**Categorising Genetic Metabolic Profiles Using tSNE Clustering**

**Project Overview**

Individuals with mental illnesses have a higher risk and prevalence of cardiovascular and metabolic diseases than the general population, resulting in a significantly reduced life expectancy (by 10-20 years). This project focuses on understanding the relationship between mental illness and cardiometabolic conditions by automating the classification of individuals based on genetic variations using the t-distributed Stochastic Neighbor Embedding (tSNE) method.

This work builds on previous research that identified three groups of individuals with different cardiovascular and metabolic disease profiles, using manual classification methods. However, manual classification is not suitable for scaling or clinical application. This project aims to automate this process by implementing tSNE for dimensionality reduction and clustering, improving scalability and applicability in clinical settings.

**Objectives**

* Automate the grouping of metabolically distinct individuals based on genetic variations using tSNE.
* Develop a scalable pipeline for applying tSNE to UK Biobank data.
* Compare tSNE results to previous methods like Multidimensional Scaling (MDS).
* Demonstrate method transferability across datasets to assess the effectiveness and adaptability.

**Data Sources (IMPROVE and UK Biobank)**

IMPROVE: 3,700 individuals of European ancestry at high risk of cardiovascular disease (CVD).

UKB2: 22,000 individuals of UK ancestry from the general population.

UKB4: 22,000 individuals of European ancestry.

**Data Preparation**

Handle missing data using complete case analysis, mean imputation, or K-nearest neighbor imputation methods.

**Methodology**

* Principal Component Analysis (PCA) for data preprocessing
* Implement tSNE with PCA initialisation to automate the grouping of individuals based on genetic variations.
* Compare the results with previous MDS-based approaches.

**Discussion & Conclusion**

* Effectiveness: tSNE, especially with PCA initialisation, is effective for grouping individuals and scaling to larger datasets.
* Challenges: Automating tSNE requires careful adjustment of variables, and further research is needed to optimise for larger datasets.
* Validation: The IMPROVE study supports the validity of the groupings, confirming the accuracy of this approach.
* Flexibility: The method shows promise for adaptation across different datasets, indicating its potential for broader application.

**Future Research**

* Identify optimal parameters for tSNE in large-scale datasets.
* Explore clinical applicability of tSNE groupings in healthcare settings.

**Data Availability**

The datasets used in this project are available upon request from the original sources.