Automating the Identification of Metabolically Distinct Groups of Individuals Using tSNE

Background

Individuals with serious mental illnesses have higher rates of cardiovascular and metabolic disease than the general population, which can lower life expectancy by 10-20 years.[1] Understanding these increased risks is critical for identifying those who are most vulnerable to cardiometabolic illnesses and developing more effective treatments. Previous research has shown that genetic variations linked to mental illnesses and cardiometabolic disorders can categorise individuals into three distinct groups. However, the manual classification methods used limit scalability and clinical application.

Aim

- To design and implement a method for automating the classification of metabolically distinct individuals based on genetic variations.
- To assess method transferability by applying it to different datasets to evaluate its effectiveness and adaptability.

Method

Data Sources

- IMPROVE study (3700 high CVD risk individuals from Europe)
- UK Biobank

Data Cleaning and Preparation

- Handled using complete case analysis, mean imputation, and K-nearest neighbour methods.

Data Processing

- Dimensionality Reduction using Principal Component Analysis

tSNE Clustering Technique

- Max_iter (maximum iterations): 1000 – 2000
- Theta (learning rate): 0.1
- Perplexities (list of perplexity values): 30 – 50
- pcaDims (PCA dimensions): 15 – 50
- figWidth (figure width in pixels): 2000
- pointSize (Size of points in the plot): 0.5
- textSize (size of text in the plot): 5
- num_threads: (number of threads to use): 0

Results

Figure 1: PCA result for IMPROVE

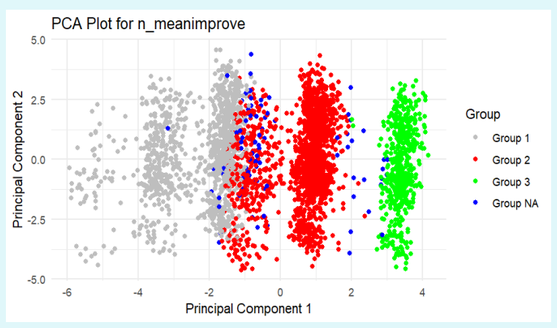


Figure 2: tSNE result for IMPROVE

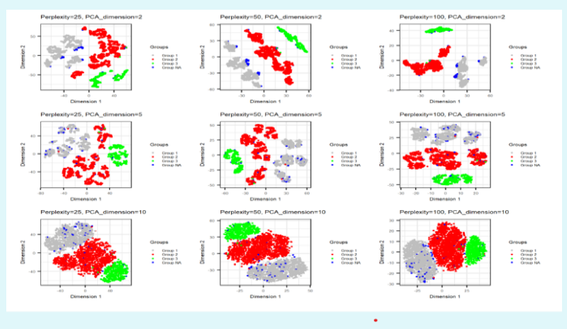


Figure 3: tSNE for UKB pilot data

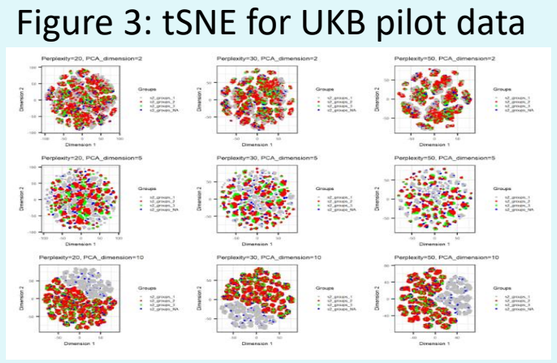


Figure 4: tSNE for UKB pilot data (3D)

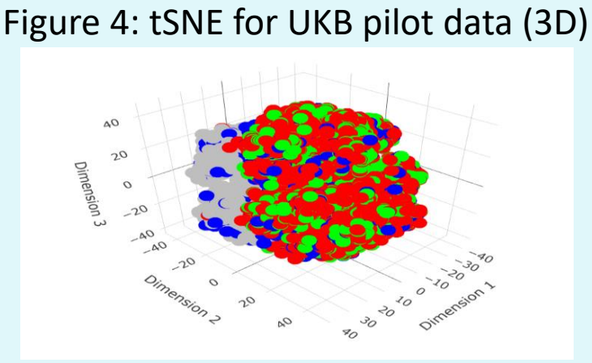
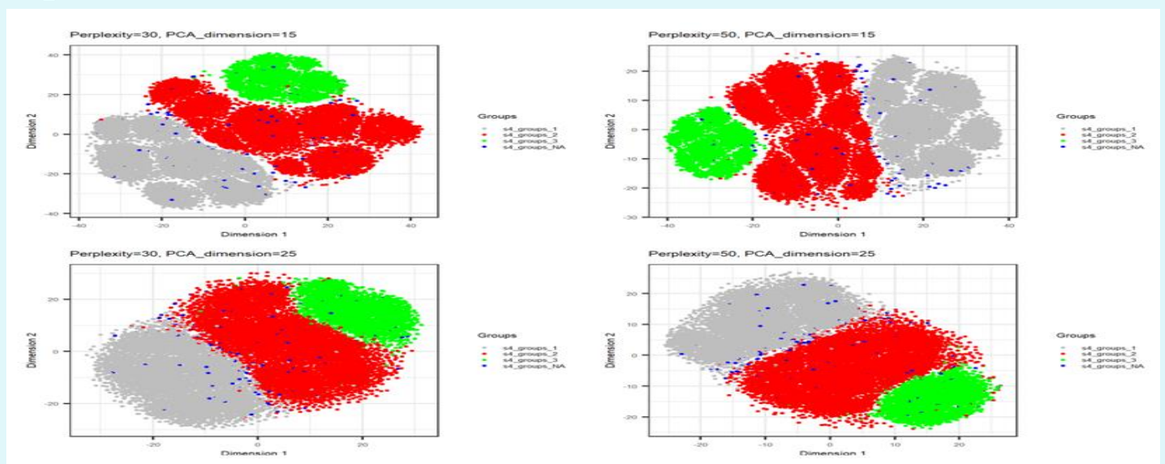


Figure 5: tSNE for more randomized UKB data



Discussion and Conclusion

- Using t-SNE with PCA initialisation works well for grouping individuals and is scalable for larger datasets. Automating t-SNE is challenging, as it requires careful adjustments to variables
- The results from the IMPROVE study support the groupings seen in previous research, confirming the accuracy of this method
- This method can be applied to different datasets, showing its flexibility

Recommendations for Further Research

- More research is needed to identify the most appropriate parameters for t-SNE when applied to larger datasets
- Continued research is required to explore the clinical applicability of t-SNE groupings for real-world use in healthcare settings

Acknowledgments

- UK Biobank Participants
- School of Health and Wellbeing, University of Glasgow

Author

- **Oluwatobi Oni**
- **LinkedIn:** https://www.linkedin.com/in/oluwatobioni/
- **Supervisor:** Rona Strawbridge, Frederick Ho

References:

1. Strawbridge RJ, Johnston KJA, Bailey MES, Baldassarre D, Cullen B, Eriksson P, deFaire U, Ferguson A, Gigante B, Giral P, Graham N, Hamsten A, Humphries SE, Kurl S, Lyall DM, Lyall LM, Pell JP, Pirro M, Savonen K, Smit AJ, Tremoli E, Tomainen TP, Veglia F, Ward J, Sennblad B, Smith DJ. The overlap of genetic susceptibility to schizophrenia and cardiometabolic disease can be used to identify metabolically different groups of individuals. Sci Rep. 2021 Jan 12;11(1):632.; PMCID: PMC7804422.