Alzheimer's in 16x16 Words: *A Novel Transformer-based Framework for the Interpretation of Gene Expression Data*

Albert Guo
School of Computing Science
Simon Fraser University
Burnaby, Canada
albert_guo@sfu.ca

Megan Fowler
School of Computing Science
Simon Fraser University
Burnaby, Canada
megan_fowler_2@sfu.ca

Kay C. Wiese
School of Computing Science
Simon Fraser University
Burnaby, Canada
wiese@cs.sfu.ca

***Abstract*—Alzheimer's disease (AD) is a growing global health concern, and early diagnosis is crucial for effective treatment. In this study, a novel method was developed for diagnosing AD by analyzing gene expression data from blood samples using a modified Vision Transformer (ViT) neural network architecture. The study utilized four publicly available Alzheimer's datasets and applied preprocessing techniques and linear discriminant analysis (LDA) to enhance the model's performance. The modified ViT was trained to analyze gene expression data and was compared with traditional convolutional neural networks (CNNs) and other approaches. The model achieved an average accuracy of 0.884 and demonstrated substantial potential for interpreting gene expression data related to AD. Notably, it outperformed existing state-of-the-art (SOTA) models and approaches in terms of both accuracy and area under the curve (AUC). The findings suggest that the modified ViT, combined with a robust preprocessing pipeline, can significantly improve the accuracy of AD diagnosis. This has the potential to facilitate early interventions, which are critical for managing AD and could be a powerful tool for healthcare providers. Additionally, the study establishes a foundation for further optimization and exploration of transformer architectures in the context of genomic diagnosis.**

Keywords—Machine Learning, Alzheimer's disease, Transformers, Gene Expression.

I. Introduction

A. Background

Over 55 million people worldwide suffer from the effects of AD or other dementias [1]. This number is expected to grow as the global population ages. Although some treatments, such as *Lecanemab* and *Aducanumab*, are somewhat effective in slowing down the progression of AD, they are effective only in the early stages of the disease [2][3]. Early diagnosis of AD, however, remains especially challenging due to the lack of any singular metric or biomarker able to predict it effectively.

B. Purpose

With the increase in life expectancy and improved access to healthcare, the incidence of aging-related diseases such as AD will continue to increase. The purpose of this project is to develop an accurate diagnostic method for AD using blood gene expression data. Accurate diagnosis of AD is imperative for administering treatments early on, improving patients' quality of life, slowing down AD progression, and reducing healthcare costs. The goal of this project is to create an effective method of AD diagnosis relying on *transformers*, which is capable of outperforming existing SOTA models as well as the accuracy of clinical diagnosis, which currently sits at 77% [4].

II. Methods

A. Datasets

Four publicly available and anonymized Alzheimer's datasets were used for this project: AddNueroMed1 (ANM1), AddNueroMed2 (ANM2), GSE140829, and the Alzheimer's Disease Neuroimaging Initiative (ADNI) [5][6].

B. Preprocessing

- Normalization was performed to scale the features of the dataset, ensuring that each feature contributes equally to the analysis.
- Kernel-based LDA was utilized to reduce the dimensionality of the dataset, which improves the computational efficiency and enhances the performance of the classifier.
- Noise was artificially injected into the data to improve the robustness of the model by making it less sensitive to small fluctuations in input.

C. Network Architecture

A modified Vision Transformer (ViT) architecture [7] was trained to analyze gene expression data in a 16x16 grid format, and its performance was compared to traditional convolutional neural networks (CNNs) and other approaches.

III. Results

The results of the study were very promising and demonstrated the substantial potential of transformers for interpreting gene expression data. Testing of the model achieved an average accuracy of 0.884 across 5 runs, significantly outperforming current (SOTA) models, as seen in Table 1.

Furthermore, the proposed model significantly outperformed Lee & Lee (2020) [12] on all datasets. The

highest performance was achieved on ANM1, beating the previous SOTA model by 0.045 points as seen in Table 2.

TABLE 1

Study	*Method*	*Accuracy*	*AUC*
El-Gawady et al. (2022) [9]	Multiple Feature Selection + SVM	0.690	0.690
Güçkıran et al. (2019) [10]	LASSO + SVM	0.764	0.850
Sharma et al. (2019) [8]	DeepInsight (tSNE + CNN)	0.670	0.743
Kalkan et al. (2022) [11]	LDA-based imaging + CNN	0.842	0.875
Proposed Model	Iterative LDA-based imaging + ViT	**0.884**	**0.951**

TABLE 2

Dataset	*Lee & Lee [12] (AUC)**	*Proposed Model (AUC)*	*Proposed Model (Accuracy)*
ANM1	0.874	**0.919**	0.857
ANM2	0.804	**0.813**	0.764
ADNI	0.657	**0.669**	0.703

**Accuracy values were not reported in Lee & Lee*

IV. Discussion

A. Contribution

The study shows that employing a modified ViT in conjunction with a robust preprocessing pipeline can significantly improve the accuracy of AD diagnosis based on gene expression data. Such a model is able to outperform existing SOTA models and the accuracy of clinical diagnosis. This can lead to earlier interventions, which are critical for the management and treatment of AD.

B. Future Work

Future work can focus on optimizing the transformer architecture and preprocessing techniques. Additionally, exploring integration with other biomarkers and clinical data for a more comprehensive diagnosis will be explored. Further investigation of different network architectures, more comprehensive data collection, and further investigation of alternative methods will be conducted. As well, further investigation of various alternative data sets and cross-validation of such data will be conducted.

V. Conclusion

This project demonstrates the potential for employing modified Vision Transformers with a robust preprocessing pipeline to diagnose Alzheimer's disease using gene expression data. The model outperforms existing approaches [11][12], notably CNNs, and holds promise for early detection and improved patient outcomes. Given the globally aging population and ever-increasing rates of Alzheimer's, effective ways to combat AD have never been of greater importance. With innovative treatments entering the market, doctors need an effective and efficient way to diagnose AD. With this powerful tool, healthcare providers would be able to diagnose accurately using only a simple blood sample from a patient and an inexpensive gene expression bead chip. Many more patients could be quickly and accurately treated, allowing for early treatment of AD. This project also provides a useful transformer framework for the diagnosis and management of other diseases through the novel approach to interpreting gene expression data. This could help in the diagnosis and treatment of other diseases more generally. As gene expression data is very hard to interpret by humans, a sophisticated transformer-based approach has the potential to open up new frontiers in medicine by allowing for greater diagnostic accuracy and the ability to process and understand more data than possible with traditional methods.

References

[1] World Health Organization, "Dementia," World Health Organization, Mar. 15, 2023. https://www.who.int/news-room/fact-sheets/detail/dementia

[2] C. H. van Dyck *et al.*, "Lecanemab in Early Alzheimer's Disease," *New England Journal of Medicine*, vol. 388, no. 1, Nov. 2022, doi: https://doi.org/10.1056/nejmoa2212948.

[3] J. Sevigny *et al.*, "The antibody aducanumab reduces Aβ plaques in Alzheimer's disease," *Nature*, vol. 537, no. 7618, pp. 50–56, Aug. 2016, doi: https://doi.org/10.1038/nature19323.

[4] M. N. Sabbagh, L.-F. Lue, D. Fayard, and J. Shi, "Increasing Precision of Clinical Diagnosis of Alzheimer's Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data," *Neurology and Therapy*, vol. 6, no. Suppl 1, pp. 83–95, Jul. 2017, doi: https://doi.org/10.1007/s40120-017-0069-5.

[5] S. Sood *et al.*, "A novel multi-tissue RNA diagnostic of healthy aging relates to cognitive health status," *Genome Biology*, vol. 16, no. 1, Sep. 2015, doi: https://doi.org/10.1186/s13059-015-0750-x.

[6] D. Nachun *et al.*, "Systems-level analysis of peripheral blood gene expression in dementia patients reveals an innate immune response shared across multiple disorders," Dec. 2019, doi: https://doi.org/10.1101/2019.12.13.875112.

[7] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, Oct. 2020, Available: https://arxiv.org/abs/2010.11929

[8] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Scientific Reports*, vol. 9, no. 1, Aug. 2019, doi: https://doi.org/10.1038/s41598-019-47765-6.

[9] A. El-Gawady, M. A. Makhlouf, B. S. Tawfik, and H. Nassar, "Machine Learning Framework for the Prediction of Alzheimer's Disease Using Gene Expression Data Based on Efficient Gene Selection," *Symmetry*, vol. 14, no. 3, p. 491, Feb. 2022, doi: https://doi.org/10.3390/sym14030491.

[10] Kivanc Guckiran, İsmail Cantürk, and Lale Ozyilmaz, "DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO," pp. 115–121, Apr. 2019, doi: https://doi.org/10.19113/sdufenbed.453462.

[11] H. Kalkan, U. M. Akkaya, G. Inal-Gültekin, and A. M. Sanchez-Perez, "Prediction of Alzheimer's Disease by a Novel Image-Based Representation of Gene Expression," *Genes*, vol. 13, no. 8, p. 1406, Aug. 2022, doi: https://doi.org/10.3390/genes13081406.

[12] T. Lee and H. Lee, "Prediction of Alzheimer's disease using blood gene expression data," *Scientific Reports,* vol. 10, no. 1, Feb. 2020, doi: https://doi.org/10.1038/s41598-020-60595-1.