AZT1D: A Real-World Dataset for Type 1 Diabetes

Saman Khamesian, Asiful Arefeen, Bithika M. Thompson, Maria Adela Grando and Hassan Ghasemzadeh

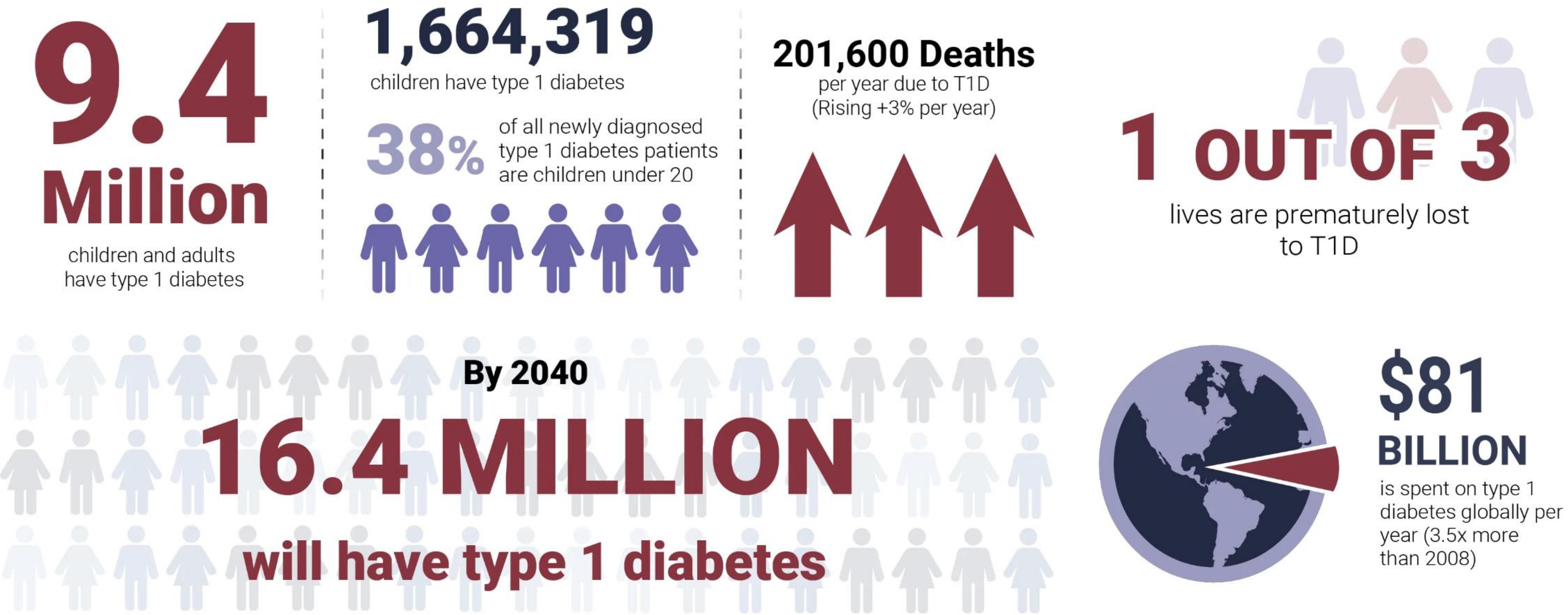
Saman Khamesian
PhD Student – Computer Science
School of Computing and Augmented Intelligence
Graduate Research Associate - College of Health Solutions
Embedded Machine Intelligence Lab (EMIL)
Arizona State University
Phoenix, AZ, USA

IEEE-EMBS BSN'25
Los Angeles, CA, USA

What is Type 1 Diabetes (T1D)?

- T1D is a lifelong **autoimmune** condition — the body can’t produce insulin.
- Requires **constant self-management** through blood sugar checks and insulin dosing.
- It is usually diagnosed in childhood or early adulthood and **requires lifelong care**.
- Without proper control, it can lead to serious health issues, like heart and kidney complications.

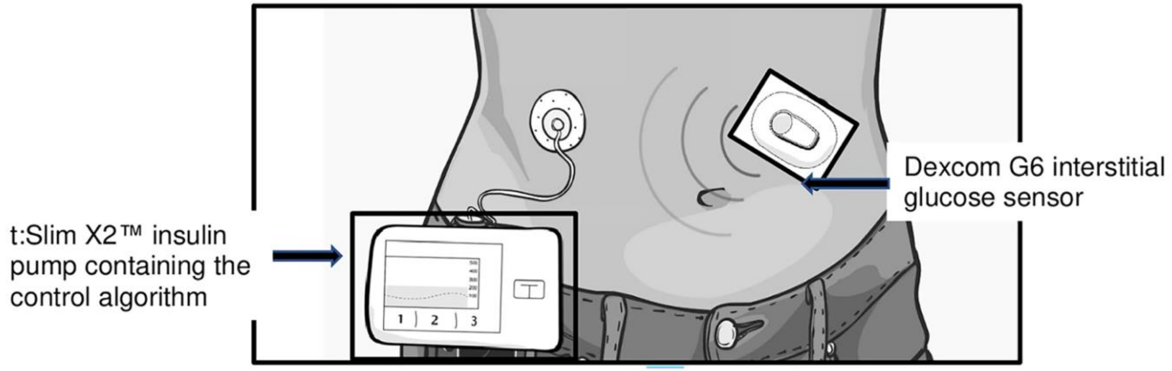
The Global Burden of T1D



Source: T1D Index and International Diabetes Foundation

Toward Automated Diabetes Management

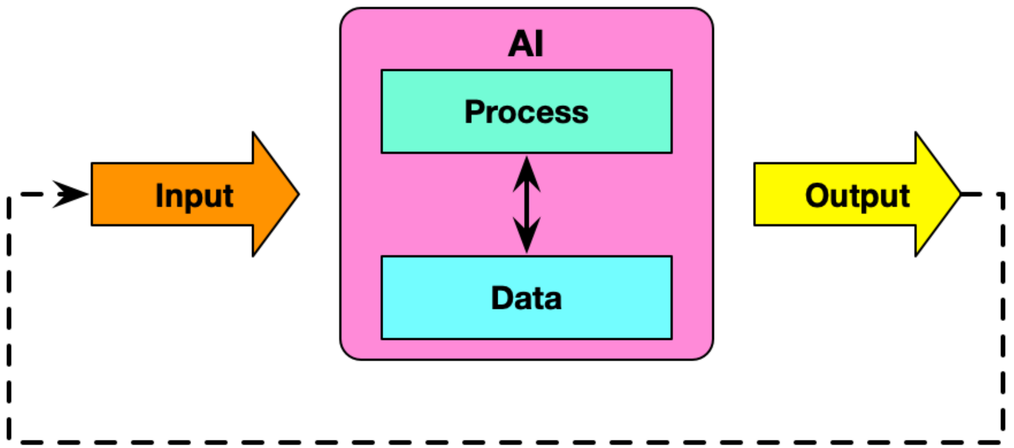
- Advances in **diabetes technology** have improved care.
- One example is the **Continuous Glucose Monitor (CGM).**
- A CGM has a tiny sensor under the skin, often on the arm or stomach.
- It sends data to a screen, like on a phone, so people can see their levels.



Why We Need Real-World T1D Data?

- But to truly **personalize treatment**, we need **data-driven** strategies
- Many emerging methods—like:
  - Therapy optimization
  - Digital twin modeling
  - Counterfactual analysis

  rely on machine learning and AI



- These methods need **high-quality**, **real-world data** to be effective

Existing Datasets

- Most publicly available T1D datasets (e.g., OhioT1DM) are:
  - **Simulated** or **synthetic**
  - **Small-scale** (few patients, short durations)
  - Lacking important **behavioral context** (device mode, carb intakes, etc.)
- **This limits:**
  - Generalizability of models
  - Clinical applicability
  - Progress in AI-driven diabetes management

Introducing AZT1D Dataset

Scan to Download!

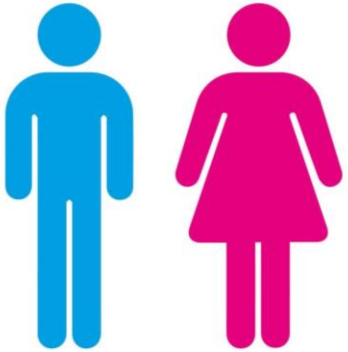
- **AZT1D** dataset was created to fill a critical gap in T1D research.
- Enables personalized, **AI-driven analysis and modeling.**
- Developed in **collaboration with the Mayo Clinic**.
- Includes data from **25 individuals with T1D**.
- Study conducted in **Scottsdale, AZ (Dec 2023 – Apr 2024)**

Introducing AZT1D Dataset

Dexcom G6 Pro
Tandem t:slim X2

Glucose Values
5-min intervals
~320k entries
~26.7k hours total

Purely observational
Collected during routine care
No intervention



12 Males
13 Females

Data Processing

- Tandem pump data came in **two formats:**

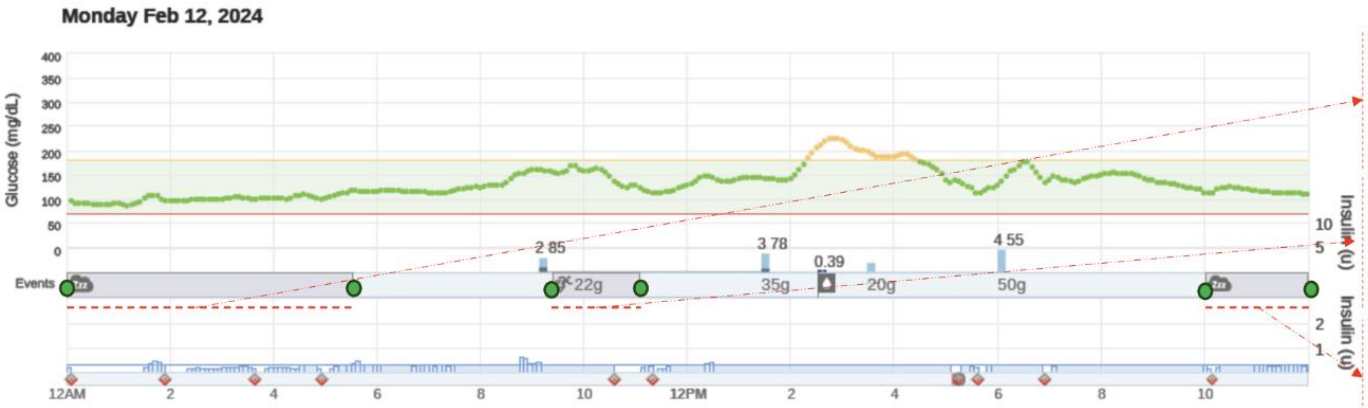
  → **CSV (glucose, bolus, carbs)**

  → **PDF (basal rates, device modes)**

- Data integration was not straightforward due to format differences.
- Timestamps of carb and bolus events were aligned with glucose readings.
- Basal and mode data were extracted from PDFs using **OCR-based** methods.

Data Processing

Extracting Device Modes (Sleep, Exercise, Normal)



	datetime	DeviceMode
0	2024-02-12 00:00:00	sleep
1	2024-02-12 00:05:00	sleep
...	...	...
65	2024-02-12 05:25:00	sleep
66	2024-02-12 05:30:00	sleep
0	2024-02-12 09:24:00	exercise
1	2024-02-12 09:29:00	exercise
...	...	...
19	2024-02-12 10:59:00	exercise
20	2024-02-12 11:04:00	exercise
0	2024-02-12 22:00:00	sleep
1	2024-02-12 22:05:00	sleep
...	...	...
22	2024-02-12 23:50:00	sleep
23	2024-02-12 23:55:00	sleep

Above Target   Below Target   Calibration

	12AM	1	2	3	4	5	6	7	8	9	10	11	12PM	1	2	3	4	5	6	7	8	9	10	11
Feb 11,2024 - Sunday																								
Glucose (mg/dL)											199					185			179	295				
Carbs (g)																12			20					
Bolus (u)					0.65					0.76	0.80					0.91 1.53			0.78 2.72	2.00			0.72	
Basal Total Delivered (u)	0.241	0.435	0.890	0.846	1.102	0.900	0.975	0.872	0.433	0.812	1.214	0.811	0.623	0.183	0.725	1.231	0.358	0.075	1.228	0.825	0.618	0.525	0.996	0.710
Basal Profile Setting (u/hr)	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
Avg. CGM (mg/dL)	91	94	121	130	161	167	141	115	110	113	188	163	129	91	114	172	139	81	169	272	174	113	138	129
# Readings	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12

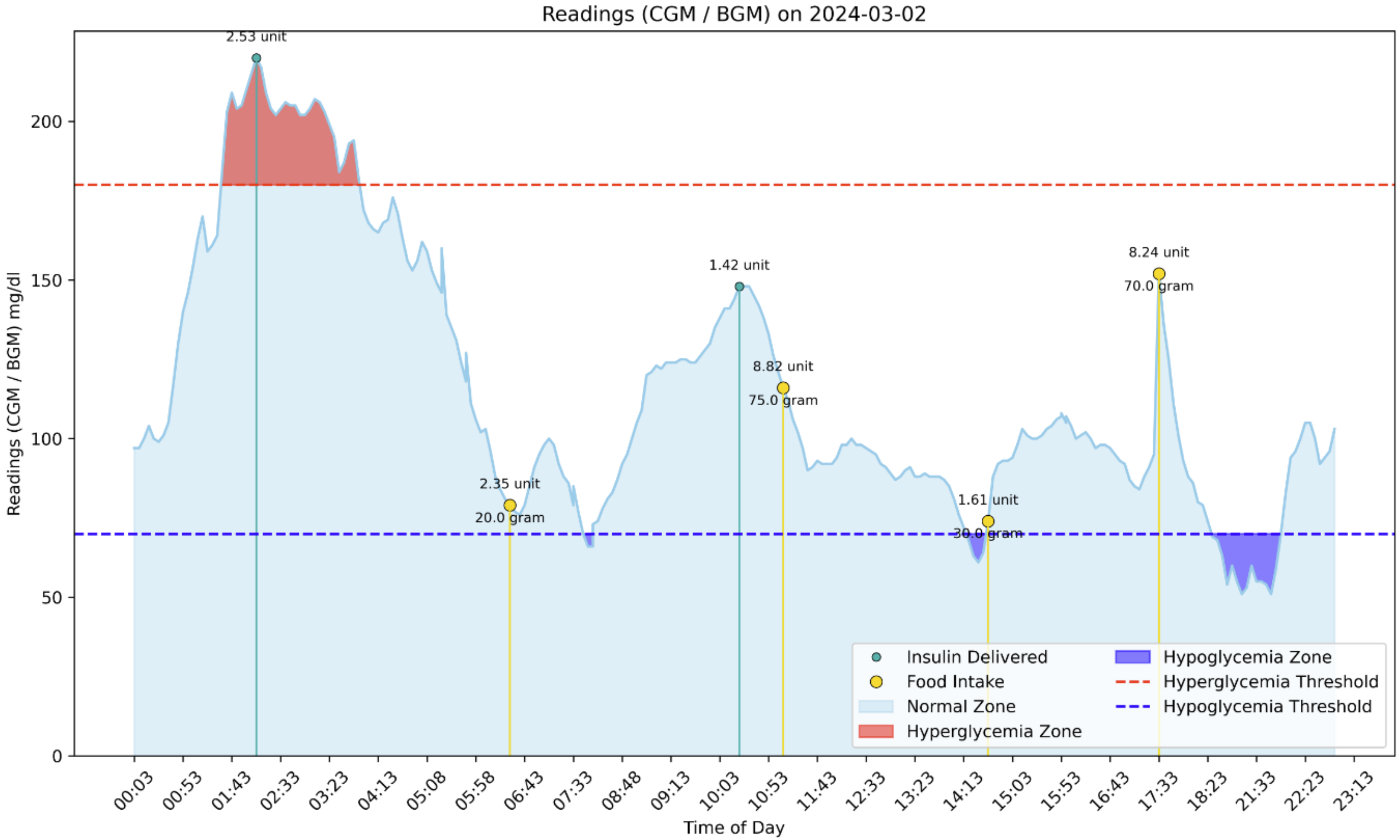
	Timestamp	Basal
0	2024-02-11 00:00:00	0.241
1	2024-02-11 00:05:00	0.241
2	2024-02-11 00:10:00	0.241
3	2024-02-11 00:15:00	0.241
4	2024-02-11 00:20:00	0.241
...	...	...
283	2024-02-11 23:35:00	0.710
284	2024-02-11 23:40:00	0.710
285	2024-02-11 23:45:00	0.710
286	2024-02-11 23:50:00	0.710
287	2024-02-11 23:55:00	0.710

Extracting Basal Insulin

Dataset Features

- Each participant's data includes time-aligned records across several key variables:
  - `EventDateTime` : Timestamp of each record
  - `DeviceMode` : Regular, sleep, or exercise
  - `BolusType` : Standard, correction, or automatic
  - `Basal` : Hourly basal insulin delivery (units)
  - `CorrectionDelivered` : Portion of insulin for correction
  - `TotalBolusInsulinDelivered` : Total bolus dose (units)
  - `FoodDelivered` : Portion of insulin for food coverage
  - `CarbSize` : Meal carbohydrate amount (grams)
  - `CGM` : Blood glucose level (mg/dL)

Daily Records Example



Use Cases

- Its structure makes it ideal for building and testing data-driven tools in T1D care.
- For example:
  - **Blood glucose prediction** using real-world time-series data
  - **Personalized treatment recommendation** based on daily patterns
  - **Counterfactual reasoning** to explore “what-if” scenarios
  - **Evaluation** of automated insulin delivery (AID) systems
  - **Temporal analysis** of insulin, meals, and glucose trends

Use Case: GLIMMER

Scan to View!

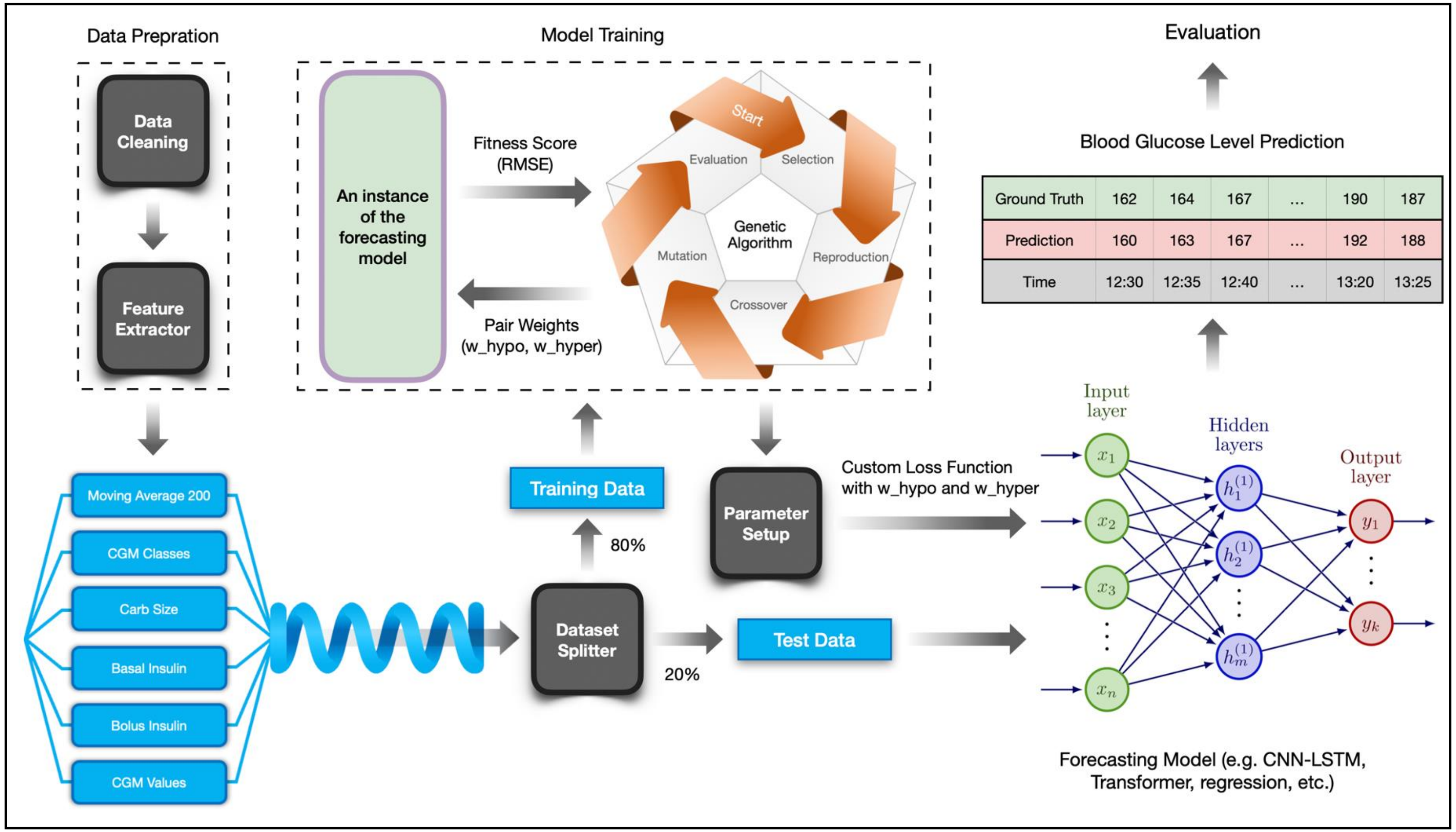
- **GLIMMER:** Glucose Level Indicator Model with Modified Error Rate.
- Developed a machine learning model for blood glucose forecasting.
- Introduced a **custom loss** to **prioritize dysglycemic regions.**
- Tuned loss weights using a **genetic algorithm.**
- Trained on multi-week real-world data from AZT1D.
- Achieved **25% lower RMSE**, **31% lower MAE** than baseline.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

➢ y_i and $\hat{y}_i$ are the actual and predicted CGM values

Use Case: GLIMMER

Scan to View!



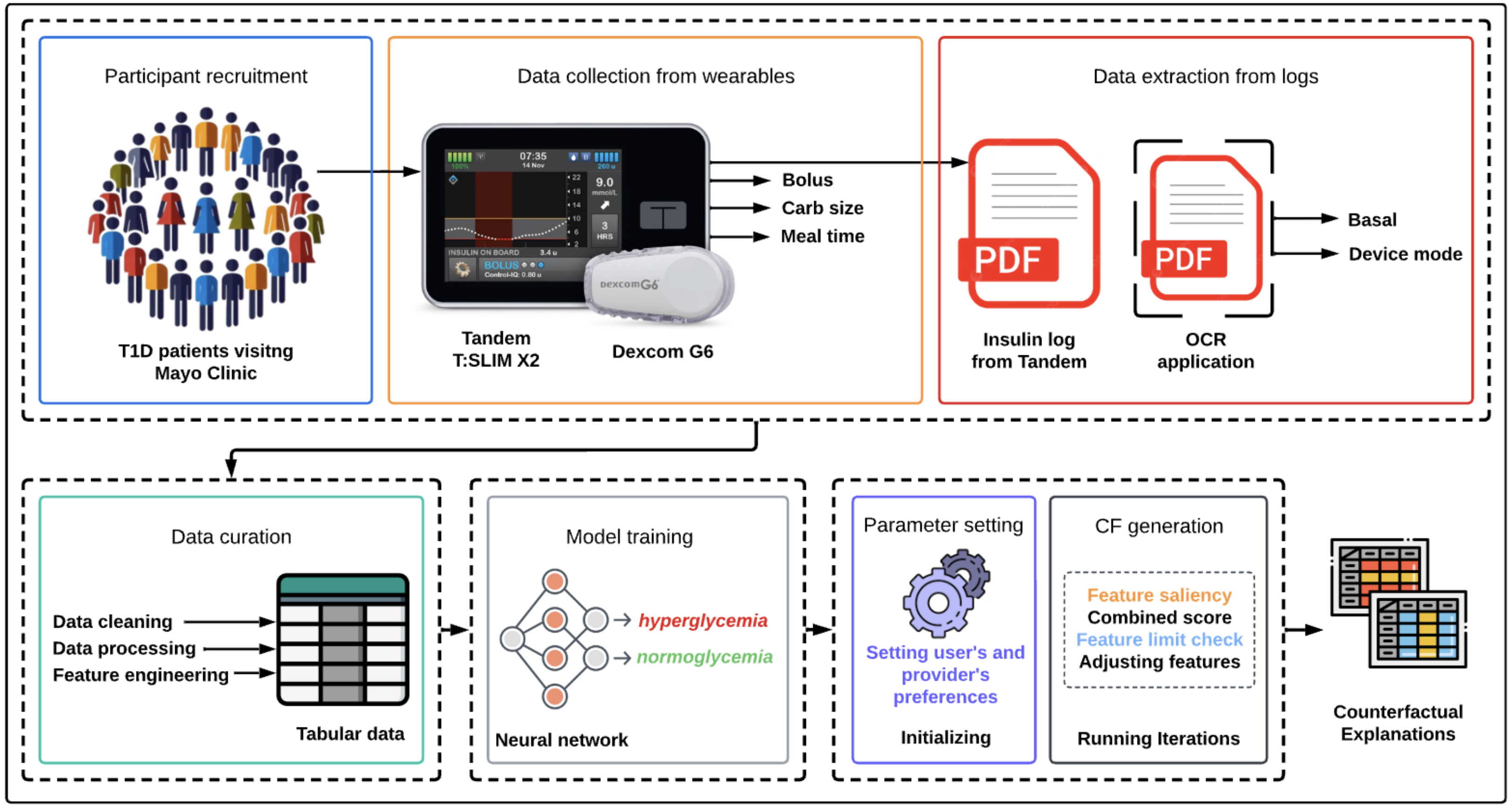
Use Case: GlyMan

Scan to View!

- **GlyMan:** Glycemic management using patient-centric counterfactuals.
- Built a model to generate **personalized "what-if" suggestions**.
- Aimed to help **reduce hyperglycemia** through **behavior changes**.
- Learned patterns from real-world data in AZT1D.
- Considered patient constraints like **minimal effort or disruption**.
- Achieved **76.6% valid explanations**, **86% effective recommendations**.

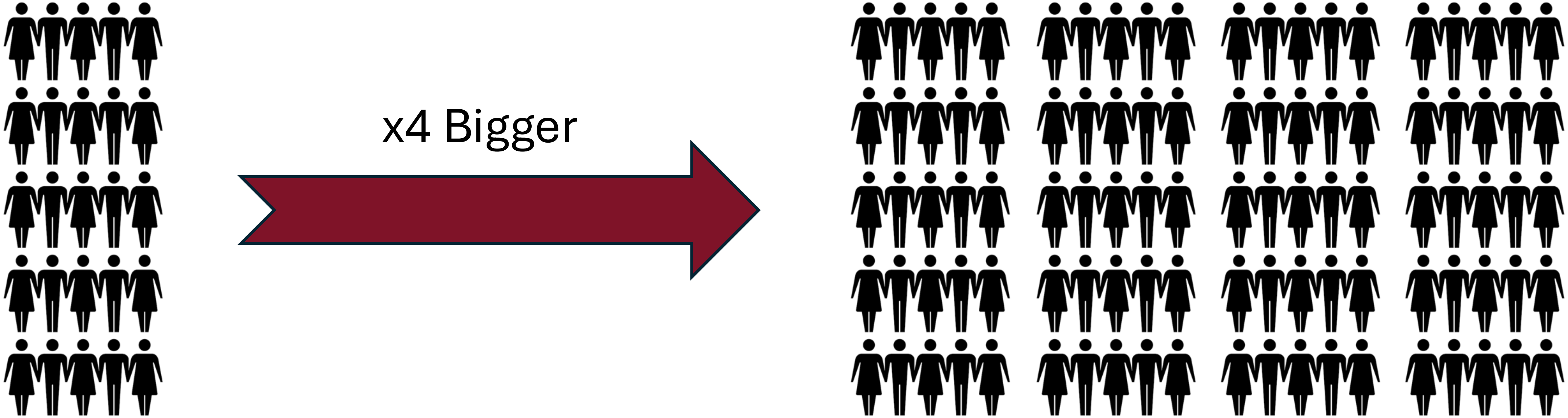
Use Case: GlyMan

Scan to View!



Future Work

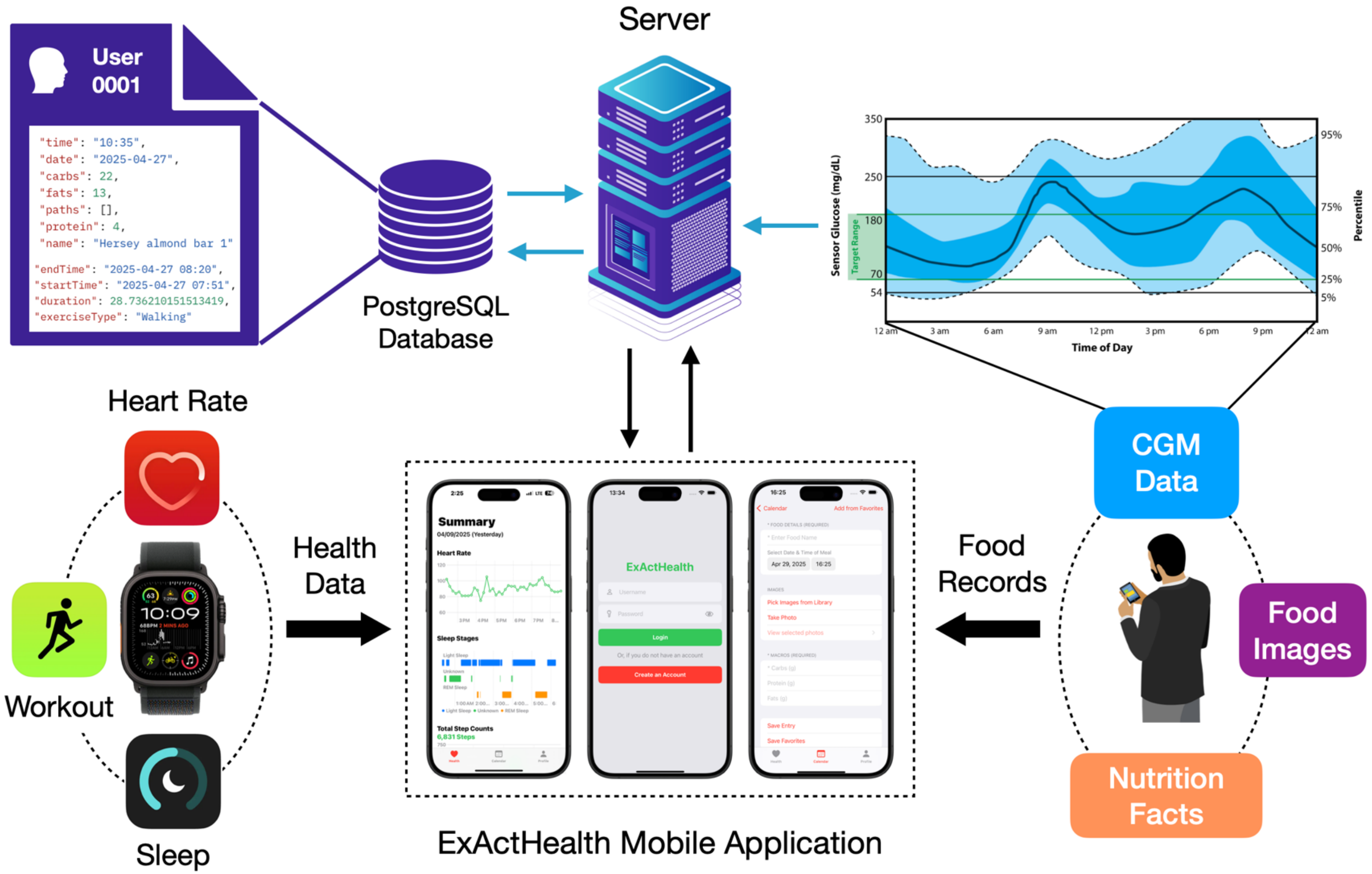
- We aim to expand the dataset to include up to **100 patients** and release **AZT1D v2**.



Ongoing Project: ExActHealth

- **ExActHealth:** Advanced Multimodal Dataset for T1D Research.
- Combines **mobile app** and **smartwatch data.**
- Includes **20 patients**, monitored for **one month**.
- **Logs detailed food intake**: time, portion, and macronutrients.
- Captures **food images** for each meal/snack.
- Tracks **heart rate**, **step count**, and **sleep stages.**
- Designed to study the impact of **lifestyle** on glucose control.

Ongoing Project: ExActHealth



Arizona State University

EMBEDDED MACHINE INTELLIGENCE LAB

Director

Hassan Ghasemzadeh

PhD Students

Abdullah Mamun

Asiful Arefeen

Reza Rahimi Azghan

Nooshin Taheri Chatrudi

Shovito Barua Soumma

Pegah Khorasani

Saman Khamesian

Eric Junyoung Kim

Ebrahim Farahmand

MS Students

Aashritha Machiraju

Undergraduate Students

Suraj Puvvadi

Embedded Machine Intelligence Lab (EMIL)

Email: Hassan.Ghasemzadeh@asu.edu

Website: ghasemzadeh.com

Thank You!