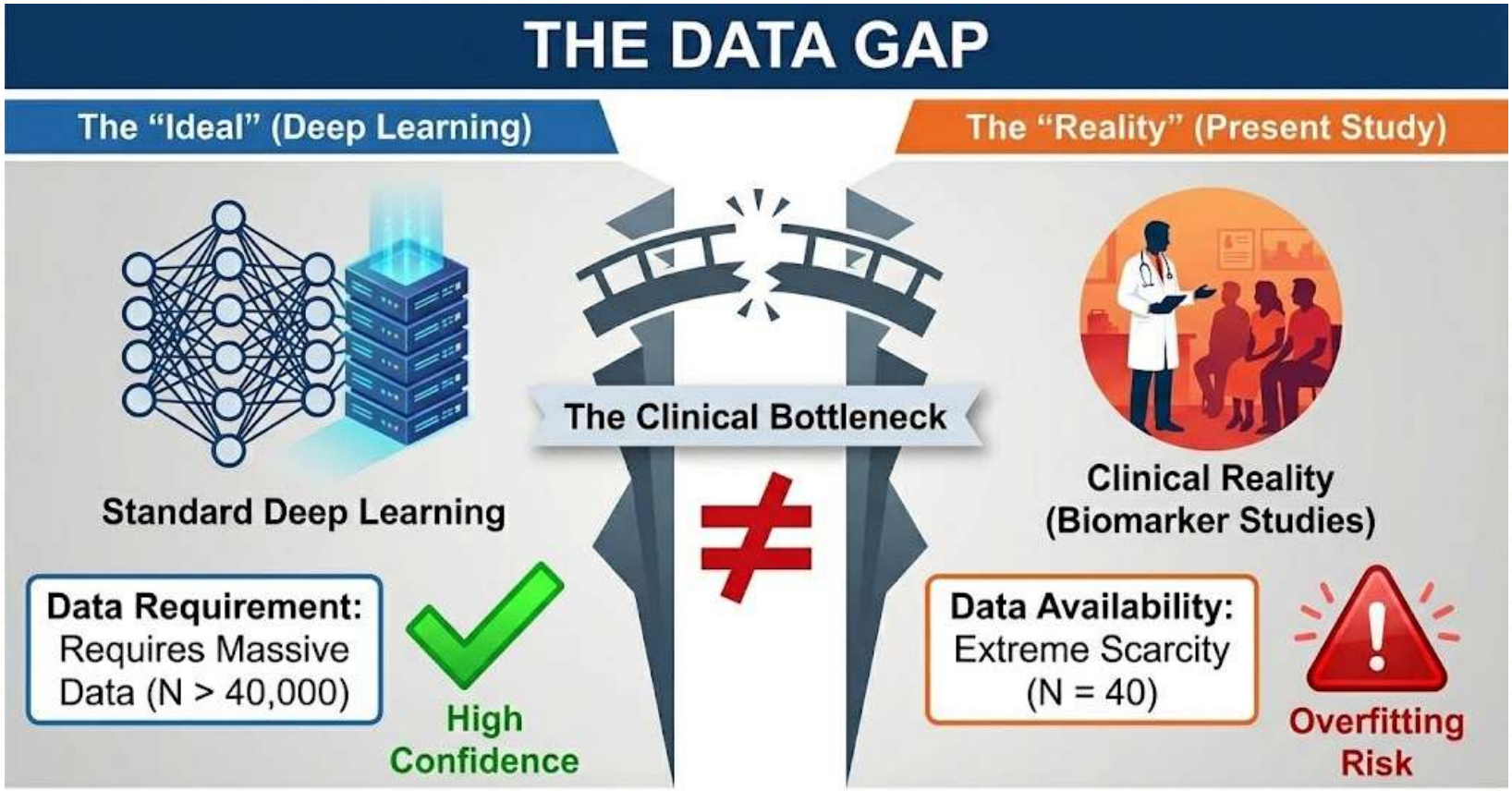
Data Augmentation for Cognitive Biomarker Prediction – Solving the Small Sample Size Problem in Clinical Research

Aaron Dumont

CMPS 7010 – Research Seminar

December 4, 2025

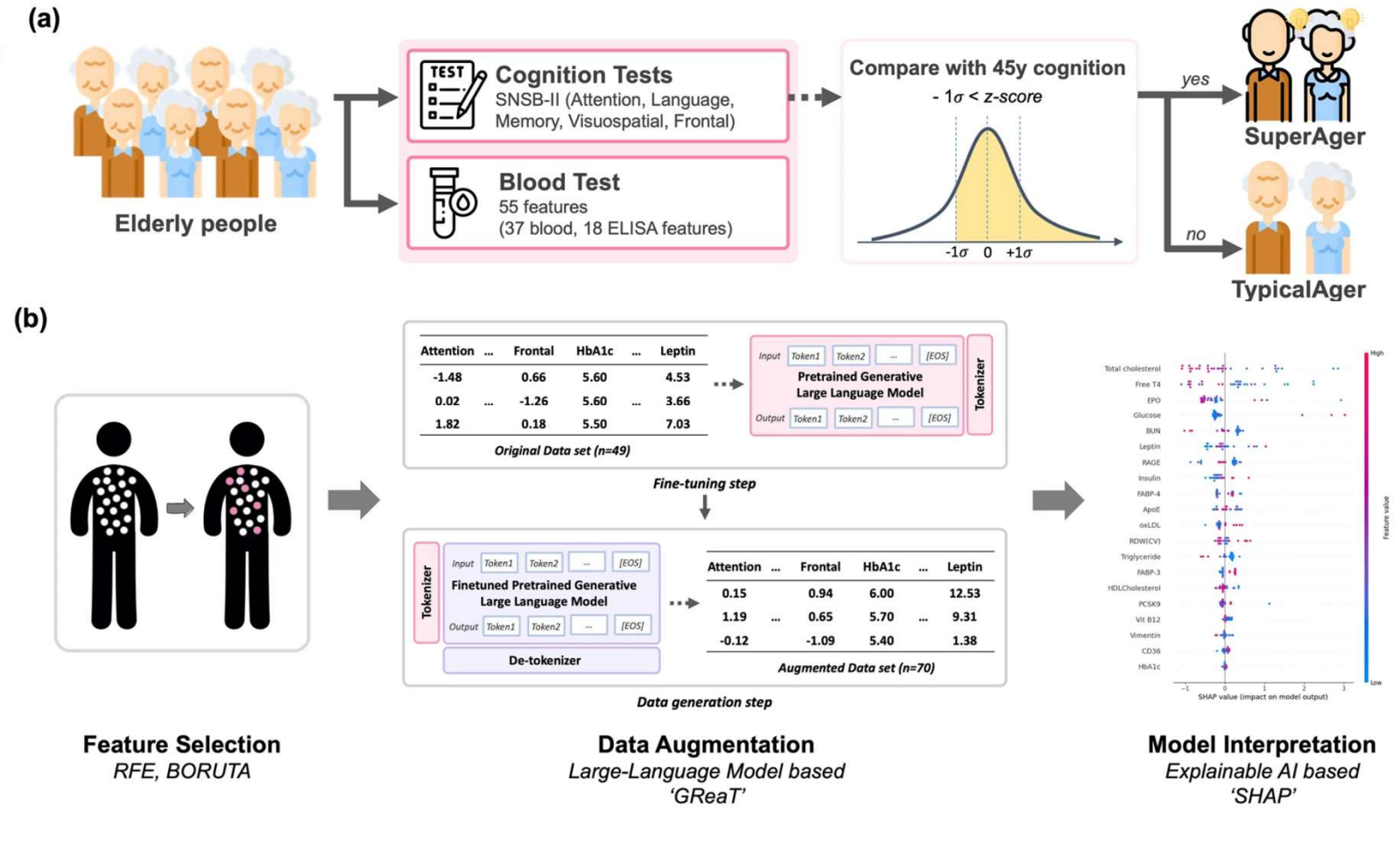
The Problem: Data Scarcity in Clinical Research



Bridging the Gap: Addressing Data Scarcity in Clinical Biomarker Research

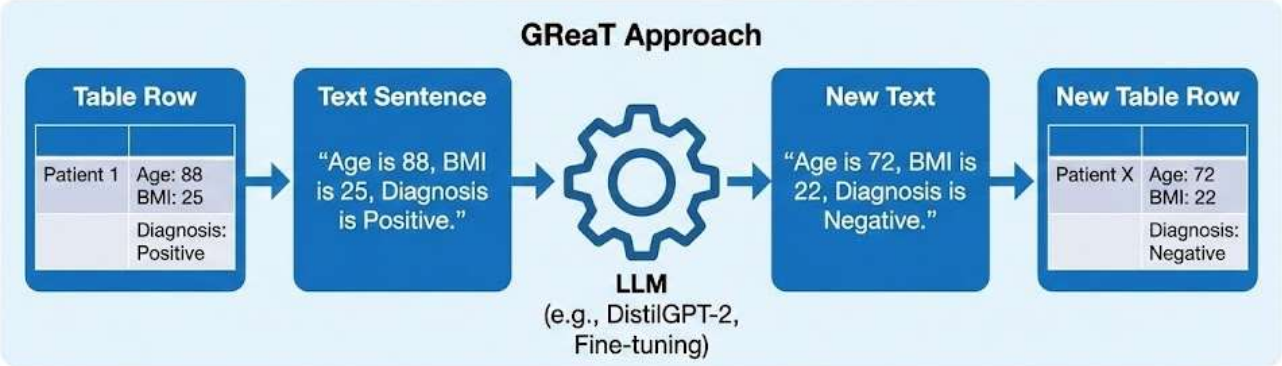
Lee HB et al. *Scientific Reports*, 2025

Blood Biomarkers and SuperAger Status



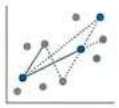
Generating New, Synthetic Patients to Increase Sample Size

The Solution: GReaT (Generative Realistic Tabular Data)



- ✓ GReaT Approach: Treats patient data as language.
- ✓ Captures complex, non-linear dependencies between features (Age, BMI, biomarkers).
- ✓ Allows LLM to "write" new, realistic patient profiles.

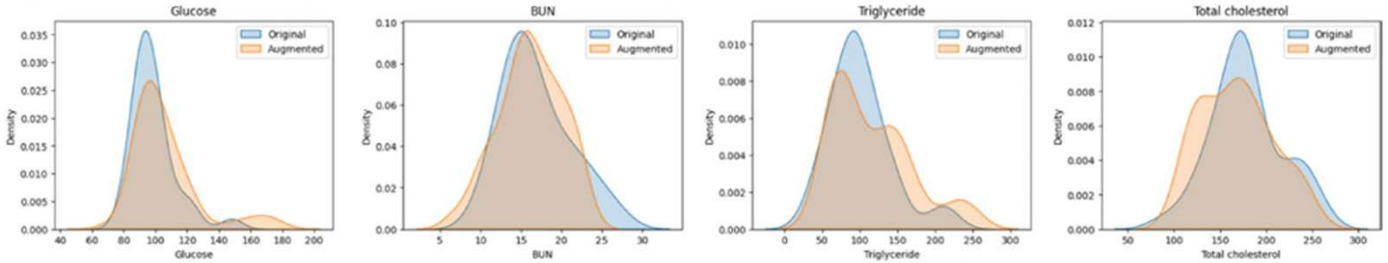
VS. Traditional Augmentation (SMOTE)



Interpolates points geometrically.

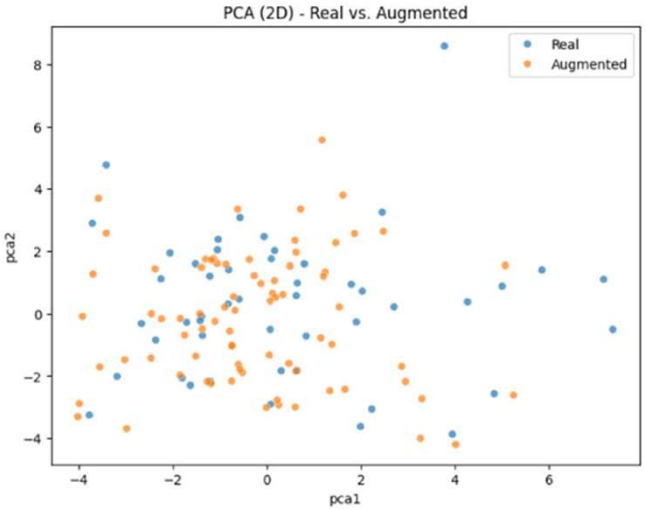
Validating the Synthetic Data

A – Kernel Density Estimation Plots



KDE plots show distribution similarity between real (blue) and augmented (orange) data across SuperAger status and key biomarkers -> substantial overlap indicates augmented data preserves statistical properties of the original data

B – Principal Component Analysis (PCA)



Demonstrates the 2D distribution of real and augmented samples. Similar scatter patterns confirm the augmented data maintains the same variance structure as the original data

Lee HB et al. *Scientific Reports*, 2025

The Present Replication Setup

Methodology: "Starving" the Model & LLM Rescue

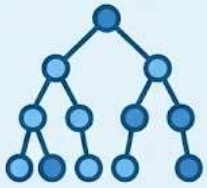
The Constraint ("Starvation Mode")

- **Source:** Alzheimer's Disease Dataset (N=2,149).
- **Action:** Simulated "Rare Disease" Scenario.
- **Result:** Isolated N=40 Total Subjects.
- **Split:** 32 Training (Real) / 8 Testing (Real)

The Rescue (Generative Augmentation)

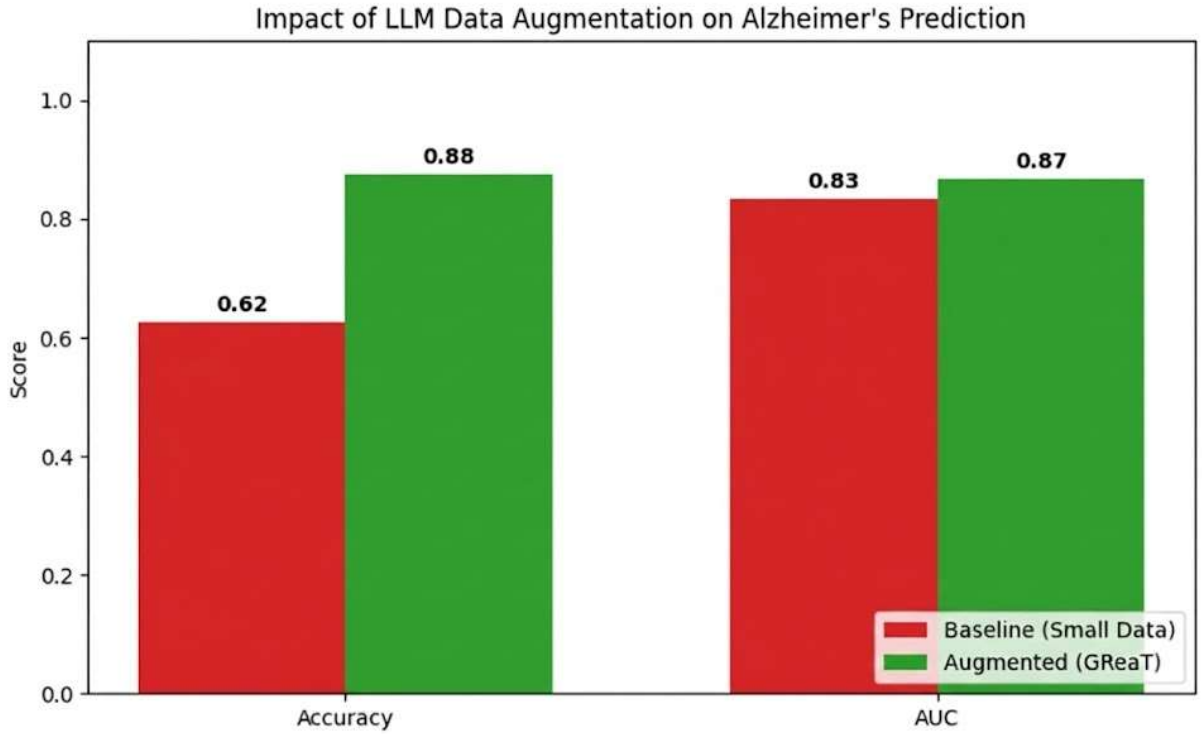
- **Engine:** GReaT Framework (DistilGPT-2).
- **Training:** Fine-tuned for 150 Epochs on the 32 real samples.
- **Generation:** Created 100 Synthetic Patients.
- **Balancing:** Enforced 50/50 split (Healthy/Disease) to correct class imbalance.

Evaluation (Random Forest)



- **Classifier:** Random Forest (Selected for robustness over XGBoost).
- **Experiment A (Baseline):** Trained on 32 Real samples.
- **Experiment B (Augmented):** Trained on 132 Samples (32 Real + 100 Synthetic).
- **Validation:** Tested on the same isolated 8 samples.

Results – Data Augmentation

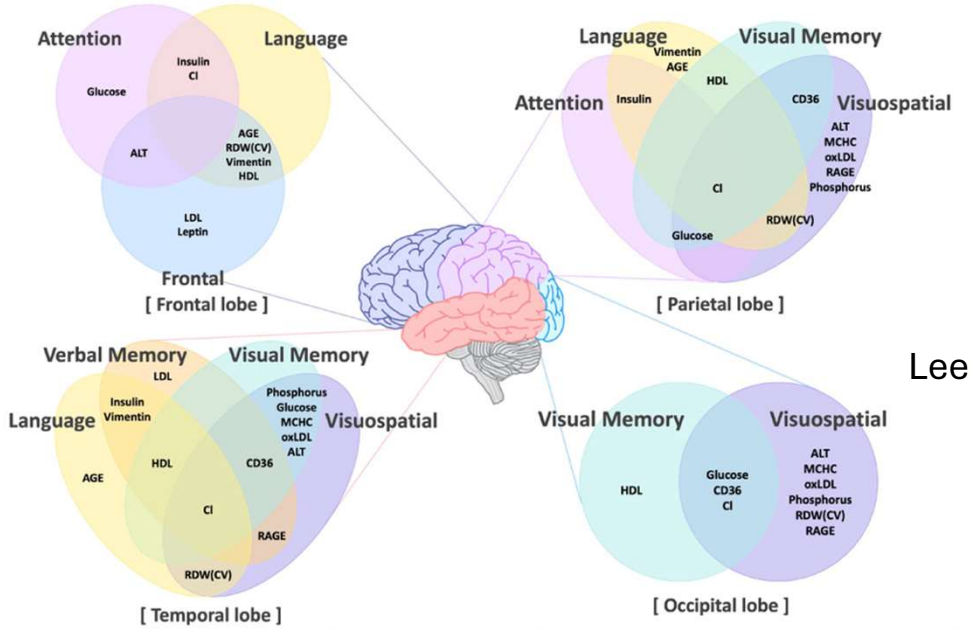


Baseline (n=40)
Augmented (n=140)

- Generative augmentation stabilizes models in data-scarce scenarios
- **This replicates a key finding from Lee et al. (2025)**

Clinical Significance & Discussion

- Easily accessible, simple and scalable blood biomarkers can be used to identify cognitive status
- Beyond math, such biomarkers may be relevant for certain dimensions of cognitive assessment mapping to specific brain regions



Lee HB et al. *Scientific Reports*, 2025

Limitations

- Small test sample
- “Seed bias” due to training on a small “seed” of only 32 patients
- Possible hallucinations with synthetic data generation using an LLM
- Limitation to tabular data only (no genetics or imaging)

Conclusions & Future Work

- We were able to replicate the ability of synthetic data augmentation using the GReaT LLM model to improve predictive ability of blood biomarkers from an initial small patient sample
- This suggests enormous future potential for clinical studies with small sample size
- Future work: Apply this to multi-omics data such as presented in Zhou et al. (2025) discussed previously


THANK
YOU

References

1. Siegel NT et al. Do transformers and CNNs learn different concepts of brain age? *Human Brain Mapping* 46:e70243, 2025
2. Lee HB et al. Machine learning based prediction of cognitive metrics using major biomarkers in SuperAgers. *Scientific Reports* 15:18735, 2025
3. Zhou S et al. A novel sequence-based transformer model architecture for integrating multi-omics in preterm birth risk prediction. *Digital medicine* 8:536, 2025
4. Gong W et al. Optimising a simple fully convolutional network for accurate brain age prediction in the Pac 2019 challenge. *Frontiers in Psychiatry.* 12:627996, 2021
5. Peng H et al. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis.* 68:101871, 2021
6. Garo-Pascal M et al. Brain structure and phenotypic profile of superagers compared with age-matched older adults. *Lancet Healthy Longev* 4:e374-e385, 2023.
7. Harrison TM et al. Superior memory and higher cortical volumes in unusually successful cognitive aging. *J Int Neuropsychol Soc* 18: 1081-1085, 2012.
8. Chaudhary, K., et al. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 24:1248-1259, 2017.
9. Huang, S., et al. More is better: Recent progress in multi-omics data integration methods. *Front Genet* 8:84, 2017.
10. Subramanian, I., et al. High-throughput sequencing of pooled samples to determine community-level microbiome diversity. *Ann Epidemiol* 39:63-68, 2019.