2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-8472-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CBMS61543.2024.00089

COMET: Constrained Counterfactual Explanations for Patient Glucose Multivariate Forecasting

Zhendong Wang, Isak Samsten, Ioanna Miliou, and Panagiotis Papapetrou
Department of Computer and Systems Sciences
Stockholm University, Stockholm, Sweden
{zhendong.wang, samsten, ioanna.miliou, panagiotis}@dsv.su.se

***Abstract*—Applying deep learning models for healthcare-related forecasting applications has been widely adopted, such as leveraging glucose monitoring data of diabetes patients to predict hyperglycaemic or hypoglycaemic events. However, most deep learning models are considered black-boxes; hence, the model predictions are not interpretable and may not offer actionable insights into medical practitioners' decisions. Previous work has shown that counterfactual explanations can be applied in forecasting tasks by suggesting counterfactual changes in time series inputs to achieve the desired forecasting outcome. This study proposes a generalized multivariate forecasting setup of counterfactual generation by introducing a novel approach, COMET, which imposes three domain-specific constraint mechanisms to provide counterfactual explanations for glucose forecasting. Moreover, we conduct the experimental evaluation using two diabetes patient datasets to demonstrate the effectiveness of our proposed approach in generating realistic counterfactual changes in comparison with a baseline approach. Our qualitative analysis evaluates examples to validate that the counterfactual samples are clinically relevant and can effectively lead the patients to achieve a normal range of predicted glucose levels by suggesting changes to the treatment variables.**

***Index Terms*—time series forecasting, blood glucose prediction, counterfactual explanations, deep learning**

I. Introduction

Deep learning (DL) models have shown promising performance in several machine learning (ML) tasks, including time series forecasting [1]. Such models can utilize large amounts of historical time series measurements, and due to their capacity in learning non-linear relations across multiple time series variables, they can provide robust and reliable forecasts [2]. The latter is essential for critical application areas, such as healthcare. More specifically, DL models, such as attention-based architectures, have been applied to forecasting daily patient arrivals and length of stays in the ICU using historical time series from electronic health records (EHRs) [3]. Furthermore, for a better understanding of predicting patients' vital measurements, variations of recurrent neural networks (RNNs) and recent transformer-based models were applied to address patient glucose forecasting and aortic pressure forecasting problems [4], [5]. Although DL models can utilize the power of learning from a large amount of medical patient data for better performance, these models are often considered "black-boxes"; hence, it is challenging to interpret the modelling process and understand the forecasting outcome.

In healthcare, closely tracking glucose levels is a crucial task, especially for patients struggling with conditions such as type 1 diabetes mellitus (T1DM) patients. These patients currently rely on continuous glucose monitoring (CGM) devices and automated insulin delivery to reduce the risk of hyperglycaemic or hypoglycaemic events [6]. In comparison with traditional clinical care, using ML in modelling CGM can significantly help diabetes patients gain a better understanding of predicting abnormal glucose events like hyperglycaemia, and also to provide better planning, like insulin dosages [7]. Moreover, recent advances in DL forecasting models have gained popularity in blood glucose forecasting tasks through data-driven DL-based approaches [4], [8], [9]. Also, recent work has applied XAI techniques, like Shapley additive explanations (SHAP), into abnormal glycaemic event predictions to explain the importance of patient features in an ML regression setup [10]. Nevertheless, most ML and DL models either lack interpretability or the provided explanations are not actionable (e.g., adjusting the treatment variables), which potentially causes clinical practitioners not to trust the predictions.

Counterfactual explanations provide model-agnostic explanations by suggesting how to modify the original input sample so that the prediction of the trained ML model is changed to a more desirable outcome [11]. In a recent approach, called ForecastCF [12] employs different black-box DL time series forecasting models (including N-Beats [13]) and applies upper and lower constraints to the forecasting horizon so that the forecasted values fall within a desired band. Nonetheless, ForecastCF is only applicable to univariate time series forecasting as it directly modifies the historical values of the target time series without taking into consideration any additional variables that occur concurrently and could potentially affect the progression of the observed (target) variable. Moreover, to propose actionable counterfactuals in a multivariate setup, additional constraints should be imposed on the actions suggested by the counterfactual explainer.

Towards this direction, in this study, we focus on a multivariate time series forecasting setup, with the objective to define meaningful and medically relevant counterfactual explanations. Specifically, we focus on the "multivariate-to-univariate" blood glucose forecasting scenario [9], [14]. Furthermore, we define three domain-specific constrain mecha-

This work was funded in part by the Digital Futures EXTREMUM project https://www.digitalfutures.kth.se/research/collaborative-projects/extremum/.

2372-9198/24/$31.00 ©2024 IEEE
DOI 10.1109/CBMS61543.2024.00089

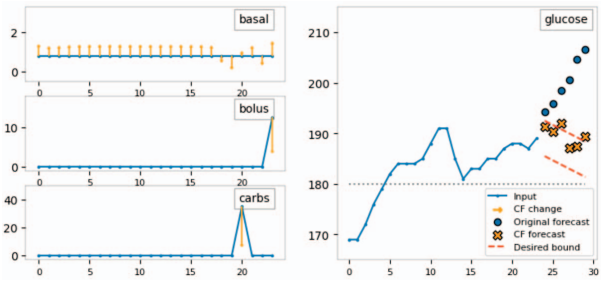


Fig. 1. CF example of a diabetes patient from OhioT1DM. The yellow arrows highlight the counterfactual changes, while the original input time series are shown as blue lines. Given the desired forecasting constraints (red-dotted lines), we would like the counterfactual forecasting values (yellow x-points) to fall within the band following the trend towards the normoglycaemia range.

nisms designed explicitly for CGM monitoring controls for type 1 diabetes patients: 1) *Clipping constraint*; 2) *Activity temporal constraint*; and 3) *Historical values constraint*. With the additional constraint mechanisms, our proposed approach, referred to as COMET, can generate counterfactual samples that are clinically relevant, while following the forecasting trend of avoiding hyperglycaemic and hypoglycaemic events.

Example. In Fig. 1, we illustrate our proposed approach using a diabetes patient example from OhioT1DM [15]. Using the 2-hour historical measurements (i.e., 24 timesteps) of glucose, basal, bolus insulins, and carbohydrate intakes, the forecasting model predicts that the patient will suffer from a hyperglycemic event (i.e., glucose values above 180 mg/dL) continuously in the next 30 minutes (i.e., 6 timesteps). Given that the objective is to impose a decreasing trend on the future glucose forecasted values, towards normoglycemia (i.e., below 180 mg/dL), COMET can provide counterfactual explanations to adjust the planned basal and bolus insulin amounts, but also reduce the carbohydrate intakes (highlighted in yellow arrows in Fig. 1). With these suggested counterfactual changes, the patient is predicted to have forecasted glucose values (yellow x-points) following the trend of the desired counterfactual bound (red-dotted lines). Given the actions on the treatment plan suggested by the counterfactual explainer, medical practitioners can not only understand the degree to which these treatment factors can support future glucose predictions, but also learn what actions could be used to improve the glucose level stability in the future plan.

Related work. Recent research has proposed various *DL models for forecasting*, such as variants of RNN-based models like gated recurrent unit (GRU), long short-term memory (LSTM), and the integration of attention mechanisms [2]. In addition, transformer models with self-attention mechanisms, such as Autoformer and Informer, were proposed to address both univariate and multivariate forecasting problems with competitive performance [16]. More recently, various DL forecasting approaches have been applied in the clinical domain of glucose forecasting. In such an approach, WaveNet was adopted to forecast future glucose levels with a modified dilated CNN structure [17]; RNN-based models were applied to forecast univariate CGM measurements by explicitly modelling the distribution of future glucose values [4]. Furthermore, transfer learning has been applied to predict diabetes patients' glucose levels by using a global pre-trained model and fine-tuning, but also incorporating additional covariates (such as insulin and carbohydrate intakes) in a multivariate forecasting setup [8], [9]. Regarding *model explainability*, DL models usually consist of complex model structures, and it can be challenging to provide intrinsic model interpretations like traditional statistical methods (e.g., ARIMA) [18]. Recently, N-Beats was proposed to utilize internal trend and seasonality stacks in a DL architecture to explain the forecasting outcome of univariate time series [13]; however, the interpretations were static and remained difficult to reflect the whole prediction process. On the other hand, post-hoc explanation methods have been applied to provide model explanations for black-box ML models, e.g., adopting LIME to show the feature importance of specific timesteps [19]. Counterfactual explanations have been applied in time series classification to explain how to modify an input sample to achieve a desired target outcome using different techniques like local time series tweaking, instance-based and gradient-based perturbations [20]–[22]. More recently, counterfactual explanations have been adopted for time series forecasting [12]. To the best of our knowledge, this work has not been generalized to multivariate forecasting and has not been applied to real-world forecasting scenarios.

Contributions. The main contributions of this paper can be summarized as follows: (1) we provide a generalized problem formulation to counterfactual generation for multivariate time series forecasting, where we incorporate both the desired forecast outcome and local temporal constraints in the optimization; (2) we propose COMET, a novel algorithm for multivariate time series counterfactual generation that is based on gradient optimization, and additionally employs three constraint mechanisms for generating domain-specific counterfactuals: a clipping mechanism, an activity temporal constraint, and a historical values constraint; (3) we conduct an empirical evaluation on two datasets, OhioT1DM (real-world patients) and SimGlucose (simulated), and then separate them into hyperglycemia and hypoglycemia patient groups for counterfactual evaluation against a baseline; and (4) we demonstrate that COMET obtains clinically relevant and actionable counterfactual explanations, following domain-specific constraints and the desired glucose forecasting trend of moving towards normoglycemia.

II. Methodology

A. Problem formulation

Let $\boldsymbol{X} =< \boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}, \boldsymbol{x}_n >$ be a multivariate time series observed at time point n, where we refer to n as the *back horizon*. Each $\boldsymbol{x}_i \in \boldsymbol{X}$ comprises m time series variables (or dimensions), i.e., $\boldsymbol{x}_i \in \mathbb{R}^{m \times n}$. More specifically, $\boldsymbol{X} = \{\boldsymbol{y}, \boldsymbol{\mathcal{Z}}\}$ is a concatenation of a target variable $\boldsymbol{y} =< y_1, ..., y_{n-1}, y_n >$ and a set of d variables $\boldsymbol{\mathcal{Z}} =< \boldsymbol{z}_1, ..., \boldsymbol{z}_{n-1}, \boldsymbol{z}_n >$, where $y_i \in \mathbb{R}^{1 \times n}$ and each $\boldsymbol{z}_i \in \mathbb{R}^{d \times n}$, i.e., $m = 1 + d$. Consider a

black-box time series forecasting model $f(\cdot)$ that predicts the next T values (*forecasting horizon*) of variable $\boldsymbol{y}$, which we denote as $\hat{\boldsymbol{y}} =< \hat{y}_{n+1}, \hat{y}_{n+2}, ..., \hat{y}_{n+T} >$, using $\boldsymbol{X}$, i.e.,

$$f(< \boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}, \boldsymbol{x}_n >) =< \hat{y}_{n+1}, \hat{y}_{n+2}, ..., \hat{y}_{n+T} > \ .$$

Given a lower and upper bound set of constraints for each time point in the forecasting horizon, i.e., $\boldsymbol{\alpha} = \{\alpha_1, \dots \alpha_T\}$ and $\boldsymbol{\beta} = \{\beta_1, \dots \beta_T\}$, respectively, our objective is to generate a counterfactual sample $\boldsymbol{X'} \in \mathbb{R}^{m \times n}$, such that the forecasted values $\hat{\boldsymbol{y}}' = f(\boldsymbol{X}')$ fall within the range defined by the two constraints at each time point, i.e., $\alpha_j \leq \hat{y_i'} \leq \beta_j$, $\forall \hat{y_i'} \in \hat{\boldsymbol{y}}', i \in [n+1, n+T], j \in [1, T]$. More specifically, we study the following problem:

Problem (***Constrained counterfactuals for time series forecasting***). *Given $f(\cdot)$, a time series sample $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, the two constraint vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we want to define a time series counterfactual $\boldsymbol{X'} \in \mathbb{R}^{m \times n}$, such that $f(\boldsymbol{X'})$ falls within the range defined by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the counterfactual $\boldsymbol{X'}$ is constrained by an $m \times n$ binary mask $\boldsymbol{C} = [\boldsymbol{c}_1, \dots, \boldsymbol{c}_n]$, with each $\boldsymbol{c}_i$ being an m-dimensional binary vector, i.e.,*

$$\boldsymbol{X'} = \underset{\boldsymbol{X^*}}{\arg\min} ||f(\boldsymbol{X^*}) - \boldsymbol{\alpha}|| + ||\boldsymbol{\beta} - f(\boldsymbol{X^*})|| + \boldsymbol{C}^\intercal |(\boldsymbol{X} - \boldsymbol{X^*})/\boldsymbol{X}| \ .$$

In this work, we study a particular scenario of Problem 1 to forecast blood glucose values (i.e., target variable $\hat{\boldsymbol{y}}$) for the next T timesteps, where the input $\boldsymbol{X}$ consists of glucose values (i.e., $\boldsymbol{y}$) and treatment measurements like insulin and meal (i.e., $\boldsymbol{\mathcal{Z}}$) from the previous n timesteps. Then our goal is to find the counterfactual $\boldsymbol{X'}$, such that the forecasting values $\hat{\boldsymbol{y}}'$ fall within the desired bounds $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ and the binary mask $\boldsymbol{C}$ can encourage changes at specific timesteps.

B. COMET: COnstrained counterfactual explanations for patient glucose MultivariatE forecasTing

Our solution is to apply gradient-based perturbation on the input sample $\boldsymbol{X}$ to reach the desired forecasting outcome, as shown in Algorithm 1. More specifically, COMET first adopted the loss function defined in earlier work [12], as one of the objectives during the counterfactual generation:

$$L_f(\boldsymbol{X^*}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{v} \odot (||f(\boldsymbol{X^*}) - \boldsymbol{\alpha}|| + ||\boldsymbol{\beta} - f(\boldsymbol{X^*})||) \ , \quad (1)$$

where $\boldsymbol{v}$ is a binary masking vector to check whether the forecasting value $\hat{y_i^*}$ at each timestep i falls within the lower and upper bounds between α_i and β_i (if so, then $v_i = 0$, and vice versa). We additionally propose three constraint mechanisms, as described below:

1) Clipping constraint: This constraint acts as the $Clip(\cdot)$ function after each gradient update step, ensuring each x_i^* at timestep i from the counterfactual $\boldsymbol{X}^*$ are within the constraint value range (defined with a minimum $\boldsymbol{\rho}$, and a maximum $\boldsymbol{\phi}$): $x_i^* = min(\boldsymbol{\rho}, max(x_i^*, \boldsymbol{\phi}))$. The constraint range [$\boldsymbol{\rho}$, $\boldsymbol{\phi}$] can be customized by a domain expert in the context of personalized medicine. We set $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$ to the minimum and maximum values of each feature from training data in the experiment.

2) Activity temporal constraint: By introducing the activity temporal constraint, we can incorporate the constraint vector $\boldsymbol{c}$ in the loss calculation to encourage counterfactual changes when there is any planned activity (e.g., meal and exercise). We apply the $ActivityTemporalConstriant(\boldsymbol{X^*})$ function to obtain constraint value c_i at timestep i as below:

$$c_i = \begin{cases} 0, & \text{if } x_i > \gamma, x_i \in \boldsymbol{X^*} \\ 1, & \text{otherwise,} \end{cases}$$

where γ can be a user-defined threshold value; in our experimental setup, we apply an activity threshold $\gamma = 0$.

3) Historical values constraint: This imposes constraints on the counterfactual samples to become closer to the historical value inputs. During the generation process, we iteratively update the constraint vector c_i at timestep i using the $HistValueConstraint(\boldsymbol{X^*}, \boldsymbol{\mathcal{G}})$ function:

$$c_i = \begin{cases} 0, & |\{|x_i^* - \boldsymbol{\mathcal{G}_j}| \leq \delta \colon j \in \{1, \dots, |\boldsymbol{\mathcal{G}}|\}\}| > 0 \\ 1, & \text{otherwise,} \end{cases}$$

where δ is the tolerance of the closeness between x_i^* and $\boldsymbol{\mathcal{G}_j}$ from the set of the historical values; we define $\delta = 0.001$ in the experiment.

In both constraint 2) and 3), $c_i = 0$ indicates that we encourage the change to happen at timestep i, and vice versa. We define the loss of the local-temporal constraint as:

$$L_c(\boldsymbol{X^*}, \boldsymbol{X}, \boldsymbol{C}) = \boldsymbol{C}^\intercal |(\boldsymbol{X} - \boldsymbol{X^*})/\boldsymbol{X}| \ , \quad (2)$$

Furthermore, we balance Eq. 1 and Eq. 2 using a forecast margin parameter w in the final optimization function:

$$L(\boldsymbol{X^*}, w, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{C}) = w \cdot L_f(\boldsymbol{X^*}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + (1 - w) \cdot L_c(\boldsymbol{X^*}, \boldsymbol{X}, \boldsymbol{C}) \ , \quad (3)$$

where we choose $w = 0.9$ in the experiment so that the loss optimization will focus on the forecast margin. Finally,

Algorithm 1: COMET counterfactual search

input : Time series $\boldsymbol{X}$, differentiable forecaster $f(\cdot)$, lower and upper bounds [$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$], clipping ranges [$\boldsymbol{\rho}$, $\boldsymbol{\phi}$], historical value set $\boldsymbol{\mathcal{G}}$, forecast margin weight w, learning rate η, maximum iteration max_iter

output: Counterfactual $\boldsymbol{X'}$ with desired outcome

$\boldsymbol{X^*} \leftarrow \boldsymbol{X}$
$\hat{\boldsymbol{y}}^* \leftarrow f(\boldsymbol{X^*})$
$\boldsymbol{C} \leftarrow ActivityTemporalConstriant(\boldsymbol{X^*})$
$loss \leftarrow L(\boldsymbol{X^*}, w, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{C})$
$t \leftarrow 0$
while $(\hat{\boldsymbol{y}}^* > \boldsymbol{\beta} \vee \hat{\boldsymbol{y}}^* < \boldsymbol{\alpha}) \wedge (t < max\_iter)$ **do**
  $\boldsymbol{X^*} \leftarrow AdamOptimize(\boldsymbol{X^*}, loss, \eta)$
  $\boldsymbol{X^*} \leftarrow Clip(\boldsymbol{X^*}, \boldsymbol{\rho}, \boldsymbol{\phi})$
  $\hat{\boldsymbol{y}}^* \leftarrow f(\boldsymbol{X^*})$
  $\boldsymbol{C} \leftarrow HistValueConstraint(\boldsymbol{X^*}, \boldsymbol{\mathcal{G}})$
  $loss \leftarrow L(\boldsymbol{X^*}, w, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{C})$
  $t \leftarrow t + 1$
$\boldsymbol{X'} \leftarrow \boldsymbol{X^*}$
return $\boldsymbol{X'}$

we incorporate an 'only_change' parameter to partition the multivariate input $\boldsymbol{X}$ into static and changeable variables, excluding the target variable $\boldsymbol{y}$. Throughout the perturbation process, the static variables remain unaffected by the gradients; while the changeable variables utilize the gradients to generate the counterfactual following the optimization function L. Without loss of generalizability, determining a static variable is equivalent to setting the constraint vector $\boldsymbol{C}$ to $-\inf$ at every timestep (i.e., $\boldsymbol{C} = [-\inf, -\inf, \ldots, -\inf]$) for the specific variable. By default, we run the experiment with all the variables as changeable except the target variable.

III. Empirical Evaluation

A. Data preparation

We conducted our experiments on two datasets: a simulated dataset from the FDA-approved UVA/PADOVA type 1 diabetes simulator [23] (referred to as *SimGlucose*) and a real-world diabetes patient dataset [15] (named *OhioT1DM*). *SimGlucose* includes CGM measurements with external variables like insulin dosages and carbohydrate intakes; we generated over 40,000 simulated recordings for ten adult patients, sampling over a week's time interval following the implementation[1]. *OhioT1DM* contains over 60,000 measurements for twelve type-1 diabetes patients collected over 8 weeks by Ohio University. Following previous works [9], [14], we extracted the most relevant clinical features, including CGM measurements, basal (slow-acting) insulin, bolus (fast-acting) insulin, carbohydrate intakes and exercise intensity.

Specifically, the training data was divided into 80/20 for the training and validation sets, while the standalone test data (i.e., marked by unique patient IDs in OhioT1DM) was assessed to report the model performance. Then, we applied min-max normalization using the training set before further segmenting them into back horizon and horizon sequences (i.e., the inputs and outputs) using a rolling origin approach [18]. We discarded any sequences with missing values after the segmenting. Finally, during the counterfactual evaluation, we separated the test set into hyperglycaemia and hypoglycaemia patient groups to investigate their differences effectively.

B. Experimental setup

In the experiment, we employed two main DL forecasting models in the evaluation: (1) a 2-layer GRU model with 100 units at each layer and a linear output; and (2) a WaveNet model[2] with a skip connection and the filter size set to 256. Early stopping was applied to prevent over-fitting, and the learning rate was fixed at 0.0001. Specifically, the glucose forecasting task involved using the previous 2 hours' patient measurements to forecast glucose values in the following 30 minutes. For the OhioT1DM dataset, this setup was equivalent to 24 timesteps for the back horizon and 6 timesteps for the horizon; while for SimGlucose, the back horizon and the horizon were set to 40 and 10 timesteps, respectively.

[1]See https://github.com/jxx123/simglucose.

[2]TFTS package: https://github.com/longxingtan/time-series-prediction.

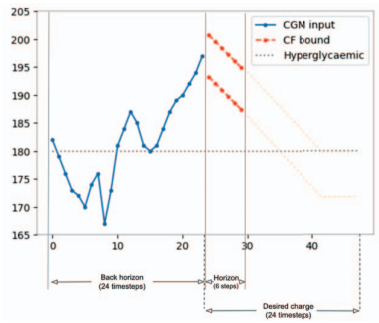


Fig. 2. Example of the counterfactual bound generation for a patient from OhioT1DM. Given an input time series of the back horizon (the blue line), and a desired center as the hyperglycaemic threshold in 24 timesteps, we follow the desired trend to define the upper and lower bounds (red-dotted lines) aligning with the forecasting horizon (i.e., 6 timesteps).

Furthermore, we conducted experiments with the counterfactual generation by incrementally incorporating additional constraints to help guide the generation process. Consequently, we had three variants of COMET: (1) COMET-C: only included the clipping constraint; (2) COMET-A: included both clipping constraint and activity temporal constraint; and (3) COMET-H: included all three mechanisms (i.e., with historical values constraint). We also compared the original ForecastCF [12] as the baseline (i.e., without additional constraints). As a side note, when generating the upper and lower bounds, since many type-1 diabetes patients often had glucose levels outside the normoglycemia range between 180 and 70 [6], we decided to generate constraint bounds by following a trend to reach the desired center in 2 hours (e.g., the next 24 timesteps for OhioT1DM). Then, we trimmed the bound vectors to match the forecasting horizon length (e.g., 6 timesteps). The process of bound generation can be visualized in Fig. 2. The desired center was set to 170 and 80 for hyperglycaemia and hypoglycaemia patient groups, respectively. We set the learning rate to 0.001, the maximum iteration to 100, and the fraction of the generated bound width parameter to 0.5 or 1 (depending on the test groups) from an empirical search. For reproducibility, our experimental code is publicly available at https://github.com/zhendong3wang/explain-glucose-forecasting-counterfactuals.

Evaluation metrics. We evaluated the DL models using two metrics: symmetric mean absolute percentage error (*sMAPE*) and root mean squared error (*RMSE*), where lower scores indicated better predictive performance. For the counterfactual evaluation, we adopted four commonly used metrics [12], [20], [21]: validity, proximity, compactness and stepwise validity AUC. More concretely, *validity* is defined as the ratio of successful forecasting values that fall within the range; while *stepwise validity AUC* calculates the area under the curve for validity at each cumulative step. The higher scores are more desirable for both metrics. *Proximity* calculates the average Euclidean distances between the original and counterfactual samples, where a lower score is desired. *Compactness* measures how close the counterfactual samples are to the original samples, with a higher score indicating better performance.

Table I. Forecasting model performance for OhioT1DM and SimGlucose. The best score for each dataset is highlighted in bold.

Dataset	Model	*sMAPE*	*RMSE*
OhioT1DM	GRU	**5.669**	13.835
	WaveNet	5.810	**13.697**
SimGlucose	GRU	**5.114**	**8.256**
	WaveNet	6.645	10.374

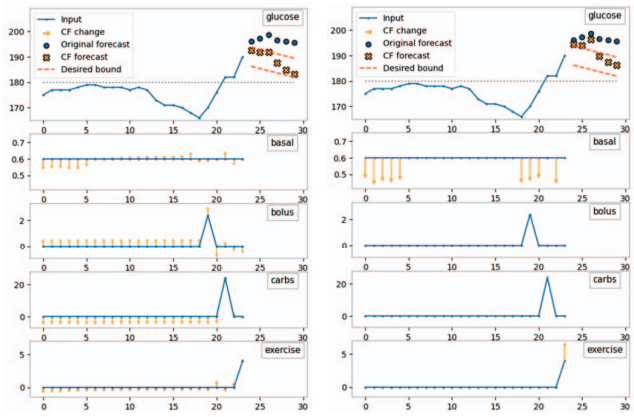
C. Results

We first compared the predictive performance of two DL models, as depicted in Table I. We observed that the GRU model obtained optimal sMAPE scores for both OhioT1DM and SimGlucose; while WaveNet outperformed GRU in RMSE for the OhioT1DM dataset.

Table II shows the counterfactual evaluation results for hyperglycemia and hypoglycemia patient groups from OhioT1DM and SimGlucose datasets, on GRU and WaveNet models separately. We observed that the baseline ForecastCF obtained optimal validity in most cases, while COMET-C got better validity scores in two specific scenarios for hypoglycemia patients. In comparison, we found that our proposed COMET-H approach achieved the best proximity and compactness scores across most hyperglycemia and hypoglycemia patients for two datasets. This suggests that the constraint mechanisms can support the generation of counterfactual samples that are proximate to the original samples and make more sparse counterfactual changes, while the baseline ForecastCF produces counterfactuals that diverge from the original samples and hence become non-realistic changes (i.e., there is a trade-off between proximity/compactness and validity).

When comparing the proposed COMET-C with the baseline ForecastCF across all the patient groups, we could observe that the validity and stepwise AUC slightly dropped while the compactness scores became much higher. This was because the *clipping mechanism* restricted the counterfactual generation with the possible range for each variable to prune out-of-range values from the counterfactual generation. Furthermore, we found that the proximity and compactness of COMET-A significantly improved compared with COMET-C. This indicated that the *activity temporal constraint* mechanism discouraged unreasonable changes in patient cases with missing planned activity (e.g., carbohydrate intake, or bolus insulin), such that the counterfactual samples became more proximate and compact. Finally, COMET-H further improved the proximity and compactness scores. With the additional *historical values constraint* mechanism to encourage the counterfactuals to follow historical treatment values, this approach produced the most reasonable counterfactual changes based on the historical values, which led to better proximity and compactness; although it sacrificed the validity performance with forecasting values. In addition, we found that the hypoglycemia patient group obtained relatively lower validity than hyperglycemia (particularly for SimGlucose), which suggested that generating counterfactual samples for hypoglycemia patients was generally more challenging in our experiment setup.

Examples. In Fig. 3, we show counterfactual examples of one specific OhioT1DM patient with a forecasted hyperglycemic event in the next 30 minutes (i.e., 6 timesteps). For the baseline example (Fig. 3a), we observed that there were many negative counterfactual suggestions for basal, carbs and exercise measurements, which were out-of-range values; in comparison, our proposed COMET-H approach (Fig. 3b) suggested all the counterfactual values to fall within the normalized range from the observed variables. This indicated that the *clipping mechanism* could effectively remove unreasonable suggestions during the counterfactual generation. Furthermore, we found that the COMET-H example provided more sparse suggestions (i.e., decreasing the basal volume at the beginning and end of the monitoring window, and increasing exercise intensity at the end), which differed from the baseline example that suggested changing the values for almost every timestep. This was because the additional constraint mechanisms of COMET-H enabled learning from historical values and only recommending changes if there were planned activities in the counterfactual generation. Although we could observe that the validity ratio of the proposed approach was 50% (i.e., only 3 out of 6 forecasting values fell within the desired bounds), the example demonstrated a forecasting trend toward decreasing the patient's glucose values and leading to a lower risk of hyperglycemia [6]. This indicated that our proposed approach could generate counterfactuals following clinically relevant constraints and provide relatively valid prediction trends towards the desired patient outcome.



(a) Baseline (b) COMET

Fig. 3. Examples of generated counterfactuals from baseline, COMET-H using a patient example from OhioT1DM. The blue lines show the original time series inputs with five measurements, and the orange arrows demonstrate the counterfactual suggestions. The red-dotted curves show the forecast constraints that the predicted values (yellow points) of the counterfactuals are more desired to fall into; the blue points are the original forecasted values.

IV. Conclusions

We proposed COMET, a novel model-agnostic counterfactual explanation method, to generate counterfactual explanations for multivariate forecasting. We incorporated three domain-specific constraints for diabetes patients in glucose forecasting, by suggesting effective counterfactual changes for both hyperglycaemia and hypoglycaemia patients to adapt

Table II. Summary of counterfactual evaluation metrics for OhioT1DM and SimGlucose datasets with hyperglycemia and hypoglycemia patent groups, using GRU and WaveNet models. The best score for each group is highlighted in bold.

			GRU				WaveNet			
Dataset	**Patient group**	**CF model**	*Valid.*	*Proxi.*	*Compa.*	*Step AUC*	*Valid.*	*Proxi.*	*Compa.*	*Step AUC*
OhioT1DM	Hyperglycemia	ForecastCF	**0.945**	0.193	0.264	**0.783**	**0.955**	0.159	0.758	**0.790**
		COMET-C	0.867	0.349	0.558	0.706	0.900	0.251	0.844	0.746
		COMET-A	0.675	0.347	0.808	0.526	0.718	0.192	0.935	0.569
		COMET-H	0.585	**0.173**	**0.910**	0.459	0.658	**0.150**	**0.955**	0.521
	Hypoglycemia	ForecastCF	**0.497**	0.284	0.049	**0.188**	0.525	0.162	0.721	**0.196**
		COMET-C	0.490	0.345	0.247	0.175	**0.530**	0.189	0.829	0.171
		COMET-A	0.207	0.396	0.775	0.066	0.238	0.111	0.937	0.074
		COMET-H	0.165	**0.173**	**0.909**	0.046	0.145	**0.042**	**0.973**	0.055
SimGlucose	Hyperglycemia	ForecastCF	**0.984**	**0.056**	0.368	**0.853**	**0.964**	0.117	0.889	**0.823**
		COMET-C	0.858	0.108	0.776	0.758	0.873	0.095	0.935	0.720
		COMET-A	0.732	0.117	0.935	0.624	0.727	0.009	**0.999**	0.457
		COMET-H	0.726	0.116	**0.935**	0.618	0.727	**0.008**	**0.999**	0.457
	Hypoglycemia	ForecastCF	0.115	0.143	0.213	0.041	**0.163**	0.285	0.856	**0.019**
		COMET-C	**0.215**	0.065	0.928	**0.062**	0.094	0.171	0.955	0.009
		COMET-A	0.155	0.031	0.998	0.052	0.103	0.020	**0.999**	0.004
		COMET-H	0.155	**0.031**	**0.998**	0.052	0.103	**0.018**	**0.999**	0.004

towards the desired forecasting outcome. Our experiments with two diabetes patient datasets showed that the proposed approach outperformed the baseline ForecastCF in terms of proximity and compactness while maintaining reasonable validity. Furthermore, our qualitative analysis of the patients demonstrated that COMET could suggest more plausible changes in the treatment plan to develop patient situations towards a desired glucose range. Future work can incorporate clinical experts in assessing the effectiveness and relevance of the generated counterfactuals in glucose forecasting; in general, our proposed approach can be extended into other forecasting applications by involving domain-specific constraints.

ACKNOWLEDGMENT

We thank Ohio University for granting approval for accessing the OhioT1DM dataset; all data contributors signed informed consent with Ohio University and were fully de-identified in accordance with 45 CFR 164.514(b).

REFERENCES

[1] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, 2021.

[2] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, 2021.

[3] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and Diagnose: Clinical Time Series Analysis Using Attention Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[4] 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*, 2018.

[5] E. Huang, R. Wang, U. Chandrasekaran, and R. Yu, "Aortic Pressure Forecasting With Deep Learning," in *Computing in Cardiology*, 2020.

[6] American Diabetes Association Professional Practice Committee, "6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes—2024," *Diabetes Care*, vol. 47, pp. S111–S125, Dec. 2023.

[7] G. Noaro, G. Cappon, M. Vettoretti, G. Sparacino, S. D. Favero, and A. Facchinetti, "Machine-Learning Based Model to Improve Insulin Bolus Calculation in Type 1 Diabetes Therapy," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 1, pp. 247–255, Jan. 2021.

[8] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–13, Jul. 2021.

[9] R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki, and H. Suominen, "Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network," in *IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021.

[10] C. Duckworth, M. J. Guy, A. Kumaran, A. A. O'Kane, A. Ayobi, A. Chapman, P. Marshall, and M. Boniface, "Explainable Machine Learning for Real-Time Hypoglycemia and Hyperglycemia Prediction and Personalized Control Recommendations," *Journal of Diabetes Science and Technology*, vol. 18, no. 1, pp. 113–123, Jan. 2024.

[11] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electronic Journal*, 2017.

[12] Z. Wang, I. Miliou, I. Samsten, and P. Papapetrou, "Counterfactual Explanations for Time Series Forecasting," in *IEEE International Conference on Data Mining (ICDM)*, Dec. 2023, pp. 1391–1396.

[13] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *International Conference on Learning Representations*, 2020.

[14] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "LSTMs and Neural Attention Models for Blood Glucose Prediction: Comparative Experiments on Real and Synthetic Data," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019.

[15] C. Marling and R. Bunescu, "The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020," *CEUR workshop proceedings*, vol. 2675, pp. 71–74, Sep. 2020.

[16] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021.

[17] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction." in *KHD@ IJCAI*, 2018, pp. 64–78.

[18] F. Petropoulos and et al., "Forecasting: theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, Jul. 2022.

[19] U. Schlegel, D. L. Vo, D. A. Keim, and D. Seebacher, "TS-MULE: Local Interpretable Model-Agnostic Explanations for Time Series Forecast Models," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2021, pp. 5–14.

[20] I. Karlsson, J. Rebane, P. Papapetrou, and A. Gionis, "Explainable Time Series Tweaking via Irreversible and Reversible Temporal Transformations," *IEEE International Conference on Data Mining (ICDM)*, 2018.

[21] E. Delaney, D. Greene, and M. T. Keane, "Instance-Based Counterfactual Explanations for Time Series Classification," in *Case-Based Reasoning Research and Development*, 2021.

[22] Z. Wang, I. Samsten, I. Miliou, R. Mochaourab, and P. Papapetrou, "Glacier: guided locally constrained counterfactual explanations for time series classification," *Machine Learning*, vol. 113, no. 3, Mar. 2024.

[23] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA Type 1 Diabetes Simulator," *Journal of Diabetes Science and Technology*, vol. 8, pp. 26–34, Jan. 2014.