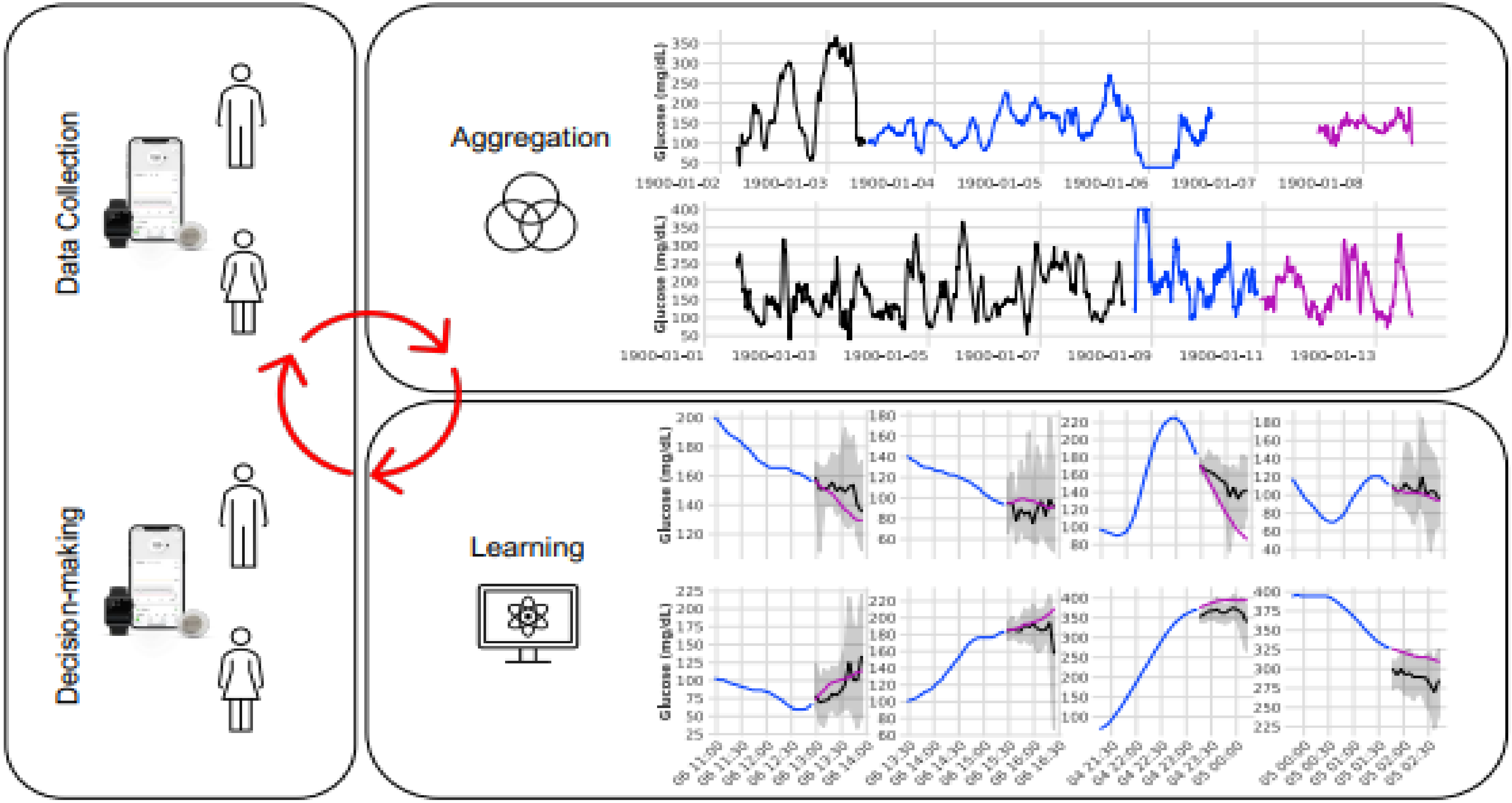
GlucoBench: A Benchmark for Evaluating Blood Glucose Forecasting Models

Introduction

- Blood glucose forecasting is essential for diabetes management.
- Current methods often lack standardized benchmarks and reproducibility.
- GlucoBench provides a comprehensive benchmark framework with:
  - Multiple real-world datasets
  - Diverse model implementations (statistical + deep learning)
  - Unified preprocessing and evaluation pipeline


Data Collection
Decision-making
Aggregation
Learning
Glucose (mg/dL)

Related Work

Previous approaches focused on:

- Small-scale datasets
- Custom evaluation methods
- Inconsistent preprocessing

GlucoBench addresses these gaps by:

- Providing 5 public datasets
- Enabling fair comparison through uniform preprocessing and metrics
- Including baseline and state-of-the-art models

Datasets Used

- **Datasets**: Weinstock, Colas, Dubosson, Hall, iGlu
- Characteristics:
  - Continuous glucose monitoring (CGM) data
  - Varying sampling frequencies and subject counts
- Data includes: timestamps, glucose levels, and contextual covariates (age, BMI, etc.

Table 2: Demographic information (average) for each dataset before (Raw) and after pre-processing (Processed). CGM indicates the device type; all devices have 5 minute measurment frequency.

Dataset	Diabetes	CGM	# of Subjects		Age		Sex (M / F)	
	Overall	Overall	Raw	Processed	Raw	Processed	Raw	Processed
Broll et al. (2021)	Type 2	Dexcom G4	5	5	NA	NA	NA	NA
Colás et al. (2019)	Mixed	MiniMed iPro	208	201	59	59	103 / 104	100 / 100
Dubosson et al. (2018)	Type 1	MiniMed iPro2	9	7	NA	NA	6 / 3	NA
Hall et al. (2018)	Mixed	Dexcom G4	57	56	48	48	25 / 32	NA
Weinstock et al. (2016)	Type 1	Dexcom G4	200	192	68	NA	106 / 94	101 / 91

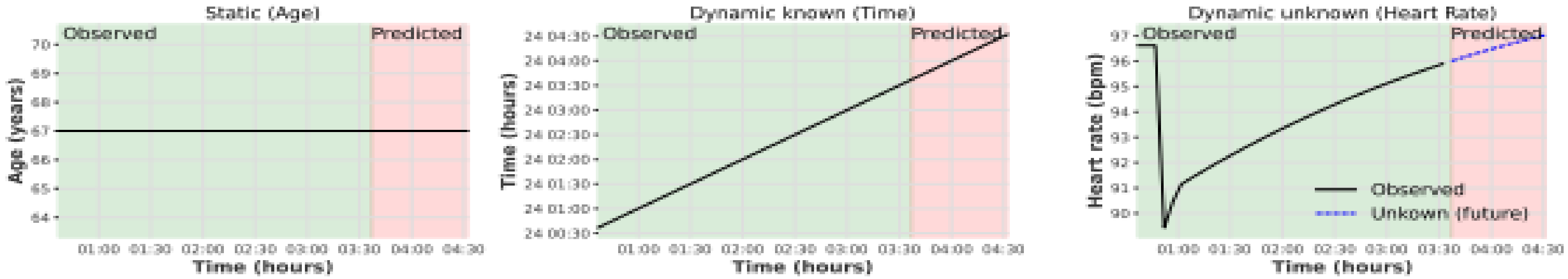


Figure 2: An illustration of static (Age), dynamic known (Date), and dynamic unknown (Heart Rate) covariate categories based on data from Hall et al. (2018) and Dubosson et al. (2018).

Table 3: Interpolation parameters for datasets.

Parameters	Broll	Colas	Dubosson	Hall	Weinstock
Gap threshold (minutes)	45	45	30	30	45
Minimum length (hours)	20	16	20	16	20

Preprocessing Pipeline

- Implemented using DataFormatter class
- Steps:
  - Data cleaning and NA removal
  - Interpolation and resampling
  - Feature encoding and scaling
  - Split into Train / Validation / Test / OOD
- Consistent preprocessing ensures fair model evaluation

Models and Benchmarks

Statistical Models:

Linear Regression

ARIMA

Machine Learning Models:

XGBoost

NHITS

Deep Learning Models:

Transformer

TFT (Temporal Fusion Transformer)

Latent ODE

Gluformer (novel)

Classical and Tree-based Models

Model	Description
ARIMA	Traditional time-series model, used in early glucose forecasting.
Linear Regression	Simple baseline, separate model for each time step t=1…Tt = 1 \dots Tt=1…T.
XGBoost	Gradient-boosted decision trees. Trained separately for each future time step.

Deep Learning Models

Transformer	Encoder-decoder	Standard auto-regressive attention model.
TFT	RNN + Attention	Quantile-based model; supports static/dynamic covariates.
NHiTS	Hierarchical Interpolation	Frequency-domain deep model for long-term patterns.
Latent ODE	RNN + ODE	Encodes into latent space, evolves with ODE, then decodes.
Gluformer	Probabilistic Transformer	Uses mixture distributions for uncertainty modeling.

Glucose Prediction Benchmark

Comprehensive evaluation framework for glucose prediction models

Two Main Tasks:

- Predictive Accuracy
- Uncertainty Quantification

Task 1 - Predictive Accuracy Overview

Objective:

- How accurately does the model predict future glucose values?

Input/Output:

- Input: Historical glucose values (x)
- Output: Predictions (ŷ) for future time window T

Key Insight:

- Why Median over Average?
- Error values (RMSE, MAE) are right-skewed with large outliers
- Median provides more robust performance estimates

Task 1 - Evaluation Metrics

RMSE (Root Mean Squared Error):

Penalizes larger errors more heavily

Formula: $RMSE = \sqrt{(1/T \times \Sigma(\text{actual} - \text{predicted})^2)}$

MAE (Mean Absolute Error):

Measures average magnitude of errors

Formula: $MAE = 1/T \times \Sigma|\text{actual} - \text{predicted}|$

Statistical Approach:

Both metrics calculated across all prediction windows

Median values reported for robustness against outliers

Task 2: Uncertainty Quantification

Core Question:

- How confident is the model about its predictions?
- Does predicted uncertainty match real uncertainty?

Two Evaluation Methods:

1. Log-Likelihood

- For: Probabilistic models only
- Measures: How well predicted distribution explains actual data
- Goal: Higher is better

2. Calibration Error (ECE)

- For: All models with confidence intervals
- Measures: Accuracy of predicted confidence bands
- Goal: Lower is better

Task 1 - Evaluation Metrics

Step-by-Step Process:

- Choose confidence levels (e.g., 50%, 70%, 90%)
- Count actual predictions within confidence bands
- Calculate calibration error

Formula:

$$Cal_t = \Sigma(\text{predicted\_confidence} - \text{actual\_coverage})^2$$

Multi-step Approach:

- Calculated marginally (per time step)
- Comprehensive multi-step prediction evaluation

Results Summary

Best ID performance: **XGBoost / NHITS**

Best OOD performance: **Gluformer**

Statistical models (like ARIMA) underperform on OOD

Transformer-based models generalize better but are slower

Accuracy	Broll		Colas		Dubosson		Hall		Weinstock	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ARIMA	**10.53**	**8.67**	5.80	4.80	13.53	11.06	8.63	7.34	13.40	11.25
Linear	11.68	9.71	**5.26**	**4.35**	**12.07**	**9.97**	7.38	6.33	13.60	11.46
Latent ODE	14.37	12.32	6.28	5.37	20.14	17.88	**7.13**	**6.11**	13.54	11.45
Transformer	15.12	13.20	6.47	5.65	16.62	14.04	7.89	6.78	**13.22**	**11.22**
Uncertainty	Lik.	Cal.	Lik.	Cal.	Lik.	Cal.	Lik.	Cal.	Lik.	Cal.
Gluformer	**-2.11**	**0.05**	**-1.07**	0.14	**-2.15**	**0.06**	**-1.56**	**0.05**	**-2.50**	0.08
TFT	–	0.16	–	**0.07**	–	0.23	–	0.07	–	**0.07**

Limitations

- Datasets still limited to specific regions and demographics
- No real-time or multi-horizon forecasting support
- Potential room for additional physiological or lifestyle covariates
- Training deep models requires high compute (GPU recommended

Conclusion & Future Work

- Provides a **comprehensive benchmark** for glucose forecasting with:
    - Public datasets
    - Standardized tasks
    - Baseline and advanced models
- Highlights importance of:
    - Dataset size
    - Population heterogeneity
    - Test conditions (ID vs. OD, day vs. night)
    - Covariate availability and quality

- Future work:
    - Incorporate synthetic data
    - Add multimodal signals (e.g., insulin, meals)
    - Expand to real-time systems