Glucose Spike Prediction from Meal Macros

An Exploratory Data Analysis

Directed Reading Program 2025, **Viggy Vanchinathan**

Mentor: Andrew Henrichson

Johns Hopkins University

Agenda

- Diabetes & Glucose Monitoring – introduction
- CGMacros Dataset
  - Study methodology
  - Example signal
  - Glucose spikes
- The Problem
- Strategies for Prediction
  - ???
  - ???
  - ???
- Success?

Glucose Monitoring

Abbott FreeStyle Libre Pro CGM

Dexcom G6 Pro CGM

CGMacros Dataset

Database | Open Access

CGMacros: a scientific dataset for personalized nutrition and diet monitoring

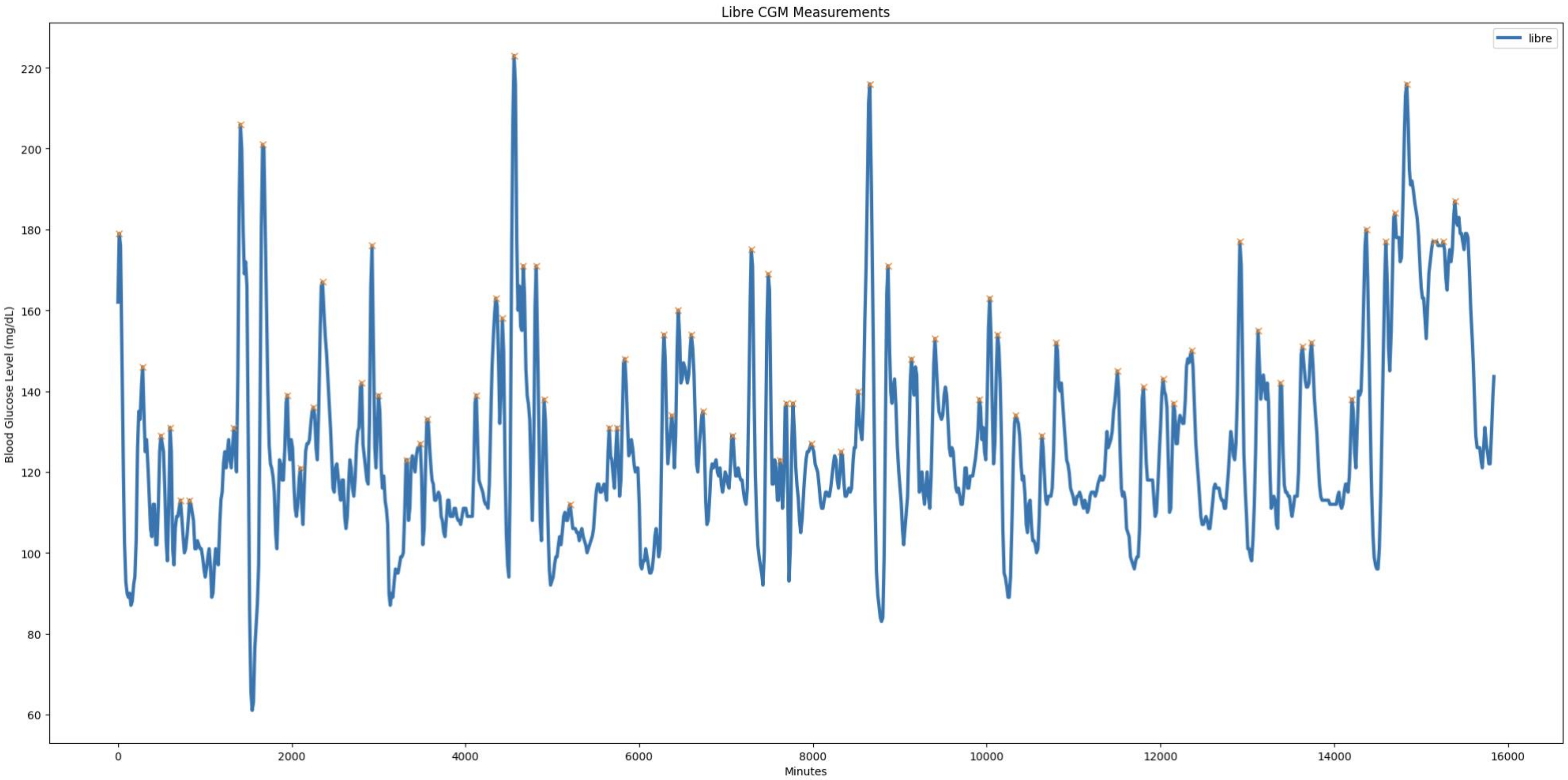
Ricardo Gutierrez-Osuna, David Kerr, Bobak Mortazavi, Anurag Das

Published: Jan. 28, 2025. Version: 1.0.0

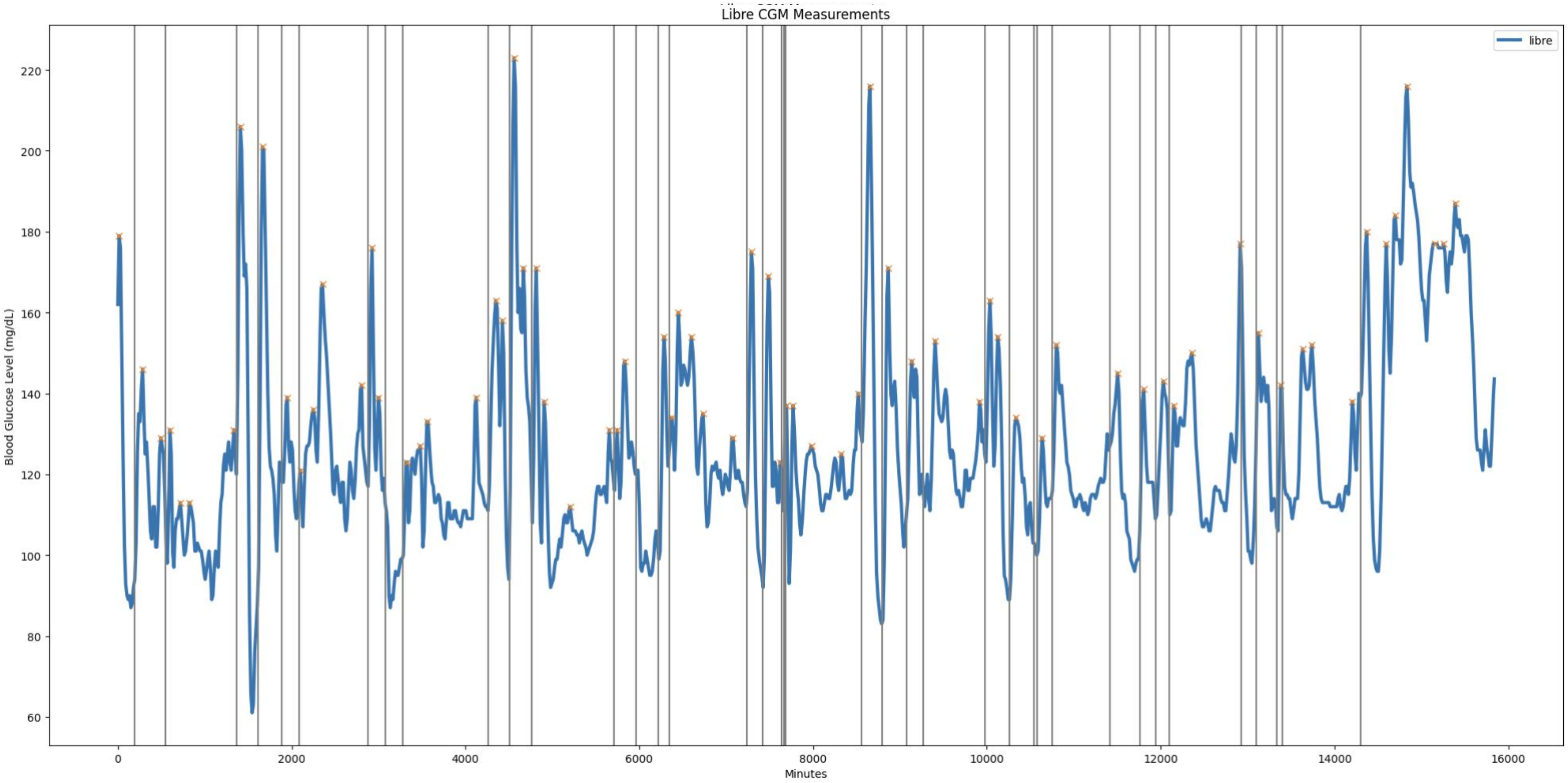
Multimodal dataset containing: meal macronutrients, photographs of food, physical activity, patient health parameters from blood analysis, gut microbiome profiles of 45 study participants (15 healthy adults, 16 with prediabetes, 14 with type 2 diabetes)

→ Continuous glucose measurement over 10 day time span with standardized meals

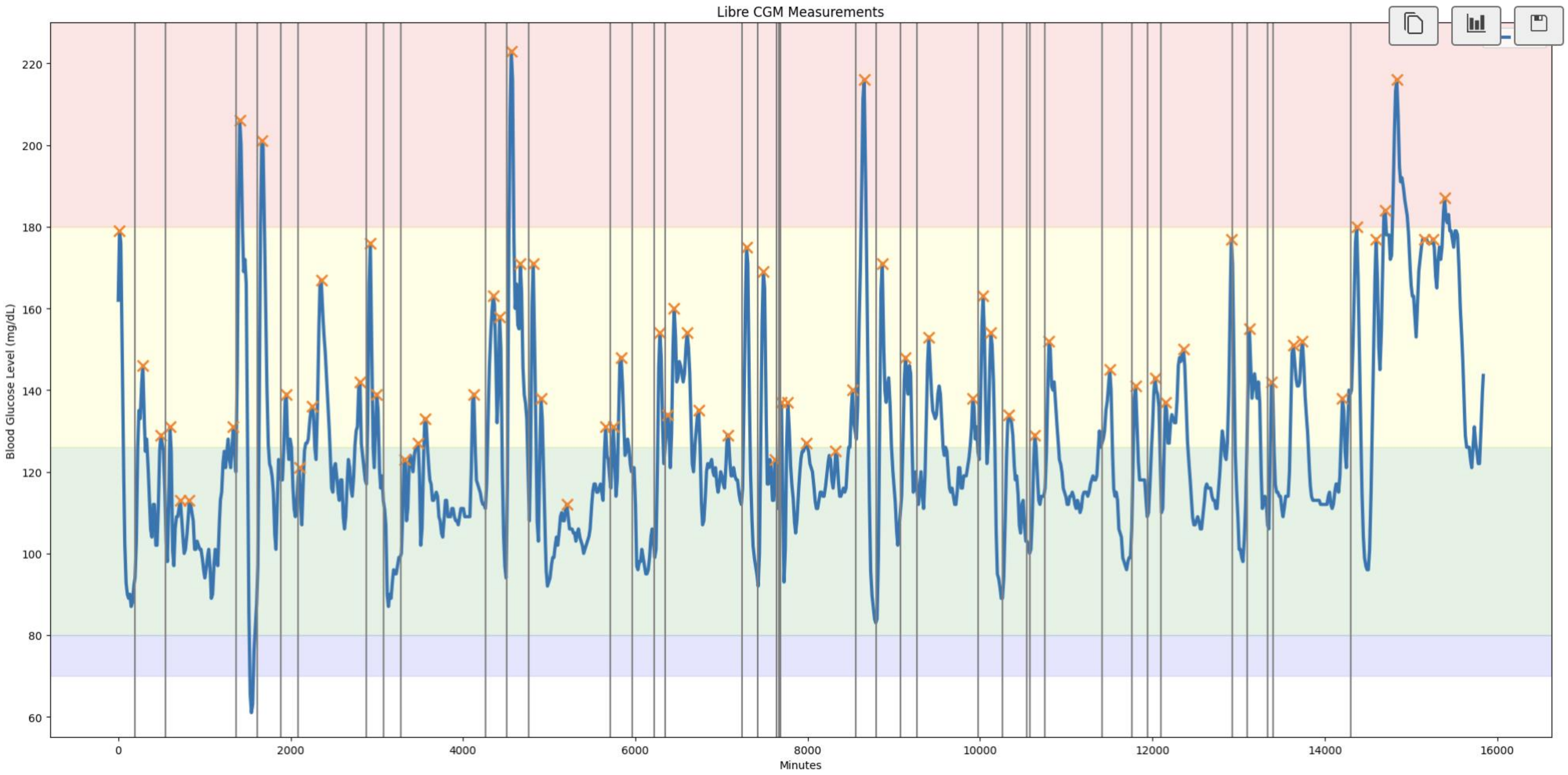
Continuous Time-Series GM



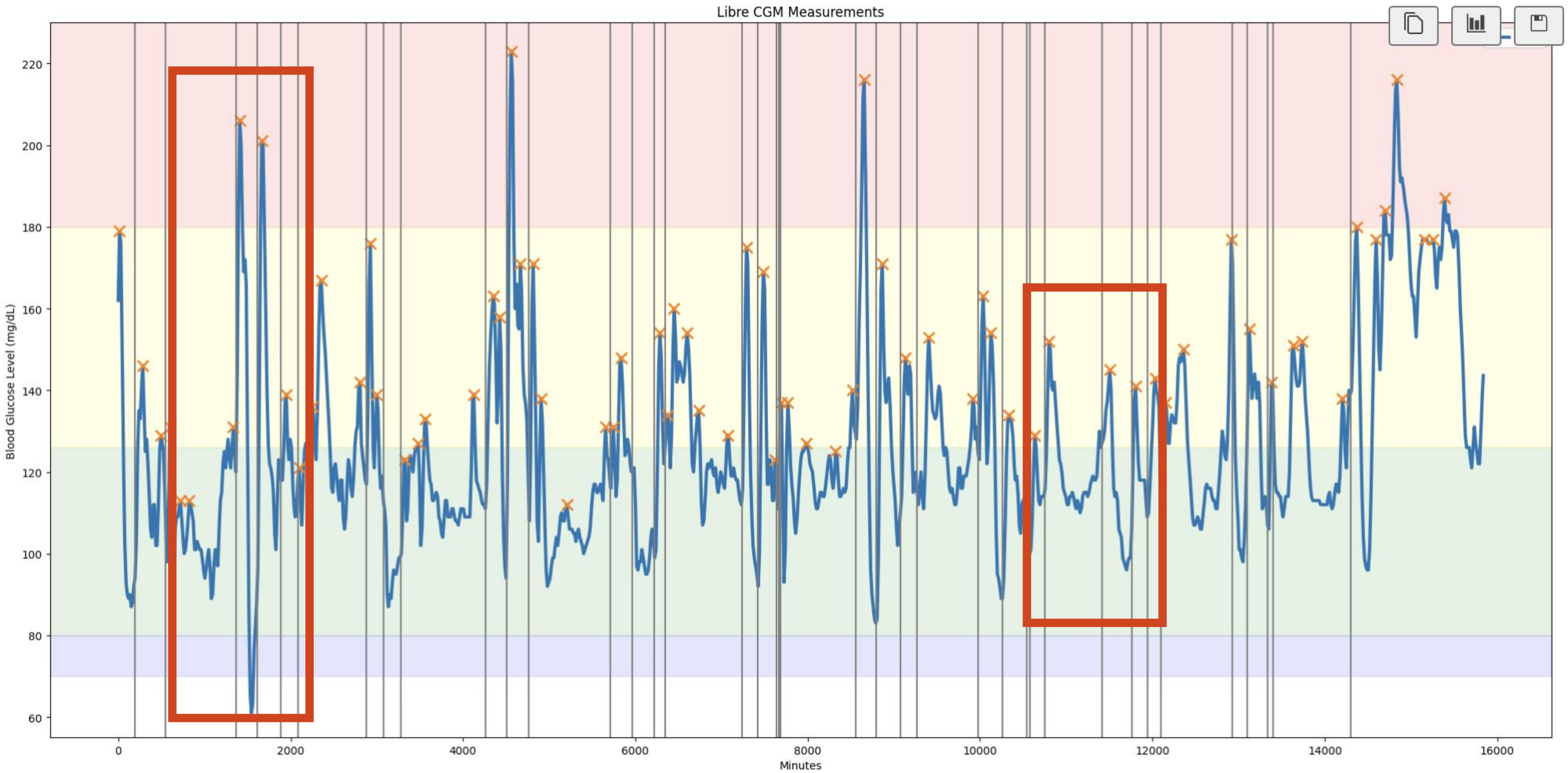
Continuous Time-Series GM



Continuous Time-Series GM

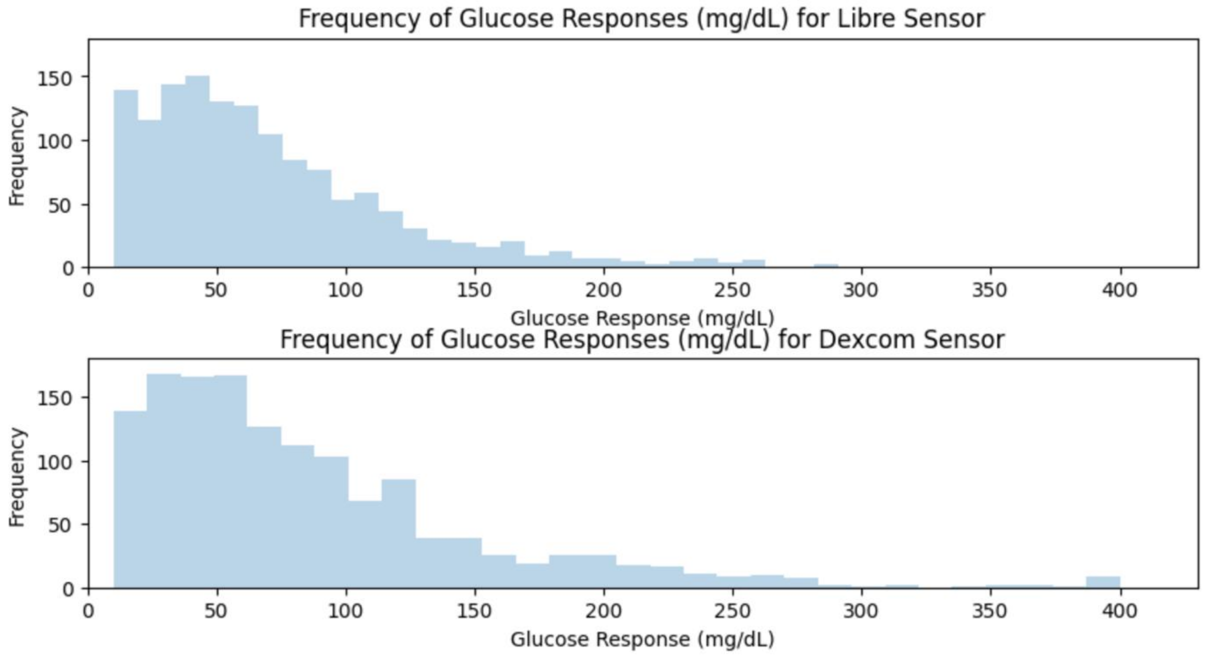


Continuous Time-Series GM



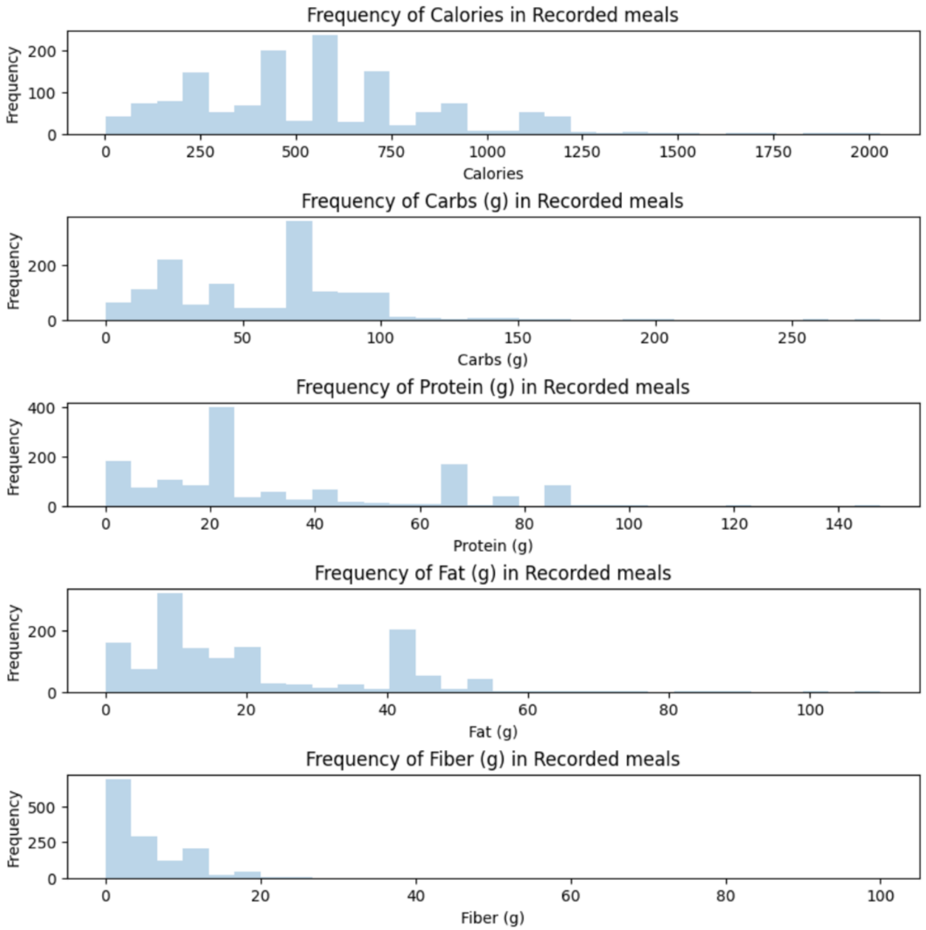
Finding Peaks

NOT the overall
height of the peak!



r = 0.735

The Problem… (postprandial spike)



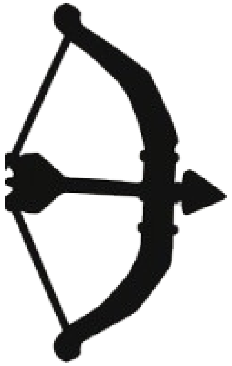
$$X \rightarrow y$$

Calories, Fat, Protein, Fiber, Carbs…

Severity of glucose spike

… the artillery!

???



???

???

First Pass

Linear Methods

???

???

Naïve Approach: Regression

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 = y$$

Carbs Protein Fats Fiber Calories

Want $\boldsymbol{w}$ to minimize Mean Squared Error (OLS)

Naïve Approach: Regression

$$w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 = y$$

Carbs Protein Fats Fiber Calories

69.764 0.493 -0.074 0.024 -1.126 0.010

$R^2 = 0.0364$ 🙁

Second Pass

Linear Methods

Nonlinear Methods

???

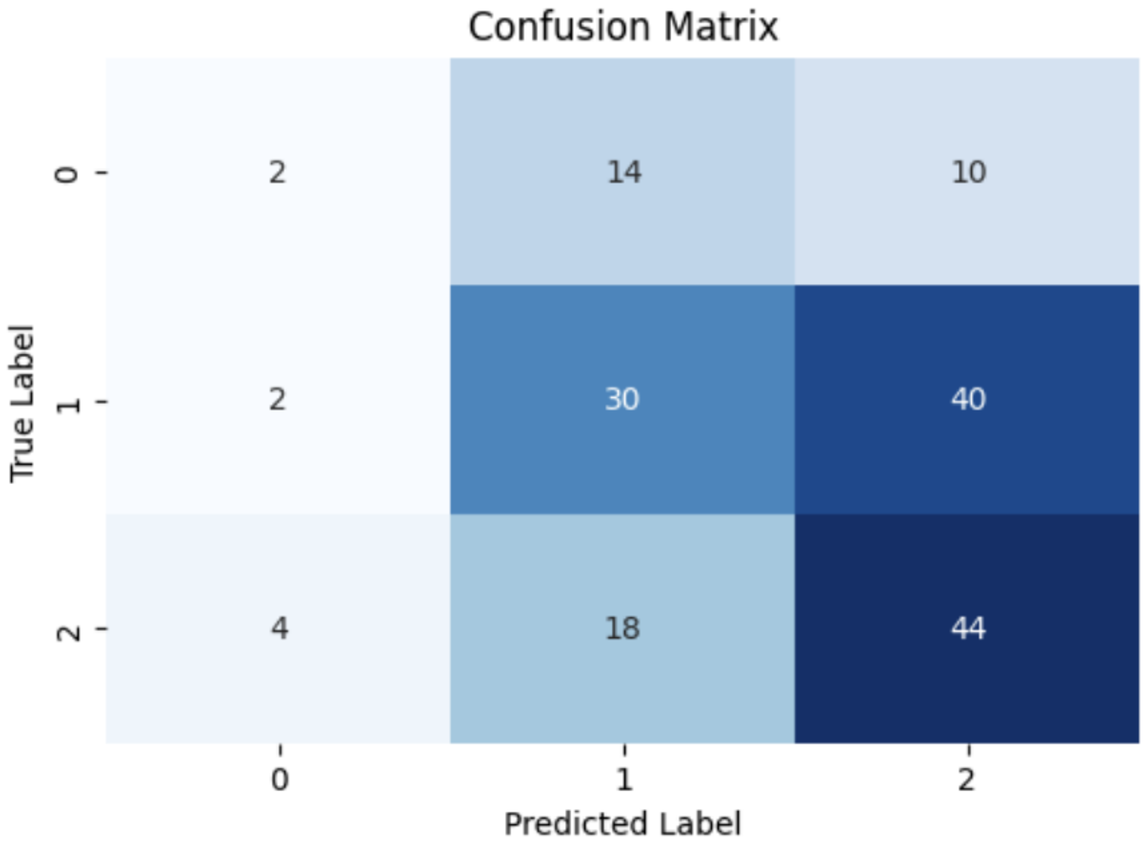
XGBoost: Gradient Boosted Decision Trees

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

- Discretized data into three stages
  - <50,
  - 50-100
  - >100
- Hopefully easier task!

XGBoost: Gradient Boosted Decision Trees

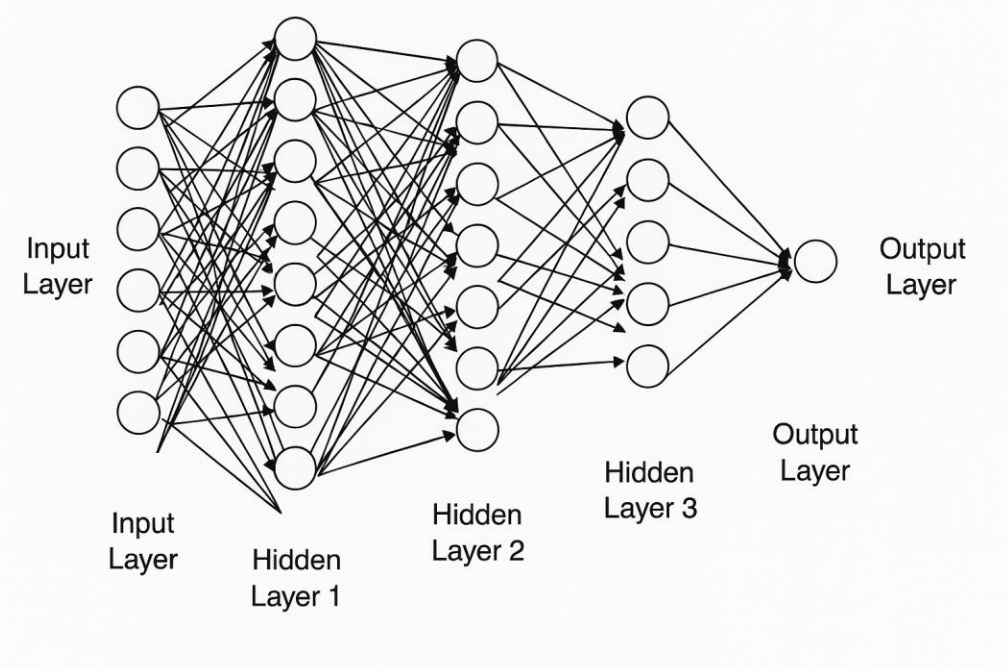
Accuracy:
0.44



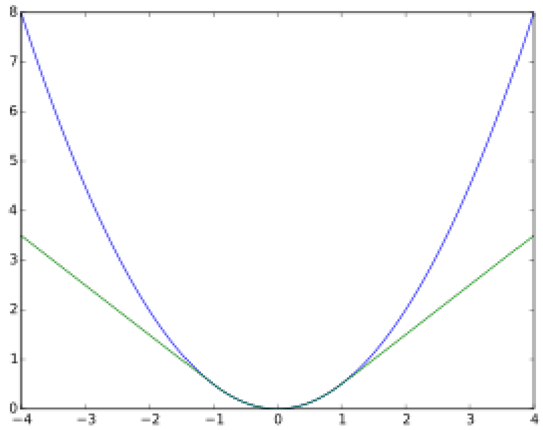
Third Pass

Linear Methods	Nonlinear Methods	Neural Methods
		

Multilayer Perceptron



$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

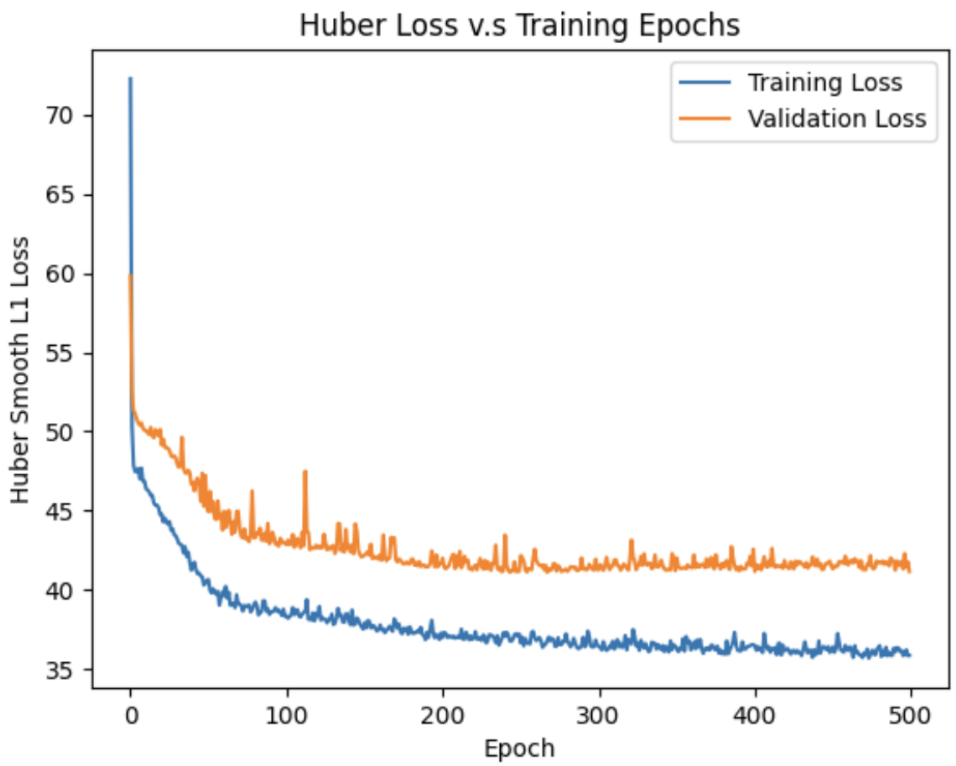


Want $\boldsymbol{w}^1, \boldsymbol{w}^2, \boldsymbol{w}^3$ to minimize Huber Loss

Multilayer Perceptron

Hyperparameters:

- Huber Smoothed L1 Loss
- Adam Optimizer
- Epoch size: 500
- Batch size: 64



MAE: 38.46 mg/dL

Conclusions & Future Directions

Conclusions

- Glucose response is highly dependent on the participant response characteristics, sensitivity
- Nonlinear dynamics between macronutrient composition and glucose spike value
- Limitations due to confounding effect of unspecified variables (e.g. insulin amount)

Extension

- Incorporating multimodal analysis:
  - Meals pictures/consumption -> macro estimation -> glucose spike prediction (famous use case)
- Time/Sequence dependent models (LSTM, HMM)
- More mechanistic models (dependent on glucose response modeling)

???