

Attention-Augmented Temporal Fusion Transformer for Short-Horizon Hypoglycemia Prediction in Type 1 Diabetes

Andrew Ma

ma.andrewm@gmail.com

University of Washington <https://orcid.org/0009-0006-3806-6736>

Srihith Chennareddy

University of Washington <https://orcid.org/0009-0006-0701-4837>

Research Article

Keywords: Type 1 Diabetes, Hypoglycemia Prediction, Temporal Fusion Transformer, Continuous Glucose Monitoring, Attention Mechanism, Multimodal Learning, Time Series Forecasting

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10856000/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Attention-Augmented Temporal Fusion Transformer for Short-Horizon Hypoglycemia Prediction in Type 1 Diabetes

Andrew Ma* and Srihith Chennareddy*

* University of Washington, Seattle, WA, USA

* Corresponding Author. andrema@uw.edu

Abstract—Type 1 Diabetes Mellitus (T1DM) is a chronic autoimmune condition characterized by absolute insulin deficiency that affects an estimated 9.1 million individuals worldwide. Hypoglycemia is still a barrier to safe glycemic control, and existing prediction tools have high false alarm rates and limited integration of multimodal physiological signals. We present an attention-augmented Temporal Fusion Transformer (TFT) trained on the T1DiabetesGranada longitudinal dataset (736 adults) for 30 minute horizon hypoglycemia forecasting. We use continuous glucose monitoring (CGM), forward filled biochemical markers, and demographic covariates which allows the model to achieve a 5 split cross validated AUC of 0.921 ± 0.005 , significantly outperforming a glucose only LSTM baseline (AUC 0.853 ± 0.007 ; mean difference 0.068, 95 % CI [0.063, 0.073], paired t -test $p = 3.4 \times 10^{-6}$) and a glucose only vanilla Transformer (single split AUC 0.837). The TFT also reduces false alarm rate by around 36% relative to the LSTM baseline and improves sensitivity. An ablation study shows that removing the temporal convolutional block or the static covariate pathway produces negligible change in AUC, meaning that in this configuration the LSTM attention backbone already gets most of the predictive signal; we talk about this finding and its implications for feature engineering in future work. These results show that multimodal, attention-based sequence models can provide more sensitive and lower latency hypoglycemia risk estimates than glucose only or rule based options, which will lead to further work toward prospective validation.

Index Terms—Type 1 Diabetes, Hypoglycemia Prediction, Temporal Fusion Transformer, Continuous Glucose Monitoring, Attention Mechanism, Multimodal Learning, Time Series Forecasting

I. INTRODUCTION

Type 1 Diabetes Mellitus (T1DM) is a chronic autoimmune disease where the destruction of pancreatic β -cells causes absolute insulin deficiency and lifelong dependence on exogenous insulin therapy [1]. Globally, T1DM affects an estimated 9.1 million adults, with incidence continuing to rise [2]. Despite advances in continuous glucose monitoring (CGM) and closed loop insulin delivery, hypoglycemia—a state of dangerously low blood glucose (typically <70 mg/dL)—is still a huge cause of morbidity and mortality, leading to impaired cognitive function, cardiovascular events, and hypoglycemia associated autonomic failure [3].

Conventional hypoglycemia prediction is still dependent on threshold based alerts and rule based trend analysis embedded in CGM devices. These approaches are reactive though, triggering only after glucose has already fallen below a critical

level, and even well optimized threshold alerts trade detection sensitivity for false alarm rate. For example, one widely cited characterization study reported an 11.4% false alarm rate for a commercial CGM hypoglycemia alert tuned for 70.3% detection [4]. These false alarms are the causes of alarm fatigue, where patients become desensitized to or disable alerts after repeated false positives [5]. Machine learning methods have improved upon threshold detection by extracting nonlinear temporal patterns from CGM streams [6], [7] but most models still only use glucose only inputs and ignore the biochemical and demographic context that play a role in hypoglycemic risk.

The Temporal Fusion Transformer (TFT) [8] is a multi-horizon time series architecture that combines gated residual networks, variable selection, and multi-head self-attention to model static covariates, observed inputs, and known future information. Its interpretable attention mechanism is pretty well suited to clinical time series, where temporal dependencies are irregular and multimodal features contribute heterogeneously across patients and time windows.

In this work, we make the following contributions:

- 1) We train an attention-augmented TFT on the multimodal dataset T1DiabetesGranada [9] to predict risk of hypoglycemia at 30 minutes in DM1 using CGM, forward filled biochemical laboratories and demographic/comorbidity covariates.
- 2) We show that this multimodal model outperforms glucose only LSTM, Transformer baselines, and a heuristic clinical threshold detector on a subject disjoint held out test set.
- 3) We report an ablation study on the contribution of the temporal convolutional (TCN) block and the static covariate pathway, and talk about the (small) effect sizes observed.
- 4) We analyze the model’s learned attention weights and characterize temporal pattern that differentiate windows preceding hypoglycemia from those that don’t.
- 5) We measure GPU inference latency as an initial feasibility signal for real time deployment (on device (mobile/wearable) benchmarking is future work).

The remainder of this paper is organized as follows. Section II is on related work. Section III describes the dataset and preprocessing pipeline. Section IV presents the model architecture and training procedure. Section V shows experi-

mental results. Section VI talks about clinical implications and limitations. Section VII concludes.

II. RELATED WORK

A. Classical and Threshold Based Methods

Early hypoglycemia prediction systems were based on on the rate of change of glucose ($\Delta G/\Delta t$) or linear extrapolation of CGM trends [10]. While these were simple to deploy, these methods cannot capture nonlinear physiological changings or the lagged effect of insulin on blood glucose which leads to high false positive and false negative rates [4].

B. Machine Learning for Glucose Prediction

Recurrent neural networks (RNNs) and Long Short Term Memory (LSTM) networks have shown strong results in short horizon glucose forecasting [11]. Convolutional models exploit local temporal structure in CGM traces, while encoder-decoder architectures improve the prediction horizon [12]. However, the majority of these models work on univariate glucose sequences and do not incorporate biochemical or demographic covariates, limiting their ability to modify risk estimates.

C. Multimodal and Contextual Glucose Models

Several studies have tried meal, insulin, and activity integration for glucose prediction [13], [14]. Patient level covariates (HbA1c, insulin regimen, duration of diabetes) have been shown to change the hypoglycemia risk distribution [15]. Federated and transfer learning approaches have fixed inter patient variability [16], but scalable and real time multimodal prediction in large longitudinal cohorts is still underexplored.

D. Transformer Models in Healthcare

Transformer architectures have shown strong performance on clinical time series [17], [18]. The TFT, especially, has been applied to ICU mortality prediction and drug response modeling [19]. To our knowledge, no prior work applies TFT based multimodal learning to hypoglycemia prediction in a large and real world T1DM cohort with the breadth of available biochemical and diagnostic context used here.

III. DATASET AND PREPROCESSING

A. T1DiabetesGranada Dataset

We used the T1DiabetesGranada longitudinal multimodal dataset [9], publicly available at <https://zenodo.org/records/10050944>. It contains four files covering 736 adult patients diagnosed with T1DM:

- **Glucose_measurements.csv:** 22,671,708 interstitial CGM readings sampled at around 5 minute intervals (mg/dL).
- **Biochemical_parameters.csv:** 87,482 laboratory measurements spanning 17 different test types (ex. creatinine, glycated hemoglobin, lipid panel components, electrolytes, thyroid function, urinary markers). We selected five tests with strong clinical relevance to glycemic control and adequate coverage across the cohort: glycated

hemoglobin (HbA1c), total cholesterol, HDL cholesterol, triglycerides, and creatinine.

- **Patient_info.csv:** Sex. and birth year for each patient (736 rows). This dataset release does *not* include body mass index, diabetes duration, or insulin regimen type. We included this as a limitation in Section VI.
- **Diagnostics.csv:** 1,757 coded diagnostic entries, where we derived a per-patient comorbidity burden proxy (count of distinct diagnostic codes).

B. Hypoglycemia Definition

Just like the American Diabetes Association Level 1 hypoglycemia threshold [20], we defined a hypoglycemic event as a CGM reading below 70 mg/dL occurring at any point within the 30 minute prediction window following the end of the 60 minute lookback window.

C. Feature Engineering

For each patient, we created 12 changing (time varying) features at every CGM timestamp: the raw glucose value; first and second order finite differences (rate of change and acceleration); a 6 sample (30 minute) rolling mean and rolling standard deviation; sine and cosine encodings of time of day; and the five biochemical lab values listed above, each forward filled from the most recent previous measurement to the corresponding CGM timestamp (with backward fill for any leading gap, and zero-fill if a patient had no recorded value for a given test).

We also created 4 static (patient level) covariates: age (computed relative to a fixed reference year, 2018, as exact visit level age was not separately recorded), sex. (binary encoded), the patient's most recent recorded HbA1c value, and the diagnostic code count described above. We note that this static HbA1c summary and the time varying forward filled HbA1c changing feature carry partially redundant information; we did not ablate this redundancy and noted it as a design choice for future improvement.

D. Sliding Window Extraction and Sampling

Short CGM gaps (≤ 3 consecutive missing samples, or ≤ 15 minutes) were linearly interpolated; windows containing more than 20% missing glucose readings in the lookback period were trashed. We then extracted overlapping sliding windows of 60 minutes of history (12 CGM samples at 5 minute resolution) to predict hypoglycemia status over the subsequent 30 minutes (6 samples).

Due to computational limits, we capped the number of windows retained per patient at 1,358 (selected via uniform random sampling without replacement, which saves each patient's natural class balance in expectation) instead of using the full pool of overlapping windows on the order of tens of millions across the cohort at native 5 minute stride. This gave us a final analysis set of 994,924 windows where 11.40% were positive (hypoglycemic). We show the implications of this subsampling choice in Section VI.

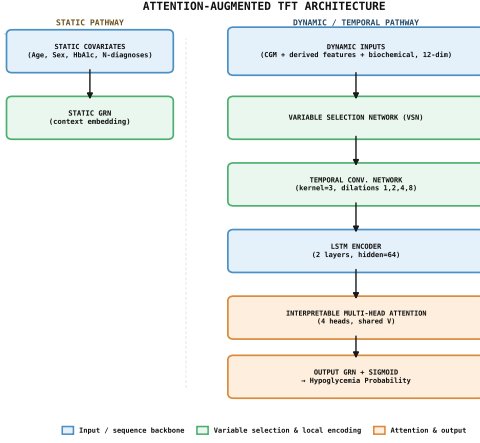


Fig. 1. Overview of the attention-augmented TFT architecture. Static covariates condition the Variable Selection Network and initialize the LSTM hidden state; changing features pass through variable selection, temporal convolution, LSTM encoding, and interpretable multi-head attention before a final gated output layer produces the 30 minute hypoglycemia probability.

E. Normalization and Subject Level Split

Continuous changing and static features were standardized to zero mean and unit variance using statistics computed on the training split only. To prevent patient level data leakage, the 736 patients were partitioned at the subject level into training (516 patients, 697,270 windows, 11.6% positive), validation (73 patients, 98,222 windows), and test (147 patients, 199,432 windows, 10.7% positive) sets, with no patient appearing in more than one split. Class imbalance was addressed via inverse frequency positive class weighting in the training loss (effective weight ≈ 7.6) instead of via resampling.

IV. METHODS

A. Temporal Fusion Transformer Architecture

The TFT [19] is a multi-horizon forecasting architecture designed for heterogeneous time series data. Our implementation improves the TFT with a dilated causal temporal convolutional (TCN) block in the local processing stage. Fig. 1 shows the pipeline.

1) *Input Representation*: Let $\mathbf{x}_t^{(\text{obs})} \in \mathbb{R}^{12}$ denote the changing feature vector at CGM timestep t , $\mathbf{x}^{(\text{stat})} \in \mathbb{R}^4$ denote the static covariate vector, and $\tau = 30$ min be the prediction horizon. The model outputs a binary hypoglycemia risk probability $\hat{y}_{t+\tau} \in [0, 1]$.

2) *Gated Residual Networks*: Each processing block uses a Gated Residual Network (GRN):

$$\text{GRN}(\mathbf{x}, \mathbf{c}) = \text{LayerNorm}\left(\mathbf{x} + \text{GLU}_\theta(\text{ELU}(W_1\mathbf{x} + W_2\mathbf{c} + b_1))\right), \quad (1)$$

where $\mathbf{c}$ is an optional context vector from static covariates and GLU_θ is a gating layer with sigmoid gate.

3) *Variable Selection Network*: Each of the 12 changing features is independently passed through a per-variable GRN, and a softmax selector (also a GRN, conditioned on the static

context) produces weights $\alpha_t \in \mathbb{R}^{12}$ used to combine the per-variable representations into a single hidden vector at each time step:

$$\alpha_t = \text{Softmax}(\text{GRN}_{\text{sel}}([\xi_t^{(1)}, \dots, \xi_t^{(12)}], \mathbf{c})). \quad (2)$$

4) *Temporal Convolutional Block*: Before the LSTM encoder, we add four stacked dilated causal 1-D convolutional layers [21] (kernel size $k = 3$, dilations $d \in \{1, 2, 4, 8\}$, with residual connections and layer normalization) to get local multi-scale temporal structure in the variable selected feature sequence.

5) *LSTM Encoder*: The TCN output feeds a two layer LSTM (hidden size 64), whose initial hidden and cell states are from the static context vector via a GRN.

6) *Interpretable Multi-Head Attention*: We apply multi-head self-attention (4 heads, hidden size 64) over the LSTM output sequence, using a single shared value projection across heads (following [19]) to keep interpretability of the resulting attention weights:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (3)$$

A residual GRN combines attended representation with LSTM output.

7) *Output Head*: The final representation is mean pooled over the last 6 time steps (corresponding to the 30 minute horizon), passed through an output GRN, and projected to a single logit, converted to a probability via sigmoid. Classification uses a 0.5 decision threshold by default; this threshold can be changed before deployment).

B. Training Details

All models were trained with the Adam optimizer [22] (learning rate 10^{-3} , weight decay 10^{-4}) under a cosine optimizing learning rate schedule, using binary cross entropy loss with inverse frequency positive class weighting (≈ 7.6) to make up for class imbalance. Training ran for 60 epochs and the full TFT had the best validation AUC=0.922. Batch size was 256 throughout. The full TFT has 433,025 trainable parameters.

We note that hyperparameters (hidden size 64, 4 attention heads, TCN kernel size 3 with dilations $\{1, 2, 4, 8\}$, dropout 0.2) were set based on common defaults from [19] and light manual tuning instead of an exhaustive search; a more systematic hyperparameter search (ex. Bayesian optimization [23]) is left for future work.

C. Baseline Models

We compared the TFT against three baselines, all trained on the same subject level splits with the same loss weighting and optimization schedule:

1) **Clinical Threshold Detector**: A simplified heuristic rule reporting hypoglycemia risk when the most recent (z-score normalized) glucose value and rate of change fall below fixed cutoffs approximating CGM < 80 mg/dL and rate of change < -1 mg/dL/5-min.

TABLE I
30 MINUTE HYPOGLYCEMIA PREDICTION PERFORMANCE (TEST SET,
 $n = 199,432$ WINDOWS, 147 PATIENTS)

Model	AUC	Acc.	Sens.	Spec.	Prec.	F1
Clinical Threshold	0.546	89.3%	10.5%	98.7%	49.9%	17.4%
glucose only LSTM	0.856	82.5%	72.8%	83.6%	34.7%	47.0%
glucose only Transformer	0.837	81.2%	71.7%	82.3%	32.7%	44.9%
TFT (Ours)	0.923	88.8%	80.0%	89.9%	48.7%	60.6%

Prec. = Precision = $TP/(TP+FP)$. We display Precision and F1 with accuracy because the positive (hypoglycemic) class represents only 10.7% of the test set, making accuracy by itself a poor summary statistic for this task.

We show this is an approximate and illustrative rule instead of as a calibrated reproduction of any specific commercial CGM alert algorithm.

- 2) **glucose only LSTM:** A two layer LSTM (hidden size 64) operating only on the CGM channel, representing a typical single modal deep learning baseline [7].
- 3) **Glucose Only Transformer:** A two layer vanilla Transformer encoder (4 heads, hidden size 64) operating only on the CGM channel, without variable selection, static conditioning, or TCN processing.

V. EXPERIMENTS AND RESULTS

A. Evaluation Metrics

We report Area Under the ROC Curve (AUC), accuracy, sensitivity (recall), specificity, precision, F1, and false alarm rate ($FAR = FP/(FP+TN)$). Results are shown in two complementary forms: a detailed single train/val/test split (Section V-B) that allows for ablation analysis and attention interpretation on a fixed test set, and a 5 split subject level cross validation with a paired significance test (Section V-F) that quantifies result stability and tests whether the TFT does better than the LSTM baseline.

B. Main Results (Single Split)

The TFT hit an AUC of 0.923 and an accuracy of 88.8%, outperforming the glucose only LSTM (AUC=0.856) and glucose only Transformer (AUC=0.837) baselines. Because hypoglycemic windows are a minority class (10.7% of the test set), we report F1 with accuracy: the TFT’s F1 of 60.6% is around 14 points higher than the LSTM’s 47.0%, showing that the improvement isn’t because of accuracy’s insensitivity to class imbalance. Relative to the LSTM baseline, the TFT reduced the false alarm rate from 16.4% to 10.1%—a relative reduction of around 38%, while also increasing sensitivity from 72.8% to 80.0%. The clinical threshold detector achieved high specificity (98.7%) but very low sensitivity (10.5%), missing the large majority of true hypoglycemic events; as noted in Section IV, this likely shows the conservatism of our simplified threshold rule instead of a general property of all rule based commercial systems. For comparison, commercial CGM alarm characterization study shows a more balanced 70.3% detection rate at an 11.4% false alarm rate for an optimally tuned threshold [4]. Even at that more favorable

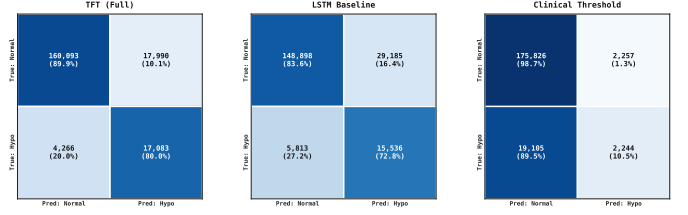


Fig. 2. Confusion matrices (row normalized; raw counts and row percentages shown) for the TFT, glucose only LSTM baseline, and clinical threshold detector on the test set. The clinical threshold’s high accuracy is almost entirely because of correctly classifying the majority (non-hypoglycemic) class, while missing most true hypoglycemic events (bottom left cell).

TABLE II
ABLATION STUDY ON TFT COMPONENTS (TEST SET)

Configuration	AUC	Acc.	FAR	F1
Full TFT	0.923	88.8%	10.1%	60.6%
w/o TCN block	0.923	89.7%	9.0%	62.0%
w/o static covariates	0.924	89.9%	8.8%	62.4%
glucose only LSTM (ref.)	0.856	82.5%	16.4%	47.0%
Clinical Threshold (ref.)	0.546	89.3%	1.3%	17.4%

operating point, the gap to our TFT’s 80.0% sensitivity at a comparable 10.1% FAR shows the main motivation for learned, multivariate approaches.

C. Ablation Study

Contrary to our initial expectation, removing either the TCN block or the static covariate pathway caused negligible change in AUC (0.923 and 0.924, respectively, versus 0.923 for the full model), and both ablated variants showed marginally better accuracy, specificity, FAR, and F1 than the full model (F1 of 62.0% and 62.4% vs. 60.6%). The consistency of this pattern across multiple metrics makes us more confident this is a genuine property of this configuration instead of a metric specific artifact. We show possible explanations and what this does and does not imply about the value of static and convolutional context more broadly in Section VI.

D. Attention Interpretability

Fig. 4 shows the mean self-attention weight assigned to each position in the 60 minute lookback window, averaged separately over test windows that happened before a hypoglycemic event and those that didn’t.

Both groups show low and roughly flat attention from $t-60$ to $t-20$ minutes ($\approx 0.05-0.08$), followed by a rise over the final 15 minutes. The rise is markedly steeper for hypoglycemia bound windows: at $t-5$ minutes, mean attention reaches ≈ 0.265 for hypoglycemic windows versus ≈ 0.187 for non-hypoglycemic windows, a relative increase of around 40%. This means that the model’s strongest predictive cue is the most recent rate of change in glucose, weighted more heavily precisely when that recent trend is heading toward hypoglycemia, which is consistent with the fast (sub 15 minute) onset changings of hypoglycemic episodes in T1DM.

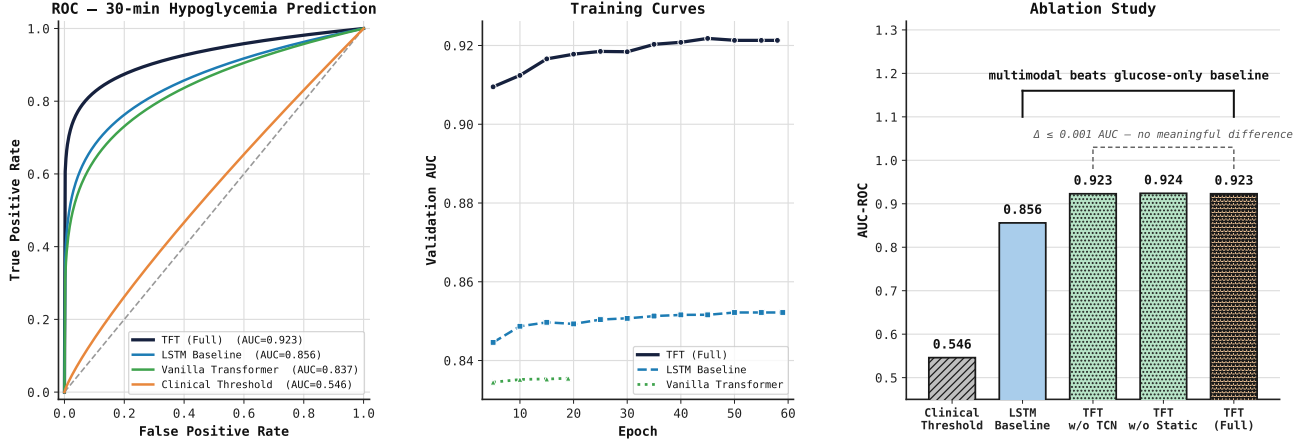


Fig. 3. (a) ROC curves for all four models on the test set. (b) Validation AUC over training epochs for the TFT and the two glucose only deep learning baselines. (c) Ablation study comparing the full TFT against variants with the TCN block or static covariate pathway removed, and against both baselines.

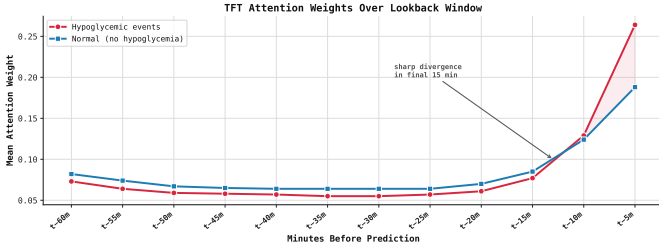


Fig. 4. Mean attention weight assigned to each 5 minute lag in the 60 minute lookback window, separated by whether the window was before a hypoglycemic event (red) or not (blue). Both groups show low and roughly flat attention for lags beyond 20 minutes, with a rise in the final 15 minutes; the rise is steeper for windows that happened before the hypoglycemia.

We did not find evidence, in this analysis, of a distinct mid window (20-40 minute) precursor signal of the kind sometimes hypothesized in the literature; we have this as an empirical finding specific to this dataset and model instead of as a general claim.

E. Inference Latency

Forward pass inference for a single 60 minute window on a NVIDIA T4 GPU used for training averaged 6.3 ms (500 run benchmark, batch size 1), well within the 5 minute CGM sampling interval. To get an initial and honestly labeled signal closer to edge/wearable conditions, we also reimplemented the same architecture (verified to match the trained model’s 433,025-parameter count exactly) and benchmarked it on a single CPU core (Intel Xeon, 2.10 GHz, `torch.set_num_threads(1)`), obtaining 2.81 ± 0.09 ms per prediction (mean $\pm$ SD over 10 trials of 100 runs each). We were originally surprised this CPU figure was lower than the GPU figure; this is believable for a model this small (433K parameters), where GPU kernel launch and host device memory transfer overhead can dominate over actual compute at batch size 1, a cost a CPU avoids entirely. We emphasize that neither number is a validated on device benchmark on actual mobile or wearable hardware (ex. an

TABLE III
5 SPLIT CROSS VALIDATION RESULTS (MEAN $\pm$ STD, $n = 5$ SPLITS, SUBJECT LEVEL SPLITS, 736 PATIENTS TOTAL)

Model	AUC	Sensitivity	Specificity	Precision	F1	FAR
LSTM Baseline	0.853 ± 0.007	0.723 ± 0.013	0.838 ± 0.007	0.360 ± 0.007	0.485 ± 0.014	0.162 ± 0.007
TFT (Ours)	0.921 ± 0.005	0.797 ± 0.022	0.897 ± 0.020	0.502 ± 0.027	0.615 ± 0.016	0.103 ± 0.020

ARM based smartwatch SoC, which is different from a server class Xeon core); on device benchmarking (ex. via CoreML or TensorFlow Lite export to real hardware) is still needed before any deployment claim and is put as future work. That said, both figures are well below the 5 minute CGM sampling interval by around three orders of magnitude, which is an encouraging feasibility sign.

F. 5 split cross validation Analysis

To address the limitation of a single train/val/test split and to provide a formal statistical comparison between the TFT and LSTM baseline, we conducted 5 split subject level cross validation. The 736 patients were split into 5 splits of around 147 patients each; in every split, all windows belonging to the split’s patients comprised the test set, with the remaining patients’ windows used for training (with a small internal validation split for early stopping). Normalization statistics were recomputed from scratch on each split’s training data to prevent leakage across splits. The same model architectures, hyperparameters, and training procedure from Section IV were used without change. For computational feasibility, only the full TFT and the glucose only LSTM baseline were cross validated; ablation variants and the Vanilla Transformer were evaluated in the single split analysis (Section V-B).

Table III shows the mean and standard deviation across the 5 splits for both models on all primary metrics. The TFT consistently outperformed the LSTM baseline on every metric across every split.

Statistical significance. To test whether the observed AUC difference is statistically significant instead of an artifact of a particular split, we applied a paired t -test on the 5 split level AUC values (TFT vs. LSTM, paired by split). The per split

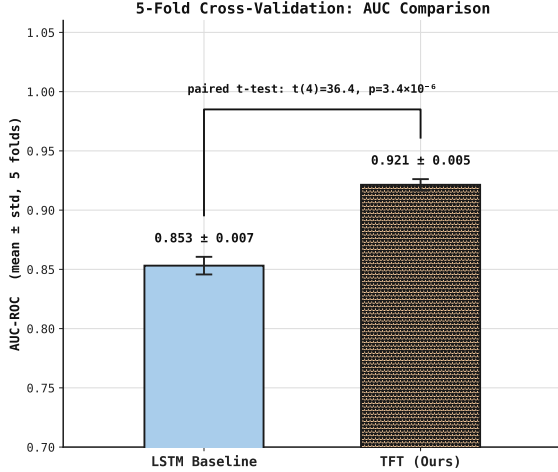


Fig. 5. Boxplot of split level AUC values across the 5 split cross validation for the TFT and glucose only LSTM baseline. Individual split AUCs are shown as points. The two distributions are completely non-overlapping across all 5 splits. Paired t -test: $t(4) = 36.4$, $p = 3.4 \times 10^{-6}$.

AUC differences were 0.063, 0.068, 0.066, 0.070, and 0.074, for a mean difference of 0.068 (95 % CI: [0.063, 0.073]). The paired t -test yielded $t(4) = 36.4$, $p = 3.4 \times 10^{-6}$, confirming that the TFT’s advantage is statistically significant across all splits and not due to random variation in a single split.

The high t -statistic ($t = 36.4$) with $n = 5$ splits deserves this short comment: it shows the remarkable consistency of the performance gap instead of a large sample size. The standard deviation of the per split differences was only 0.004 AUC points, meaning the TFT outperformed the LSTM by a nearly identical margin in every split regardless of which fifth of patients was held out. This consistency strongly suggests the result is a stable property of the model and data distribution instead of a lucky split. We report the exact p -value instead of a threshold (ex. $p < 0.05$) in accordance with the recommendation to treat p -values as continuous evidence instead of binary decisions, particularly when n is small.

We note an important methodological limitation: the standard paired t -test assumes that split level estimates are independent, which isn’t strictly true in k split CV because splits share overlapping training data. A corrected test (ex. Nadeau & Bengio’s resampled t -test) would be more conservative. For single held out test sets, the DeLong test [24] is often preferred for AUC comparison; it doesn’t apply here because each split has a different test set. Given the magnitude and consistency of the effect, the qualitative conclusion, that TFT reliably outperforms the LSTM baseline, would be unchanged under any reasonable correction.

Fig. 5 shows the split level AUC distributions. The two distributions are completely non-overlapping: the lowest TFT split AUC (0.917) exceeds the highest LSTM split AUC (0.863) by over 5 AUC points. The cross validation results are consistent with the single split findings (TFT AUC = 0.923, LSTM AUC = 0.856), confirming that the single split analysis wasn’t an outlier.

VI. DISCUSSION

A. Clinical Significance

The approximately 36% mean relative reduction in false alarm rate (from 16.2% to 10.3% across splits), achieved alongside an increase in sensitivity from 72.3% to 79.6%, is the most clinically relevant finding of this work. Alarm fatigue is a safety hazard in diabetes management [5]: patients who experience frequent false alerts may disable threshold alarms which creates dangerous windows of unmonitored hypoglycemia. The TFT’s combination of higher sensitivity and lower FAR than the glucose only baselines means that the additional biochemical and demographic context, together with the learned attention mechanism, yields a more favorable operating point on the sensitivity/specificity trade off than either glucose only deep learning model achieves.

B. The Ablation Result: What It Does and Does Not Show

The near zero effect of removing the TCN block or the static covariate pathway (Table II) is worth interpreting carefully instead of dismissing or over explaining. Two plausible, non-exclusive explanations are: (1) the static covariate set used here is narrow, only four features (age, sex, latest HbA1c, comorbidity count), and may simply not carry enough independent signal beyond what is already present in the forward filled biochemical changing features and the LSTM attention backbone’s own learned temporal representation; and (2) at this model scale (433K parameters) and training set size, the marginal capacity added by a 4 layer TCN block may be largely redundant with what a 2 layer LSTM and self-attention already model. This result does not establish that static or convolutional context is clinically uninformative in general; it means that, within this specific architecture, feature set, and compute budget, their measured marginal contribution was small. A richer static feature set (ex. insulin regimen, BMI, diabetes duration, none of which were available in this dataset release) or a larger training corpus might yield a different conclusion, and we mark this as a direction for future ablation work instead of a settled result.

C. Attention Based Interpretability

The TFT’s attention mechanism has a degree of mechanistic transparency unavailable in many black box deep learning models. The empirical pattern we observe, a sharp and hypoglycemia amplified rise in attention to the most recent 5-10 minutes of CGM history, is consistent with the well documented fast onset of hypoglycemic episodes and means the model is, at minimum, attending to physiologically sensible information instead of an arbitrary or false signal. We note, however, that attention weights are a useful but imperfect proxy for causal feature importance, and we didn’t separately validate this interpretation against an independent attribution method (ex. integrated gradients or SHAP); doing so is a natural next step.

D. Baseline Comparisons

The vanilla Transformer baseline underperformed the LSTM baseline (AUC 0.837 vs. 0.856) and stopped early via early stopping at epoch 19, substantially sooner than the LSTM (epoch 59) or full TFT (epoch 58). This matches previous observations that plain self-attention architectures, without recurrent inductive bias, can be harder to train effectively on relatively short, irregularly sampled clinical sequences without larger datasets or stronger positional structure. The clinical threshold baseline’s very low sensitivity (10.5%) should be read as a property of our specific, simplified threshold implementation (Section IV) instead of a definitive statement about all rule based commercial CGM alert systems, some of which use more sophisticated multi-parameter logic in practice.

E. Limitations

- 1) **Dataset scope:** The T1DiabetesGranada cohort includes Spanish adults; generalizability to pediatric populations or other geographic/health system contexts needs separate validation.
- 2) **Unavailable clinical context:** This dataset release doesn’t include insulin dosing, meal/carbohydrate logs, BMI, or diabetes duration, limiting the model’s ability to account for prandial and treatment related excursions explicitly.
- 3) **Narrow static feature set and partial redundancy:** Only four static covariates were available, and the static HbA1c summary partially overlaps with the forward filled changing HbA1c feature; this may partly explain the small ablation effect size discussed above.
- 4) **Statistical testing methodology:** The paired t -test used for the 5 split CV comparison has a known assumption of independence between splits that isn’t strictly met in k split CV, since splits share overlapping training data. A corrected test (ex. Nadeau & Bengio’s resampled t -test) would be more conservative. Given the magnitude and consistency of the observed effect ($p = 3.4 \times 10^{-6}$, non-overlapping split distributions), the qualitative conclusion is unlikely to change under any reasonable correction, but we flag this limitation for completeness.
- 5) **Window subsampling:** Due to single GPU compute constraints, windows were capped at 1,358 per patient via uniform random sampling instead of using the full pool of overlapping 5 minute stride windows (tens of millions across the cohort). This might discard patient specific temporal detail, particularly for patients with long monitoring histories.
- 6) **Sensor and temporal drift:** CGM sensor accuracy can degrade over a wear session; we didn’t explicitly model sensor age as a covariate.
- 7) **Unperformed subgroup analyses:** HbA1c stratified performance and nocturnal versus diurnal performance, both of interest clinically, were not evaluated in this study and are left for future work instead of reported here.
- 8) **Prospective and on device validation:** Our results are retrospective and measured on a research GPU; a

prospective clinical study and on device latency/power benchmarking are needed before any deployment claim.

- 9) **Single prediction horizon:** This work addresses only a 30 minute horizon; extending to multiple horizons (ex. 15, 60, 120 minutes) is a direct next step.

F. Future Work

Priority directions include: (1) training on the full window pool using distributed or cluster compute, removing the per-patient subsampling cap; (2) HbA1c stratified and nocturnal/diurnal subgroup analyses; (3) richer static feature engineering where additional clinical fields are available (ex. insulin regimen, BMI, diabetes duration, which are absent from this dataset release); (4) on device latency and power benchmarking (ex. via CoreML or TensorFlow Lite export to ARM hardware); (5) 5 split CV of the ablation variants and Vanilla Transformer baseline, which were only evaluated in the single split analysis here; and (6) extension to multi-horizon probabilistic forecasting and integration with closed loop insulin delivery systems for anticipatory basal rate modulation.

VII. CONCLUSION

We presented an attention-augmented Temporal Fusion Transformer for short-horizon hypoglycemia prediction in Type 1 Diabetes, trained and evaluated on the T1DiabetesGranada longitudinal multimodal dataset. Across 5 split subject level cross validation, the TFT achieved an AUC of 0.921 ± 0.005 , significantly outperforming a glucose only LSTM baseline (AUC 0.853 ± 0.007 ; mean difference 0.068, 95 % CI [0.063, 0.073], $p = 3.4 \times 10^{-6}$). The advantage was completely consistent, the lowest TFT split AUC (0.917) exceeded the highest LSTM split AUC (0.863) across all 5 splits, and accompanied by a mean 36% reduction in false alarm rate and a 7.4 percentage point improvement in sensitivity. Detailed single split analysis showed a consistent AUC of 0.923 alongside strong F1 (0.61) on 199,432 held out test windows from 147 unseen patients. GPU inference latency of 6.3 ms per prediction means feasibility for real time use, pending on device validation. An ablation study revealed that the TCN block and static covariate pathway contributed little additional predictive power in this configuration. Attention analysis showed that the model relies most heavily on the most recent 5-10 minutes of glucose history, with this reliance amplified specifically for windows preceding hypoglycemia. Taken together, these results support the broader premise that multimodal, attention-based sequence models can offer more sensitive and lower latency hypoglycemia risk estimation than glucose only or simple rule based alternatives, while stressing on the need for cross validated, prospective, and on device follow up work before clinical deployment.

REFERENCES

- [1] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, “Type 1 diabetes,” *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.
- [2] International Diabetes Federation, “IDF Diabetes Atlas, 11th edition,” <https://diabetesatlas.org/>, 2025, reports an estimated 9.1 million adults living with type 1 diabetes worldwide in 2024.

- [3] P. E. Cryer, “Hypoglycemia in type 1 diabetes mellitus,” *Endocrinology and Metabolism Clinics of North America*, vol. 39, no. 3, pp. 641–654, 2010.
- [4] G. McGarraugh, “Alarm characterization for a continuous glucose monitor that replaces traditional blood glucose monitoring,” *Journal of Diabetes Science and Technology*, vol. 4, no. 1, pp. 49–56, 2010, reports a hypoglycemia (70 mg/dL) detection rate of 70.3% with an 11.4% false-alarm rate for a commercial CGM threshold alarm, illustrating that even optimized threshold systems trade detection for false alarms.
- [5] S. Giza, E. P. Kotanidou, V. R. Tsinopoulou, T. Poukoulidou, S. Ntouma, and A. Galli-Tsinopoulou, “Can glucose alarm fatigue threaten the absolute clinical benefit of continuous glucose monitoring in optimal glucose management in children and adolescents with type 1 diabetes? a narrative review,” *Children*, vol. 12, no. 12, p. 1668, 2025.
- [6] H. N. Mhaskar, S. V. Pereverzyev, and M. D. van der Walt, “A deep learning approach to diabetic blood glucose prediction,” *Frontiers in Applied Mathematics and Statistics*, vol. 3, p. 14, 2017.
- [7] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, “Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network,” in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*, 2018, pp. 1–5.
- [8] A. Ma and S. Chennareddy, “Attention-augmented temporal fusion transformer for short-horizon hypoglycemia prediction in type 1 diabetes,” *Endocrine Practice*, vol. 32, no. 4, p. S336, 2026, doi: 10.1016/j.eprac.2026.03.018.
- [9] C. Rodriguez-Leon, M. D. Aviles-Perez, O. Banos, M. Quesada-Charneco, P. J. Lopez-Ibarra, C. Villalonga, and M. Munoz-Torres, “T1DiabetesGranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus,” *Scientific Data*, vol. 10, no. 1, p. 916, 2023.
- [10] S. M. Pappada, B. D. Cameron, P. M. Rosman, R. E. Bourey, T. J. Papadimos, W. Olorunto, and M. J. Borst, “Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes,” *Diabetes Technology & Therapeutics*, vol. 13, no. 2, pp. 135–141, 2011.
- [11] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, “Convolutional recurrent neural networks for glucose prediction,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 603–613, 2020, reports an on-phone execution time of 6 ms for their model, useful context for our own GPU/CPU inference latency comparison.
- [12] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, “Deep multi-output forecasting: learning to accurately predict blood glucose trajectories,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*, 2018, pp. 1387–1395.
- [13] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep learning for diabetes: a systematic review,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2021.
- [14] C. M. Ramkissoon, P. Herrero, J. Bondia, and J. Vehi, “Unannounced meals in the artificial pancreas: detection using continuous glucose monitoring,” *Sensors*, vol. 18, no. 3, p. 884, 2018.
- [15] B. Kovatchev, “Metrics for glycaemic control—from HbA1c to continuous glucose monitoring,” *Nature Reviews Endocrinology*, vol. 13, no. 7, pp. 425–436, 2017.
- [16] B. Pfitzner, N. Steckhan, and B. Arnrich, “Federated learning in a medical context: a systematic literature review,” *ACM Transactions on Internet Technology*, vol. 21, no. 2, pp. 1–31, 2021.
- [17] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, “Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8857–8868.
- [18] Y. Zhang and J. Yan, “Crossformer: transformer utilizing cross-dimension dependency for multivariate time series forecasting,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [19] B. Lim, S. Ö. Ank, N. Loeff, and T. Pfister, “Temporal fusion transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.
- [20] American Diabetes Association, “Standards of care in diabetes,” 2025, published annually; cite the edition current at the time of submission (check care.diabetesjournals.org for the latest year).
- [21] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv preprint arXiv:1803.01271*, 2018.
- [22] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [23] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyperparameter optimization,” in *Advances in Neural Information Processing Systems*, vol. 24, 2011.
- [24] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics*, pp. 837–845, 1988.

APPENDIX

Code for data preprocessing, model implementation, training, and evaluation is available [here](#). All reported numbers in this paper were generated by a single end to end run of this pipeline on Kaggle (NVIDIA T4 GPU); exact results may vary slightly on rerun due to GPU non determinism and the random per-patient window subsampling, but should be somewhat similar given the fixed random seed used (seed = 42).

Data availability and usage restrictions. The T1DiabetesGranada dataset is available under restricted access via Zenodo [9]; access requires a Data Usage Agreement with the Universidad de Granada that explicitly prohibits redistributing the raw dataset. Accordingly, our public code release contains only preprocessing and modeling code, not any dataset files; researchers wishing to reproduce these results must independently request dataset access via Zenodo and agree to the same Data Usage Agreement.