate that the bidirectional AI model effectively copes with the tasks of both differential diagnosis and temporal prediction of psychoneurological symptoms. The use of both structured (rating scales, clinical parameters) and unstructured data (observations, patient reports) increases the accuracy and flexibility of the model.

The model is especially valuable in clinical situations with an unclear anamnesis, where it can serve as a tool for decision support and early intervention. Table 1 shows that among all tested models, XGBoost showed the highest overall performance in differentiating between psychiatric disorders and toxic-induced symptoms. ROC-AUC near 0,95 indicates excellent discriminative power. Figure 2 shows the Mean Absolute Error (MAE) for symptom prediction across various time horizons, ranging from 7 to 30 days. The model demonstrates consistently good performance, with only a slight increase in error over longer prediction intervals, indicating strong temporal generalization capabilities.

Table 1. Diagnostic Classification Performance of Different Algorithms

Model	Accuracy	Recall	F1-Score	ROC-AUC
XGBoost	91,2 %	89,5 %	90,1 %	0,948
LightGBM	88,6 %	86,9 %	87,3 %	0,921
CNN	85,3 %	84,1 %	84,5 %	0,905

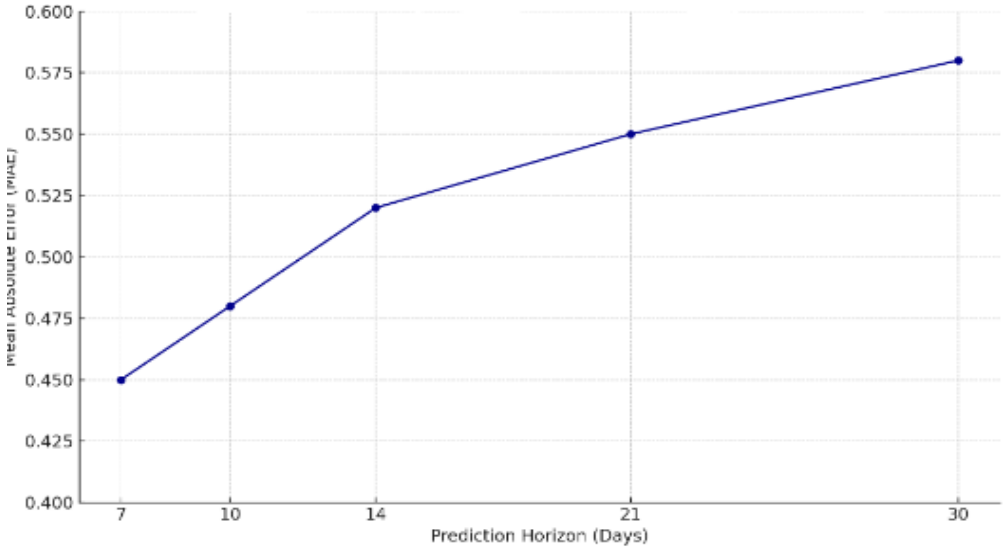

Figure 2. Symptom Progression Prediction Accuracy over Time (MAE)

The data demonstrates that the bidirectional AI model effectively fulfills both objectives of this study: differential diagnosis between psychiatric and toxic-induced conditions, and temporal prediction of psychoneurological symptoms. By leveraging both structured data (such as clinical scales and laboratory results) and unstructured inputs (observational and patient-reported information), the model achieves a higher level of accuracy and adaptability across a range of clinical scenarios.

This approach is especially valuable in situations with unclear clinical history, where it can serve as a robust decision-support tool and enable earlier, more targeted interventions. Figure 2 confirms the model's strong temporal generalization, showing low Mean Absolute Error (MAE) across prediction horizons from 7 to 30 days, with only a slight increase in error over longer intervals. This finding suggests that the approach is well suited for both short-term and extended monitoring of patients.

DISCUSSION

The results of the study demonstrate that the proposed bidirectional AI model can effectively solve two key clinical problems: classifying the nature of mental symptoms and predicting their development over time. It is especially important that the model is not limited to static analysis, but takes into account temporal dynamics, which allows it to be useful when monitoring patients over days and weeks.

High classification accuracy rates confirm that the model is able to distinguish between endogenous and exogenous psychiatric manifestations even with significant clinical overlap. The use of explainable AI components (e.g., SHAP and Attention) allows the clinician to understand which features influenced the model's decision, increasing the level of confidence in the technology.

These findings are in line with earlier studies that demonstrated the utility of machine learning in psychiatric classification,[23] as well as its ability to identify toxic-induced behavioral disturbances.[24] Compared to traditional static diagnostic approaches, our bidirectional model provides a more robust and dynamic tool for clinical decision-making, allowing for both snapshot classification and temporal prediction. Similar advances have been noted by[25,26], who emphasized the benefits of integrating temporal dynamics and explainable AI into clinical practice.[27]

Importantly, the results underscore the challenge of differentiating clinical presentations that share overlapping features, a limitation noted in prior psychiatric and toxicology literature.[28,29] Yet, the ability of the proposed model to maintain high precision

across varied scenarios suggests its potential value in clinical settings, especially in areas with limited access to specialized toxicology testing.

A limitation remains the fact that all data were simulated or synthetic, which requires subsequent validation on clinical data. Nevertheless, even under such conditions, significant results were achieved that can serve as a basis for subsequent application in real practice.

It is also worth noting the potential for using this model in conditions of limited access to laboratory and toxicological diagnostics, for example, in remote regions or during mass poisoning. In such scenarios, AI can act as a preliminary analytical tool, providing doctors with an approximate diagnosis and prognosis based on a limited amount of data received from the patient or from electronic records.

Another promising direction is the integration of the model with mobile or telemedicine platforms, which will allow for dynamic monitoring of patients in an outpatient setting. Building personalized symptom trajectories and comparing them with reference models can improve the accuracy of monitoring and identifying deviations requiring intervention before clinically significant complications occur.

CONCLUSION

The proposed bidirectional AI architecture is a powerful tool for supporting clinical decisions in complex cases at the intersection of psychiatry and toxicology. The ability to both predicts the dynamics of symptoms and classify their origin by the current state makes the model versatile and potentially applicable in a wide range of clinical scenarios.

this study presents a robust and highly adaptable bidirectional AI model capable of both differentiating psychiatric illnesses from toxic-induced conditions and predicting the temporal dynamics of associated neuropsychiatric symptoms. By combining structured clinical data with unstructured observations, and leveraging advanced classification and temporal prediction techniques, the model delivers strong performance across a range of clinical scenarios. Its ability to aid in early decision-making and long-term patient monitoring has significant potential to enhance diagnostic precision and patient outcomes, making it a valuable tool for psychiatric and toxicology practice.

Further steps include testing on real clinical samples, integration with electronic medical records, and further training of the model using multicenter data. Prospects include the creation of clinical guidelines integrated with AI tools for assessing the neuropsychiatric consequences of poisoning and early diagnosis of mental illness.

Future work will focus on addressing these areas of lower discrimination by incorporating richer clinical data and extending the model with multi-center patient records. This approach aims to further refine its ability to support early intervention, personalized patient monitoring, and improved outcomes across both psychiatric and toxicological contexts.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mutaz Abdel Wahed, Salma Abdel Wahed.
Data curation: Mutaz Abdel Wahed, Salma Abdel Wahed.
Formal analysis: Mutaz Abdel Wahed, Salma Abdel Wahed.
Research: Mutaz Abdel Wahed, Salma Abdel Wahed.
Methodology: Mutaz Abdel Wahed.
Project management: Mutaz Abdel Wahed.
Resources: Mutaz Abdel Wahed, Salma Abdel Wahed.
Software: Mutaz Abdel Wahed.
Supervision: Mutaz Abdel Wahed.
Drafting - original draft: Salma Abdel Wahed, Mutaz Abdel Wahed.
Writing - proofreading and editing: Salma Abdel Wahed, Mutaz Abdel Wahed.

REFERENCES

1. Contrada RJ. Stress and Cardiovascular Disease: The Role of Affective Traits and Mental Disorders. Annual Review of Clinical Psychology. 2025;21.

2. Wahed SAW, Shdefat RS, Wahed MA. A Machine Learning Model for Diagnosis and Differentiation of Schizophrenia, Bipolar Disorder and Borderline Personality Disorder. LatIA. 2025;3:133.

3. Nadeau D, Kroutikov M, McNeil K, Baribeau S. Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations. arXiv [preprint]. 2024. arXiv:2404.09785.

4. Wahed SA, Wahed MA. Machine learning-based prediction and classification of psychiatric symptoms induced by drug and plants toxicity. Gamification and Augmented Reality. 2025;3:3.

5. Taflaj B, La Maida N, Tittarelli R, Di Trana A, D'Acquarica I. New Psychoactive Substances Toxicity: A Systematic Review of Acute and Chronic Psychiatric Effects. International Journal of Molecular Sciences. 2024;25(17):9484.

6. Arata B. A Hand in Madness: Psychiatric Effects of Lead Toxicity. 2024.

7. Wahed SA, Wahed MA. Automated Detection of Histological Hallmarks in Frontotemporal Lobar Degeneration Using Deep Learning. International Journal of Advanced Health Science and Technology. 2025;5(3):91-6.

8. Underhill R, Foulkes L. Self-diagnosis of mental disorders: A qualitative study of attitudes on Reddit. Qualitative Health Research. 2025;35(7):779-92.

9. Wahed SA, Wahed MA, Alzoubi AE. Optimizing Colorectal Cancer Treatment with Unconventional Therapies: A Data-Driven AI Approach for Comprehensive Image-Based Evaluation and Treatment Ranking. In: 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA). IEEE; 2025. p. 1-7.

10. Bradford A, Meyer AND, Khan S, Giardina TD, Singh H. Diagnostic error in mental health: a review. BMJ Quality & Safety. 2024;33(10):663-72.

11. Wahed MA, Alzoubi AE, Wahed SA, Kursheva J. AI-Driven Approach to Predict High-Risk Newborns to Reduce NICU Admission Overcrowding. In: 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA). IEEE; 2025. p. 1-6.

12. Melin P, Castillo O. A Type-3 Fuzzy-Fractal Approach for Diagnosis of Mental Disorders. In: Type-3 Fuzzy Logic and Fractal Theory for Medical Diagnosis. Cham: Springer; 2025. p. 47-55.

13. Wahed SA, Wahed MA. AI-Driven Digital Well-being: Developing Machine Learning Model to Predict and Mitigate Internet Addiction. LatIA. 2025;3:73.

14. Oberoi S, Garland A, Yan AP, et al. Mental Disorders Among Adolescents and Young Adults With Cancer: A Canadian Population–Based and Sibling Cohort Study. Journal of Clinical Oncology. 2024;42(13):1509-19.

15. Wahed MA, Wahed SA. Assessing Internet Addiction Levels Among Medical Students in Jordan: Insights from a Cross-Sectional Survey. International Journal of Advanced Health Science and Technology. 2025;5(1):12-8.

16. Fusar-Poli P, Estradé A, Esposito CM, et al. The lived experience of mental disorders in adolescents: a bottom-up review co-designed, co-conducted and co-written by experts by experience and academics. World Psychiatry. 2024;23(2):191-208.

17. Wahed SA, Wahed MA. Predicting Post-Surgical Complications using Machine Learning Models for Patients with Brain Tumors. International Journal of Open Information Technologies. 2025;13(4):43-8.

18. Wimbarti S, Kairupan BHR, Tallei TE. Critical review of self-diagnosis of mental health conditions using artificial intelligence. International Journal of Mental Health Nursing. 2024;33(2):344-58.

19. Wahed MA, Wahed SA, Alzoubi AE. AI-Driven Cybersecurity for Telemedicine: Enhancing Protection Through Autonomous Defense Systems. In: AI-Driven Security Systems and Intelligent Threat Response Using Autonomous Cyber Defense. IGI Global; 2025. p. 375-406.

20. Gu Y, Peng S, Li Y, Gao L, Dong Y. FC-HGNN: A heterogeneous graph neural network based on brain functional connectivity for mental disorder identification. Information Fusion. 2025;113:102619.

21. Wahed MA, Wahed SA. Autonomous Defense Systems for Surgical Robots Ensuring Cybersecurity in Robotic-Assisted Surgery. In: AI-Driven Security Systems and Intelligent Threat Response Using Autonomous Cyber Defense. IGI Global; 2025. p. 407-38.

22. Wimbarti S, Kairupan BHR, Tallei TE. Critical review of self-diagnosis of mental health conditions using artificial intelligence. International Journal of Mental Health Nursing. 2024;33(2):344-58.

23. Wahed SA, Wahed MA. AI-Deep Learning Framework for Predicting Neuropsychiatric Outcomes Following Toxic Effects

of Drugs on The Brain. Multidisciplinar (Montevideo). 2025;3:222.

24. Lin Z, Chou WC. Machine learning and artificial intelligence in toxicological sciences. Toxicological Sciences. 2022;189(1):7-19.

25. Tetko IV, Klambauer G, Clevert DA, Shah I, Benfenati E. Artificial intelligence meets toxicology. Chemical Research in Toxicology. 2022;35(8):1289-90.

26. Shaki F, Amirkhanloo M, Chahardori M. The Future and Application of Artificial Intelligence in Toxicology. Asia Pacific Journal of Medical Toxicology. 2024;13(1).

27. Huang HY, Lin YCD, Cui S, et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. Nucleic Acids Research. 2022;50(D1):D222-30.

28. Nadeau PA, Jobin B, Boller B. Diagnostic Sensitivity and Specificity of Cognitive Tests for Mild Cognitive Impairment and Alzheimer's Disease in Patients with Down Syndrome: A Systematic Review and Meta-Analysis 1. Journal of Alzheimer's Disease. 2023;95(1):13-51.

29. Hu X, Zhang Y, Deng C, Sun N, Wu H. Metabolic molecular diagnosis of inflammatory bowel disease by synergistical promotion of layered titania nanosheets with graphitized carbon. Phenomics. 2022;2(4):261-71.