**MSc Project Progress Update**

## **Student: **James Oluwafemi Adeshina****

## **Project Title**: Early Prediction of Diabetic Complications Using Multi-Modal Deep Learning

Supervisor: **Dr. Oluwarotimi Samuel**

## **Date**: 1st July, 2025

## Progress Since Last Meeting (18th June 2025):

Since our last meeting, I have completed a comprehensive data coverage analysis on the labelled diabetic cohort (N=10,807) extracted from the MIMIC-IV dataset. We anchored all complication labels to true diabetes diagnosis dates and applied condition-specific follow-up windows (e.g., 30–730 days for retinopathy and cardiovascular; 90–365 days for neuropathy, nephropathy, and foot ulcer). Clinical feedback was incorporated to ensure the prevalence metrics matched known NHS benchmarks.

Lab test coverage was evaluated across 10 key biomarkers (glucose, A1C, creatinine, albumin, albumin-creatinine ratio, LDL, HDL, triglycerides, BUN, CRP). Coverage analysis showed that at least 3–6 labs have sufficient representation (≥100+ patients) per complication, enabling viable modelling. However, BUN and CRP had zero coverage across all complications and have been excluded.

In parallel, I explored the eICU Collaborative Research Database for potential external testing. After a thorough structural review, I concluded that the dataset lacks sufficient diabetes diagnosis anchoring and complication-specific longitudinal records, making it unsuitable for reliable external validation of our model.

We finalised the primary datasets for modelling as:

**patients.csv.gz**, **admissions.csv.gz**, **diagnoses\_icd.csv.gz**, d\_**icd\_diagnoses.csv.gz** – for label generation

**labevents.csv.gz**, **d\_labitems.csv.gz** – for biomarker features

## Challenges Encountered:

1. Initial pipelines suffered from misaligned anchoring (first admission vs. diagnosis), leading to inflated or clinically invalid prevalence rates.
2. ICD code contamination (e.g., including non-diabetic cardiovascular disease) required rigorous clinical filtering using diabetes-specific codes.
3. Follow-up filtering was too strict, excluding >80% of diabetics and creating survivorship bias.
4. The eICU dataset was found unsuitable for external testing due to missing temporal markers for diabetic complications.
5. Memory Errors in PySpark
6. **Sparse Lab Coverage:** Key biomarkers like BUN and CRP had no usable data within prediction windows