# **2025 Graduation Project Midterm Progress Report**

**Division:** AI and Data Science  
**Project Title:** A Multimodal Digital Twin for Type 2 Diabetes Patients  
**Team Name:** SheCodes  
**Team Members:**   
**Advisor:** Prof. Song Giltae

## **1. Requirement & Constraint Updates**

The initial concept aimed to create a digital twin system for Type 2 Diabetes (T2D) patients using structured datasets. By the midterm phase, key updates include:

* Replacement of less informative datasets (Pima) with more realistic, structured clinical data (OMOP format)
* Integration of hospital and national health survey data
* Addition of explainability tools and interactive interface via Streamlit

## **2. Design Refinement & Architecture Changes**

### **Original Design:**

* Binary classification model (MLP)
* Simple data from EHR and Pima datasets
* No interface or explainability

### **Current Design:**

* Full pipeline with 6 input branches:  
  EHR, Lifestyle, Clinical, CDC, Hospital, Wearable *(planned)*
* Enhanced preprocessing with dynamic input shapes
* Streamlit-based UI + SHAP explanation
* Improved data balance using CDC 50-50 split

## **3. Updated Project Timeline**

| **Period** | **Tasks Completed** |
| --- | --- |
| June–July | Data integration, model rewriting, UI creation |
| July 18 | Submit midterm report & evaluation form |
| August | Integrate wearable data + expand SHAP interface |
| September | Final model training + user scenario simulation |
| October | Presentation + SW registration |

## **4. Member Contribution**

| **Name** | **Role** | **Progress** |
| --- | --- | --- |
| Pak Elina | Model architecture, data processing, Streamlit UI |  |

## **5. Current Progress and Results**

### **Datasets in Use:**

| **Dataset** | **Features Used** | **Notes** |
| --- | --- | --- |
| diabetes.csv (EHR) | 8 clinical vars incl. Outcome | Original base dataset |
| diabetes-2.csv | Age, Frame, Waist, Hip | Lifestyle-based metrics |
| OMOP Clinical (4 csv) | Gender, Birth Year, Avg Obs, Avg Meas, Condition Count | Replaced Pima |
| CDC BRFSS | Health indicators (10 vars) | Balanced for target labels |
| Hospital (UCI) | Admissions, diagnosis codes | After label encoding |
| *(Planned)* Wearable | Blood glucose time series | Will extract trend features |

**Insert Figure: Dataset Summary Table or Pie Chart of Source Contribution**

**Insert PNG: Data type classification result from data\_analysis.py (e.g., binary/numeric breakdowns)**

**Multimodal Model Architecture**

ehr\_branch -> FC(32) → ReLU → BN

lifestyle\_branch-> FC(32) → ReLU → BN

clinical\_branch -> FC(32) → ReLU → BN

cdc\_branch -> FC(32) → ReLU → BN

hospital\_branch -> FC(32) → ReLU → BN

(Concat all) → FC(64) → Dropout → FC(1) → Sigmoid

* Final classifier trained on 80/20 train-test split
* Regularization added: BatchNorm, Dropout
* Custom MultimodalT2DPredictor model implemented

### **Streamlit UI Functionality**

* Collects inputs: Age, Gender, Waist, Hip, Birth Year, Health indicators, etc.
* Real-time inference with prediction and risk level (Low / Warning / High)
* Integrated SHAP explainer to show which features influenced prediction

**Insert Screenshot: Streamlit Web Interface showing prediction result**

**Insert Screenshot: Streamlit form with filled example values**

### **SHAP Explainability**

Implemented via shap\_explainer.py:

* Bar chart visualization of top features influencing individual risk
* Supports user-level understanding of outcomes

**Insert PNG: SHAP Feature Importance (streamlit\_app\_shap.png or similar)**

### **Automated Dataset Analysis Tool**

Script: data\_analysis.py

* Automatically scans all datasets for:
  + Missing values
  + Data types (binary, numeric, categorical)
  + Statistical summaries

**Insert PNG: Missing values heatmap or table from dataset analysis**

**Insert PNG: Summary stats per dataset**

## **6. Mentor Feedback Response**

| **Feedback** | **Action Taken** |
| --- | --- |
| Include more realistic clinical data | Replaced Pima with OMOP-format clinical records |
| Add explainability | SHAP integration and risk threshold mapping |
| Enable interaction with model | Built Streamlit app for inputs and visual outputs |

## **7. Technical Challenges and Solutions**

| **Issue** | **Fix** |
| --- | --- |
| NaN-based crashes during scaling or scoring | Replaced with column mean using fillna(df.mean()) |
| Streamlit errors due to unsubmitted forms | Added st.form\_submit\_button() |
| Model loading error (size mismatch) | Synced model layers with exact input feature size |
| Low AUC in early models | Improved architecture, replaced unhelpful datasets |

## **8. Future Plans**

| **Item** | **Goal Description** |
| --- | --- |
| **Wearable Integration** | Preprocess and use CGM-based glucose variability |
| **SHAP Simulation** | “What if” feature for adjusting input and seeing effect |
| **Backend Upgrade** | Convert to Gradio or FastAPI-based app |
| **Model Tuning** | Try ensemble, LightGBM or attention-based models |
| **Responsive UI** | Mobile-friendly layout and input optimization |

## **9. Conclusion**

This project has evolved significantly from a simple diabetes predictor to a modular, explainable AI system that simulates multimodal health data. Realistic clinical datasets, government health indicators, and hospital records are integrated in a single pipeline.

The Streamlit interface allows user-level simulation, and SHAP ensures decisions are interpretable. By the final presentation, we aim to include wearable data and advanced simulation features