UCS1001

Assignment 2

*Critical Reader Response to "***Early detection of mental health disorders using machine learning models using behavioral and voice data analysis***"*

Done by:

Student Name	Student ID
Muhammad Hasif Bin Mohd Faisal	2500619

Declaration:

I/we hereby pledge that this <------------------------------------ > task is not plagiarised and has been written wholly as a result of my/our own research and compilation of information.

Signed:

Dated: 12/10/2025

The article “Early Detection of Mental Health Disorders Using Machine Learning Models Incorporating Behavioural and Voice Data Analysis” by Sunil Kumar Sharma et al. (2025), published in *Scientific Reports*, presents NeuroVibeNet, a multimodal machine-learning framework designed to identify early signs of mental health disorders. The authors claim that by combining behavioural and voice data through advanced pre-processing, feature reduction, and ensemble modelling, the system can effectively distinguish normal from pathological states, achieving about 99% accuracy and outperforming existing baseline models.

Although the study’s multimodal framework is technically strong, its reliance on limited, non-diverse benchmark datasets, and the absence of longitudinal or real-world clinical validation, compromise its generalisability and practical applicability in healthcare.

Acknowledged but unresolved challenges

Following the discussion on dataset limitations, it is important to note that Sharma et al. (2025) explicitly acknowledge these challenges in their problem statement, emphasising the need for “high-quality and diverse datasets” and “clinical validation”; yet their model relies only on narrow benchmarks without external testing.

Limited benchmark datasets and clinical diversity

Sharma et al. (2025) build NeuroVibeNet on two publicly available datasets: the Mental Disorder Classification dataset from Kaggle for behavioural features, and the MODMA dataset for voice features. According to dataset documentation, the Kaggle dataset comprises around 120 participants with 17 recorded symptom variables (Kaggle, n.d.), while the MODMA audio subset includes about 52 participants, of whom 23 had depression and 29 were controls (Cai et al., 2022). Such modest sample sizes limit statistical power, as psychiatric machine-learning studies usually require hundreds to thousands of cases for precision psychiatry (Bzdok & Meyer-Lindenberg, 2018). Moreover, research shows that limited datasets can inflate accuracy while failing to generalise, as models may overfit to spurious correlations (Geirhos et al., 2020) or label imbalances (He & Garcia, 2009), instead of capturing meaningful patterns (Flint et al., 2021). Such biases also raise fairness concerns, since under-represented groups may receive less accurate predictions, potentially exacerbating disparities in mental healthcare. While small, public datasets can serve as proof-of-concept tools given recruitment barriers, they remain far from deployment readiness.

Lack of longitudinal and real-world validation

Longitudinal validation, typically spanning follow-up periods of six months to several years across hundreds of participants (Dinga et al., 2018; Iniesta et al., 2016), evaluates whether predictive accuracy remains stable as patients' symptoms evolve across settings (Onnela, 2021).

Sharma et al. (2025) relied solely on the Kaggle and MODMA datasets, with no external or temporal validation. As previous research shows, many AI models that perform well on constrained datasets fail to replicate their results in heterogeneous populations (Nagendran et al., 2020). Addressing this limitation requires substantial investment and coordination, and cost-sharing consortia alongside transparent governance can promote fair benefit distribution across healthcare systems. The World Health Organization (WHO, 2023) similarly warns that unvalidated mental-health AI risks undermining safety, equity, and trust. This raises not only methodological but also ethical concerns, since deploying inadequately validated tools risks undermining patient trust and may compromise safety if errors go undetected in clinical practice.

Mitigation strategies for reliable and equitable AI

Given the limitations identified above, the next steps should be pragmatic and testable. Larger, more representative cohorts can be built through multicentre collaborations and privacy-preserving data sharing, with federated learning offering a practical route to leverage data without centralisation (Rieke et al., 2020). Secure enclaves, data-use agreements, GDPR/HIPAA-aligned governance, and phased cost-sharing pilots provide workable models (Javed et al., 2024; Kalkman et al., 2019). These measures curb overfitting, demonstrate stability across settings, and reveal subgroup performance gaps for remediation via reweighting or targeted recruitment (Lee et al., 2018). Systems should align with clinical workflows and remain transparent through regular audits and updates, consistent with WHO guidance on safe and equitable AI (WHO, 2023). Future regulatory changes or shifts in social attitudes towards AI may also require recalibration or retraining of such systems to remain trustworthy and equitable. These strategies align with broader calls in mental-health AI for transparent, scalable, and clinically validated systems.

In summary, NeuroVibeNet's technical strengths are offset by limited, non-diverse datasets and a lack of longitudinal validation, restricting claims of clinical reliability. These limitations are interdependent, reflecting broader issues in mental-health AI where strong in-lab performance rarely translates to real-world practice (Lee et al., 2018; WHO, 2023; Rieke et al., 2020; Kalkman et al., 2019). A credible path forward combines diverse cohorts, longitudinal follow-up, and multicentre trials under transparent governance to safeguard fairness, safety, and equity (WHO, 2023).

References

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Cai, H., Yuan, Z., Gao, Y., Sun, S., Li, N., Tian, F., … Hu, B. (2022). A multi-modal open dataset for mental-disorder analysis. *Scientific Data, 9*, 178. https://doi.org/10.1038/s41597-022-01211-x

Dinga, R., Schmaal, L., Penninx, B. W. J. H., van Tol, M. J., Veltman, D. J., van Velzen, L., Mennes, M., van der Wee, N. J. A., & Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage: Clinical, 22*, 101796. https://doi.org/10.1016/j.nicl.2019.101796

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D. M. A., Emden, D., Winter, N. R., Leenings, R., Eickhoff, S. B., Kircher, T., Krug, A., Nenadic, I., Arolt, V., Clark, S., Baune, B. T., Jiang, X., Dannlowski, U., & Hahn, T. (2021). Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology, 46*(8), 1510–1517. https://doi.org/10.1038/s41386-021-01020-7

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine, 46*(12), 2455–2465. https://doi.org/10.1017/S0033291716001367

Javed, H., Muqeet, H. A., Javed, T., Rehman, A. U., & Sadiq, R. (2024). Ethical frameworks for machine learning in sensitive healthcare applications. *IEEE Access, 12*, 16233–16254. https://doi.org/10.1109/ACCESS.2023.3340884

Kaggle. (n.d.). *Mental disorder classification dataset.* Kaggle. https://www.kaggle.com/datasets/cid007/mental-disorder-classification

Kalkman, S., Mostert, M., Gerlinger, C., van Delden, J. J. M., & van Thiel, G. J. M. W. (2019). Responsible data sharing in international health research: A systematic review of principles and norms. *BMC Medical Ethics, 20*(1), 21. https://doi.org/10.1186/s12910-019-0359-9

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A. P., … McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders, 241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ, 368*, m689. https://doi.org/10.1136/bmj.m689

Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology, 46*(1), 45–54. https://doi.org/10.1038/s41386-020-0771-3

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., … Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1

Sharma, S. K., Alutaibi, A. I., Khan, A. R., Tejani, G. G., Ahmad, F., & Mousavirad, S. J. (2025). Early detection of mental health disorders using machine learning models incorporating behavioral and voice data analysis. *Scientific Reports, 15*, 16518. https://doi.org/10.1038/s41598-025-00386-8

World Health Organization. (2023). *Ethics and governance of artificial intelligence for health: WHO guidance.* https://www.who.int/publications/i/item/9789240029200