Transfer Learning for Pediatric Glucose Forecasting

Alain Ryser[*1] ALAIN.RYSER@INF.ETHZ.CH
Chuhao Feng[*1]
Tobias Scheithauer[1]
Marc Pfister[4,5]
Marie-Anne Burckhardt[5,6]
Sara Bachmann[5,6]
Alexander Marx[†1,2,3]
Julia E. Vogt[†1,7]
[1] *Department of Computer Science, ETH Zurich, Zürich, Switzerland*
[2] *Department of Statistics, Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, TU Dortmund University, Dortmund, Germany*
[3] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[4] *Pediatric Pharmacology and Pharmacometrics, University Children's Hospital Basel, Basel, Switzerland*
[5] *Department of Clinical Research, University of Basel, Basel, Switzerland*
[6] *Pediatric Endocrinology and Diabetology, University Children's Hospital Basel, Basel, Switzerland*
[7] *SIB Swiss Institute of Bioinformatics, Ecublens, Switzerland*

Abstract

Effective blood glucose forecasting is crucial for detecting events such as hypo- or hyperglycemia in people with diabetes, yet remains challenging in domains with only small, heterogeneous datasets, such as in the pediatric field. We present GluTFT, a novel transfer learning approach that allows leveraging models pretrained on publicly available adult diabetes datasets for pediatric glucose forecasting. We systematically evaluate multiple transfer learning strategies, including zero-shot prediction and fine-tuning across the entire dataset as well as specific subgroups of participants. Our extensive experiments reveal that GluTFT excels on the pretraining datasets and significantly outperforms baseline methods when fine-tuned. To validate the clinical relevance of our approach, we evaluate Parkes Error Grids, demonstrating the quality of GluTFT's blood glucose forecasts and its potential for enhancing clinical decision-making for pediatric diabetes.

Keywords: Multi-Horizon Forecasting, Transfer Learning, Transformer, Type 1 Diabetes, Time Series

Data and Code Availability We make use of the publicly available dataset CITY (Laffel et al., 2020). The OhioT1DM dataset (Marling and Bunescu, 2020) can be obtained for research purposes by following the instructions described by Marling and Bunescu (2020). Further, we use the DiaCamp dataset (Marx et al., 2023), for which access is restricted. Access requests should be directed to the authors of the dataset as stated in their manuscript. We provide the code for our experiments in the supplementary material and will make it publicly available upon acceptance.

Institutional Review Board (IRB) This work does not require IRB approval.

1. Introduction

Type 1 diabetes mellitus (T1D) is a chronic endocrine disorder characterized by the body's inability to produce insulin, a hormone crucial for regulating blood glucose levels (Association, 2014). If left untreated, this condition leads to elevated blood glucose (hyperglycemia). Typically, people with diabetes are treated with exogenous insulin, which helps to maintain glucose levels in a normal range. However, administering such treatment often leads to events of low blood glucose levels (hypoglycemia). Both hyperglycemic and hypoglycemic events lead to undesirable symptoms. In particular, unrecognized nocturnal hy-

[*] Equal contribution
[†] Shared last

poglycemia in children with T1D (Bachmann et al., 2016) poses a threat, causing severe damage such as loss of consciousness, seizures, or in extreme cases even sudden cardiac death (Tu et al., 2008; Abraham et al., 2022). Hence, proper management of insulin therapy and, ideally, prediction of such future events (Leutheuser et al., 2024), is essential.

Recently, the publication of diabetes datasets, such as OhioT1DM (Marling and Bunescu, 2020), which was released in the context of the blood glucose level prediction (BGLP) challenge (Bach et al., 2020), or CITY (Laffel et al., 2020), has drawn a lot of attention to forecasting blood glucose levels from continuous blood glucose measurements (CGM). Originally, some of the most successful deep methods for BGLP were N-BEATS (Oreshkin et al., 2020), which won the BGLP challenge, and its extension N-HITS (Challu et al., 2023). Recently, the success of autoregressive models, especially transformers (Vaswani et al., 2017), has led to more accurate forecasting of time series data in various fields (Li et al., 2021; Zhou et al., 2021; Ramos-Pérez et al., 2021). The development of the Temporal Fusion Transformer (TFT, Lim et al., 2021) is of particular interest here due to the transformer architecture's potential for transfer learning (Gruetzemacher and Paradice, 2022) and TFT's ability to perform time series forecasting with different data types, allowing us to leverage additional information, e.g., about food intake or insulin treatment (Marx et al., 2023; Zhu et al., 2023a). Apart from TFT, a wide range of deep-learning methods have been applied to the BGLP setting, which we review in Section 2.

While we can successfully leverage the properties of advanced machine-learning methods for larger datasets such as the OhioT1DM and CITY datasets, these models may not be applicable for smaller studies that investigate sensitive populations or consider non-standard study settings. Marx et al. (2023), for instance, investigate a dataset (DiaCamp) containing samples of children with T1D. Participants took part in a summer sports camp where they conducted daily physical activities, which is known to directly affect the blood glucose levels (Romijn et al., 1993; Riddell et al., 2017) leading to an increased complexity in the prediction problem. Due to the limited data availability, transformer models could not outperform vanilla recurrent neural networks in this setting. To leverage these powerful models for BGLP in smaller studies, we thus propose a transfer learning approach.

Contributions In this work, we propose GluTFT, a transfer learning approach (cf. Figure 1) for forecasting blood glucose levels in small-scale, heterogeneous datasets. In particular, we demonstrate how to pretrain large models on publicly available datasets such as OhioT1DM and CITY and fine-tune them on the DiaCamp dataset. Our model based on the TFT architecture simultaneously predicts the blood glucose level on three forecasting horizons based on past observed and plannable future variables (activity, food intake, and insulin injections). By leveraging the multi-horizon quantile loss (Wen et al., 2017), our predictions further produce confidence intervals, making predictions more useful in clinical practice. In summary, our contributions are the following:

- To the best of our knowledge, we are the first to investigate transfer learning for pediatric blood glucose forecasting, introducing a feature alignment and fine-tuning strategy.
- Our model exhibits competitive performance on the challenging OhioT1DM and CITY datasets.
- We conduct extensive experiments, investigating transfer learning approaches using different datasets, zero-shot prediction, and various fine-tuning approaches on groups of participants for blood glucose forecasting on the DiaCamp dataset.
- A qualitative evaluation of our approach through Parkes Error Grids, presenting the applicability of our approach to clinical practice.

2. Related Work

Research on predicting blood glucose levels can be divided into two main approaches: mathematical modelling (Visentin et al., 2018; Deichmann and Kaltenbach, 2023) and data-driven approaches, which are the focus of this paper.

The release of the OhioT1DM dataset (Marling and Bunescu, 2020) has triggered extensive studies on blood glucose forecasting by the machine learning community. Various classical shallow machine learning methods have been investigated and evaluated on the OhioT1DM dataset, including ridge regression, SVMs, random forests, XGBoost, autoregressive models, GPs, and unscented Kalman filter (McShinsky and Marshall, 2020; De Bois et al., 2022), as well as ensemble methods (Saiti et al., 2020). Subsequently, deep learning models have been developed and applied. Li et al. (2020) utilized a convolutional neural network (CNN) as the feature extractor in a

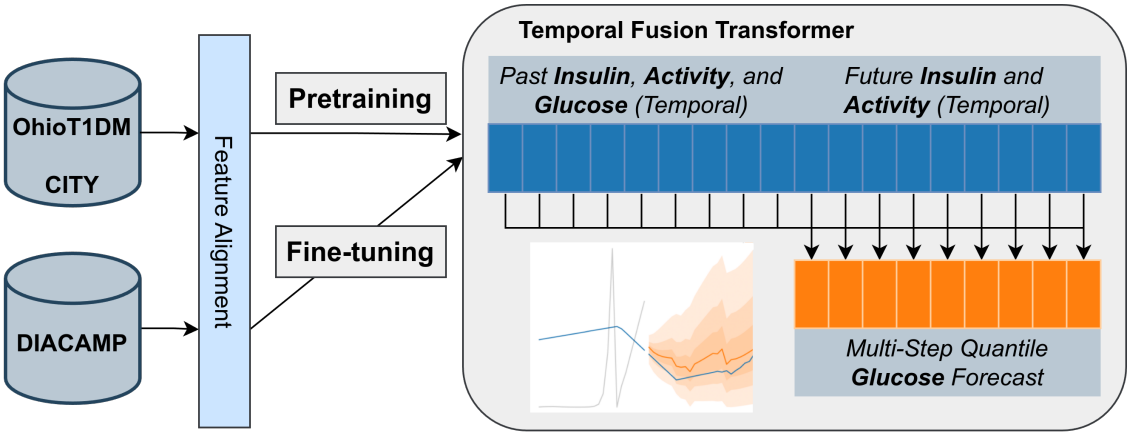


Figure 1: Overview of our GluTFT approach: First, we align the features of the OhioT1DM and CITY datasets to be compatible with DiaCamp. Then we pretrain the Temporal Fusion Transformer with i) past temporal features like insulin intake, activity like sports or meals, and glucose (over a two-hour time window) and ii) plannable future features such as insulin injections or activities. We minimize the quantile loss for multi-step ahead prediction (2 hours). Last, we fine-tune our model on DiaCamp and produce accurate multi-step quantile glucose forecasts at inference time (illustration in the middle), where blue represents ground-truth glucose measurements, orange corresponds to quantile predictions, and the gray line illustrates the temporal importance weights of the time-series given by the attention scores of the model.

deep-learning model to achieve leading accuracy in glucose prediction and implemented their algorithm on an Android mobile phone, inspiring further studies with CNNs (Zhu et al., 2023b; Jaloli et al., 2022). Zhu et al. (2020b) exploited a generative adversarial network (GAN) to predict future glucose levels based on past continuous glucose measurements, meal ingestion, and insulin injections. Rabby et al. (2021) proposed a novel glucose prediction approach that combines stacked long short-term memory (LSTM) and deep recurrent neural network (RNN) and corrects CGM sensor error with the Kalman smoothing technique. Shuvo and Islam (2023) also utilized stacked LSTMs in their method to form shared hidden layers that learn generalized features from all participants. In contrast, Kalita and Mirza (2022) introduced a multi-layer deep neural framework with gated recurrent units (GRU) to predict the future glucose level. Further, Zhu et al. (2020a) applied the idea of dilation (Yu and Koltun, 2016) in their RNN to construct a dilated RNN that gains a much larger receptive field in terms of neurons aiming at capturing long-term dependencies. Yu et al. (2021) explored deep transfer learning for forecasting blood glucose in Type 2 Diabetes patients using a combination of convolutional and recurrent layers.

More recently, research building upon transformer architectures, e.g., to quantify uncertainty (Sergazinov et al., 2023), or the Temporal Fusion Transformer (TFT) architecture (Lim et al., 2021) have shown promising results. While Zhu et al. (2023a) evaluate TFT on the OhioT1DM dataset and demonstrate that it can be executed on an edge device, Marx et al. (2023) utilize the capabilities of TFT to model both static and temporal information and benchmark its performance on the DiaCamp dataset, a small dataset including data from children with T1D. For adult data, the concept of pretraining has been applied by Rubin-Falcone et al. (2020), the winning entry of the blood glucose forecasting challenge on the OhioT1D dataset, who pretrain a slightly modified version of the N-BEATS architecture (Oreshkin et al., 2020) on the proprietary Tidepool dataset (Neinstein et al., 2016).

In contrast to the above works, we investigate TFT in a transfer learning setting. In particular, we pretrain TFT on publicly available adult T1D datasets OhioT1DM (Marling and Bunescu, 2020) and CITY (Laffel et al., 2020) and evaluate transfer learning strategies for the DiaCamp dataset, comparing training from scratch with both zero-shot learning and fine-tuning. Further, we allow for the incorporation of planned future insulin dos-

ing, meal intake, and physical exercise and demonstrate the feasibility of our approach using Parkes Error Grids. Finally, we compare the results for different participant groups based on insulin delivery method and participant-specific fine-tuning in Appendices C.3 and C.4.

3. Methods & Datasets

In the following, we provide an overview of the datasets we analyze in this paper in Section 3.1 and introduce the GluTFT in Section 3.2.

3.1. Datasets

Here, we provide details about the three datasets CITY, OhioT1DM, and DiaCamp. Table 1 contains an overview of the available features for each dataset that are relevant to this study. We provide more details about each feature, sample counts, and dataset metadata in Appendix A.

CITY Laffel et al. (2020) collected the CITY dataset to investigate whether continuous glucose monitoring improves glycemic control compared to standard blood glucose monitoring in adolescents and young adults. The dataset contains CGM data from 153 participants with T1D aged 14 to 24 years. The study participants have been diagnosed with diabetes for at least one year and use either an insulin pump or multiple daily insulin injections to treat their diabetes. The CITY dataset contains CGM measurements recorded over two weeks for each participant.

OhioT1DM This dataset was introduced by Marling and Bunescu (2020) to promote and facilitate research in blood glucose level prediction. The OhioT1DM dataset is the first publicly available dataset containing CGM, insulin, physiological sensor, and self-reported life-event data, such as physical exercise or meal intake, for 12 adult participants with T1D over eight weeks. All participants in the OhioT1DM dataset use an insulin pump for diabetes treatment. In particular, our paper uses temporal data about glucose levels, insulin injections, carbohydrate intake, and exercise.

DiaCamp The DiaCamp dataset was previously used to investigate whether the success of deep-learning algorithms on adult data transfers to children (Marx et al., 2023). The dataset contains data of 14 children aged 7 to 16 who were diagnosed with T1D for at least six months before the study. Participants take insulin treatment consisting of multiple daily injections (MDI) or continuous subcutaneous insulin infusion (CSII). Each participant attended a one-week day camp, where clinicians controlled activities, insulin treatment, and nutrition and recorded insulin doses, carbohydrate intake, physical activities, and symptoms of hypoglycemia. DiaCamp contains similar temporal features as the OhioT1DM dataset, potentially allowing us to pretrain on OhioT1DM and fine-tune on the smaller DiaCamp dataset. However, temporal features like insulin are not properly aligned. For instance, while OhioT1DM contains *basal* (continuous injections) and *bolus* (single doses, e.g., before meals) insulin features, DiaCamp records which type of insulin (*fast* or *slow* acting) has been injected. In addition, the OhioT1DM dataset contains data from an adult population, whereas participants in DiaCamp are all children, making the dataset much more heterogeneous. GluTFT allows us to align the two feature sets in both datasets to successfully perform transfer learning from OhioT1DM to DiaCamp (see Section 3.2.2).

Table 1: Available Features for each dataset. OhioT1DM contains different temporal features than DiaCamp. We develop a feature alignment strategy to pretrain on OhioT1DM and fine-tune on DiaCamp.

	CITY	OhioT1DM	DiaCamp
timestamp	✓	✓	✓
glucose	✓	✓	✓
basal/bolus insulin	-	✓	-
fast/slow insulin	-	-	✓
meal (calories and types)	-	✓	-
food intake (slow/fast/mixed)	-	-	✓
sport	-	✓	✓

3.2. Methods

This section introduces GluTFT and describes our feature-alignment and fine-tuning strategy.

3.2.1. Glucose Temporal Fusion Transformer

We introduce the glucose temporal fusion transformer (GluTFT). GluTFT builds on the temporal fusion transformer (TFT), an attention-based deep neural network architecture for multi-horizon forecasting,

including an interpretability mechanism to analyze temporal feature importance (Lim et al., 2021).

TFT is an encoder-decoder transformer that can take three types of inputs: static, temporal past, and known temporal future inputs. As GluTFT does not rely on static features, we ignore them for the remainder of this section. To extract representations from past and future temporal inputs, TFT introduces a so-called variable selection network (VSN), removing unnecessary noisy inputs and indicating feature importance through the weights of the VSN. The TFT uses separate VSNs for each type of input. In addition, TFT introduces the interpretable multi-head attention layer, which uses a modified version of the self-attention of transformer-based architectures (Li et al., 2019; Vaswani et al., 2017). The interpretable multi-head attention layer provides explainability through the attention weights and allows us to learn long-ranged temporal relationships across different time steps (see Appendix C.1). For more details on the specific architecture of TFT, see Lim et al. (2021).

GluTFT uses the multi-horizon quantile loss (Wen et al., 2017) for training. Given the past and future temporal features for a timepoint t, GluTFT learns to jointly estimate future timepoints up to τ_{max} points into the future. We call this time range the prediction horizon. Further, instead of point predictions, our model produces quantile forecasts $\hat{X}_q^{(t)} := (\hat{\mathbf{x}}_q^{(t)}, \ldots, \hat{\mathbf{x}}_q^{(t+\tau_{max})})$, where $\hat{\mathbf{x}}_q^{(t)}$ is a prediction for quantile $q \in \mathcal{Q}$ at timestep t. To learn these quantiles for a given ground-truth horizon $\mathcal{X}_{t:} := (\mathbf{x}^{(t)}, \ldots, \mathbf{x}^{(t+\tau_{max})})$, we jointly minimize the quantile loss summed across all quantile outputs $\hat{X} := \{\hat{X}_q^{(t)}\}_{q\in\mathcal{Q}}$ as follows:

$$\mathcal{L}(\mathcal{X}_{t:}, \hat{X}^{(t)}) = \sum_{q\in\mathcal{Q}} \sum_{\tau=1}^{\tau_{max}} QL\left(\mathbf{x}^{(t+\tau)}, \hat{\mathbf{x}}_q^{(t+\tau)}, q\right)$$

$$QL(\mathbf{x},\ \hat{\mathbf{x}},\ q) = q(\mathbf{x} - \hat{\mathbf{x}})_+ + (1-q)(\hat{\mathbf{x}} - \mathbf{x})_+$$

$$(x)_+ = \max(0, x)$$

For blood glucose forecasting, GluTFT leverages the past temporal features on *meal intake*, *physical activity*, *insulin injection*, and *glucose level*. Further, as many features correspond with plannable occurrences, GluTFT includes future inputs of *meal intake*, *physical activity*, and *insulin injection* to produce more accurate blood glucose predictions.

3.2.2. Alignment & Fine-tuning

In the following, we summarize the feature alignment strategy of GluTFT to perform transfer learning for blood glucose forecasting. We pretrain GluTFT separately on the CITY dataset with only glucose data and on the OhioT1DM dataset and then fine-tune GluTFT on the DiaCamp dataset.

In contrast to prior works, we propose the usage of *plannable* future features in the forecasting problem. Typically, T1D participants can decide and plan when to inject insulin, have meals, and do sports. We thus include these features as known temporal future inputs in GluTFT.

To effectively leverage all available data during transfer learning, we have to align the temporal features of OhioT1DM and DiaCamp during pretraining and fine-tuning. For CSII participants in DiaCamp, instead of basal insulin, the dataset contains a feature for fast-acting insulin, recording hourly insulin doses with occasional spikes before meals. We map the hourly fast insulin feature to the basal insulin injection rate to align with the hourly basal insulin injection rate in OhioT1DM. On the other hand, we infer the bolus insulin for CSII participants from the spikes in fast insulin before meals. Since MDI participants do not have an insulin pump, they get a high dosage of slow-acting insulin twice a day. To infer the hourly basal insulin injection rate, we thus divide the amount of injected slow insulin by the number of hours until the next slow insulin injection. We get bolus insulin for MDI participants directly from the recorded fast insulin doses. We present an example of our insulin feature alignment strategy in Appendix A.

While OhioT1DM contains more detailed information about the type of food and carbohydrate intake, the DiaCamp contains only information on carbohydrates classified into the three broad categories: *fast*, *mixed*, and *slow* food, signaling how fast the respective food intake raises blood glucose levels. To properly align these features, we redefine OhioT1DM's meal-type *HypoCorrection* to the category *fast*. We set the rest of OhioT1DM's categories to the new category *mixed/slow* food, which combines DiaCamp's *mixed* and *slow* food categories. The physical activity features are already aligned, as both datasets record physical exercise in minutes.

Table 2: Results on OhioT1DM, where we report the mean ± standard deviation across participants for 30, 60, and 120 minute prediction horizons. We take baselines from the literature, which report single-seed results with standard deviations across participants. In contrast, we run each experiment with 5 seeds and average each seed's mean and standard deviations across participants for comparability. We report the RMSE in *mmol/l*, bold the **best**, and underline the second best results. Metrics that have not been reported by baselines imported from the literature are marked with a "—". Except for $RMSE_{30}$, GluTFT is the best or second best model in all metrics/horizons.

Model	$RMSE_{30}$	$RMSE_{60}$	$RMSE_{120}$	$MAPE_{30}$	$MAPE_{60}$	$MAPE_{120}$
Identity (De Bois et al., 2022)	1.57 ± 0.13	2.29 ± 0.16	3.20 ± 0.20	13.51 ± 2.27	20.37 ± 3.87	30.16 ± 4.94
LSTM (De Bois et al., 2022)	1.14 ± 0.12	1.83 ± 0.15	2.64 ± 0.17	9.24 ± 2.19	16.99 ± 3.57	24.69 ± 5.34
SVR (De Bois et al., 2022)	1.12 ± 0.13	1.79 ± 0.13	2.61 ± 0.15	9.08 ± 2.12	15.38 ± 3.42	23.45 ± 4.96
GP (De Bois et al., 2022)	1.11 ± 0.13	1.78 ± 0.14	2.57 ± 0.16	9.16 ± 2.16	15.92 ± 3.80	24.65 ± 5.69
FCNN (Zhu et al., 2023b)	1.04 ± 0.14	1.73 ± 0.20	—	—	—	—
TFT_{Edge} (Zhu et al., 2023a)	1.06 ± 0.14	1.80 ± 0.21	—	—	—	—
MTL-LSTM (Shuvo and Islam, 2023)	$\mathbf{0.89 \pm 0.15}$	1.72 ± 0.24	$\mathbf{2.25 \pm 0.29}$	—	—	—
N-BEATS	0.98 ± 0.14	1.69 ± 0.23	2.5 ± 0.32	$\mathbf{8.17 \pm 1.85}$	15.03 ± 3.49	24.34 ± 5.55
N-HiTS	1.20 ± 0.46	1.81 ± 0.51	2.55 ± 0.48	9.25 ± 1.64	14.85 ± 2.57	22.47 ± 3.66
GluTFT	1.01 ± 0.16	$\mathbf{1.65 \pm 0.22}$	2.4 ± 0.30	8.39 ± 1.67	$\mathbf{14.33 \pm 2.69}$	$\mathbf{21.81 \pm 3.67}$

4. Experiments

In the following, we introduce our baselines (Section 4.1) and empirically evaluate GluTFT by comparing our new feature aligned GluTFT against the state-of-the-art (SOTA) on the OhioT1DM dataset (Section 4.2), demonstrating how different pretraining strategies influence the performance when finetuning on the DiaCamp dataset (Section 4.3) and presenting qualitative results in the form of Parkes Error Grids to demonstrate the practical applicability of our method (Section 4.4).

4.1. Baselines

The main baselines for our experiments are **N-BEATS** and **N-HiTS**. N-BEATS (Oreshkin et al., 2020) is a deep neural network architecture focusing on univariate time-series forecasting. This baseline can only perform univariate forecasting, we thus train it on glucose data only. The N-HiTS (Challu et al., 2023) model extends N-BEATS by introducing multi-rate data sampling and multi-scale interpolation, gaining significant accuracy improvements and reducing computational complexity. N-HiTS and GluTFT, by default, support multivariate time series, we thus use features as described in Section 3.2.2. Some experiments contain the univariate versions N-$HiTS_{gluc}$ and $GluTFT_{gluc}$, trained on glucose data only, as a reference. In addition, we consider the following baselines for the OhioT1DM dataset. The **Identity** baseline uses the last observed glucose value as a constant prediction across the horizon. Further, we consider classical machine learning methods such as a Gaussian process (**GP**), support vector regression (**SVR**), and an **LSTM** only trained on glucose data (De Bois et al., 2022). Finally, we also consider the more recent methods **FCNN** (Zhu et al., 2023b), an attention-based recurrent neural network model, $\mathbf{TFT_{Edge}}$ (Zhu et al., 2023a), an instantiation of TFT on an edge device without feature alignment, and **MLT-LSTM** (Shuvo and Islam, 2023), a stacked LSTM architecture consisting of a joint network and individual neural network heads for each participant.

4.2. Evaluation on OhioT1DM

Before we show the capabilities of GluTFT on the transfer learning task from adult datasets to DiaCamp, we compare our proposed feature alignment and training strategy to the state-of-the-art on the OhioT1DM dataset (Marling and Bunescu, 2020). We follow the standard pre-processing pipeline and the pre-defined train and test splits to align our results to prior work. We train and evaluate GluTFT, N-BEATS, and N-HITS (with our aligned features) for 5 independent seeds and compare them to additional SOTA results obtained from the literature. For each seed, we compute the mean and standard deviation of the corresponding metric of the results per participant and, for consistency with baselines from the literature, report the mean of both statistics across all seeds in Table 2.

Table 3: Results on DiaCamp for different pretraining strategies and baselines, where we report the mean ± standard deviation across participants for 30, 60, and 120 minute prediction horizons over 5 seeds. *Random* means the model is initialized with random weights, and *zero-shot* means we evaluate directly after pretraining. We report the RMSE in (mmol/l), bold the **best**, and underline the second best result. Except for MAPE_{120}, GluTFT pretrained on OhioT1DM or CITY is the best and second best performing model for all metrics/horizons.

Initialization	Model	RMSE_{30}	RMSE_{60}	RMSE_{120}	MAPE_{30}	MAPE_{60}	MAPE_{120}
	Identity	1.96	2.81	3.74	20.22	28.58	40.3
Random	N-BEATS	1.73 ± 0.01	2.43 ± 0.04	2.97 ± 0.03	17.84 ± 0.41	24.82 ± 0.4	33.76 ± 0.73
	$\text{N-HiTS}_{\text{gluc}}$	1.77 ± 0.01	2.49 ± 0.02	3.05 ± 0.02	18.03 ± 0.26	25.09 ± 0.39	32.6 ± 0.41
	N-HiTS	1.75 ± 0.01	2.38 ± 0.01	2.97 ± 0.03	18.26 ± 0.23	24.74 ± 0.31	32.42 ± 0.36
	$\text{GluTFT}_{\text{gluc}}$	1.75 ± 0.05	2.47 ± 0.03	3.07 ± 0.02	17.5 ± 0.49	24.67 ± 0.14	32.65 ± 0.22
	GluTFT	1.69 ± 0.02	2.31 ± 0.04	2.87 ± 0.08	17.28 ± 0.25	23.49 ± 0.6	**30.17 ± 0.76**
CITY (0-shot)	N-BEATS	2.48	2.74	3.35	21.29	27.26	36.88
	$\text{N-HiTS}_{\text{gluc}}$	1.9	2.92	4.17	19.17	27.75	41.18
	N-HiTS	1.78	2.6	3.18	18.88	26.47	34.57
	$\text{GluTFT}_{\text{gluc}}$	1.76	2.57	3.17	18.51	26.03	34.43
	GluTFT	1.79	2.57	3.14	18.86	26.44	34.29
OhioT1DM (0-shot)	N-BEATS	1.77	2.52	3.05	18.5	26.7	36.37
	$\text{N-HiTS}_{\text{gluc}}$	1.77	2.53	3.09	18.0	26.32	34.51
	N-HiTS	1.87	2.58	3.15	19.67	27.01	35.4
	$\text{GluTFT}_{\text{gluc}}$	1.69	2.47	3.04	17.73	25.48	33.54
	GluTFT	1.87	2.63	3.22	19.72	27.75	35.58
CITY	N-BEATS	1.65 ± 0.03	2.36 ± 0.03	2.96 ± 0.03	17.29 ± 0.21	24.78 ± 0.3	33.46 ± 0.39
	$\text{N-HiTS}_{\text{gluc}}$	1.76 ± 0.03	2.49 ± 0.02	3.05 ± 0.02	17.41 ± 0.07	24.69 ± 0.09	32.86 ± 0.14
	N-HiTS	1.81 ± 0.02	2.46 ± 0.02	3.07 ± 0.02	18.7 ± 0.22	25.24 ± 0.2	33.19 ± 0.56
	$\text{GluTFT}_{\text{gluc}}$	1.60 ± 0.00	2.32 ± 0.0	2.98 ± 0.0	16.68 ± 0.03	23.85 ± 0.05	32.28 ± 0.05
	GluTFT	1.60 ± 0.00	2.29 ± 0.0	**2.84 ± 0.01**	16.85 ± 0.1	23.46 ± 0.05	30.64 ± 0.1
OhioT1DM	N-BEATS	1.63 ± 0.01	2.34 ± 0.02	2.94 ± 0.02	17.23 ± 0.21	24.57 ± 0.19	33.26 ± 0.35
	$\text{N-HiTS}_{\text{gluc}}$	1.69 ± 0.01	2.44 ± 0.01	2.98 ± 0.01	17.04 ± 0.04	24.77 ± 0.17	32.61 ± 0.28
	N-HiTS	1.71 ± 0.01	2.31 ± 0.02	2.87 ± 0.03	17.78 ± 0.17	24.37 ± 0.25	31.58 ± 0.34
	$\text{GluTFT}_{\text{gluc}}$	1.61 ± 0.00	2.36 ± 0.00	2.94 ± 0.0	16.38 ± 0.02	23.64 ± 0.07	31.39 ± 0.07
	GluTFT	**1.59 ± 0.00**	**2.24 ± 0.01**	2.85 ± 0.01	**16.29 ± 0.05**	**23.05 ± 0.07**	30.54 ± 0.18

We observe that GluTFT ranks second, outperforming all baselines except MLT-LSTM, which, however, utilizes participant-specific neural network heads, which we cannot reuse for transfer learning. In particular, it is encouraging that GluTFT outperforms TFT_{Edge} for both time horizons considered by (Zhu et al., 2023a). Based on these results, we conclude that GluTFT is best suited for the transfer learning setting.

To further support these results, we provide a similar analysis for GluTFT, N-BEATS, and N-HITS on the CITY dataset in Appendix C.2.

4.3. Pediatric Glucose Forecasting

In this experiment, we demonstrate the capabilities of our transfer learning approach, comparing different baselines and model initialization strategies on DiaCamp. We keep the trivial *Identity* baseline and further compare N-BEATS, $\text{N-HiTS}_{\text{gluc}}$, N-HiTS, $\text{GluTFT}_{\text{gluc}}$, and GluTFT when training from scratch (*Random* Initialization), pretraining on CITY or OhioT1DM and performing *zero-shot* forecasting, and *fine-tuning* with these pretrained models on the DiaCamp dataset before evaluating on the test set. We present the results of this experiment in Table 3. Clearly, models pretrained on the OhioT1DM dataset perform much better than training them from scratch. On the other hand, our baselines do not seem to benefit much from pretraining on the CITY dataset. In contrast, our GluTFT exhibits superior performance over training from scratch, demonstrating that our feature alignment approach works across datasets. Interestingly, zero-shot predictions with GluTFT are very close to models trained from scratch for the 30 minute horizon, whereas 60 and 120 minutes still lack behind, indicating that larger pretraining datasets could potentially generalize even to smaller, more heterogeneous diabetes subpopulations. In general, GluTFT demonstrates the strongest performance of all methods, independent of the initialization strategy. Hence, we conclude that with a proper feature alignment strategy, pretraining and fine-tuning a model can lead

Table 4: Percentage of predictions within each zone for Parkes error grids on DiaCamp. All models have been pretrained on OhioT1DM

Model	A	B	C	D	E
PH 30min					
N-BEATS	75.86%	20.90%	2.69%	0.52%	0.03%
N-HiTS	74.34%	22.15%	3.12%	0.37%	0.03%
GluTFT	78.31%	19.07%	2.29%	0.34%	0%
PH 60min					
N-BEATS	61.63%	31.32%	6.57%	0.49%	0%
N-HiTS	60.95%	33.06%	5.32%	0.67%	0%
GluTFT	62.69%	31.81%	5.04%	0.46%	0%
PH 120min					
N-BEATS	48.00%	42.22%	7.58%	2.20%	0%
N-HiTS	48.46%	42.56%	7.64%	1.28%	0.06%
GluTFT	49.77%	41.98%	7.00%	1.25%	0%

to superior performance in glucose forecasting, even for smaller, more specialized datasets.

We further investigate the effect of transfer learning to participant subgroups with (MDI/CSII) and participant-specific fine-tuning in Appendices C.3 and C.4.

4.4. Parkes Error Grids

While RMSE and MAPE provide a quantitative impression from a machine learning perspective, we also want to provide a more clinically motivated analysis of our results. In this section, we present Parkes Error Grids (PEG; Parkes et al., 2000) for our predictions, which are a standard evaluation measure for methods focusing on glucose forecasting. When plotting ground truth glucose levels against predictions, PEGs act as a supplement indicating each prediction's clinical quality. The grids classify prediction errors into five zones "A"-"E", where "A" signals a correct forecast, whereas "E" represents critical failure. That is, zone "E" specifies the case in which the model predicts hyperglycemia while the true blood glucose level will be very low (hypoglycemia). Thus an intervention of a doctor to prevent the falsely predicted hyperglycemia can have serious consequences. We provide the medical interpretation of each zone in Appendix C.5.

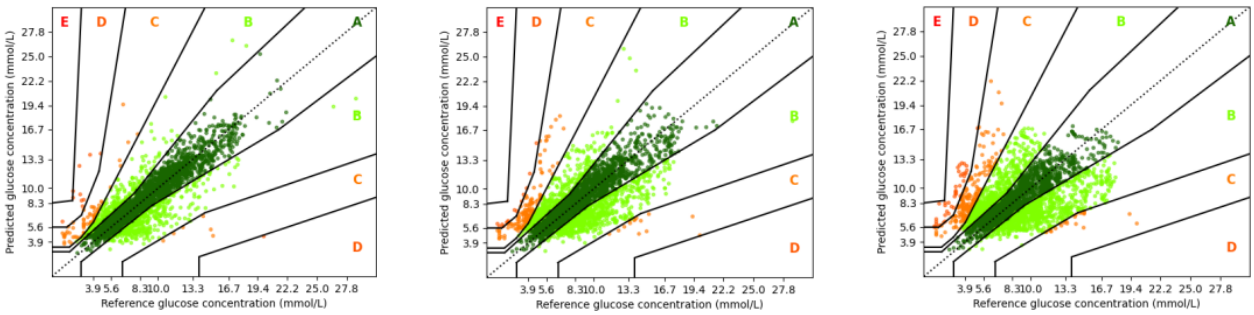
We create Parkes error grids for GluTFT, N-BEATS and N-HiTS pretrained on OhioT1DM and evaluated on DiaCamp. We present the Parkes plots for prediction horizons 30, 60, and 120 minutes for GluTFT in Figure 2 and supplement the corresponding plots for the baselines in Appendix C.5. In addition, we list the percentage of predictions categorized to zones "A" to "E" for all approaches in Table 4. For GluTFT, we can observe that 97.38% of all predictions are classified to zones "A" and "B" whereas only 2.29% are in zone "C," 0.34% in zone "D" and none are classified to zone "E" — for a prediction horizon of 30 minutes. For larger prediction horizons, the predictions become less accurate: for a prediction horizon of two hours, we observe 7% in zone "C" and 1.25% and 0% in zones "D" and "E," respectively. When inspecting the PEGs in Figure 2, we can see that GluTFT (similar to the baselines, cf. Appendix C.5) tends to overestimate the glucose levels rather than underestimate them. Overall, GluTFT performs best on these metrics whereas N-BEATS and N-HiTS, which are also pretrained on OhioT1DM, perform slightly worse.

5. Conclusion

In this study, we investigated whether transfer learning helps the problem of forecasting blood glucose levels in children. In extensive experiments, we first presented state-of-the-art performance of GluTFT on the OhioT1DM dataset. By fine-tuning our pretrained models, we found that large, publicly available data of adults with T1D helps us improve forecasting for blood glucose levels in the much more difficult population of children, even when only limited data is available. We further investigated different transfer learning approaches, demonstrating the potential of zero-shot predictions and the benefit of fine-tuning with our GluTFT. Finally, we presented evidence of the applicability of GluTFT in clinical practice by investigating the quality of our forecasts with Parkes Error Grids.

Limitations & Future Work Our current method allows for reliable prediction of short to mid-time horizon predictions of blood glucose; however, performance on longer time horizons could still be improved. Additionally, better alignment and incorporation of static features, along with access to larger pretraining datasets, could enhance the overall model performance.

In future work, we plan to investigate incorporating prior knowledge about T1D into our fine-tuning strategy by leveraging mathematical models (Berger and Rodbard, 1989; Kobayashi et al., 1983), allowing us to enrich the changes of the *basal* insulin for MDI,



(*a*) PH 30 (*b*) PH 60 (*c*) PH 120

Figure 2: Parkes Error Grids for GluTFT pretrained on OhioT1DM and fine-tuned and evaluated on DiaCamp for increasing prediction horizons (PH).

improving the alignment between CSII and MDI participants. Further, we want to investigate the effect of larger datasets, e.g., it could be beneficial to get access to datasets like T1DEXI (Riddell et al., 2023). Finally, we are interested in applying our models to build Hypo-/Hyperglycemia detectors, leveraging our forecasts of future blood glucose levels.

Ethics Statement Ethical concerns of our approach include potential misinterpretation by healthcare professionals, especially given its reduced reliability at longer prediction horizons, which could lead to suboptimal treatment decisions. This is particularly concerning when forecasting blood glucose in children, where prediction errors may have serious long-term health consequences. However, our model's interpretability mechanism and uncertainty estimates enhance transparency and help clinicians make more informed decisions.

Acknowledgments

We would like to thank Chuhao Feng for his dedicated work on this project as part of his master's thesis. We further thank the StimuLoop grant #1-007811-002 and the Vontobel Foundation for supporting the research of Alain Ryser. We thank the Schweizerische Diabetesstiftung (SDS) and Stiftung Pro UKBB for their support of the Diacamp Project and the Freiwillige Akadamische Gesellschaft Basel (FAG) for supporting Sara Bachmann. Marie-Anne Burckhardt was supported by a research fellowship provided by the Research Fonds (Special Program "Nachwuchsförderung Klinische Forschung") of the University of Basel and a Young Investigator Grant of the Swiss Society for Endocrinology and Diabetes (SSED). Alexander Marx was funded through ETH Zurich for part of the project.

References

Mary B Abraham, Beate Karges, Klemen Dovc, Diana Naranjo, Ana Maria Arbelaez, Joyce Mbogo, Ganesh Javelikar, Timothy W Jones, and Farid H Mahmud. Ispad clinical practice consensus guidelines 2022: Assessment and management of hypoglycemia in children and adolescents with diabetes. *Pediatric diabetes*, 23(8):1322, 2022.

American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 37 (Supplement_1):S81–S90, 2014.

Kerstin Bach, Razvan C. Bunescu, Cindy Marling, and Nirmalie Wiratunga, editors. *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, volume 2675 of *CEUR Workshop Proceedings*, 2020. CEUR-WS.org.

Sara Bachmann, Melanie Hess, Eva Martin-Diener, Kris Denhaerynck, and Urs Zumsteg. Nocturnal hypoglycemia and physical activity in children with diabetes: new insights by continuous glucose monitoring and accelerometry. *Diabetes Care*, 39(7): e95–e96, 2016.

Markus Berger and David Rodbard. Computer Simulation of Plasma Insulin and Glucose Dynamics After Subcutaneous Insulin Injection. *Diabetes Care*,

12(10):725–736, November 1989. ISSN 0149-5992, 1935-5548. doi: 10.2337/diacare.12.10.725.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i6.25854.

Stephen Colagiuri. Glycated haemoglobin (hba1c) for the diagnosis of diabetes mellitus—practical implications. *Diabetes research and clinical practice*, 93 (3):312–313, 2011.

Maxime De Bois, Mounîm A. El Yacoubi, and Mehdi Ammi. GLYFE: review and benchmark of personalized glucose predictive models in type 1 diabetes. *Medical & Biological Engineering & Computing*, 60 (1):1–17, January 2022. ISSN 0140-0118, 1741-0444. doi: 10.1007/s11517-021-02437-4.

Julia Deichmann and Hans-Michael Kaltenbach. Model predictive control to account for prolonged changes in insulin requirements following exercise in type 1 diabetes. *Journal of Process Control*, 129:103042, September 2023. ISSN 09591524. doi: 10.1016/j.jprocont.2023.103042.

Ross Gruetzemacher and David Paradice. Deep transfer learning & beyond: Transformer language models in information systems research. *ACM Computing Surveys (CSUR)*, 54(10s):1–35, 2022.

Mehrad Jaloli, William Lipscomb, and Marzia Cescon. Incorporating the Effect of Behavioral States in Multi-Step Ahead Deep Learning Based Multivariate Predictors for Blood Glucose Forecasting in Type 1 Diabetes. *BioMedInformatics*, 2 (4):715–726, December 2022. ISSN 2673-7426. doi: 10.3390/biomedinformatics2040048.

Deepjyoti Kalita and Khalid B. Mirza. LS-GRUNet: Glucose Forecasting Using Deep Learning for Closed-loop Diabetes Management. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–6, Mumbai, India, April 2022. IEEE. ISBN 978-1-66542-168-3. doi: 10.1109/I2CT54291.2022.9824867.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Tetsuro Kobayashi, Shinji Sawano, Tokuji Itoh, Kinori Kosaka, Hiroki Hirayama, and Yasuji Kasuya. The Pharmacokinetics of Insulin After Continuous Subcutaneous Infusion or Bolus Subcutaneous Injection in Diabetic Patients. *Diabetes*, 32(4):331–336, April 1983. ISSN 0012-1797, 1939-327X. doi: 10.2337/diab.32.4.331.

Lori M. Laffel, Lauren G. Kanapka, Roy W. Beck, Katherine Bergamo, Mark A. Clements, Amy Criego, Daniel J. DeSalvo, Robin Goland, Korey Hood, David Liljenquist, Laurel H. Messer, Roshanak Monzavi, Thomas J. Mouse, Priya Prahalad, Jennifer Sherr, Jill H. Simmons, R. Paul Wadwa, Ruth S. Weinstock, Steven M. Willi, Kellee M. Miller, and CGM Intervention in Teens and Young Adults with T1D (CITY) Study Group. Effect of Continuous Glucose Monitoring on Glycemic Control in Adolescents and Young Adults With Type 1 Diabetes: A Randomized Clinical Trial. *JAMA*, 323(23):2388, June 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.6940.

Heike Leutheuser, Marc Bartholet, Alexander Marx, Marc Pfister, Marie-Anne Burckhardt, Sara Bachmann, and Julia E Vogt. Predicting risk for nocturnal hypoglycemia after physical activity in children with type 1 diabetes. *Frontiers in Medicine*, 11:1439218, 2024.

Kezhi Li, John Daniels, Chengyuan Liu, Pau Herrero, and Pantelis Georgiou. Convolutional Recurrent Neural Networks for Glucose Prediction. *IEEE Journal of Biomedical and Health Informatics*, 24 (2):603–613, February 2020. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2019.2908488.

Liang Li, Yuewen Jiang, and Biqing Huang. Long-term prediction for temporal propagation of seasonal influenza using transformer-based model. *Journal of biomedical informatics*, 122:103894, 2021.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In Hanna M. Wallach, Hugo Larochelle, Alina

Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5244–5254, 2019.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, October 2021. ISSN 01692070. doi: 10.1016/j.ijforecast.2021.03.012.

Cindy Marling and Razvan Bunescu. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. *CEUR workshop proceedings*, 2675:71–74, September 2020. ISSN 1613-0073.

Alexander Marx, Francesco Di Stefano, Heike Leutheuser, Kieran Chin-Cheong, Marc Pfister, Marie-Anne Burckhardt, Sara Bachmann, and Julia E. Vogt. Blood glucose forecasting from temporal and static information in children with T1D. *Frontiers in Pediatrics*, 11:1296904, December 2023. ISSN 2296-2360. doi: 10.3389/fped.2023.1296904.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Richard McShinsky and Brandon Marshall. Comparison of Forecasting Algorithms for Type 1 Diabetic Glucose Prediction on 30 and 60-Minute Prediction Horizons. In *KDH@ECAI*, 2020.

Aaron Neinstein, Jenise Wong, Howard Look, Brandon Arbiter, Kent Quirk, Steve McCanne, Yao Sun, Michael Blum, and Saleh Adi. A case study in open source innovation: developing the tidepool platform for interoperability in type 1 diabetes management. *Journal of the American Medical Informatics Association*, 23(2):324–332, 2016.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

J L Parkes, S L Slatin, S Pardo, and B H Ginsberg. A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. *Diabetes Care*, 23(8):1143–1148, August 2000. ISSN 0149-5992, 1935-5548. doi: 10.2337/diacare.23.8.1143.

Md Fazle Rabby, Yazhou Tu, Md Imran Hossen, Insup Lee, Anthony S. Maida, and Xiali Hei. Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction. *BMC Medical Informatics and Decision Making*, 21(1):101, December 2021. ISSN 1472-6947. doi: 10.1186/s12911-021-01462-5.

Eduardo Ramos-Pérez, Pablo J Alonso-González, and José Javier Núñez-Velázquez. Multi-transformer: A new neural network-based architecture for forecasting s&p volatility. *Mathematics*, 9 (15):1794, 2021.

Michael C Riddell, Ian W Gallen, Carmel E Smart, Craig E Taplin, Peter Adolfsson, Alistair N Lumb, Aaron Kowalski, Remi Rabasa-Lhoret, Rory J McCrimmon, Carin Hume, et al. Exercise management in type 1 diabetes: a consensus statement. *The lancet Diabetes & endocrinology*, 5(5):377–390, 2017.

Michael C. Riddell, Zoey Li, Robin L. Gal, Peter Calhoun, Peter G. Jacobs, Mark A. Clements, Corby K. Martin, Francis J. Doyle Iii, Susana R. Patton, Jessica R. Castle, Melanie B. Gillingham, Roy W. Beck, Michael R. Rickels, T1DEXI Study Group, Michael C. Riddell, Michael R. Rickels, Roy W. Beck, Jessica R. Castle, Peter Calhoun, Mark A. Clements, Francis J. Doyle, Robin L. Gal, Melanie B. Gillingham, Peter G. Jacobs, Zoey Li, Corby K. Martin, Susana R. Patton, Deniz Dalton, Laura E. Bocchino, Lindsey C. Beaulieu, Steven Bell, Brian Bugielski, Gabriela Cardenas Villamil, Ellis L. Greene, Russell Guzzetta, Anna Mazzuoccolo, Kyle Shaver, Karissa Neubig, Suna Onengut-Gumuscu, Stephen S. Rich, Wei-Men Chen, Joe Pinsonault, Joseph Leitschuh, Sos Oganessian, Eleonora Aiello, Matt Heelan, Lisa Sanesanong, Brian "Mooose" Rivera, Harpreet Gill, Avinash Kollu, Earl Glynn, Brent Lockee, and Mitchell Barnes. Examining the Acute Glycemic Effects of Different Types of Structured Exercise Sessions in Type 1 Diabetes in a Real-World Setting: The Type 1 Diabetes and Exercise Initiative (T1DEXI).

Diabetes Care, 46(4):704–713, April 2023. ISSN 0149-5992, 1935-5548. doi: 10.2337/dc22-1721.

Johannes A Romijn, EF Coyle, LS Sidossis, A Gastaldelli, JF Horowitz, E Endert, and RR Wolfe. Regulation of endogenous fat and carbohydrate metabolism in relation to exercise intensity and duration. *American Journal of Physiology-Endocrinology And Metabolism*, 265(3): E380–E391, 1993.

Harry Rubin-Falcone, Ian Fox, and Jenna Wiens. Deep residual time-series forecasting: Application to blood glucose prediction. *KDH@ ECAI*, 20:105–109, 2020.

Kyriaki Saiti, Martin Macaš, Lenka Lhotská, Kateřina Štechová, and Pavlína Pithová. Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus. *Computer Methods and Programs in Biomedicine*, 196:105628, November 2020. ISSN 01692607. doi: 10.1016/j.cmpb.2020.105628.

Renat Sergazinov, Mohammadreza Armandpour, and Irina Gaynanova. Gluformer: Transformer-based personalized glucose forecasting with uncertainty quantification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Md Maruf Hossain Shuvo and Syed Kamrul Islam. Deep Multitask Learning by Stacked Long Short-Term Memory for Predicting Personalized Blood Glucose Concentration. *IEEE Journal of Biomedical and Health Informatics*, 27(3):1612–1623, March 2023. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2022.3233486.

Emily Tu, Stephen M Twigg, Johan Duflou, and Christopher Semsarian. Causes of death in young australians with type 1 diabetes: a review of coronial postmortem examinations. *Medical journal of Australia*, 188(12):699–702, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Roberto Visentin, Enrique Campos-Náñez, Michele Schiavon, Dayu Lv, Martina Vettoretti, Marc Breton, Boris P Kovatchev, Chiara Dalla Man, and Claudio Cobelli. The UVA/Padova type 1 diabetes simulator goes from single meal to single day. *Journal of Diabetes Science and Technology*, 12(2):273–281, 2018.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Xia Yu, Tao Yang, Jingyi Lu, Yun Shen, Wei Lu, Wei Zhu, Yuqian Bao, Hongru Li, and Jian Zhou. Deep transfer learning: a novel glucose prediction framework for new subjects with type 2 diabetes. *Complex & Intelligent Systems*, pages 1–13, 2021.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Taiyu Zhu, Kezhi Li, Jianwei Chen, Pau Herrero, and Pantelis Georgiou. Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes. *Journal of Healthcare Informatics Research*, 4(3): 308–324, September 2020a. ISSN 2509-4971, 2509-498X. doi: 10.1007/s41666-020-00068-2.

Taiyu Zhu, Xi Yao, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Blood Glucose Prediction for Type 1 Diabetes Using Generative Adversarial Networks. In *KDH@ECAI*, 2020b.

Taiyu Zhu, Tianrui Chen, Lei Kuangt, Junming Zeng, Kezhi Li, and Pantelis Georgiou. Edge-Based Temporal Fusion Transformer for Multi-Horizon Blood Glucose Prediction. In *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, Monterey, CA, USA, May 2023a. IEEE. ISBN 978-1-66545-109-3. doi: 10.1109/ISCAS46773.2023.10181448.

Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Personalized Blood Glucose Prediction for Type 1 Diabetes Using Evidential Deep Learning and Meta-Learning. *IEEE Transactions on Biomedical Engineering*, 70(1):193–204, January 2023b. ISSN 0018-9294, 1558-2531. doi: 10.1109/TBME.2022.3187703.

Appendix A. Datasets

We provide a more comprehensive overview of the different features of CITY, OhioT1DM, and DiaCamp used in this study in Table 5. Further, Table 6 contains the total number of measurements and Table 7 the number of measurements per participant for each of the datasets. We also provide some additional metadata for OhioT1DM and DiaCamp in Tables 8 and 9 respectively. Finally, we visualize our feature alignment strategy for insulin on an example from an MDI and a CSII participant from DiaCamp in Figure 3.

Table 5: Overview of the different features of our datasets that were used for our experiments.

Dataset	Feature	Description
CITY	<PtID>	The unique participant ID number.
	<DeviceDtTm>	The time at which a CGM measurement is made, including date, hour, minute, and second.
	<Value>	The glucose measurements made by CGM devices.
OhioT1DM	<timestamp>	The time at which a measurement is made, including date, hour, minute, and second.
	<glucose_level>	The CGM data, recorded every 5 minutes.
	<basal>	The rate at which basal insulin is continuously infused.
	<temp_basal>	A temporary basal insulin rate that supersedes the participant's normal basal rate.
	<bolus>	Insulin delivered to the participant, typically before a meal or when the participant is hyperglycemia.
	<meal>	The self-reported time and type of a meal, plus the participant's carbohydrate estimate for the meal.
	<exercise>	Time and duration, in minutes, of self-reported exercise.
DiaCamp	<timestamp>	The time at which a measurement is made, including date, hour, minute, and second.
	<glucose>	The CGM data recorded every 5 minutes or the SMBG data.
	<fast_insulin>	Short-acting insulin that takes about 30 minutes to work fully.
	<slow_insulin>	Long-acting insulin that provides a full day of insulin coverage.
	<fast_food>	Carbohydrates with fast absorption rate, e.g. glucose tablets or orange juice.
	<slow_food>	Carbohydrates with slow absorption rate, e.g. whole grain or fatty products.
	<mixed_food>	Carbohydrates with mixed absorption rate, i.e. full meals.
	<sport>	Physical activity recorded in minutes.

Table 6: Number of participants and total CGM counts after preprocessing.

Dataset	**Participants**	**Measurements**
OhioT1DM	12	187'584
City	143	4'594'161
DiaCamp	14	22'624

Appendix B. Training Details

In the following, we provide additional details about the software we are using (Appendix B.1), the GluTFT's hyperparameters (Appendix B.2), and training efficiency (Appendix B.3).

B.1. Software

In our experiments, we use the *PyTorch Forecasting* library version 1.0.0 to implement and train N-BEATS, N-HiTS, and GluTFT. Further, we create Parkes Error Grids with the Python library *methcomp* version 1.0.1.

B.2. Hyperparameters

We provide the choice of hyperparameters of GluTFT in Table 10. In our experiments, we use data from two hours to predict glucose levels for the next two hours. After preprocessing according to Marx et al. (2023), the sampling period of glucose measurements is five minutes, meaning we have $\frac{120}{5} = 24$ data points for every

Table 7: The number of samples in train, validation, and test splits of each dataset. For the CITY dataset, we provide the number of samples per split since the dataset contains data from 153 participants. For OhioT1DM and DiaCamp we provide samples per participant. The CITY dataset is about 22.5 times larger than the OhioT1DM dataset, and the OhioT1DM dataset is about seven times larger than the DiaCamp dataset. Participant 017 in the DiaCamp dataset dropped out of the study, which is why this time-series is shorter than its counterparts.

Dataset	Participant ID	Train Set	Validation Set	Test Set
CITY	All	2,998,186	701,830	703,236
OhioT1DM	540	10,905	2,181	3,043
	544	10,540	2,109	3,113
	552	9,228	1,846	3,927
	559	10,048	2,010	2,853
	563	10,895	2,180	2,668
	567	11,260	2,253	2,848
	570	9,656	1,932	2,857
	575	10,900	2,181	2,695
	584	11,020	2,205	2,972
	588	10,902	2,181	2,857
	591	10,610	2,122	2,824
	596	11,338	2,268	2,980
	All	127,302	25,468	35,637
DiaCamp	004	1,152	288	288
	005	1,152	288	288
	006	1,152	288	288
	007	1,152	288	288
	008	1,152	288	288
	009	1,152	288	288
	010	1,152	288	288
	011	1,152	288	288
	013	1,152	288	288
	014	1,152	288	288
	015	1,152	288	288
	016	1,152	288	288
	017	748	187	187
	020	1,152	288	288
	All	15,724	3,931	3,931

Table 8: Metadata for participants in the OhioT1DM Dataset (Marling and Bunescu, 2020). Half of the complete OhioT1DM dataset was first released in 2018 for the first blood glucose level prediction (BGLP) challenge, and the second half was released in 2020 for the second BGLP challenge.

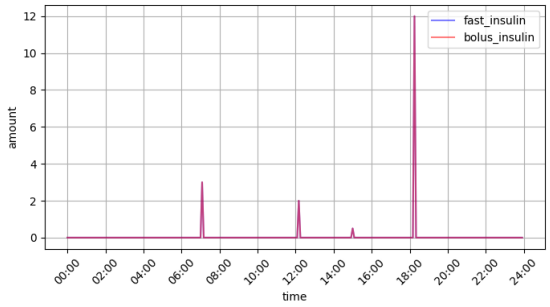
ID	Gender	Age	Pump Model	Sensor Band	Cohort
540	male	$20-40$	630G	Empatica	2020
544	male	$40-60$	530G	Empatica	2020
552	male	$20-40$	630G	Empatica	2020
567	female	$20-40$	630G	Empatica	2020
584	male	$40-60$	530G	Empatica	2020
596	male	$60-80$	530G	Empatica	2020
559	female	$40-60$	530G	Basis	2018
563	male	$40-60$	530G	Basis	2018
570	male	$40-60$	530G	Basis	2018
575	female	$40-60$	530G	Basis	2018
588	female	$40-60$	530G	Basis	2018
591	female	$40-60$	530G	Basis	2018

Table 9: Statistics across participants in the DiaCamp Dataset (Marx et al., 2023). Nine out of the 14 participants are male. HbAlc is identified as a special hemoglobin used as an objective measure of glycemic control (Colagiuri, 2011), and an HbAlc of 6.5% is recommended as the cut point for diagnosing diabetes.

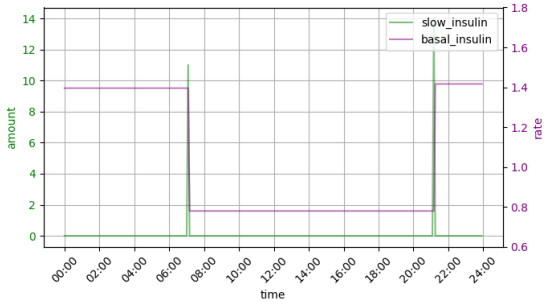
	Mean $\pm$ Std	Range
Age (years)	11.2 ± 2.1	$[7.5, 13.9]$
BMI (kg/m^2)	19.5 ± 4.2	$[13.2, 27.7]$
Weight (kg)	45.1 ± 16.5	$[21.0, 77.8]$
Height (cm)	149.4 ± 14.1	$125.3, 171.0$
Duration of Diabetes (years)	3.8 ± 2.8	$[0.5, 9.5]$
HbAlc (%)	7.2 ± 0.8	$[5.1, 8.5]$
Basal Insulin (%)	46.1 ± 12.5	$[26.2, 69.4]$
Total Daily Calculated Insulin (U/kg/d)	0.90 ± 0.39	$[0.31, 1.65]$

Table 10: Hyperparameters for TFT.

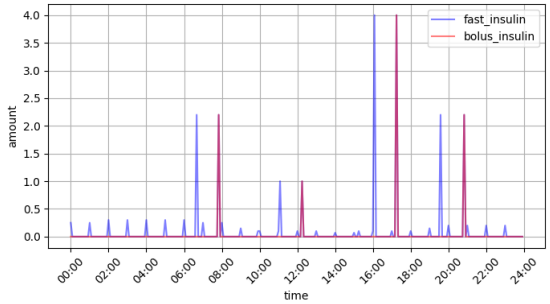
Hyperparameter	Value
max_encoder_length	24
max_prediction_length	24
hidden_size	16
lstm_layers	1
dropout	0.2
attention_head_size	4
output_size	7
quantiles	$[0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98]$



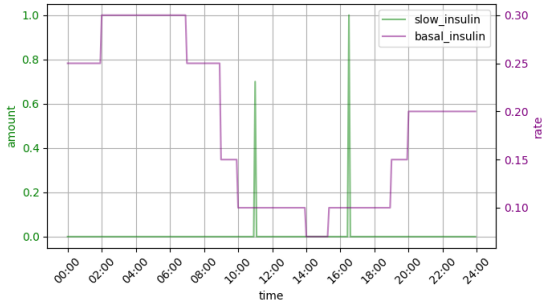
(*a*) Fast acting insulin to Bolus insulin for MDI.



(*b*) Slow acting insulin to Basal insulin for MDI.



(*c*) Fast acting insulin to Bolus insulin for CSII.



(*d*) Slow acting insulin to Basal insulin for CSII.

Figure 3: Example showing the transformation from fast/slow acting insulin to bolus/basal insulin for a time series of an MDI (a,b) and a CSII (c,d) participant.

two hours. Hence, we set *max_encoder_length* and *max_prediction_length* of the transformer architecture to 24. Further hyperparameters specifying the network architecture are *hidden_size*, *lstm_layers*, *dropout*, and *attention_head_size*. Additionally, *output_size* and *quantiles* correspond to the outputs of our model and the quantile loss used during training.

Table 11: Hyperparameters for Training.

Hyperparameter	Value
learning_rate (training)	0.001
learning_rate (pretraining)	0.001
learning_rate (fine-tuning)	0.0001
early_stop_patience	16
early_stop_min_delta	0.0001
max_epochs	1,000
batch_size	256
optimizer	Adam
gradient_clip_val	0.1

Training The training hyperparameters for GluTFT can be found in Table 11. We set the *learning_rate* for fine-tuning to 0.0001, i.e., smaller than that for training and pretraining, to diminish the effect of catastrophic forgetting (McCloskey and Cohen, 1989) during the fine-tuning phase. Further, we apply early stopping with a patience of 16 epochs and a cutoff of 0.0001. We train all models with a batch size of 256 with the Adam optimizer (Kingma and Ba, 2015). To improve training stability, we additionally apply gradient clipping (Zhang et al., 2020) with a cutoff of 0.1.

Baselines For our baseline models N-BEAT (Oreshkin et al., 2020) and N-HiTS (Challu et al., 2023), we keep hyperparameters at the default defined in the pytorch-forecasting library. We train N-BEATS with the RMSE and N-HiTS with the quantile loss (Wen et al., 2017).

B.3. Training Efficiency

Table 12: Number of GPU hours for a single run of each of our GluTFT experiments.

Experiment	#GPU Hours
Pretrain OhioT1DM	2
Pretrain CITY	8
Fine-tune DiaCamp	0.5

Table 13: Total number of parameters for the different models used in our experiments. M represents million, and K represents thousand.

Model	#Parameters
N-BEATS	1.7M
$\text{N-HiTS}_{\text{gluc}}$	943K
N-HiTS	1.3M
$\text{GluTFT}_{\text{gluc}}$	17.4K
GluTFT	27.8K

All our experiments were conducted on single *RTX2080Ti* GPUs. We provide an overview of the computing hours in Table 12. Additionally, Table 13 contains the number of parameters of the different models used in our experiments.

Appendix C. Further Experiments

In the following, we provide additional ablations, showcasing the interpretability of predictions (Appendix C.1), evaluating GluTFT on CITY similar to Section 4.4 (Appendix C.2), providing results when only fine-tuning on a subset of DiaCamp (Appendices C.3 and C.4), and additional insights into the Parkes Error Grids (Appendix C.5).

C.1. Interpretable Predictions

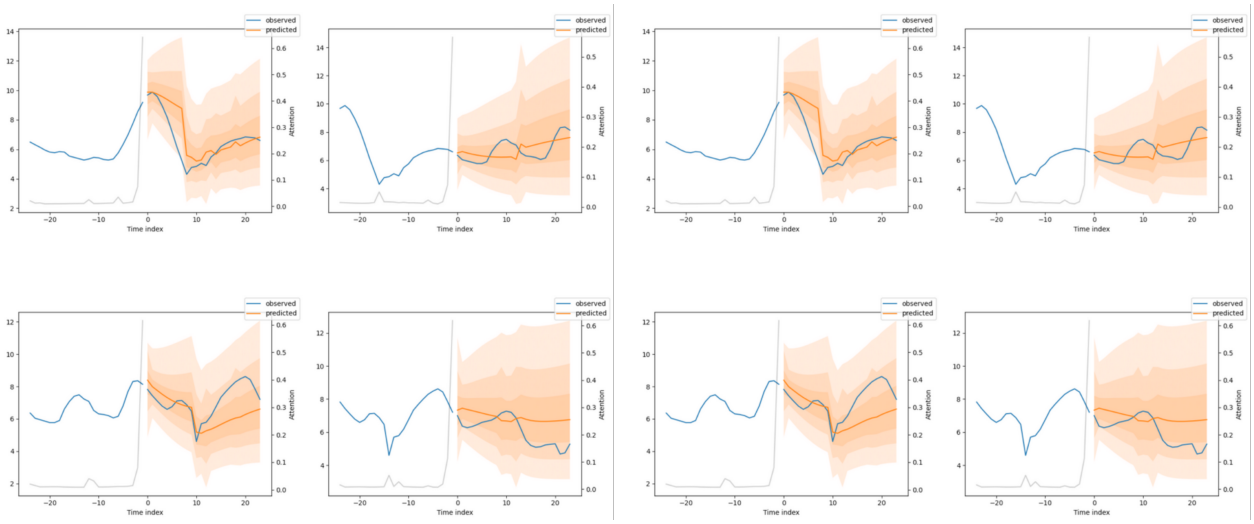


Figure 4: Example predictions of GluTFT. Gray lines illustrate the temporal importance weights, and orange regions correspond to the uncertainty inferred from the quantile predictions.

In Figure 4, we provide some example predictions of GluTFT, demonstrating the interpretability feature of the architecture. The gray lines illustrate the temporal importance weights, which show the influence of different parts of the time-series for the forecast and come from the interpretable multi-head attention layers of the model. The orange regions signal the uncertainty of our forecast coming from the quantile predictions that are learned through the quantile loss introduced in Section 3.2.1.

C.2. Evaluation on CITY

In this ablation, we evaluate GluTFT on the CITY dataset similar to Section 4.2. We preprocess our data the same way as with OhioT1DM. Note that CITY only contains CGM data without temporal information on insulin injections, meals, or physical activity. We thus compare only the models N-BEATS, N-HiTS$_{\text{gluc}}$, and GluTFT$_{\text{gluc}}$. While all models perform relatively similar, GluTFT is almost always the best or second-best performing model, despite only having access to CGM data.

C.3. MDI and CSII Fine-tuning

In DiaCamp, we have two groups of participants, some with insulin pumps (CSII) and some others with multiple daily insulin injections (MDI), as described in Section 3.1. We are thus interested to see how our

Table 14: Results on CITY, where we report the mean ± standard deviation across participants for 30, 60, and 120 minute prediction horizons. For consistency with Table 2, we average the mean and standard deviations of each seed across participants. We report the RMSE in (mmol/l), bold the **best result**, and underline the second best result. Apart from $RMSE_{120}$, GluTFT is always the best or second-best performing model.

Model	$RMSE_{30}$	$RMSE_{60}$	$RMSE_{120}$	$MAPE_{30}$	$MAPE_{60}$	$MAPE_{120}$
N-BEATS	1.24 ± 0.36	**2.21 ± 0.66**	**3.39 ± 1.01**	9.43 ± 3.88	17.95 ± 8.02	30.43 ± 14.18
N-HiTS$_{gluc}$	1.23 ± 0.36	2.22 ± 0.69	3.4 ± 1.06	**9.02 ± 3.53**	16.97 ± 7.2	28.8 ± 12.99
GluTFT$_{gluc}$	**1.23 ± 0.35**	2.21 ± 0.68	3.4 ± 1.08	9.02 ± 3.58	**16.84 ± 7.14**	**28.35 ± 12.58**

results change when fine-tuning and evaluating each group separately. The results of this ablation are in Table 15. We see that, surprisingly, MDI participants seem to perform better on average, despite the fact that CSII participants should have a more controlled blood glucose level, as blood glucose is constantly monitored and adjusted through the insulin pump. While we see that CSII participants are, on average, worse than MDI, most of the difference seems to stem from participant 008, who seems to have a very unforeseeable blood glucose level. For instance, $RMSE_{30}$ decreases to a mean value of 1.47 when excluding participant 008. In total, we see that the results do not considerably change in comparison to Table 3, indicating that our feature alignment strategy successfully unifies the features of OhioT1DM with CSII and MDI participant in DiaCamp.

Table 15: Results of GluTFT when fine-tuning on MDI and CSII participants separately. For each metric, we provide mean and standard deviation. The best result at each horizon is highlighted in **bold** while the worst result is highlighted in *italics*. The means and standard deviations of each group are calculated across the mean of the respective participants.

Type	ID	$RMSE_{30}$	$RMSE_{60}$	$RMSE_{120}$	$MAPE_{30}$	$MAPE_{60}$	$MAPE_{120}$
MDI	005	1.37 ± 0.01	1.63 ± 0.01	1.96 ± 0.01	12.02 ± 0.09	16.76 ± 0.18	22.79 ± 0.17
	007	1.20 ± 0.01	2.15 ± 0.01	3.27 ± 0.03	10.8 ± 0.04	20.57 ± 0.11	33.10 ± 0.28
	009	1.36 ± 0.01	2.56 ± 0.01	*3.97 ± 0.01*	10.08 ± 0.04	19.47 ± 0.07	29.68 ± 0.16
	010	1.89 ± 0.00	2.64 ± 0.01	2.66 ± 0.03	15.36 ± 0.03	20.36 ± 0.05	23.70 ± 0.13
	011	1.83 ± 0.00	2.57 ± 0.00	3.15 ± 0.01	15.57 ± 0.13	21.94 ± 0.28	30.29 ± 0.25
	013	2.24 ± 0.00	*2.98 ± 0.00*	3.89 ± 0.01	13.96 ± 0.02	23.20 ± 0.07	35.16 ± 0.10
	017	1.13 ± 0.01	1.89 ± 0.01	2.45 ± 0.03	12.21 ± 0.05	20.16 ± 0.06	28.19 ± 0.40
	020	1.15 ± 0.01	2.10 ± 0.02	3.13 ± 0.04	**09.50 ± 0.06**	17.54 ± 0.19	27.26 ± 0.44
CSII	004	1.91 ± 0.02	2.85 ± 0.02	3.82 ± 0.02	28.13 ± 0.32	40.58 ± 0.27	50.09 ± 0.24
	006	1.53 ± 0.01	2.13 ± 0.01	2.33 ± 0.01	13.32 ± 0.07	20.43 ± 0.09	24.07 ± 0.13
	008	*2.99 ± 0.01*	2.85 ± 0.01	2.84 ± 0.01	*52.24 ± 0.15*	*51.96 ± 0.15*	*52.95 ± 0.27*
	014	1.75 ± 0.01	2.46 ± 0.02	2.74 ± 0.02	15.64 ± 0.06	22.13 ± 0.15	24.17 ± 0.21
	015	1.22 ± 0.01	1.83 ± 0.00	2.28 ± 0.02	15.55 ± 0.12	23.69 ± 0.10	28.10 ± 0.14
	016	**0.94 ± 0.00**	**1.28 ± 0.01**	**1.43 ± 0.01**	09.84 ± 0.04	**13.09 ± 0.08**	**13.36 ± 0.19**
$Mean_{MDI}$		1.52 ± 0.41	2.31 ± 0.45	3.06 ± 0.69	12.44 ± 2.32	20.00 ± 2.12	28.77 ± 4.26
$Mean_{CSII}$		1.72 ± 0.71	2.23 ± 0.62	2.57 ± 0.79	22.45 ± 15.85	28.65 ± 14.58	32.12 ± 15.83
$Mean_{Total}$		1.61 ± 0.55	2.28 ± 0.50	2.85 ± 0.75	16.73 ± 11.22	23.71 ± 10.19	30.21 ± 10.44

C.4. Per Participant Fine-tuning

CGM data can vary quite dramatically from subject to subject, especially for children. We thus test how the results change when fine-tuning a pretrained model on each participant separately. In this experiment, we

pretrain GluTFT on OhioT1DM and then fine-tune the model on each subject separately by training on the first 4 days, i.e., 1152 samples, of the measurement (except for participant 007 where we only have 748 samples for training). Then, we predict the last day of the time series of each subject using their respective, fine-tuned model. We present the results of these experiments in Table 16.

Table 16: Results of GluTFT when fine-tuning and evaluating the model separately for each participant. For each metric, we provide mean and standard deviation. The best result at each horizon is highlighted in **bold** while the worst result is highlighted in *italics*. The mean and standard deviation are calculated across the means of the participants.

ID	$RMSE_{30}$	$RMSE_{60}$	$RMSE_{120}$	$MAPE_{30}$	$MAPE_{60}$	$MAPE_{120}$
004	1.81 ± 0.00	2.68 ± 0.00	3.68 ± 0.01	24.60 ± 0.06	38.39 ± 0.08	*54.28 ± 0.17*
005	1.32 ± 0.00	1.54 ± 0.00	1.72 ± 0.01	12.69 ± 0.04	**16.78 ± 0.07**	20.79 ± 0.15
006	1.56 ± 0.01	2.26 ± 0.01	2.54 ± 0.01	12.64 ± 0.01	21.46 ± 0.04	29.41 ± 0.15
007	1.48 ± 0.00	2.91 ± 0.01	*4.23 ± 0.01*	14.57 ± 0.03	30.67 ± 0.12	49.14 ± 0.24
008	*2.82 ± 0.03*	2.91 ± 0.01	2.96 ± 0.01	*46.77 ± 0.27*	*48.28 ± 0.74*	47.97 ± 0.40
009	1.16 ± 0.00	2.33 ± 0.01	3.72 ± 0.01	**08.67 ± 0.02**	17.92 ± 0.06	26.86 ± 0.10
010	1.71 ± 0.01	2.29 ± 0.01	2.59 ± 0.05	15.19 ± 0.11	19.58 ± 0.08	26.26 ± 0.63
011	1.80 ± 0.00	2.58 ± 0.00	2.94 ± 0.01	16.98 ± 0.08	26.58 ± 0.14	31.24 ± 0.21
013	2.22 ± 0.00	*2.99 ± 0.01*	3.98 ± 0.01	15.27 ± 0.04	23.84 ± 0.10	36.53 ± 0.21
014	1.89 ± 0.00	2.62 ± 0.00	2.96 ± 0.01	15.71 ± 0.02	22.01 ± 0.06	25.33 ± 0.09
015	1.16 ± 0.00	1.70 ± 0.01	2.21 ± 0.01	13.86 ± 0.01	20.49 ± 0.03	22.43 ± 0.10
016	1.14 ± 0.01	1.79 ± 0.01	1.67 ± 0.03	11.28 ± 0.07	18.22 ± 0.06	17.92 ± 0.18
017	**1.01 ± 0.00**	**1.46 ± 0.02**	**1.40 ± 0.01**	11.52 ± 0.05	17.19 ± 0.22	**17.06 ± 0.16**
020	1.35 ± 0.02	2.30 ± 0.04	3.18 ± 0.02	11.56 ± 0.08	20.87 ± 0.19	31.18 ± 0.34
Mean	1.60 ± 0.49	2.31 ± 0.52	2.84 ± 0.88	16.52 ± 9.46	24.45 ± 9.05	31.17 ± 11.79

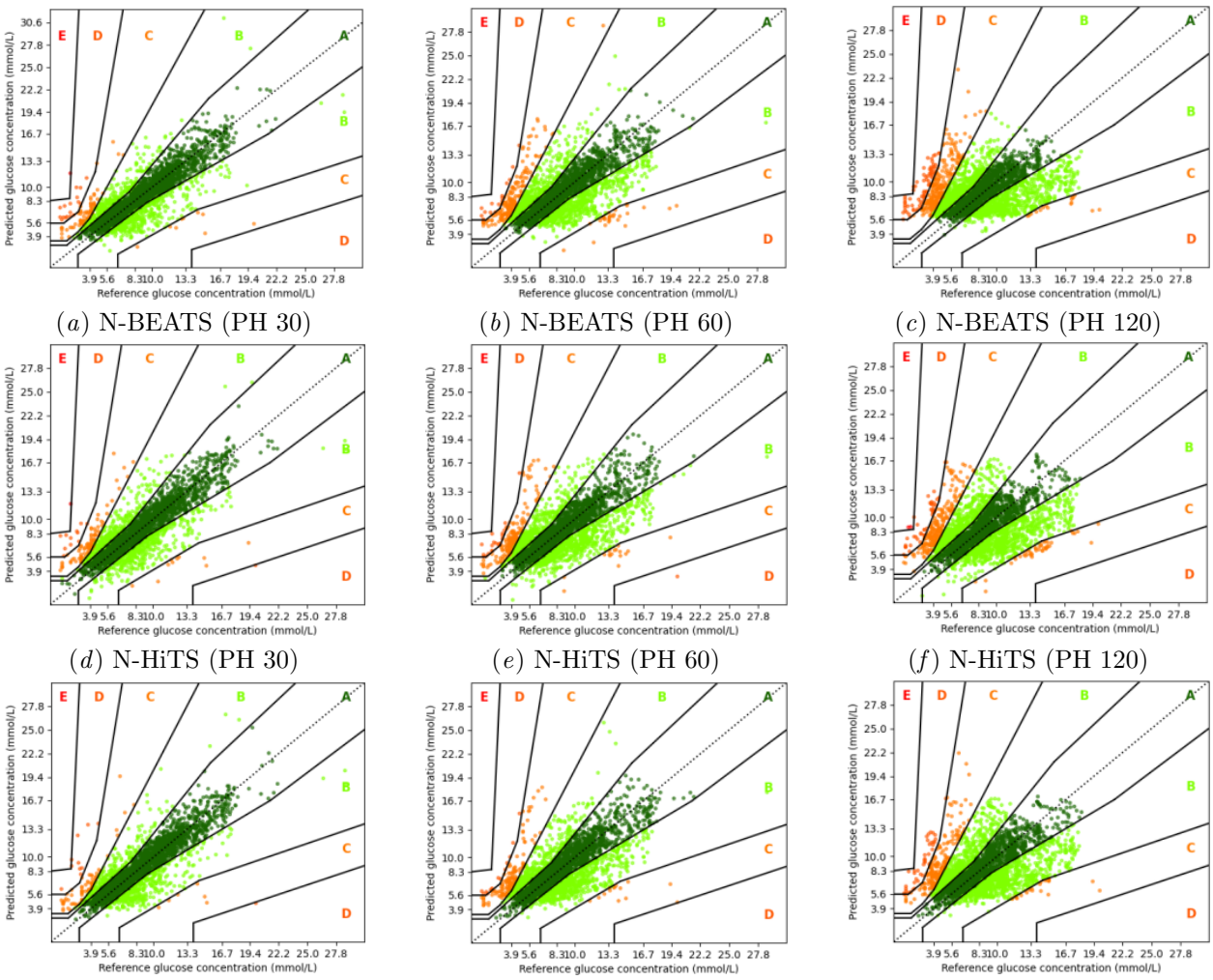
As CGM measurements are very subjective, the performance of GluTFT varies considerably among DiaCamp participants. Generally speaking, we see that GluTFT performs great for subjects 005, 009, and 017, whereas 004, 008, and 013 exhibit relatively poor performance. Similar to Appendix C.3, we find performance to be similar to fine-tuning on the whole dataset, indicating that our pretrained model cannot easily be fine-tuned toward small, specialized tasks such as single-subject glucose forecasting. Thus, it seems that in the case of a small and heterogeneous dataset such as DiaCamp, we still depend on the information of every single subject in the dataset to achieve good forecasting performance, highlighting that the pretrained model did not completely learn the general disease dynamic from the OhioT1DM dataset alone. Consequently, there is a need for larger and broader datasets for continuous glucose forecasting to create better pretrained models.

C.5. Parkes Error Grids

Table 17: Description of the medical interpretation of the different zones in Parkes Error Grids.

Zone	Description
A	No effect on clinical action
B	Altered clinical action—little with no effect on clinical outcome
C	Altered clinical action—likely to affect clinical outcome
D	Altered clinical action—could have significant medical risk
E	Altered clinical action—could have dangerous consequences

We provide the medical interpretation of each zone in the Parkes Error Grids in Table 17. In the following, we supplement the plotted Parkes error grids of N-BEATS, and N-HiTS for the experiment discussed in Section 4.4 alongside with the results for GluTFT. The exact numbers quantifying the percentage of data points falling within each zone is provided in Table 4. As discussed in the main experiments section, we can also visually observe that GluTFT has fewer predictions in zones C to E.



(a) N-BEATS (PH 30) (b) N-BEATS (PH 60) (c) N-BEATS (PH 120)

(d) N-HiTS (PH 30) (e) N-HiTS (PH 60) (f) N-HiTS (PH 120)

(g) GluTFT (PH 30) (h) GluTFT (PH 60) (i) GluTFT (PH 120)

Figure 5: Comparison of Parkes Error Grids for different model architectures pretrained on OhioT1DM and fine-tuned and evaluated on DiaCamp for increasing prediction horizons (PH).