**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**: 18th June, 2025

## Progress Since Last Meeting (2nd June 2025):

Since our last meeting, I have expanded the literature review and discovered two particularly relevant studies:

1. Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms
2. A Multi-modal Deep Learning Approach for Predicting Type 2 Diabetes Complications: Early Warning System Design and Implementation

These reinforce the value of my project focus while also helping clarify my contribution. Unlike prior work, my emphasis is on integrating **longitudinal, multi-modal data** with strong explainability, tested across **independent clinical datasets** for robustness.

## Data and Technical Progress:

 I have begun exploratory analysis of the **MIMIC-IV** dataset and confirmed that it includes detailed EHR records relevant to diabetes and associated complications.

 I have also started validating data structure and variables for suitability across modalities (lab results, medication history, diagnostic codes, etc.).

 As a proposed novelty, I plan to **train and validate the model using MIMIC-IV**, then **test it externally using the eICU Collaborative Research Database** to assess generalisability.

 I now have access to the eICU Collaborative Research Database and will begin exploring its structure and compatibility this week.

## Challenges Encountered:

Working with a large dataset like MIMIC-IV has been technically demanding, especially within a Python environment. To address performance bottlenecks, I’ve started integrating PySpark for more efficient data manipulation. Another significant challenge is navigating the clinical structure of the dataset, particularly ensuring accurate use of ICD diagnosis codes when identifying the five targeted diabetic complications. Although the dataset is large, filtering down to relevant cohorts (with valid timestamps and corresponding lab results) results in much smaller usable subsets, which may impact class distribution and model training. These issues are being addressed iteratively during exploration.