Investigating Microbiome's Role in Neuropsychiatric Disorders in Quest of Novel Therapeutics Using Computational Methods: A research proposal

Shakil Ahmed Rafi

July 3, 2024

Abstract

The gut microbiome has been extensively and fruitfully studied recently using machine learning tools, yielding novel connections between neuropsychiatric disorders and the presence and abundance of various microbiota. However, despite the recent successes, the oral microbiota has received little attention in the literature, let alone the machine learning treatment that the gut microbiome receives. *This presents a promising research gap.* The author notices that an emerging consensus for an oral-microbiome-brain exists, that several robust datasets of the oral microbiome already exist, that novel attentive transformer based models have shown incredible promise as regressors and classifier but still underused in genomic analysis, and that variational auto-encoder based clustering of the oral microbiome do not yet exist. Therefore, a growing and rather promising potential exists in the literature for this gap to be fulfilled, and proposes that this be their post-doctoral research goal.

1 Introduction

The role of human host microbiomes in neuropsychiatric diseases has been studied extensively in the literature, e.g. for reviews see [1], [2], and [3]. Furthermore they have been implicated in a number of neuropsychiatric disorders, such as ADHD, in [4], in increasing severity of autism spectrum disorders, ASD in children [5], and e.g. in Alzheimer's Disease in the elderly, in [6] and [7].

1.1 The gut microbiome in contrast to the oral microbiome

Of the different microbiomes in the human body, (e.g. gut, oral, skin, vaginal), the gut microbiome is the most extensively studied in relation to neuropsychiatric disorders, see [8]. It is not only the most studied but also the microbiome where modern machine learning techniques have been most frequently and fruitfully applied.

This leaves a research gap for other microbiomes of the human body. For instance both [1] and [9] note the dearth of literature, especially machine learning literature, with respect to the oral microbiome.

We propose, therefore, to take some of the tools, especially machine learning tools, used in gut microbiome analysis and apply them for the analysis of the oral microbiome. We propose this for four reasons:

i. The oral microbiome is yet under-studied with the machine learning tools applied to other microbiomes, as noted earlier in [1] and [9].

ii. An emerging argument exists for an oral-microbiome-brain axis, OMBA, similar to the gut-brain axis in [10], [11], [12], and [13]. The consensus seems to be that the interplay between the oral microbiome and mental health disorders is complex, and that this area still needs to be studied, see the literature review [14].

iii. Large, robust, and mature datasets exist for the oral microbiomes, such as the Human Oral Microbiome Dataset, and extended Human Oral Microbiome Dataset[15], Cultivated Oral Bacterial Genome Reference [16], and even some smaller datasets such as the U.A.E. Healthy Future Study participants of 330 Emirati citizens [17], and less specialized datasets such as FinnGen, [18].

iv. As noted in [9], the oral microbiome is readily accessible for therapeutic intervention com-

pared to the gut microbiome, allowing quick and cost-effective measures to be developed and for them to be effective.

2 Machine Learning Methods to be Applied

Our main source of inspiration will be methods already applied to the gut microbiome, we will use the framework given in [19], and explore some of the hypotheses in [9]. Particular attention will be paid to the use of `TabNet` and attentive transformer based regression and classification method and variational auto-encoders as a preprocessor to clustering. The justifications will follow.

2.1 Regression and Classification via `TabNet`

Regression in its various forms have a long history of use with predictions from microbiome, e.g. LASSO regression in using the blood microbiome to predict gut α-diversity in [20].

Indeed logistic and linear regression models have also had some use in e.g. predicting autism spectrum disorders, ASD from the oral microbiome, [21].

What the literature seems to lack is more sophisticated neural network regression and classification techniques like `TabNet` (introduced in [22]). Because genomic data is well-known to be large and yet quite sparse simpler regression methods may not be the best at this task.

`TabNet` was introduced to tackle just this kind of problem. It uses an attention-based mechanism using a transformer-based [23] architecture where-in the model learns from a sparse but wide tabular data, extracting the salient features of the dataset as it reads the data along. It updates feature importances based on new data that it reads and assigns attention to these features updating the model. This also yields interpretable results.

Attentive transformer based models have shown incredible promise in other areas of AI research such as Large Language Models and in computer vision with vision transformers and indeed continues to be the landmark paradigm for machine learning as of 2024. Indeed benchmarking with the TabZilla Benchmarking Suite shows that `TabNet` may out-perform traditional gradient-boosted decision trees in contexts where the dataset is extremely large with high dimensionality and where there is large sparseness in the tabular structure, [24].

These features make advanced attention based neural networks particularly appealing for genomic analysis where data is known to be high dimensional. Combined with the fact that the literature is already thin in the case of the oral microbiome this presents a fertile area of research.

As an added bonus `TabNet` has a robust implementation in PyTorch [25] not only as a regressor but also as a classifier, making it easy to laterally transfer from regression to classification.

We must note however the valid concern that neural network methods are thought to be less "interpretable" compared to standard statistical methods. This is largely mitigated with `TabNet`, for instance [26] notes `TabNet` to be interpretable and provides "better or comparable" results to XGBoost & GLM with predicting insurance claims. On top of that modern tools like SHAPS [27] and LIME [28], largely bridge the interpretability gap.

The author therefore proposes using `TabNet`, and the related `TabPFN` [29] for smaller datasets as a novel way of exploring the impact of the oral microbiome in neuropsychiatric disorders.

2.2 Clustering via Variational Auto-encoders

In a similar vein to the previous section on regression and classification we note that microbiome research, especially gut microbiome research has had a long and fruitful use of clustering algorithms, e.g. [30] looks at four such clustering techniques, k-Means, hierarchical, partition around medoids, and Dirichlet multinomial model.

Because genomic data often has very high dimensions it is often desirable to perform dimensionality reduction before any clustering algorithms can be used. There has been some promising applications for variational auto-encoders, VAEs, to reduce the dimensionality down and then feed the data to clustering algorithms. This technique has seen use especially in the detection of cancer e.g. [31] and [32], and in PubMed only one example that the author can find for metagenome binning, [33].

VAEs also have the added benefit of reducing noise, making the data smaller by discarding redundant features similar to principal component analysis (PCA) and related principal co-ordinate analysis (PoCA), making other, more traditional, analyses easier on the dataset.

Regardless, a research gap still exists in genomic analysis of the oral microbiome using variational auto-encoders. This gap is even more salient considering that software tools like DeepMicro [34], a suite

of tools for the use of variational auto-encoders for genomic analysis, already exist, dramatically reducing the software overhead for such an analysis.

Finally, it must be noted that the exact mechanism by which the oral microbiome affects the brain is still up for debate. Several possible mechanisms have been suggested in [9], see Figure 1. What is clear is that this mechanism could possibly be more "direct" in the sense that for the gut microbiome, any toxins released into the bloodstream must first be metabolized by the liver before it reaches the bloodstream and brain [35] whereas this is not the case for the oral microbiome [9].

This, thus means that we may expect to see stronger associations between the presence and respective α and β diversities of different microbiota and the expected predicted neuropsychiatric disorders. However this is speculative, and more review on the part of the author is needed.

This represents a further area of fruitful research for the post-doc.

3 Objective and Reason for Research

The target of our research will be to see the impact different species of microbiota on neuropsychiatric disorders. The author tentatively proposes focusing on disorders such as major depressive disorder, MDD, Alzheimer's disease, AD, and autism spectrum disorder, ASD, although more literature review will be needed to select a prime candidate disorder to focus on.

Among the many reasons the research proposal is especially focused on the oral cavity is that it is relatively straightforward to focus interventions to the oral cavity. Quick, targeted, and effective interventions to the oral microbiome exist in the form of tablets, sprays, simple diet change, and lifestyle changes such as flossing. The oral microbiome is therefore a prime and easy target in the care for neuropsychiatric disorders.

4 Tentative Timeline and Expected Contributions and Impact

The author proposes the following timeline for their post-doc assuming a two year period. The pipeline takes inspiration from the one proposed in [19], see Figure 2. Again, this is subject to change but mostly in the direction of needing less time. Also this assumes a two-year post-doc period.

i. Months 1-6: Explore the proposed datasets, dive deeper into the literature, and run preprocessing tasks including feature selection. This may include dimensionality reduction techniques like principal component analysis PCA, or singular value decomposition, SVM. The author also proposes to run models on a smaller subset of the data as trial runs.

ii. Months 7-12: Construct large scale models from the datasets, taking inspiration from existing literature. Steps include parameter tuning, benchmarking, and cross-validation.

iii. Months 13-18: Evaluate models using an appropriate metric such as RMSE, confusion matrix etc. At this stage, we may retune or reengineer features depending on the models predictive ability, or chose an alternative model altogether though this is unlikely.

iv. Months 19-24: Disseminate the findings in the form of papers, posters, and talks given.

Note that this timeline envisions a handful of models which will trained, evaluated, and retrained again, aiming for enough results to yield at-least two papers as a target.

We expect to see similar or even stronger correlations from the oral microbiome as we do with the gut microbiome, partly because of the absence of the liver metabolizing toxins that are secreted by the oral microbiome into the bloodstream [35]. Indeed, if it is seen that the same species in the oral microbiome is associated with the same neuropsychiatric disorder as its presence in the gut microbiome this bolsters the argument for the species as a causative agent in said neuropsychiatric disorder.

This research will therefore not only bolster findings from the gut microbiome but will also provide a more accessible way of altering certain microbiomes as therapeutics.

5 Conclusion

We therefore see that combined with the presence of robust and mature datasets, a thin presence in the literature, a lack of sophisticated machine learning treatment (attentive transformers, variational autoencoders) so far, and finally the potential for discovering extremely strong associations given how easily toxins enter the bloodstream from the oral cavity in

the absence of the mediating liver [35], and the ease with which therapeutic interventions can be made, the author sees an extremely fruitful avenue of research.

The author therefore proposes this as a postdoctoral research project.

References

[1] Goswami, A. *et al.* Role of Microbes in the Pathogenesis of Neuropsychiatric Disorders. *Front Neuroendocrinol* **62**, 100917 (2021). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8364482/.

[2] Hashimoto, K. Emerging role of the host microbiome in neuropsychiatric disorders: overview and future directions. *Mol Psychiatry* **28**, 3625–3637 (2023).

[3] Bonnechre, B., Amin, N. & van Duijn, C. The role of gut microbiota in neuropsychiatric diseases creation of an atlas-based on quantified evidence. *Frontiers in Cellular and Infection Microbiology* **12** (2022). URL https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2022.831666.

[4] Bull-Larsen, S. & Mohajeri, M. H. The Potential Influence of the Bacterial Microbiome on the Development and Progression of ADHD. *Nutrients* **11** (2019). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6893446/. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[5] Tomova, A. *et al.* Gastrointestinal microbiota in children with autism in slovakia. *Physiology & Behavior* **138**, 179–187 (2015). URL https://www.sciencedirect.com/science/article/pii/S0031938414005101.

[6] Yk, K. & C, S. The Microbiota-Gut-Brain Axis in Neuropsychiatric Disorders: Pathophysiological Mechanisms and Novel Treatments. *Current neuropharmacology* **16** (2018). URL https://pubmed.ncbi.nlm.nih.gov/28925886/. Publisher: Curr Neuropharmacol.

[7] Escobar, Y.-N. H., O'Piela, D., Wold, L. E. & Mackos, A. R. Influence of the Microbiota-Gut-Brain Axis on Cognition in Alzheimer's Disease. *J Alzheimers Dis* **87**, 17–31 (2022).

[8] Sorboni, S. G., Moghaddam, H. S., Jafarzadeh-Esfehani, R. & Soleimanpour, S. A Comprehensive Review on the Role of the Gut Microbiome in Human Neurological Disorders. *Clin Microbiol Rev* **35**, e0033820 (2022).

[9] Tao, K., Yuan, Y., Xie, Q. & Dong, Z. Relationship between human oral microbiome dysbiosis and neuropsychiatric diseases: An updated overview. *Behav Brain Res* **471**, 115111 (2024).

[10] Bowland, G. B. & Weyrich, L. S. The Oral-Microbiome-Brain Axis and Neuropsychiatric Disorders: An Anthropological Perspective. *Front Psychiatry* **13**, 810008 (2022).

[11] Xi, Y., Yu, M., Li, X., Zeng, X. & Li, J. The coming future: The role of the oralmicrobiotabrain axis in aroma release and perception. *Comprehensive Reviews in Food Science and Food Safety* **23**, e13303 (2024). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1541-4337.13303. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1541-4337.13303.

[12] Martnez, M. *et al.* The Role of the Oral Microbiota Related to Periodontal Diseases in Anxiety, Mood and Trauma- and Stress-Related Disorders. *Front Psychiatry* **12**, 814177 (2021).

[13] Y, M. *et al.* Did the Brain and Oral Microbiota Talk to Each Other? A Review of the Literature. *Journal of clinical medicine* **9** (2020). URL https://pubmed.ncbi.nlm.nih.gov/33260581/. Publisher: J Clin Med.

[14] Skallevold, H. E., Rokaya, N., Wongsirichat, N. & Rokaya, D. Importance of oral health in mental health disorders: An updated review. *J Oral Biol Craniofac Res* **13**, 544–552 (2023).

[15] Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**, baq013 (2010). URL https://doi.org/10.1093/database/baq013. https://academic.oup.com/database/article-pdf/doi/10.1093/database/baq013/1132285/baq013.pdf.

[16] Li, W. *et al.* A catalog of bacterial reference genomes from cultivated human oral bacteria. *npj Biofilms Microbiomes* **9**, 1–13 (2023). URL https://www.nature.com/articles/s41522-023-00414-3. Publisher: Nature Publishing Group.

[17] Human Oral Microbiota Composition Data from UAE Healthy Future Study (UAEHFS) Pilot Participants. URL https://datacatalog.med.nyu.edu/dataset/10404.

[18] FinnGen: an expedition into genomics and medicine | FinnGen. URL https://www.finngen.fi/en.

[19] Li, P., Luo, H., Ji, B. & Nielsen, J. Machine learning for data integration in human gut microbiome. *Microb Cell Fact* **21**, 241 (2022).

[20] Wilmanski, T. *et al.* Blood metabolome predicts gut microbiome -diversity in humans. *Nat Biotechnol* **37**, 1217–1228 (2019). URL https://www.nature.com/articles/s41587-019-0233-9. Publisher: Nature Publishing Group.

[21] Li, C. *et al.* A genetic association study reveals the relationship between the oral microbiome and anxiety and depression symptoms. *Front Psychiatry* **13**, 960756 (2022). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9685528/.

[22] Arik, S. . & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *AAAI* **35**, 6679–6687 (2021). URL https://ojs.aaai.org/index.php/AAAI/article/view/16826.

[23] Vaswani, A. *et al.* Attention is All you Need. In *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[24] McElfresh, D. *et al.* When do neural nets outperform boosted trees on tabular data? (2023). 2305.02997.

[25] pytorch-tabnet: PyTorch implementation of TabNet. URL https://github.com/dreamquark-ai/tabnet.

[26] McDonnell, K., Murphy, F., Sheehan, B., Masello, L. & Castignani, G. Deep learning in insurance: Accuracy and model interpretability using tabnet. *Expert Systems with Applications* **217**, 119543 (2023). URL https://www.sciencedirect.com/science/article/pii/S0957417423000441.

[27] Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[28] Ribeiro, M. T., Singh, S. & Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 11351144 (Association for Computing Machinery, New York, NY, USA, 2016). URL https://doi.org/10.1145/2939672.2939778.

[29] Hollmann, N., Mller, S., Eggensperger, K. & Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second (2022). 2207.01848.

[30] Shi, Y., Zhang, L., Peterson, C. B., Do, K.-A. & Jenq, R. R. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome* **10**, 25 (2022). URL https://doi.org/10.1186/s40168-021-01199-3.

[31] Hira, M. T. *et al.* Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep* **11**, 6265 (2021). URL https://www.nature.com/articles/s41598-021-85285-4. Publisher: Nature Publishing Group.

[32] Zhang, X. *et al.* Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 765–769 (2019).

[33] Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* **39**, 555–560 (2021).

[34] Oh, M. & Zhang, L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci Rep* **10**, 6026 (2020). URL https://www.nature.com/articles/s41598-020-63159-5. Publisher: Nature Publishing Group.

[35] Refisch, A. *et al.* Microbiome and immunometabolic dysregulation in patients with major depressive disorder with atypical clinical presentation. *Neuropharmacology* **235**, 109568 (2023). URL https://www.sciencedirect.com/science/article/pii/S0028390823001582.

6 Figures

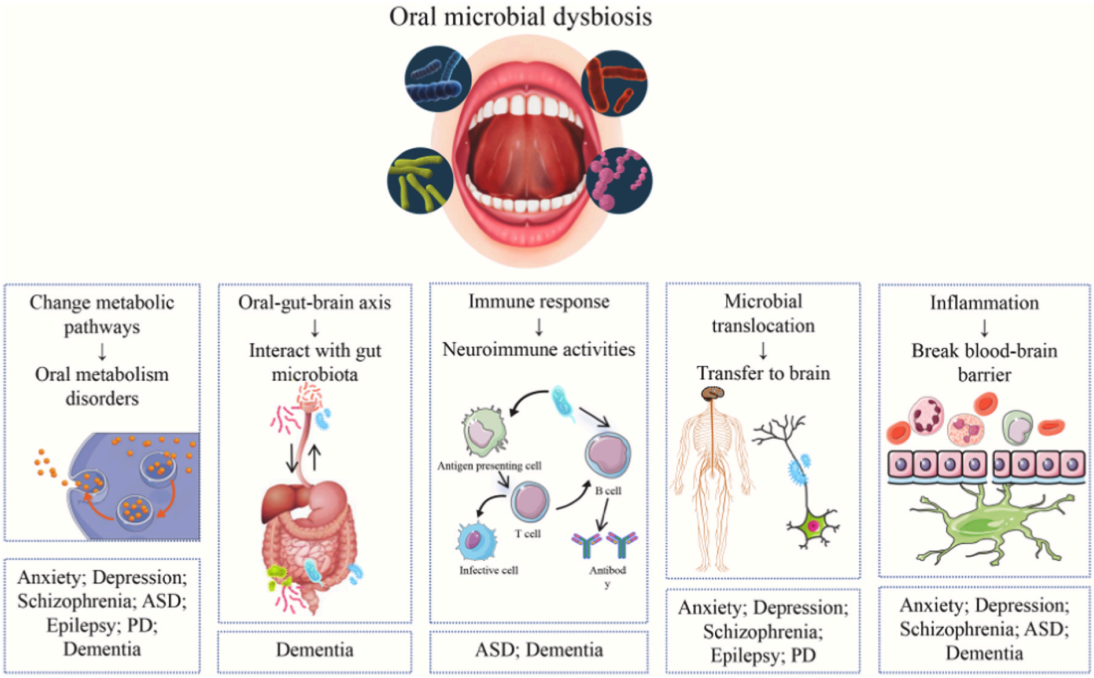

Figure 1: Proposed mechanisms by which the oral microbiome may affect in neuropsychiatric disorders. Figure copied from [9], pg. 3.

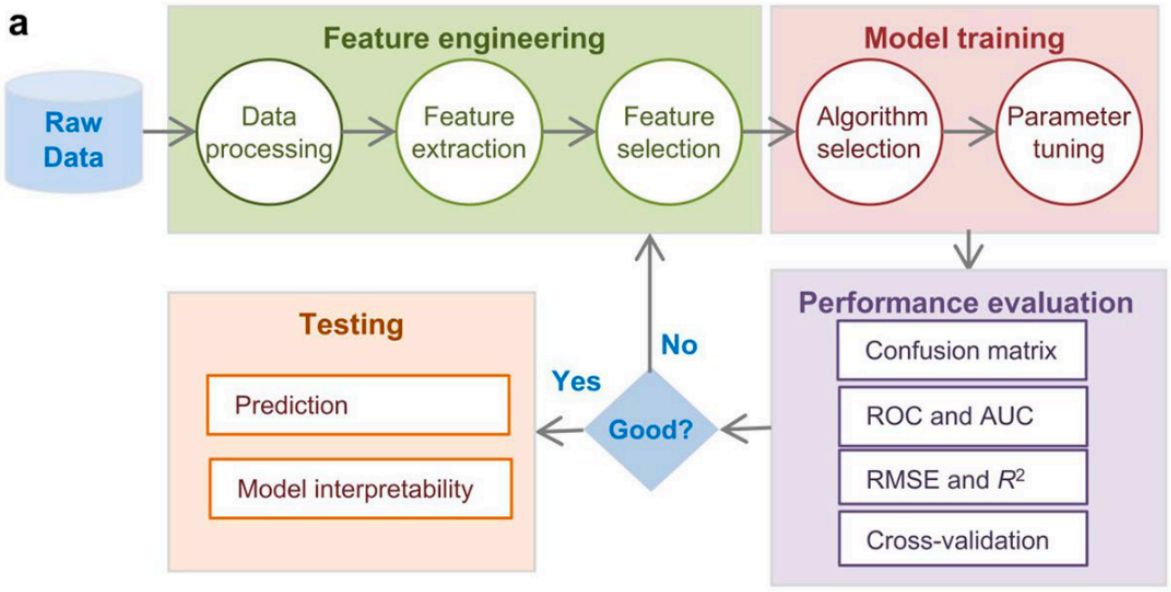

Figure 2: Proposed workflow for ML pipelines in genomic analysis, Figure taken from [19].