Universiteit Leiden

Master Computer Science

Applying Counterfactual Explanations and Multivariate Forecasting to Medical Prediction Tasks

Name:	Tomke Meyer
Student ID:	s2231086
Date:	dd/mm/2025
Specialisation:	Bioinformatics
1st supervisor:	Jan van Rijn
2nd supervisor:	Panagiotis Papapetrou

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Contents

1 Introduction

The goal of this thesis is to gain insights into the application of multivariate forecasting and counterfactual explanations to medical prediction tasks.
The increasing availability of large and complex healthcare datasets has accelerated the adoption of machine learning (ML) and deep learning (DL) techniques for clinical decision support [ASFWHY+25, MPS19]. In particular, time series forecasting plays a vital role in monitoring patient states and predicting clinical outcomes based on continuously recorded data, such as vital signs and laboratory results [KCS+20, JPS+16]. However, despite the progress in predictive accuracy, many DL models remain black boxes, limiting their interpretability and acceptance in critical healthcare environments.
Counterfactual explanations (CEs) have emerged as a promising approach to enhance model interpretability by identifying minimal and actionable changes in input features that would lead to a more desirable prediction [VDH21, Gui24]. While CEs have been widely studied in static tabular data contexts, their extension to time series forecasting—especially in multivariate medical data—is still underexplored. Existing methods for counterfactual time series mainly focus on altering historical data points, which is impractical in real-world scenarios where the past is fixed [WSMP24].
This thesis aims to address this knowledge gap by developing a novel counterfactual forecasting framework that proposes actionable changes to exogenous variables during the forecast horizon, enabling dynamic and interpretable decision-making in healthcare. We leverage and integrate classical statistical models and state-of-the-art deep learning architectures to learn the relationships between target variables and exogenous inputs, facilitating constrained forecasting that aligns with clinical goals.
Our main contributions are summarized as follows:

- We introduce a new counterfactual time series forecasting method that modifies exogenous variables in the forecast horizon to achieve desired constrained outcomes.
- We incorporate various forecasting models—including SARIMAX, OLS, GRU, and NBEATSx—to learn interpretable relationships between exogenous inputs and target variables, ensuring actionable insights.
- We validate our approach on two healthcare applications, glucose level prediction in diabetes management and heart failure with preserved ejection fraction (HFpEF), demonstrating its practical utility.

[**TODO:** more thesis introduction, what is the goal, what is the motivation] The global population is both growing and ageing, which is resulting in a rise in chronic diseases. This is resulting in a demand for more efficient, personalised, and proactive healthcare solutions [HSL+23]. Previously, treatment was only provided once symptoms had appeared, and treatment plans were very general. Nowadays, this approach is no longer sufficient, as patients are in need of a more personalised approach to their care. Technological advancements in the field of healthcare have had a significant impact on healthcare providers and patients, with statistical methods being used to predict outcomes. This approach struggles with the complex nature of clinical, demographic and molecular factors that influence the disease progression, leading to Machine Learning (ML), Deep Learning (DL), Artificial Intelligence (AI), and Big Data Analytics becoming increasingly popular fields within the medical and health science domains [ASFWHY+25]. Recent healthcare has been characterised by an increased need for data-driven approaches, with the care process being driven by the flow of data between patients and doctors, and the sharing of decisions, instructions

and information amongst care providers. The role that data and information play in decision making and provision of healthcare, has only increased with the growing digitisation of healthcare. This results in great amounts of data, which enables the implementation of advanced analytical methods, including ML and AI, to derive valuable and actionable insights. These insights are essential in supporting decision-making processes, ensuring high quality patient care, responding to real-time situations and ultimately reducing mortality. Especially ML becomes more and more relevant in healthcare applications, including predictive analysis, treatment optimisation, and patient monitoring [MPS19]. ML algorithms can potentially be used to improve diagnostic accuracy, as well as support early disease detection and prediction. Other applications include, analysing medical imaging data, such as X-rays and MRIs, to detect signs of cancer or neurological disorders, which allows for early diagnosis, as well as a personalised treatment plan for each patient. These uses show the potential ML has for both research and clinical trials, and support the improvement of healthcare overall [RNZ17], [JHS+22].

Time series analysis uses an ordered sequence of data points recorded over time, usually at regular intervals, to understand patterns, trends, and relationships within the data over time. This time series data also allows for forecasting or predicting future values, based on historical observations. Time series forecasting plays a crucial role in a number of different applications. Some applications include its usage in finance, for predicting stock prices and market trends [MSG14], in meteorology for weather and climate forecasting [KS20], in transportation, for effective traffic flow forecasting [LBF13] and especially recently in healthcare. In healthcare, time series forecasting is increasingly relevant, as patient data such as heart rate, blood pressure, glucose levels, and laboratory results are being collected continuously or in regular intervals. Precise forecasting of these medical time series can allow for early detection of negative outcomes, support the personalisation of treatment plans, and optimise healthcare management overall [KCS+20], [JPS+16].

[**TODO:** Extend, add citations] Generally, time series forecasting consists of finding temporal dependencies and trends in the data. Depending on the nature of the time series, such as for example linear or non-linear, stationary or non-stationary, and univariate or multivariate, requires different forecasting methods. These different forecasting methods can be categorised into three subgroups, statistical models, such as ARIMA and ETS, ML models, such as Linear Regression and SVRs, and DL models, such as RNNs, LSTMs, GRUs as well as transformer-based models Autoformer or DLinear.
Especially the latter models have shown promising results in recent years, achieving very accurate predictions. Despite these recent developments, many deep-learning based forecasting systems are considered "black-box"-models, as it is challenging to interpret and understand both the modelling process and the forecasting outcome. This is particularly problematic in healthcare applications, where especially interpretability is very important, as clinicians need to understand the reasoning behind the predictions to make informed decisions. One approach to tackle this problem is working with counterfactual explanations, which aim to identify small changes in the input variables needed to change a model's forecast to a more desirable outcome.
In healthcare, time series forecasting has been applied to a number of different tasks, especially the analysis of electronic health records (EHRs). These contain information about a patient's health over time, allowing predictions about the disease progression,

treatment efficacy or patient mortality. For example, forecasting glucose levels in diabetic patients or predicting heart failure with preserved ejection fraction (HFpEF) can aid early interventions and optimise treatment plans.
Traditional time series forecasting focuses mainly on predicting future values given historical observations, but modifying past values is not feasible in real-world settings, especially in healthcare applications. Instead, a more practical approach is to explore how changing exogenous variables during the forecast horizon could lead to a desired outcome. This approach allows continuous monitoring of patients, making it possible to dynamically adjust treatment plans to lead towards more optimal results.

1.1 Diabetes

Diabetes is one of the most prevalent chronic diseases in the world, with it being a leading cause of death and disability. According, to the World Health Organisation (WHO), around 830 million people worldwide suffer from diabetes, and it is the direct cause of 1.5 million deaths a year. Nowadays, it affects around 14% of adults, this number has doubled since 1990, making it a major public health problem. Diabetes is characterised by elevated levels of blood glucose, which over time seriously damages the heart, blood vessels, eyes, kidneys and nerves. Type 2 diabetes is much more common and occurs when the body either becomes resistant to insulin or does not make enough insulin. It usually has a later onset and its development can be attributed to factors such as being overweight, not getting enough exercise and genetic. Type 1 Diabetes is a chronic condition, where the pancreas only produces little to no insulin by itself. It is caused by the autoimmune destruction of pancreatic β-cells and affects 5-10% of the diabetes patients [Dia], [EMA+24a]. For patients with conditions such as type 1 diabetes mellitus (T1DM), closely tracking their glucose levels is a necessity. To reduce the risk of complications such as hyperglycaemia or hypoglycaemia, these patients rely on continuous glucose monitoring (CGM) devices and automated insulin delivery. Using machine learning (ML) to model CGM, can help to gain a better understanding of predicting abnormal glucose events and help with insulin dosage planning. By also incorporating variables such as insulin intake, carbohydrate consumption, and physical activity, a predictive model can allow timely interventions through the generation of actionable recommendations for patients or healthcare providers. This allows for more dynamic treatment based on these forecasted expected glucose trends, which can reduce the long-term risk of diabetes-related complications [EMA+24b].

1.2 Heart failure with preserved ejection fraction

Heart failure with preserved ejection fraction (HFpEF) is a prevalent and severe cardiovascular condition [BP10]. Heart failure (HF) is classified based on the left ventricular ejection fraction (LVEF) and can be split into three subcategories. Heart failure with reduced ejection fraction (HFrEF) with LVEF $\leq$40%, heart failure with mildly reduced ejection fraction (HFmrEF) with LVEF = 41-49%, and heart failure with preserved ejection fraction (HFpEF) with LVEF $\geq$50% [MMA+21]. Studies have shown that the mortality within one year is around 29%, [OHH+06], [SBA+24], with increased mortality for patients with previous heart failure hospitalisations and other comorbidities [MGL+20]. This makes HFpEF a very serious condition, where early diagnosis is key. So another

possible application of medical time series forecasting could be to try identifying early warning signs of heart failure or more specific HFpEF to suggest either lifestyle or treatment changes. Using the vital signs, as for example the heart rate and blood pressure, of a patient as well as other factors like gender and possible comorbidities, allows specific and personal monitoring of disease progression. This way, early warning signs of worsening heart condition can be identified and personalised modifications to lifestyle or medication can be suggested. Such proactive monitoring can help reduce hospital readmissions and improve patient outcomes.

1.3 Counterfactual Explanations

Counterfactual explanations (CE) are an emerging technique with the potential to improve the interpretability and explainability of ML models. The objective is to identify minimal changes to the input features that would result in a different, and typically more desirable, outcome. More specifically, they provide information such as if an input data point would be X' instead of X, then the trained ML models prediction would be Y' instead of Y, assuming outcome Y' would be more favourable [VDH21]. So counterfactuals are a type of explanation method in ML, that might help users to understand the model predictions and decisions better. This is supported by counterfactual explanations being actionable, since they suggest specific changes to alter the outcome, and intuitive for humans, since they align with human reasoning about cause and effect. Often CEs are also model-agnostic, meaning they can be applied to a variety of black-box models [Gui24]. A more specific example application of counterfactuals would be a suggestion that slightly reducing a patients systolic blood pressure could have led to a prediction of a lower cardiovascular risk. So these kind of insight can not only help clinicians understand a models decisions better, but also provide help for intervention and treatment planning.

1.4 Main Contributions

[**TODO:** Check, rewrite?] Time series forecasting can be quite useful in healthcare by predicting how well a treatment works or predicting the risk of complications, relapse or mortality. Recent studies, such as COMET [WSMP24], have been focusing on counterfactuals in time series forecasting, but these methods work by altering historical observations. Our work aims to bridge this gap by developing a counterfactual forecasting mechanism that identifies optimal changes in exogenous variables during the forecast horizon to achieve desirable outcomes. More specifically, we propose a method that learns the relationship between the forecasted targets and exogenous variables, which leads to a more effective and interpretable decision-making in healthcare.
The main contributions of this thesis can be summarised as follows:

- We propose a new model for counterfactual time series forecasting to achieve a desired constrained forecast by modifying exogenous variables within the forecast horizon.

- We incorporate existing forecasting models, such as SARIMAX, OLS, GRU and NBEATSx, for learning the relationship between exogenous variables and a target variable to ensure actionable and interpretable predictions.

- We evaluate the models on two applications in healthcare, specifically for glucose level prediction and HFpEF management, demonstrating the practical utility of our approach.

1.5 Problem Formulation

1) Let $\boldsymbol{X} := (\boldsymbol{x}_i)_{i\in\{1,\dots,n\}}$ be a time series of length n (*back horizon*), with each $x_i \in \mathbb{R}^{m+1}$ composed of the target variable $y_i \in \mathbb{R}$ and the exogenous variables

$$z_i = \begin{pmatrix} z_{1,i} \\ \vdots \\ z_{m,i} \end{pmatrix} \in \mathbb{R}^m.$$

Then $\boldsymbol{X}$ can be denoted as the combined matrix of the target vector $\boldsymbol{y} \in \mathbb{R}^{1\times n}$ and the exogenous matrix $\boldsymbol{Z} \in \mathbb{R}^{m\times n}$:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{Z} \end{pmatrix} := \begin{pmatrix} y_1 & \dots & y_n \\ z_{1,1} & \dots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{m,1} & \dots & z_{m,n} \end{pmatrix}.$$

The relationship between $\boldsymbol{y}$ and $\boldsymbol{Z}$ can be described by the function:

$$r : \mathbb{R}^{m\times n} \to \mathbb{R}^{1\times n}$$

which calculates the target vector from the exogenous matrix:

$$r(\boldsymbol{Z}) = \boldsymbol{y}.$$

2) Given a time series forecasting model f that predicts the next t values (*forecasting horizon*) of $\boldsymbol{X}$, we define the forecast as:

$$f(\boldsymbol{X}) = \boldsymbol{X'} := (x_{n+i})_{i\in\{1,\dots,t\}}.$$

Additionally, consider a lower and upper bound set of constraints for each time step in the forecasting horizon, denoted as:

$$\boldsymbol{\alpha} = (\alpha_{n+i})_{i\in\{1,\dots,t\}}, \boldsymbol{\beta} = (\beta_{n+i})_{i\in\{1,\dots,t\}}.$$

The objective is to generate a counterfactual time series sample $\boldsymbol{Z}^*$, such that $\boldsymbol{y}^* = r(\boldsymbol{Z}^*)$ satisfies the given bounds:

$$\alpha_i \le y_i^* \le \beta_i, \forall y_i^* \in \boldsymbol{y}^*, i \in \{n+1, \dots, n+t\}.$$

Summarized research objective: Given a target vector $\boldsymbol{y}$ affected by the exogenous matrix $\boldsymbol{Z}$, a forecast horizon t, the original forecasted vector $\boldsymbol{y'}$ and the original forecasted exogenous matrix $\boldsymbol{Z'}$, the goal is to modify $\boldsymbol{Z'}$ to $\boldsymbol{Z}^*$ such that the corresponding target vector $\boldsymbol{y}^*$ is within constraints $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

2 Related Work

[**TODO:** check, add citations, rewrite]

2.1 Related Forecasting Models

Recent research has explored various deep learning (DL) models for time series forecasting, including recurrent neural network (RNN)-based models such as gated recurrent units (GRU) and long short-term memory (LSTM), as well as attention-based architectures like transformers. Transformer-based models, including Autoformer and Informer, have demonstrated strong performance in both univariate and multivariate forecasting tasks by capturing long-range dependencies more effectively than traditional RNN approaches. In the clinical domain, deep learning models have been applied extensively to glucose forecasting. For instance, Deep Multi-Output Forecasting [FAJ+18] introduced a multi-step forecasting framework that explicitly models the distribution of future glucose values over a prediction horizon using a multi-output deep architecture. Similarly, WaveNet has been adapted for glucose forecasting by leveraging dilated convolutional neural networks (CNNs) to model long-term dependencies [ZLH+18]. In addition, transfer learning techniques have been employed to enhance predictive performance by fine-tuning pre-trained models on patient-specific data while incorporating exogenous covariates such as insulin dosage and carbohydrate intake [MB20].

Beyond predictive performance, explainability remains a critical challenge in deep learning-based forecasting models. Traditional statistical models, such as ARIMAX and VARIMAX, are able to quantify relationships between exogenous factors and the target variable, but their forecasting accuracy is often outperformed by DL approaches. Recent research has focused on integrating explainability into forecasting models to combine the strengths of both interpretability and predictive performance. For example, NBEATSx extends the NBEATS framework by incorporating exogenous variables into its deep architecture, enabling a more structured decomposition of trend and seasonality. However, its interpretability remains static and does not fully capture the dynamic nature of forecasting outcomes.

To address the need for explainability, counterfactual explanations have gained traction in time series analysis. Initial efforts focused on time series classification, where counterfactuals were generated through instance-based modifications and gradient-based perturbations [AALC20]. This was done by introducing a framework for generating counterfactual explanations for multivariate time series classification, identifying minimal input modifications needed to alter the model's decision, providing interpretability for high-dimensional time series models.

More recently, counterfactual explanations have been extended to time series forecasting. ForecastCF [WMSP23] proposed a deep learning-based method for generating counterfactuals in time series forecasting by identifying minimal input changes required to achieve desired prediction outcomes. Building on this, COMET [WSMP24] extended counterfactual explanations to multivariate time series forecasting, focusing on modifying exogenous variables (e.g., insulin, carbohydrates, and exercise) to generate actionable recommendations for glucose management.

In addition to counterfactual forecasting, diagnostic tools like TimeTuner [HSYZ23] have been developed to analyse how time representations influence forecasting models. By

employing counterfactual explanations, TimeTuner enables the evaluation of multivariate time series representations and their impact on model predictions.
Despite these advances, counterfactual explanations for multivariate time series analysis remain an emerging research area. While existing methods demonstrate the feasibility of generating counterfactuals for univariate forecasts, their generalisation to multivariate forecasting and real-world clinical applications remains limited. This work aims to extend on these existing methods by integrating counterfactual reasoning with multivariate forecasting models, focusing on modifying exogenous variables within the prediction horizon to provide actionable and interpretable interventions.

2.2 Related Work in Glucose Forecasting

Forecasting physiological indicators such as blood glucose levels, is crucial for managing diabetes. Time series forecasting and model explainability are becoming increasingly important in this field of medical prediction tasks, as in many others. Recent contributions to this area of research highlight the growing use of multivariate machine learning models to improve predictive accuracy and personalisation. This shows the growing need for interpretable and actionable insights, which aligns closely with the goals of this thesis.
Recent research highlights the growing development of multivariate and deep learning-based models for predicting glucose levels. For example, Kalita and Mirza [KM25] proposed a model that combines multi-head attention layers with neural basis expansion networks, capturing complex temporal and cross-feature dependencies in glucose data. Similarly, Benaida et al. [BAI25] demonstrated the effectiveness of deep learning architectures for both single- and multi-step glucose forecasting, emphasising the importance of longer-term prediction capabilities in real-world applications. These multivariate models are consistent with the focus of this thesis on leveraging multiple signals, such as past glucose levels, physiological parameters, and contextual variables, for accurate forecasting.
Personalisation has emerged as a key factor in clinical forecasting settings, as patient diversity affects model performance. Shen and Kleinberg [SK25a] addressed this issue by using incrementally retrained LSTM networks that adapt to each individual's glucose dynamics. This improves performance, even when the CGM data is limited. Lara-Abelenda et al. [LCP+25] introduced large language models (LLMs) to model personal glucose trends, highlighting the capacity of foundation models to generalise across individuals while retaining patient-specific nuances. These methods emphasise the importance of adaptive and context-aware forecasting.
Several works have also incorporated physiological signals beyond glucose levels to support multivariate forecasting. For example, Giancotti et al. [GBV+24] explored the utility of heart rate as a predictor of forecasting glucose levels in patients with type 1 diabetes, which demonstrates that multimodal data can significantly enhance predictive accuracy. Similarly, Rodríguez-Rodríguez et al. [RCR23] utilised data, such as physical activity and diet logs, to enable more holistic and personalised glycaemic forecasting.
Interpretability remains a major challenge for deep learning-based forecasting models, especially in critical fields such as medicine. In response to this, Sun and Kosmas [SK25b] combined Bayesian forecasting with expert medical knowledge to model CGM values in type 2 diabetes patients. Their framework improves both uncertainty quantification and clinician interpretability, which is an essential consideration in healthcare AI. The need for model transparency directly motivates the use of counterfactual explanations to improve

the explainability and actionability of predictive models in medical applications. While many current studies emphasise predictive accuracy, fewer address how predictions can be explained and acted upon by clinicians or patients.
Taken together, these studies reflect a shift towards data-driven, multivariate, and personalised models for medical forecasting. However, there remains a clear gap in integrating these powerful models with robust, interpretable explanations. This thesis aims to bridge this gap by combining multivariate forecasting approaches with counterfactual reasoning, to provide accurate predictions and actionable, understandable explanations, which are essential components for supporting medical decision-making and patient self-management.

2.3 Related Work in Mortality Prediction of Patients with HFpEF

Heart Failure with Preserved Ejection Fraction (HFpEF) is a complex and heterogeneous condition, characterised by diagnostic and prognostic uncertainty. This makes it a compelling use case for machine learning (ML) in clinical decision support. Recent research has applied various ML techniques to improve diagnosis, predict outcomes such as hospitalisation and mortality, and guide individualised management strategies. These efforts emphasise the increasing relevance of multivariate forecasting and the growing demand for explainable models, which are central to the objectives of this thesis.
A significant amount of research has focused on prognostic modelling using structured clinical data. For example, Hu et al. [HMH+25] developed and validated a machine learning model to predict the risk of readmission within one year for HFpEF patients, demonstrating the utility of routinely collected electronic health records (EHRs) in anticipating adverse outcomes. Similarly, McDowell et al. [MKT+24] constructed models for predicting both mortality and morbidity in HFpEF patients, showing that complex risk factors, including comorbidities and laboratory values, can be effectively integrated into predictive models. These studies emphasise the importance of leveraging multivariate data sources to forecast long-term patient outcomes.
Short-term outcome prediction has also been explored, particularly in the context of the early identification of high-risk patients. The study "Predicting 30-Day and 1-Year Mortality in HFpEF" [SBA+24] used machine learning to predict short-term mortality, which is essential for planning acute care. Other models have focused on hospitalisation prediction, using historical patient trajectories to anticipate future events. These forecasting tasks not only require accurate time series modelling but also benefit from interpretability to inform clinical decisions.
The diagnosis of HFpEF remains a challenging area due to its symptomatic overlap with other heart failure subtypes. Kavas et al. [KBKB23] developed an ML-based decision support system using photoplethysmography (PPG) signals to differentiate between HFpEF and HFrEF (Heart Failure with reserverd Ejection Fracion), demonstrating the potential of non-invasive, sensor-based diagnostics. Liao and Hung [LH+24] further extended this approach by incorporating data from a wearable patch device to enhance diagnostic precision. These works highlight the growing role of physiological signal data in heart failure classification, which directly supports multivariate modelling approaches by introducing continuous and high-frequency signals into prediction tasks.
Genomic and molecular data have also been used to support precision medicine approaches in HFpEF. Zhou et al. [ZGW+21] utilised gene expression profiles to build ML models ca-

pable of risk stratification in HFpEF patients, adding a layer of biological interpretability to purely clinical models. Although these models are powerful, they are often perceived as "black boxes," emphasising the need for explainability techniques such as counterfactual explanations to bridge the gap between model prediction and clinical insight.
Across these studies, however, the challenge of model transparency and interpretability remains largely unaddressed. Most existing models prioritise predictive performance without offering sufficient explanations for individual predictions, which is a critical issue in medical contexts where understanding why a prediction was made is often as important as the prediction itself. This thesis aims to bridge this gap by integrating counterfactual reasoning into multivariate forecasting models, offering clinicians not just a forecast, but a clear explanation of the factors driving the prediction and the minimal changes that could alter an adverse outcome.
In summary, the current research in HFpEF prediction demonstrates the power of machine learning to handle complex, multivariate data across diagnostic and prognostic applications. However, a lack of interpretability limits clinical adoption. This thesis tries to contribute to the field by combining accurate time series forecasting with interpretable, counterfactual explanations, thereby supporting more transparent and actionable decision-making in heart failure care.

3 Methodology

3.1 Algorithm

[**TODO:** update, check] The proposed algorithm is designed to generate counterfactual exogenous inputs that lead to a desired change in the forecasted target variable within a multivariate time series context. It operates by integrating a differentiable forecasting model and a counterfactual optimization process constrained by both feasibility and historical plausibility. The algorithm takes multivariate time series data as input, which is split into a target variable t (for example blood glucose level) and associated exogenous variables E (for example insulin dosage, carbohydrate intake, activity level). Several hyperparameters are defined: the learning rate η, target bounds (α, β), a clipping range (ρ, ϕ), and the maximum number of optimization iterations max_iter. Additionally, differentiable models for forecasting $f(\cdot)$ and counterfactual generation $c(\cdot)$ are provided, where f predicts the target and c estimates the outcome, given perturbed inputs. The goal is to identify a perturbed exogenous input E' that results in a forecasted target t' within the desired bounds. The algorithm consists of different stages:

- Multivariate Forecasting: The forecasting model f is applied to the input sequence to obtain predicted targets $\hat{t}^*$ and exogenous outputs $\hat{E}^*$. These serve as a reference for evaluating the effect of the perturbations.

- Sample and Bound Selection: A representative subset $\mathcal{S}$ of the test data is selected to compute statistical bounds (α, β) on the desired target outcome. These bounds guide the optimization process by defning what is considered a succesful counterfactual prediction

- Loss Initialization: A composite loss function L is initialized. It incorporates forecasting error, a regularization term to penalize unrealistic perturbations, an activity

temporal constraint and a historical similarity constraint.

- Iterative Counterfactual Optimization: Counterfactuals are generated through gradient-based optimization, with the Adam optimizer. In each iteration the loss is back propagated with respect to $\hat{E}^*$, and updates are applied. Then a clipping function ensures that the perturbed values stay within predefined limits (ρ, ϕ) and a historical similarity constraint $\mathcal{G}$ encourages solutions to be close to real historical values, to enhance realism and interpretability.
- Stopping criteria: The loop terminates once the forecasted target falls within the desired bounds or when the maximum number of iterations is reached.
- Return Output: The final optimized exogenous input E' and its associated target t' are returned as the counterfactual explanation.

This algorithmic approach provides an adjustable and interpretable method for generating actionable interventions in time series forecasting tasks. It ensures that the generated counterfactuals are both effective, as they achieve the target, and feasible, as they are within realistic operational constraints, making it a suitable method for the delicate field of healthcare. Algorithm 1 shows the pseudocode for the described method.

3.1.1 Constraints

[**TODO:** check, rewrite, improve] To ensure that the generated counterfactual trajectories are not only effective in altering predicted outcomes but also clinically feasible and temporally realistic, we introduce a set of constraint mechanisms directly into the optimization process. These are the clipping constraint, the historical value constraint, and the activity temporal constraint. Each plays a distinct role in maintaining interpretability, trustworthiness, and clinical plausibility of the counterfactuals.
These constraints are applied iteratively during optimization, guiding the perturbation of exogenous variables such that the resulting trajectories adhere to domain-specific boundaries and realistic temporal patterns.

3.1.1.1 Clipping Constraint

The clipping constraint ensures that the perturbed exogenous variables $\hat{E}^*$ remain within a realistic value range. These bounds (ρ, ϕ) are selected based on physical limits, clinical safety, or operational feasibility (e.g., insulin dosages must be non-negative and within a patient-safe maximum range).
After each optimization step, every element in $\hat{E}^*$ is clipped such that:

$$\hat{E}_*^{i,j,k} = \min(\phi_k, \max(\rho_k, \hat{E}_*^{i,j,k}))$$

for each sample i, time step j, and feature k. This ensures that the generated counterfactuals remain within a trusted domain of possible values.

3.1.1.2 Historical Value Constraint

The historical value constraint is introduced to promote plausibility by encouraging counterfactual inputs that are statistically similar to observed historical inputs. This constraint

acts as a regularization mechanism, penalizing counterfactuals that deviate too far from past observed patterns. It leverages a historical dataset $\mathcal{G}$ to compute a similarity or distance metric—often based on Euclidean or Mahalanobis distance.
Formally, a constraint term C is computed based on the deviation of $\hat{E}^*$ from the nearest historical samples:

$$C(\hat{E}^*) = \lambda \cdot \min_{E \in \mathcal{G}} \|\hat{E}^* - E\|_2$$

where λ is a hyperparameter that controls the strength of this constraint. The closer $\hat{E}^*$ is to a real historical sample, the lower the penalty.
This mechanism ensures that the generated exogenous sequences are not only valid within a static range but also dynamically aligned with real-world behavior.

3.1.1.3 Activity Temporal Constraint

The activity temporal constraint imposes structure on *when* changes to features are allowed during the time series. This is useful in scenarios where only certain time steps are clinically actionable or when domain knowledge suggests that interventions (e.g., medication, exercise, meals) are expected at specific moments.
This constraint is encoded using a binary vector $C \in \{0, 1\}^T$, where each element C_j corresponds to a time step j. A value of $C_j = 0$ allows counterfactual perturbations at that time, while $C_j = 1$ discourages them.
We define the constraint as:

$$C_j = \begin{cases} 0 & \text{if } \hat{E}^j > \gamma \\ 1 & \text{otherwise} \end{cases}$$

where γ is a user-defined threshold (e.g., $\gamma = 0$) that determines whether the feature value at time j indicates meaningful activity. This ensures changes happen primarily at relevant time steps, avoiding unrealistic shifts in periods where intervention is clinically unlikely or inappropriate.
The corresponding activity penalty is computed as:

$$C_{\text{act}}(\hat{E}^*, E) = \sum_j C_j \cdot \left| \frac{\hat{E}^j - E^j}{E^j + \epsilon} \right|$$

where ϵ is a small constant to prevent division by zero. This encourages local changes to occur only when and where they are contextually valid.

3.1.1.4 Combined Loss Function

The final optimization objective integrates all the above constraints with the primary forecasting goal. The loss function is composed of:

- **Forecast loss**: Penalizes predictions $\hat{t}^*$ that fall outside the desired outcome bounds (α, β).
- **Historical deviation penalty**: Encourages similarity to real historical trajectories.

- **Activity temporal penalty**: Controls when temporal changes can occur.

The total loss is defined as:

$$L = \text{ForecastLoss}(\hat{t}^*, \alpha, \beta) + \lambda_{\text{hist}} \cdot C_{\text{hist}}(\hat{E}^*) + \lambda_{\text{act}} \cdot C_{\text{act}}(\hat{E}^*, E)$$

Here, λ_{hist} and λ_{act} are hyperparameters that tune the importance of each constraint. The clipping constraint is applied through projection and is not explicitly included in the loss function.
These three constraints work together to guide the counterfactual generation process as the clipping constraint ensures that variable values remain in a trusted clinical range, the historical value constraint keeps the counterfactuals close to real, observed trajectories, and the activity temporal constraint encourages changes to occur at appropriate, clinically meaningful time steps. By embedding these constraints into the optimization loop, the generated counterfactuals are not only effective but also interpretable, trustworthy, and aligned with realistic clinical dynamics.

Algorithm 1 Counterfactual Hybrid Forecasting

Input: Time series data: target t, exogenous variables E, learning rate η, desired bounds (α, β), clipping range (ρ, ϕ), max iterations max_iter, differentiable forecaster $f(\cdot)$ differentiable counterfactual generator $c(\cdot)$, w, $\mathcal{G}$
Output: Counterfactual E' with desired outcome t'
$(t^*, E^*) \leftarrow (t, E)$
$(\hat{t}^*, \hat{E}^*) \leftarrow f(t^*, E^*)$
$\mathcal{S} \leftarrow$ SelectTestSamples
$(\alpha, \beta) \leftarrow$ GenerateBounds$(\mathcal{S})$
$C \leftarrow$ ActivityTemporalConstraint
$loss \leftarrow L((t^*, E^*), \alpha, \beta, (t, E), C)$
time $\leftarrow 0$
while $(\hat{t}^* > \beta$ **or** $\hat{t}^* < \alpha) \wedge (\text{time} < \text{max\_iter})$ **do**
    $\hat{E}^* \leftarrow \text{AdamOptimize}(\hat{E}^*, loss, \eta)$
    $\hat{E}^* \leftarrow \text{Clip}(\hat{E}^*, \rho, \phi)$
    $\hat{t}^* \leftarrow c(\hat{E}^*)$
    $C \leftarrow \text{HistValueConstraint}(\hat{E}^*, \mathcal{G})$
    loss $\leftarrow L(\hat{E}^*, w, \alpha, \beta, X^*, C)$
    time $\leftarrow$ time $+ 1$
end while
$(t', E') \leftarrow (\hat{t}^*, \hat{E}^*)$
return (t', E')

3.2 Data

While the proposed model has broad applicability across domains beyond healthcare, this study focuses exclusively on medical use cases. In particular, we investigate the models usefulness in optimizing treatment plans for two conditions: type 1 diabetes and heart failure with preserved ejection fraction (HFpEF). The datasets used in these experiments

contain physiological measurements and treatment-related variables, allowing for personalized forecasts and counterfactual intervention generation. To prepare the data for modelling, several preprocessing steps have been applied. First, the data is split into training, validation and test sets and normalized using min-max scaling to ensure stable and consistent model training. The data is then separated into target variables and exogenous inputs. Then, a sequence generator is employed to segment the time series into overlapping windows comprising a back horizon (historical input) and a prediction horizon (target output). This is done both for the target variable alone and for sequences that include both the target and exogenous features. Any sequences containing missing values are discarded to ensure data integrity.

3.2.1 SimGlucose

The SimGlucose dataset is generated using the FDA-approved UVA/PADOVA type 1 diabetes simulator [Xie18], a Python-based tool that models the physiological responses of individuals with type 1 diabetes. The simulator includes 30 virtual patients, comprising of 10 adults, 10 adolescents, and 10 children and produces continuous glucose monitoring (CGM) measurements along with insulin dosages and carbohydrate intake events. The dataset is generated with a predefined CGM sampling frequency and insulin pump settings based on the algorithm developed in [DMML+14]. For this study, simulated data was generated for ten adult patients over a one-week period. The blood glucose (BG) levels serve as the primary target variable, while the insulin dosage and carbohydrate intake are used as exogenous variables influencing the glucose levels. After generation, the data undergoes preprocessing steps as described above to prepare it for the forecasting and counterfactual generation. An example of the generated data is shown in Figure 1. In this example, BG denotes blood glucose levels, CHO indicates carbohydrate intake, and Insulin reflects the administered dosage. The risk index illustrates periods of hyperglycaemic or hypoglycaemic risk. The green band in the BG trace highlights the target glucose range (70–180 mg/dL), while red regions denote values that fall outside this range, corresponding to hypo- or hyperglycaemia.

3.2.2 OhioT1DM:

The OhioT1DM dataset [MB20] contains real-world glucose monitoring data collected from 12 individuals with type 1 diabetes over an eight-week period by Ohio University. Following prior research [CHN+21], [WSMP24], we extracted the most clinically relevant features for forecasting: continuous glucose monitoring (CGM) measurements, basal insulin, bolus insulin, carbohydrate intake, and physical activity. Compared to the SimGlucose dataset, OhioT1DM includes a more varied set of exogenous variables, particularly basal and bolus insulin administration, dietary intake, and physical activity data. As a real-world dataset, it presents additional challenges such as missing values and irregular sampling intervals. These are addressed using interpolation and resampling techniques to ensure consistency in the input sequences. The inclusion of diverse exogenous variables enables the development of more nuanced counterfactual interventions and supports improved forecasting performance. Figure 2 illustrates a 24-hour time window for one patient, showing the temporal relationship between glucose levels, insulin administration,

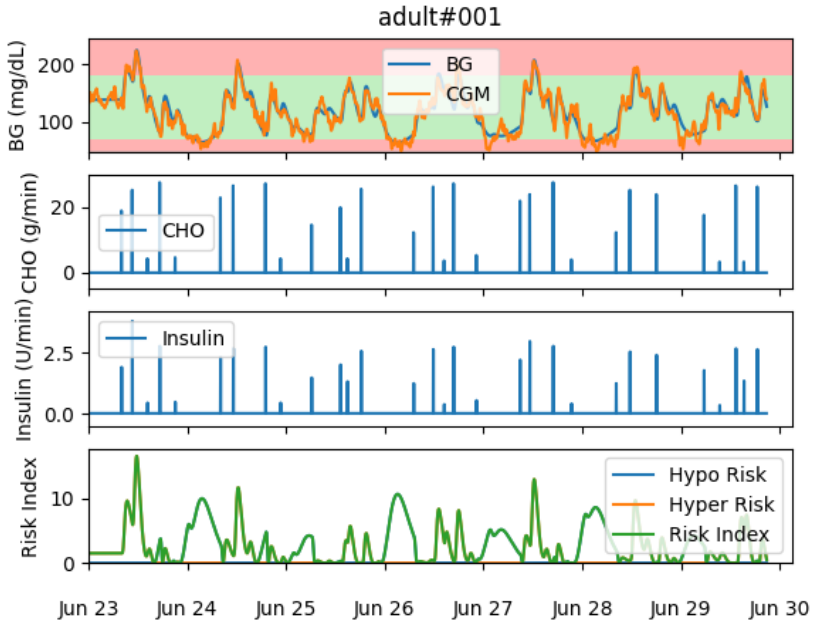


Figure 1: Simulation of an adult patient over the span of one week.

and carbohydrate consumption. The blue dotted line represents CGM-based blood glucose levels, while the black line represents the basal insulin, reflecting its slow-acting, sustained delivery throughout the day. Orange dots indicate the amount of bolus insulin, while small blue boxes indicate meals. Spikes in blood glucose often correspond to meal times (carbohydrate intake), followed by bolus insulin doses that aim to bring glucose back into the target range. This visualization exemplifies the complex dynamics and temporal dependencies that the model must capture to enable accurate and personalized glucose forecasting.

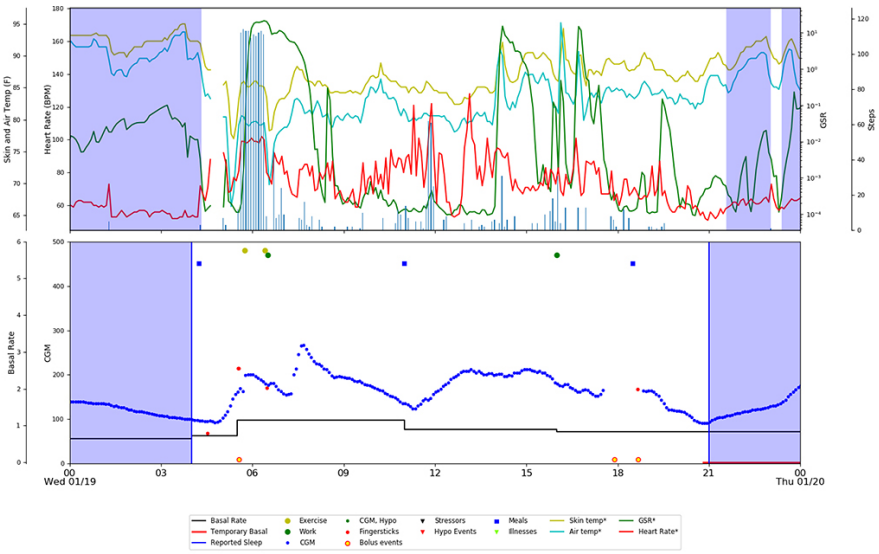


Figure 2: 24-hour measurements of one patient from the OhioT1DM Viewer [MB20].

3.2.3 MIMIC-IV:

The proposed model aims to generalize beyond diabetes forecasting to other medical applications, such as predicting disease progression in HFpEF patients. The MIMIC-III dataset [JPS+16] contains de-identified electronic health records (EHRs) of ICU patients, including vital signs (heart rate, blood pressure, oxygen saturation, etc.), medication records, and laboratory results. For this study, a subset of MIMIC-IV focusing on cardiovascular patients is utilized. The target variable, mortality risk, was divided into two groups: death within 30 days and death within one year. The exogenous variables include vital signs and laboratory values, while gender and comorbidities are used to split the data into multiple cohorts. By analysing the data in different cohorts, it is possible to get more specific and accurate counterfactual interventions.

3.2.3.1 Preprocessing

Following a preceding study[SBA+24], the International Classification of Diseases (ICD) codes were used for the initial preprocessing. The ICD-codes are a standardized international classification system used for the categorisation and encoding of diseases, symptoms, and associated health-related conditions. By using the appropriate ICD-9 and ICD-10 codes, as outlined in Table 1, the hospital admissions involving patients aged ≥ 18 with HFpEF as a primary diagnosis have been identified. Given that the diagnosis was based on ICD codes, the clinical notes were analysed in order to validate the selection of patients. This was achieved through filtering the clinical notes on mentions of the left ventricular ejection fraction (LVEF) value, with a value of 50 and above being counted as a normal LVEF value. Some clinical notes only mentioned a normal or preserved LVEF value with out a measured LVEF. These were also counted as normal LVEF values. Table 1 also shows the number of hospital admissions per diagnosis. The study sample consisted of 16122 individual hospitalizations with a suspected diagnosis of HFpEF. We had access to clinical notes for 11720 (72.7%) hospital admissions of which 4458 (38%) had an LVEF measurement reported. Of these, 3798 (85.2%) had an LVEF value $\geq 50\%$, and 400 (9%) had an LVEF $< 50\%$. An additional 260 (5.8%) admissions had mention of a normal or preserved LVEF. For these 4058 admissions, vital signs and laboratory values were available for 2432 (60%). It should be noted that in this instance, only the most recent admission of a patient who had previously been admitted with a similar diagnosis was taken into consideration. Prior admissions were incorporated into the analysis as a comorbidity, and after adding all laboratory values, vital signs and comorbidities, the resulting number of unique hospital admissions is 2113 with 1845 unique patients.
The extracted features are listed in table 2, split into four categories. These are the targets (Death within 30 days, Death within 1 year) as well as vital signs and laboratory values, comorbidities, and gender. Prior admission was included in the list of comorbidities, since it is here used as a comorbidity for the clustering and not the forecasting.

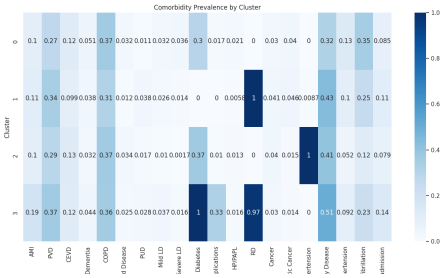
3.2.3.2 Clustering

The MIMIC data does not only include a wide range of features, but also a great number of comorbidities. Since these comorbidities can influence the chances of HFpEF, it is

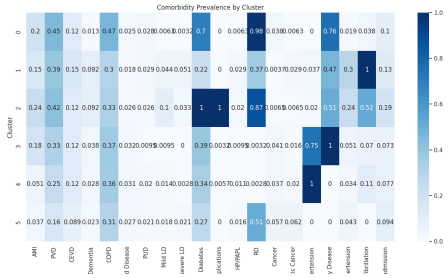
Diagnosis	ICD Code	Frequency
Unspecified diastolic (congestive) heart failure	I5030	38
Diastolic heart failure, unspecified	42830	69
Acute diastolic (congestive) heart failure	I5031	84
Acute diastolic heart failure	42381	180
Chronic diastolic (congestive) heart failure	I5032	256
Chronic diastolic heart failure	42382	453
Acute on chronic diastolic (congestive) heart failure	I5033	426
Acute on chronic diastolic heart failure	42833	623

Table 1: The corresponding Diagnosis and ICD-9 and ICD-10 codes for HFpEF.

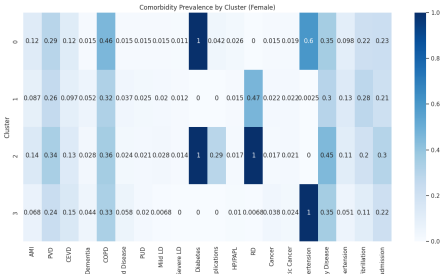
important to include these in the analysis. In this study, the comorbidities were used to cluster the patients, to get more specific counterfactuals. For this, the data was split according to patients having similar comorbidities to get factual cohorts. Another split was made by dividing the data on gender, since HFpEF is more prevalent in female patients and might need different treatment. Figure 3 shows the prevalence of different comorbidities of the MIMIC patients in the clusters, divided into four sub figures depending on the clustering coefficient and the gender. Figure 4 shows the Principal Component Analysis (PCA) of the different comorbidities, again divided into the same subgroups.



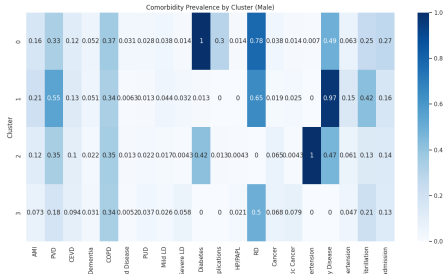
(a) Comorbidities in 4 clusters.



(b) Comorbidities in 6 clusters.

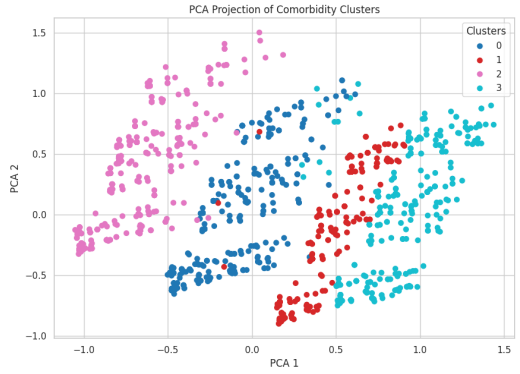


(c) Comorbidities for female patients in 4 clusters.

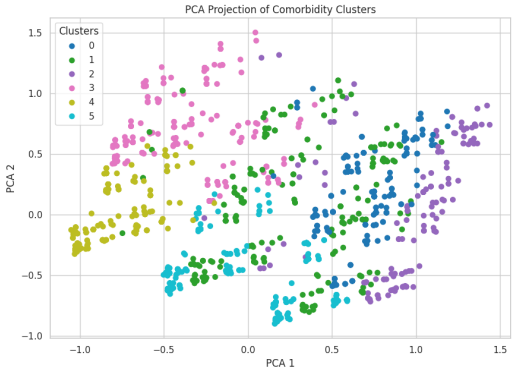


(d) Comorbidities for male patients in 4 clusters.

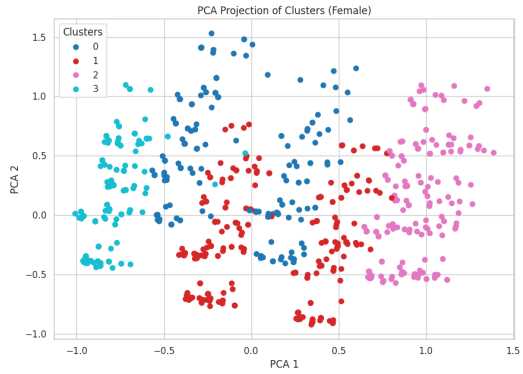
Figure 3: Prevalence of different comorbidities of the MIMIC patients in the clusters. There are 4 subgroups, clustering by comorbidity with $k = 4$, clustering by comorbidity with $k = 6$, clustering only the female patients with $k = 4$ and clustering only the male patients with $k = 4$.



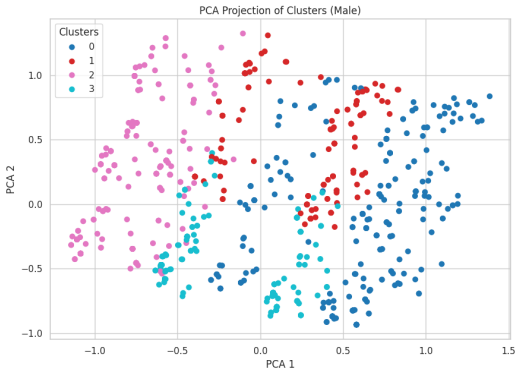
(a) PCA clustering in 4 clusters.



(b) PCA clustering in 6 clusters.



(c) PCA clustering for female patients in 4 clusters.



(d) PCA clustering for male patients in 4 clusters.

Figure 4: Principal component analysis (PCA) clustering of different comorbidities of the MIMIC patients. There are 4 subgroups, clustering by comorbidity with $k = 4$, clustering by comorbidity with $k = 6$, clustering only the female patients with $k = 4$ and clustering only the male patients with $k = 4$.

3.3 Experiments

3.3.1 Experimental Setup

The model can be split into two main parts. Initially a multivariate forecasting model is used to make a first forecast for both the target variable and the exogenous variables. This forecast is then used for the second part, where different regression models are used to change the exogenous and target variable to get the desired outcome. For the forecasting part we used GRU and NBEATS and for the second part we used four different models. Here the focus lies on using very different kinds of models, like a statistical (SARIMAX), a regression based (OLS), and two different deep learning based (GRU and NBEATSx) models.

3.3.1.1 Multivariate Forecasting:

For the multivariate forecasting task, two deep learning architectures were implemented and evaluated: (1) a 2-layer Gated Recurrent Unit (GRU) model, and (2) a 4-layer Neural Basis Expansion Analysis for Time Series (N-BEATS) model. The GRU model consisted of two stacked layers, each comprising 200 hidden units, followed by a linear output layer to produce the forecast. The N-BEATS model was configured with four fully connected layers, integrating both backcast and forecast blocks, and concluded with a linear output head to reconstruct future values. This design allows the model to capture both short-term patterns and longer-range temporal dependencies effectively.
To prevent overfitting, early stopping was employed in both architectures with a patience of 10 epochs, and training was conducted using a fixed learning rate of 0.0001. The models were trained on varying backward input horizons and forecast windows, allowing a thorough investigation of how different historical contexts influenced future prediction accuracy.
Model performance was evaluated using two standard error metrics: symmetric Mean Absolute Percentage Error (sMAPE) and Root Mean Squared Error (RMSE). These metrics were selected to balance sensitivity to outliers (RMSE) with scale-invariant accuracy (sMAPE), ensuring a robust evaluation of predictive performance. Lower values for both metrics indicated better forecasting quality. Model performance was assessed using four complementary metrics: Symmetric Mean Absolute Percentage Error (sMAPE), Root Mean Squared Error (RMSE), Accuracy, and F1 Score.

- Symmetric Mean Absolute Percentage Error (sMAPE): sMAPE measures the relative accuracy of forecasts by comparing the absolute difference between predicted and actual values to their average magnitude. It is scale-invariant, making it suitable for comparing performance across different ranges. Lower sMAPE values indicate higher predictive accuracy.

- Root Mean Squared Error (RMSE): RMSE quantifies the square root of the average squared differences between predicted and actual values. It is sensitive to large errors (outliers) and provides insight into the magnitude of typical prediction errors. Lower RMSE values reflect better model performance.

- Accuracy: Accuracy represents the proportion of correctly predicted instances (both positives and negatives) out of all predictions. It provides a straightforward measure of overall model correctness, particularly relevant in binary classification tasks.

- F1 Score: The F1 score is the harmonic mean of precision and recall. It balances the trade-off between false positives and false negatives, offering a robust evaluation for imbalanced classification scenarios. A higher F1 score indicates better balance and reliability in predicting the positive class.

These four metrics were chosen to provide a comprehensive evaluation framework. To assess the continuos forecasting performance as necessary for the OhioT1DM and SimGlucose datasets, sMAPE and RMSE are used. Accuracy and F1 Score specifically evaluate classification performance, particularly important for binary outcomes as found in the MIMIC dataset. In all cases, lower sMAPE and RMSE values, along with higher Accuracy and F1 scores, indicate better model quality.

After thorough evaluation across multiple configurations, the model with the lowest average error on all validation sets was selected as the baseline for subsequent counterfactual analysis. This selection ensured that counterfactual explanations were derived from the most reliable and accurate forecasting model available.
[**TODO:** Extend on method]

3.3.1.2 Counterfactual Generation:

For the counterfactual generation four different models have been used, all following a similar implementation. First two subsets of test samples were selected using two thresholds. This way, the samples were divided into a hyperglycemia set (blood glucose levels ≥ 180 mg/dL) and a hypoglycemia set (blood glucose levels ≤ 70 mg/dL) and then the following procedure was used:

- Bound Generation: Polynomial-based upper and lower bounds were generated for each test sample's predicted output, targeting desired glucose thresholds within a given window.
- Gradient Perturbation: For each forecasting model, the model was inverted by optimizing the exogenous inputs via gradient descent to steer the predicted output within the generated bounds.
- Loss Function: A custom loss function was used to penalize deviation from the bounds, and gradients were approximated using automatic differentiation or finite differences depending on model compatibility.

The models used for the counterfactual generation span both traditional statistical methods and deep learning architectures. [**TODO:** Extend on method, add more about mimic]

SARIMAX: The Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) model was employed to account for seasonality and external factors. Model parameters, including the order of autoregression (p), differencing (d), and moving average (q), were selected based on the Akaike Information Criterion (AIC). Exogenous inputs such as insulin dosage and carbohydrate intake were incorporated as regressors, and model parameters were optimized using maximum likelihood estimation.
OLS: Ordinary Least Squares (OLS) regression was used to examine linear relationships between exogenous variables and the target series. Feature engineering involved the creation of time-lagged variables and interaction terms between insulin and carbohydrate intake. To mitigate overfitting, regularization was applied. Additionally, linear constraints on exogenous variables were enforced post-hoc to facilitate counterfactual generation.
GRU: A Gated Recurrent Unit (GRU) network was implemented. The GRU architecture consisted of two layers, each with 128 hidden units, and dropout regularization set at 0.2 to prevent overfitting. The network was trained using the Adam optimizer with an initial learning rate of 0.001, which was reduced by a factor of 0.1 if the validation loss plateaued over five epochs. Counterfactual explanations were generated by computing gradients of the output with respect to exogenous inputs, enabling identification of minimal interventions.
NBEATSx: The NBEATSx model was utilized, extending on the standard N-BEATS architecture by incorporating exogenous variables into its forecasting blocks. The model

employed trend and seasonality decomposition through fully connected layers, and included exogenous inputs such as carbohydrate intake, insulin dosage, and physical activity. Quantile loss was used during training to capture distributional uncertainty, and feature importance analyses were performed to assess the contribution of exogenous variables to forecast accuracy.

3.3.2 Evaluation Metrics

[**TODO:** check, rewrite? unsure] To evaluate the quality of the counterfactual interventions, multiple evaluation metrics were implemented. First, traditional forecasting performance was measured using Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (sMAPE), giving a quantitative assessment of prediction accuracy across the forecast horizon. For the evaluation of the generated counterfactuals, several additional metrics were introduced. These have been applied to ensure that the newly generated data for the exogenous variables is realistic, plausible and applicable.

RMSE, sMAPE scores sum of MSE scores, normalize for horizon?

Average value of change The average value of change quantifies the magnitude of adjustments made to the exogenous variables in order to achieve the desired forecasted outcome. For this the averaged original results of the multivariate forecast were compared to the averaged results of the counterfactual generation. euclidean distance?
The fraction of values to change captures the proportion of exogenous variable entries that required modification, offering insight into the sparsity of the interventions. By evaluating for how many of the timesteps in the time series the data changes, it is possible to analyse the differences that the counterfactuals
Severity of change To further characterize the nature of the interventions, the severity of change metric measures the relative intensity of the modifications compared to the original values. outlier score, LOF-score compare original exog to changed prediction
Fitting of predictions in bounds Additionally, the fitting of predictions within bounds was evaluated, reflecting the extent to which the counterfactual predictions adhered to the predefined upper and lower target constraints. This was done using the euclidean distance of the predicted target to the bounds. how good is the prediction, one in the band better than others?
Compare exogenous variables to those of a healthy patient Finally, to ensure the plausibility of the counterfactual scenarios, the modified exogenous variable profiles were compared to reference profiles from healthy patients. This comparison provides a qualitative and quantitative check on the biological or clinical realism of the generated counterfactual forecasts.
what are the differences in exog? changes from healthy patient

4 Results

4.1 Multivariate Forecasting

[**TODO:** explain chosen horizon, back horzion] To evaluate the performance of the multivariate forecasting and to ensure a realistic basis for counterfactual generation, both the GRU and the N-BEATS models were tested across multiple datasets and forecasting configurations. Models were trained on historical segments of the data and evaluated on subsequent future segments, allowing for a robust comparison between predicted and actual values. The forecasting quality for the two diabetes datasets was assessed using two standard metrics: Symmetric Mean Absolute Percentage Error (sMAPE) and Root Mean Squared Error (RMSE). sMAPE was chosen due to its scale-invariant properties, allowing for fair comparison across variables with different magnitudes. RMSE was included to capture sensitivity to larger errors, penalizing substantial deviations between predictions and true values. Together, these metrics provide a balanced view of model performance, addressing both relative and absolute error considerations. For the MIMIC dataset, which involves binary survival prediction (death vs. survival), classification metrics were more appropriate given the nature of the target variable. Specifically, Accuracy and F1-Score were employed. Accuracy provides a general measure of how often the model's predictions match the true outcomes. However, due to the severe class imbalance in the MIMIC dataset—where death events are comparatively rare—F1-Score was also reported to better capture the trade-off between precision and recall for the minority class. This combination of metrics ensures both the overall predictive reliability and the ability to identify critical but infrequent events are adequately evaluated.

Table 3 summarizes the results across the two diabetes datasets, prediction horizons, and back horizons. For the OhioT1DM dataset, N-BEATS consistently outperformed the GRU model in both sMAPE and RMSE across all configurations. Notably, with a back horizon of 12 and a forecast horizon of 6, N-BEATS achieved a sMAPE of 6.03 compared to GRU's 6.28, and a lower RMSE of 14.15 versus 14.85 for GRU. The same pattern was observed for back horizon 6 and 24, with N-BEATS slightly outperforming GRU in all cases. For the SimGlucose dataset, N-BEATS again generally showed improved forecasting performance. While GRU slightly outperformed N-BEATS in sMAPE for the 20-step back, 5-step horizon configuration (0.3792 vs. 0.3867), N-BEATS achieved a substantially lower RMSE (0.8271 vs. 1.6770), indicating more accurate absolute predictions. This performance gap widened in the 40-back, 10-horizon configuration, where N-BEATS clearly surpassed GRU on both metrics.
[**TODO:** add mimic] Overall, these results demonstrate that the N-BEATS model generally offers superior performance across datasets and configurations, especially in terms of RMSE, making it a strong candidate for generating accurate multivariate forecasts and serving as a foundation for counterfactual analysis.

Table 3: Multivariate forecasting training metrics.

Dataset	Back Horizon	Horizon	Model	sMAPE	RMSE
OhioT1DM	12	6	GRU	6.2816	14.8471
			NBEATS	**6.0328**	**14.1500**
	24	6	GRU	6.1983	14.6856
			NBEATS	**5.8003**	**13.7677**
SimGlucose	20	5	GRU	0.3792	1.6770
			NBEATS	**0.3867**	**0.8271**
	40	10	GRU	1.5423	5.5572
			NBEATS	**1.4303**	**3.0461**

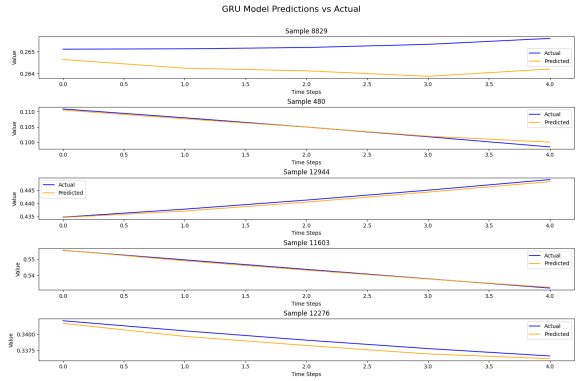
[**TODO:** rerun, recalculate results] To complement the quantitative evaluation of the multivariate forecasting models, several visualizations were created to illustrate how well the N-BEATS and GRU models track actual future values. These plots compare the predicted time series against the true values for randomly selected test samples from the datasets. Each subplot presents a single test sample, with time steps on the x-axis and the corresponding variable value on the y-axis. The actual values are shown in blue, while the predicted values are shown in orange. These plots offer a qualitative insight into the temporal alignment and amplitude accuracy of the forecasts, beyond what is captured by metrics like sMAPE or RMSE.

4.1.1 SimGlucose

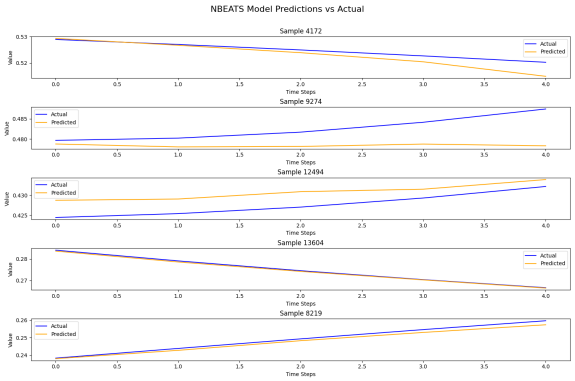
For a back horizon of 20 timesteps and a horizon of 5 timesteps, both the GRU and N-BEATS models operate on a relatively short forecast horizon, making it easier to capture the trend with high fidelity. As shown in Figure 5, N-BEATS predictions closely align with the actual values across multiple samples, reflecting low forecasting error. Minor deviations appear mostly at the end of the forecast window. While the trend direction for the GRU predictions is usually correct, the model exhibits slightly larger prediction drift compared to N-BEATS, especially in sharper transitions. Still, GRU maintains reasonable short-term accuracy. For a back horizon of 40 timesteps and a horizon of 10 timesteps, the forecast horizon is increased, making the prediction task more difficult. Figure 6 shows the N-BEATS model's ability to handle long-range dependencies. Although small prediction lags and amplitude mismatches occur, the model captures the overall progression well, preserving directionality in most sequences. When looking at the GRU predictions under the same setup, the model struggles a bit more with extended forecasts, often showing divergence from actual trajectories, particularly toward the final time steps. These deviations are consistent with the slightly higher RMSE reported for this configuration.

4.1.2 OhioT1DM

Figure 7 compares N-BEATS and GRU predictions for a back horizon of 12 timesteps and 6 future timesteps. The N-BEATS model tracks the general shape and scale of the actual values effectively. Although slight under- or over-predictions occur, especially in rising or falling trends, the model remains closely aligned overall. In contrast, the GRU predictions tend to diverge slightly more in the later steps of the forecast horizon. GRU often preserves the direction of the trend but struggles with amplitude and slope consistency, leading to

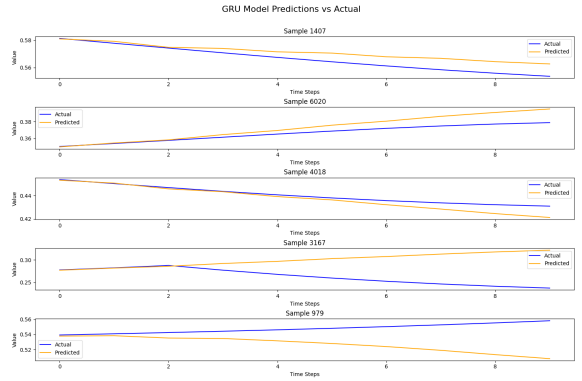


(a) Results of the multivariate forecasting using GRU.

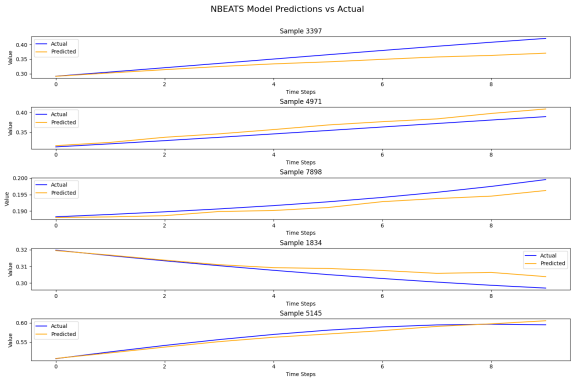


(b) Results of the multivariate forecasting using NBEATS.

Figure 5: Results of the multivariate forecasting for the SimGlucose dataset with back horizon = 20 and forecast horizon = 5, showing the accuracy of the forecasting.



(a) Results of the multivariate forecasting using GRU.



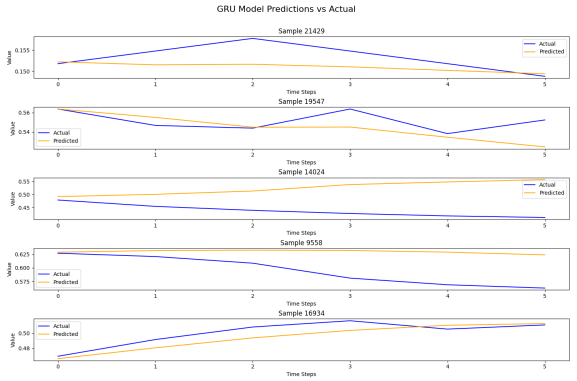
(b) Results of the multivariate forecasting using NBEATS.

Figure 6: Results of the multivariate forecasting for the SimGlucose dataset with back horizon = 40 and forecast horizon = 10, showing the accuracy of the forecasting.

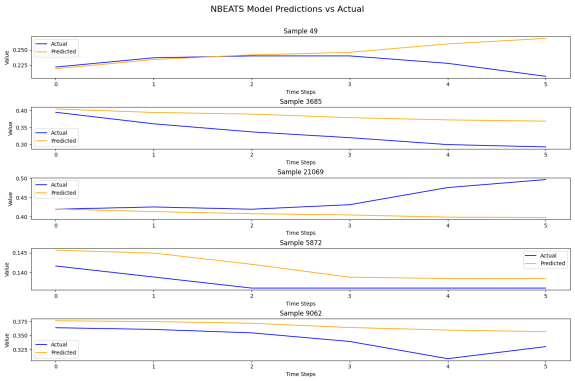
larger deviations as the prediction window progresses. With a back horizon of 24 timesteps and the same future timesteps, figure 8 depicts the performance of N-BEATS and GRU under a longer input sequence. The figure illustrates that N-BEATS maintains robust trend prediction, with only modest lag or dampening effects in some samples. The model adapts well to both upward and downward trajectories. The GRU model performs more variably under this longer back horizon. While the general trajectory is still often captured, the forecasts occasionally overreact or smooth out variations, resulting in lower precision at the end of the forecast horizon.

4.1.3 MIMIC

[**TODO:** check, rewrite] The GRU and N-BEATS models were evaluated on the MIMIC dataset for both 30-day and 1-year mortality prediction tasks. While both models achieved high overall accuracy across clusters (up to 94.8% in some subgroups), they consistently failed to detect the minority *Died* class. For example, in Cluster 0, the N-BEATS model

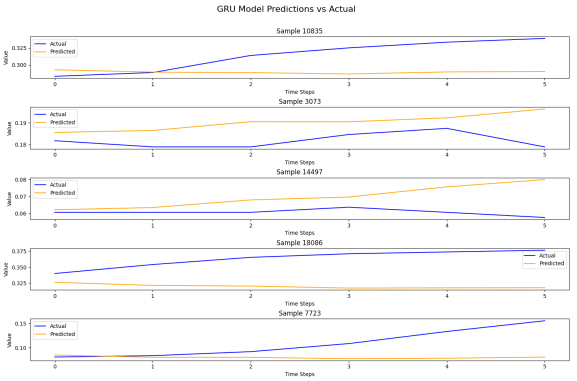


(a) Results of the multivariate forecasting using GRU.

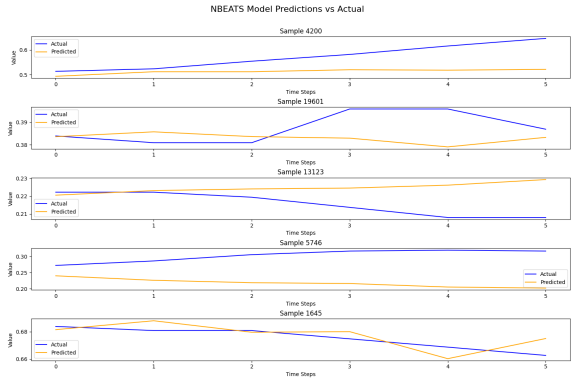


(b) Results of the multivariate forecasting using NBEATS.

Figure 7: Results of the multivariate forecasting for the OhioT1DM dataset with back horizon = 12 and forecast horizon = 6, showing the accuracy of the forecasting.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using NBEATS.

Figure 8: Results of the multivariate forecasting for the OhioT1DM dataset with back horizon = 24 and forecast horizon = 6, showing the accuracy of the forecasting.

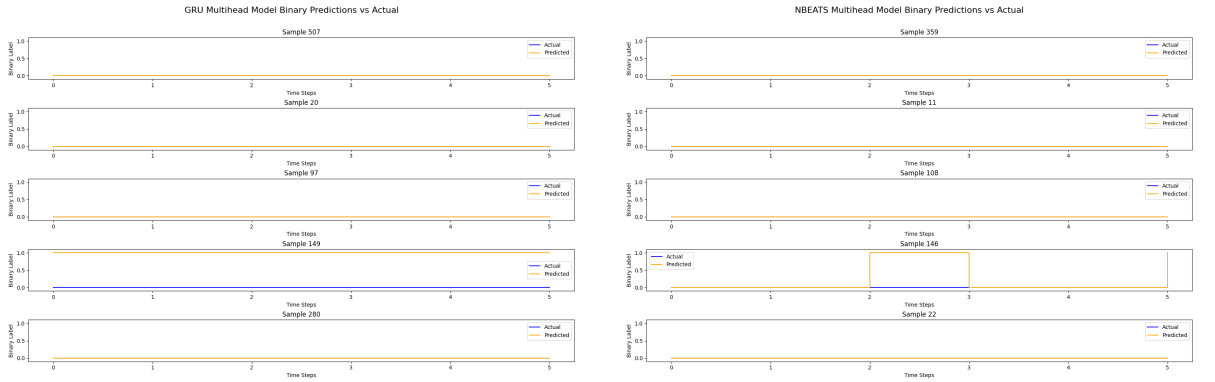
achieved an accuracy of 89.2% for 30-day mortality but reported 0.00 precision, recall, and F1-score for the *Died* class, indicating complete bias toward the majority *Survived* class (Table **??**).

Notably, some clusters such as Cluster 2 showed slightly better detection of the *Died* class. For instance, the GRU model achieved a recall of 0.50 and precision of 0.12 for the *Died* class in 30-day mortality prediction in Cluster 2, though with a lower overall accuracy of 52.9%.

These results highlight the limitations of using accuracy as a sole metric in imbalanced clinical datasets and emphasize the need for alternative evaluation strategies and model improvements targeting minority class detection.

For the MIMIC dataset, the classification performance of GRU and N-BEATS models was evaluated using accuracy, precision, recall, and F1-score, with class-wise metrics reported for both 30-day and 1-year mortality prediction tasks, as seen in table 11. For 30-day mortality, the GRU model achieved an overall accuracy of 83.5%, while the N-BEATS model reached 91.6%. Despite these high overall accuracy scores, both models failed to

correctly identify any instances of the "Died" class. Specifically, both GRU and N-BEATS models reported precision, recall, and F1-score values of 0.00 for the "Died" class, indicating that all predictions were biased toward the majority "Survived" class. This pattern persisted in the 1-year mortality prediction task, where both models again achieved high overall accuracy (90.6% for both), but with no correct predictions for the "Died" class. These results suggest that both models are biased towards the majority class, a common issue in imbalanced datasets. While overall accuracy may appear high, it does not reflect clinically important performance for detecting critical outcomes like patient death. Figure 30 compares the N-BEATS and GRU predictions for a back horizon of 13 timesteps and 6 future timesteps and target variable 30-day mortality. Both models show a strong tendency to predict the majority class, which appears to be the negative (0) class in most samples. For example, in the plots for samples where the actual labels are all positive (1), such as Sample 53 in the N-BEATS model and Sample 71 in the GRU model, the predictions are consistently negative (0). This indicates a failure of both models to identify positive cases.Neither model demonstrates effective discrimination between classes at the individual sample level, as the predicted values remain flat and do not align with the actual binary changes over the time steps. The predictions appear to be largely constant across time steps, showing little to no responsiveness to actual positive events. This behavior suggests that both the N-BEATS and GRU models are biased toward predicting the majority class, likely due to class imbalance in the dataset. Despite potentially high overall accuracy, these models fail to capture the minority positive class, resulting in poor sensitivity and F1 scores for that class.



(a) Results of the multivariate forecasting using GRU.

(b) Results of the multivariate forecasting using NBEATS.

Figure 9: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

For complete results across all clusters and subgroups, refer to Appendix A.1.

4.2 Counterfactuals

4.2.1 OhioT1DM

Figure 10 presents an example of generated counterfactuals for the OhioT1DM dataset. It shows a short sequence of 6 timesteps where the predicted blood glucose (BG) level is ap-

proximately 155 mg/dL—lower than the original range of 180–190 mg/dL. This new BG level falls within the predefined desired bounds, indicating a plausible improvement. Notably, the generated exogenous variables—such as insulin doses, carbohydrate intake, and exercise intensity—differ considerably from the original values. These differences suggest actionable changes that may help achieve improved glycemic control.

To evaluate the realism of the generated exogenous variables, we searched for pairs of samples in the dataset where both the original and counterfactual blood glucose levels were similar. By comparing their exogenous features, we can assess whether the generated values are plausible within the real data distribution, ensuring that counterfactual recommendations are actionable and clinically meaningful.

For this analysis, we identified two samples of 6 timesteps each, with a minimal Euclidean distance of 2.0295 between their blood glucose target sequences. Their exogenous variables are compared in Figure 11. The observed differences highlight the variability in exogenous features that can still lead to similar blood glucose outcomes.

Additionally, Table 5 summarizes the Euclidean distances between exogenous features for the five closest pairs of samples, using both actual and normalized values. This quantitative measure helps to contextualize the variability seen in the visual comparison and supports the feasibility of the generated counterfactuals within real-world data ranges.

4.2.2 SimGlucose

Figure 12 present the results of counterfactual forecasting for the SimGlucose dataset using a GRU-based multivariate forecasting model. The aim was to identify alternative exogenous variable trajectories—specifically carbohydrate intake and insulin administration—that lead to improved blood glucose (BG) levels within a clinically desirable range. In the upper figure the predicted exogenous variables are compared against the original sequences. Both carbohydrate intake and insulin administration in the counterfactual scenario are noticeably reduced or even set to zero during most time steps. Carbohydrate intake increases only slightly at later time steps, while insulin remains consistently absent. This suggests that the model identifies minimal or delayed interventions as sufficient to guide BG levels toward healthier ranges, reflecting a preference for conservative adjustments.

The lower figure shows the impact of these counterfactual exogenous variables on BG trajectories. While the original BG levels remain relatively stable and slightly above the desired range, the predicted BG values stay well within the predefined target bounds (highlighted in green). The predicted trajectory is also smoother and lower, indicating improved glycemic control according to the model's optimization.

Overall, these results demonstrate that the model effectively proposes simplified and realistic counterfactual adjustments in the SimGlucose environment. It prioritizes conservative intervention strategies—often minimizing both insulin administration and carbohydrate intake—while achieving BG outcomes that align more closely with clinical targets. This highlights the model's potential applicability for decision support in artificial pancreas systems or glucose management recommendations.

4.2.3 MIMIC

To evaluate how well our model can generate counterfactuals for clinical time series data, we tested several modeling approaches on the MIMIC-III dataset, focusing specifically

on patients with heart failure with preserved ejection fraction (HFpEF). The goal was to create realistic counterfactual trajectories that shift a patient's predicted outcome from death to survival (binary target = 0), by making minimal and interpretable changes to exogenous clinical variables.

Figures 13 through 16 show examples of counterfactuals generated using four different models: SARIMAX, Ordinary Least Squares (OLS), GRU, and N-BEATS. In each case, the original values are plotted in blue, and the counterfactual values—those modified to achieve a survival outcome—are shown in yellow, across six time steps.

While all models share the same goal—altering the inputs just enough to flip the predicted outcome to survival while keeping the changes realistic—they approach the task in different ways, each with its own strengths.

SARIMAX (Figure 13) leans on temporal trends and smooth transitions. The adjustments it makes tend to be gradual and follow the original signal's shape closely. It mainly modifies variables like Heart Rate, Bicarbonate, and Troponin T, all of which are clinically important in HFpEF. At the same time, it leaves others, such as BMI and Creatinine, mostly untouched—suggesting a conservative approach that focuses only on what's most relevant.

OLS (Figure 14) takes a more direct, linear route to modifying inputs. While this makes the changes easy to understand, they can sometimes appear abrupt or less physiologically natural, especially for variables that typically show nonlinear patterns (e.g., SpO or Platelet Count). OLS still gets the job done in terms of outcome flipping, but its results can be less smooth or realistic compared to the time series-specific models.

GRU (Figure 15) offers more flexibility thanks to its ability to capture complex temporal relationships. The counterfactuals it produces tend to preserve the overall shape of the original sequence while introducing focused changes to the most influential variables—like Hemoglobin, Temperature, and Heart Rate. GRU's changes are more dynamic than SARIMAX's but still maintain temporal coherence, making it well-suited for the variability found in clinical settings.

N-BEATS (Figure 16) also performs well, especially when it comes to maintaining both short- and long-term trends in the data. It introduces meaningful, smooth modifications to key features such as NT-proBNP and Systolic Blood Pressure, staying true to both clinical relevance and the natural flow of the time series. Its architecture allows for a high level of nuance in how changes are applied across time.

One notable pattern across all models is that certain features—like BMI, Creatinine, and Platelet Count—tend to remain unchanged. This suggests that the models aren't making arbitrary edits but are instead focusing on variables that actually influence the outcome. That kind of selective intervention supports interpretability and builds trust in the results.

In short, while all four models are capable of generating plausible counterfactuals that achieve the intended goal, they vary in how they balance smoothness, realism, and intervention strength. SARIMAX sticks to subtle, trend-based changes; OLS is simple and straightforward but sometimes rough; GRU introduces more flexible, targeted edits; and N-BEATS strikes a strong balance between structure and adaptability. These differences highlight the versatility of counterfactual reasoning for clinical time series and point to exciting opportunities for personalized intervention support in healthcare.

4.3 Measurements

4.3.1 Mean Absolute Change

Figure 17 presents the mean absolute change between the original and generated counterfactual time series for all four features from the OhioT1DM dataset. This metric quantifies the average magnitude of alteration introduced by the counterfactual generation process. A relatively stable decline in change over timesteps can be observed across features, indicating temporal consistency in the perturbation patterns.

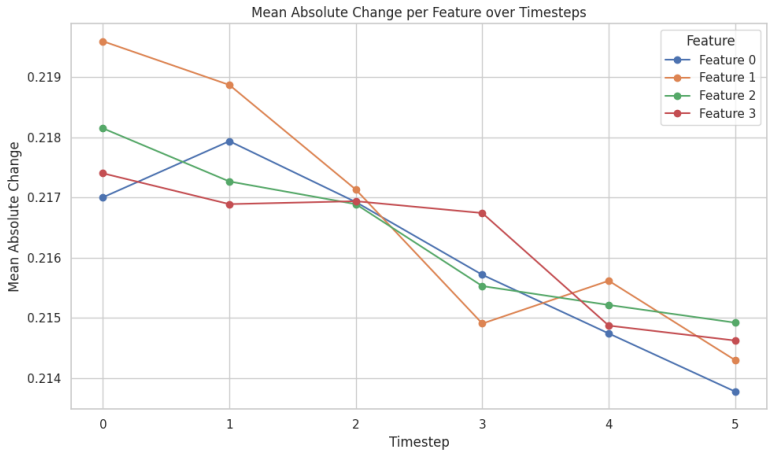


Figure 17: Mean absolute change per feature over timesteps, comparing original and counterfactual sequences.

4.3.2 Severity of Change

To better understand the extent and nature of modifications across features, Figure 18 visualizes two key indicators. The top panel shows the fraction of changed values per feature over time, revealing that modifications are sparse and localized. The bottom panel reiterates the mean absolute change, highlighting that although the fraction of altered values is low, the individual changes can still be substantial. Together, these plots provide a detailed view of how features were selectively modified.

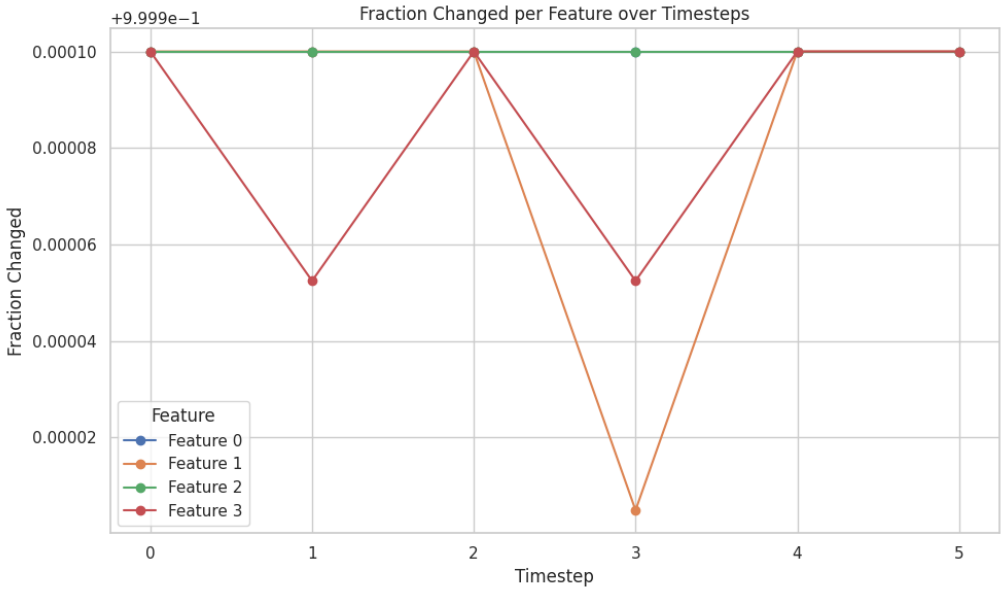
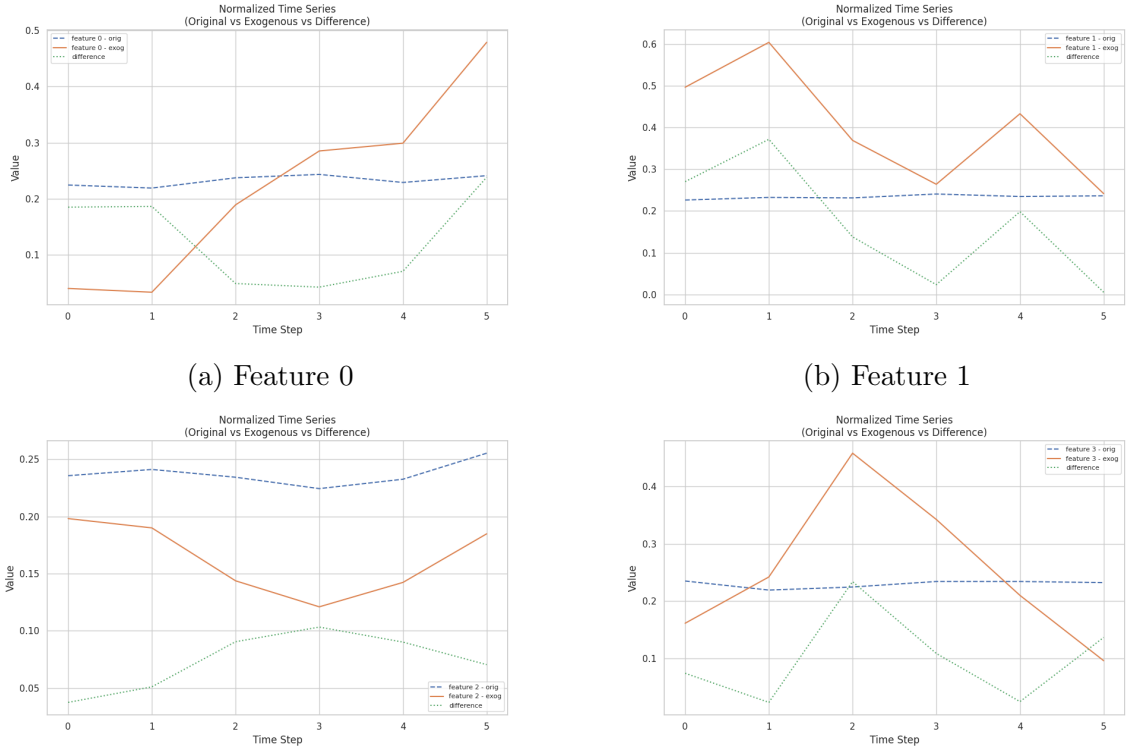


Figure 18: Top: Fraction of features changed over timesteps. Bottom: Mean absolute change per feature.

4.3.3 Visualization of Feature-wise Changes

Figure 19 shows the normalized time series for each feature, comparing the original inputs with the generated counterfactuals. The dotted line represents the absolute difference between the two series at each timestep. These visualizations help pinpoint specific regions where interventions have occurred and how significantly each feature has been affected. This is particularly useful for interpreting which variables are being targeted by the counterfactual model.



(a) Feature 0

(b) Feature 1

(c) Feature 2

(d) Feature 3

Figure 19: Comparison of original and counterfactual time series for each feature, showing normalized values and their absolute differences.

4.3.4 Coefficient of Determination

The coefficient of determination (R^2 score) is used to evaluate how well the counterfactual time series preserves the original temporal dynamics. A high R^2 score indicates that the generated data maintains the structure of the original inputs while introducing targeted modifications. This metric supports assessing the realism and fidelity of counterfactual sequences.

4.3.5 Target Prediction within Bounds

In addition to matching the temporal patterns, an important constraint for counterfactual validity is that the prediction for the target variable remains within clinically acceptable bounds. This ensures that generated scenarios do not produce unrealistic or medically implausible outcomes. Here, we verify that counterfactual predictions remain within the same classification range (e.g., hyperglycemia vs. normoglycemia) as the original data or follow a desired clinical transition.

4.3.6 Comparison to Healthy Patient

Finally, we compare the generated counterfactuals with healthy patient profiles to evaluate whether the model is steering inputs toward realistic, health-aligned trajectories. This analysis allows us to assess whether counterfactual edits reflect desirable or therapeutically meaningful interventions, rather than arbitrary perturbations that may reduce prediction error but lack interpretability.

5 Discussion

5.1 Multivariate Forecasting

5.1.1 Results

[**TODO:** überarbeiten, controllieren] The multivariate forecasting experiments provide a comprehensive understanding of how sequence models like GRU and N-BEATS behave under various dataset conditions, back horizons, and forecast horizons. Across both the OhioT1DM and SimGlucose datasets, N-BEATS consistently demonstrated superior forecasting capability, a trend reflected in both quantitative metrics (sMAPE and RMSE) and qualitative visualization analyses.

Specifically, in OhioT1DM, N-BEATS consistently outperformed GRU across all tested configurations, including back horizons of 12 and 24 timesteps and a forecast horizon of 6. For instance, with a back horizon of 12, N-BEATS achieved a sMAPE of 6.03 and RMSE of 14.15, outperforming GRU's 6.28 sMAPE and 14.85 RMSE. The results were even more pronounced at a back horizon of 24, where N-BEATS' structural advantage in processing longer sequences through its stacked feedforward blocks was evident.

Similarly, in the SimGlucose dataset, while GRU slightly outperformed N-BEATS in terms of sMAPE for the 20-back, 5-horizon setup, N-BEATS delivered a markedly lower RMSE (0.8271 vs. 1.6770), emphasizing its capacity for more accurate absolute magnitude prediction. This pattern persisted and even intensified in the more challenging 40-back, 10-horizon configuration, where N-BEATS surpassed GRU on both sMAPE and RMSE.

These findings underscore two important dynamics. First, the selection of forecasting metrics matters: RMSE's sensitivity to large deviations makes it particularly valuable in real-world healthcare settings, where under- or over-predicting magnitudes (such as blood glucose levels) can have direct clinical consequences. Second, N-BEATS' resilience against performance degradation with longer forecast horizons suggests its architectural backcast-forecast decomposition effectively mitigates the long-range dependency challenges faced by recurrent models like GRU.

Visual inspection through time series plots reinforces these quantitative conclusions. N-BEATS forecasts adhered more closely to ground truth values, especially in capturing amplitude and abrupt changes. GRU often captured general trend directions but exhibited smoothing effects and lagging amplitude response. This limitation of GRU is likely due to its reliance on recurrent structures, which may struggle to retain detailed information over longer temporal windows, particularly in irregular or sparse real-world datasets like OhioT1DM.

Dataset complexity emerged as a decisive factor as well. SimGlucose, being synthetic and structured, naturally yielded lower forecasting errors compared to the noisier, patient-specific OhioT1DM data. Nevertheless, the relative performance advantages of N-BEATS remained consistent across both contexts, reinforcing the robustness of its design not only in idealized settings but also in messy real-world scenarios.

Regarding the MIMIC dataset, a starkly different challenge surfaced. Both GRU and N-BEATS achieved high overall accuracy for 30-day and 1-year mortality prediction tasks; however, neither model successfully identified instances of the minority "Died" class. As Table 11 shows, precision, recall, and F1-scores for the minority class were all zero, highlighting a critical limitation. This underscores the importance of selecting not just appropriate models but also loss functions and evaluation strategies that account for severe

class imbalance—particularly crucial in healthcare applications where detecting rare but critical outcomes is often the primary objective.

This discrepancy between forecasting and classification tasks suggests model suitability is context-dependent. While N-BEATS offers clear advantages for continuous multivariate forecasting, its feedforward architecture may not inherently address classification imbalance challenges, especially without specific adaptation strategies such as resampling, class weighting, or focal loss. GRU shares this limitation, likely due to its similar reliance on standard optimization objectives that do not penalize class imbalance.

The choice of back horizon length was found to significantly influence model performance, particularly benefiting N-BEATS. As observed in both OhioT1DM and SimGlucose, increasing the back horizon from 12 to 24 or from 20 to 40 steps generally improved N-BEATS's predictive accuracy, while GRU showed more modest or inconsistent gains. This suggests that N-BEATS is better equipped to process extended historical contexts, offering greater flexibility when designing counterfactual generation pipelines that depend on multivariate forecasting.

The qualitative plots provide additional clarity. For both diabetes datasets, N-BEATS maintained tighter alignment with ground truth across short and extended forecast horizons. In contrast, GRU's forecasts were prone to drift and smoothing effects, especially toward the end of longer prediction windows. This visual evidence, combined with metric-based assessments, validates the selection of N-BEATS as the primary forecasting engine within the broader counterfactual framework.

When evaluating overall system design, these findings point to a few key takeaways. N-BEATS delivers more reliable magnitude predictions (lower RMSE), making it particularly suitable for tasks where precise value estimation is critical, such as blood glucose level forecasting. GRU may still be useful in contexts where capturing general trends is sufficient, especially when computational simplicity is desired. Both models require careful adjustment and augmentation when applied to classification tasks with class imbalance, as seen in MIMIC.

In conclusion, the experimental results support the adoption of N-BEATS as the default forecasting model within the counterfactual analysis pipeline. Its ability to maintain performance across various datasets, back horizons, and forecast horizons ensures that generated counterfactuals are anchored in realistic, data-consistent future trajectories. This is essential not just for accuracy, but also for the clinical interpretability and trustworthiness of the resulting recommendations.

5.1.2 Limitations

While the GRU and N-BEATS models demonstrate strong performance on aggregate evaluation metrics, a critical limitation was observed in their inability to detect the minority "Died" class. Both models predominantly focus on the majority "Alive" class, which skews performance metrics such as accuracy and mean squared error. This imbalance is especially problematic in clinical prediction settings where identifying high-risk patients is of paramount importance. The observed class imbalance inherently biases the learning process, causing the models to overlook rare but clinically significant outcomes. This limitation suggests that relying solely on conventional evaluation metrics may provide an incomplete picture of model utility in imbalanced clinical datasets. The results highlight the pressing need for incorporating alternative strategies—such as resampling techniques,

class weighting, or cost-sensitive learning—to ensure more equitable predictive performance across all outcome classes. Without such interventions, these models may fail to provide actionable insights in real-world healthcare applications, where missing minority class instances can have serious implications.

5.1.3 Future Work

Addressing the limitations outlined above should be a central focus of future research. Several promising strategies could be explored to mitigate class imbalance effects and improve minority class detection. One approach involves implementing class weighting during model training, adjusting the loss function so that misclassifying instances of the minority "Died" class incurs a higher penalty. This adjustment would compel the models to pay greater attention to these critical cases.
Additionally, resampling strategies such as Synthetic Minority Over-sampling Technique (SMOTE), random under-sampling of the majority class, or hybrid resampling methods can be applied to rebalance the training dataset. Specialized loss functions, including focal loss or those optimized for metrics like Matthews correlation coefficient, may also enhance minority class prediction by explicitly addressing class distribution disparities.
Further investigation into model architecture adjustments could offer additional improvements. For instance, integrating attention mechanisms or developing hybrid models that combine sequence learning with anomaly detection frameworks may provide more nuanced representations of minority class instances. Threshold tuning and calibration techniques could also be leveraged to fine-tune decision boundaries and improve sensitivity to minority outcomes.
In sum, while the GRU and N-BEATS models show promise on general predictive tasks, realizing their full clinical utility requires targeted strategies to enhance minority class performance. Addressing class imbalance through these techniques will be essential for developing predictive models that are not only statistically robust but also practically valuable in healthcare decision-making contexts.

5.2 Counterfactuals

The results show that generating multivariate counterfactuals for time series forecasting is not only possible but also practically useful. By adjusting exogenous variables while preserving the natural shape and rhythm of the data, the counterfactual sequences successfully reach desired target outcomes—such as keeping key variables within certain clinical ranges—without introducing unrealistic values.
Both visual and quantitative evaluations suggest that the counterfactuals stay within the bounds of what we'd expect from real patient data. When compared with naturally occurring trajectories that lead to similar outcomes, the counterfactuals often follow different paths using different combinations of features. This reflects a core characteristic of multivariate time series: multiple input patterns can lead to similar predictions due to flexibility or redundancy in the system.
We also observed that the "distance" between counterfactuals and nearby real samples helps highlight how different features can work together to drive a particular outcome. Instead of relying on a single-variable tweak, the models tend to adjust several features at once—much like how real clinical decisions often involve multiple interventions happening in tandem.

Taken together, these findings support the idea that multivariate counterfactual generation can be a valuable tool for time series forecasting tasks where interpretability and actionable insight are important. Moving forward, this work could be extended by applying domain-specific constraints, incorporating user feedback, or bringing in causal modeling to strengthen the relevance of generated interventions.

5.2.1 OhioT1DM

In the OhioT1DM results, the counterfactual sequence successfully brings BG levels down into a healthier range (around 155 mg/dL), starting from higher original values (180–190 mg/dL). This shift is achieved by changing key exogenous factors—particularly insulin dosage, carbohydrate intake, and exercise intensity. While these counterfactual adjustments differ noticeably from the original data, they remain within clinically plausible ranges, suggesting that the model identifies meaningful levers for improving glycemic control.

To evaluate how realistic these adjustments are, we compared the counterfactual inputs to real pairs of samples in the dataset with similar BG levels. These comparisons (visualized in Figure 11 and summarized in Table 5) reveal that quite different input patterns can still lead to similar BG outcomes. This variability supports the idea that there's no single "correct" intervention—multiple combinations of inputs can achieve the same result, and the counterfactuals generated here appear to fall within those bounds.

Overall, these results highlight the model's ability to generate not just effective but also flexible and interpretable recommendations. The variability observed between similar BG outcomes further reinforces the value of multivariate approaches that consider a range of plausible paths rather than relying on single-variable tweaks.

5.2.2 SimGlucose

In the SimGlucose setting, the counterfactual sequences take a slightly different approach. The model tends to propose conservative strategies, minimizing both carbohydrate intake and insulin doses to improve BG levels. In many cases, insulin is reduced to zero, and carb intake is either flat or only slightly increased in later time steps. These subtle adjustments still lead to BG predictions that fall comfortably within the desired range, showing smoother and more stable glycemic trajectories.

This behavior reflects a model that's tuned not just for accuracy but also for simplicity in intervention—something that could be especially valuable in real-world decision support tools. For example, in artificial pancreas systems or diabetes management apps, recommending minimal yet effective interventions is often preferable to aggressive changes.

Together, these findings demonstrate the potential for counterfactual generation in personalized glucose management. The models are able to suggest realistic adjustments tailored to the patient context while respecting physiological and behavioral boundaries. Whether through more varied strategies in OhioT1DM or conservative control in SimGlucose, both models showcase how counterfactual reasoning can enhance interpretability and support meaningful decision-making in clinical time series forecasting.

5.2.3 MIMIC

This study explored the use of counterfactual generation in clinical time series data, focusing on patients with heart failure with preserved ejection fraction (HFpEF) in the MIMIC-III dataset. The goal was to generate alternate trajectories for exogenous clinical variables that could change a predicted outcome from death to survival, in a way that remains realistic and interpretable.

All four modeling approaches (SARIMAX, OLS, GRU, and N-BEATS) were able to generate counterfactual sequences that followed physiological patterns while successfully flipping the model's prediction. That said, each model approached the task in different ways, showing their unique strengths and limitations.

SARIMAX, a more traditional statistical model, produced gradual, trend-consistent changes that respected the structure of the original signal. This made the counterfactuals smooth and easy to interpret, though the model's reliance on linear relationships limited its flexibility in capturing more complex or nonlinear dynamics.

OLS offered simplicity and interpretability but struggled with temporal consistency. While it could generate counterfactuals that changed the predicted outcome, some of the resulting trajectories showed abrupt changes that don't align well with physiological expectations. This could reduce trust in real-world settings where clinicians rely on coherent time series patterns.

The neural models—GRU and N-BEATS—were more adaptive and expressive. GRUs, with their ability to model time dependencies and nonlinear relationships, produced counterfactuals that stayed faithful to the original sequence while making smart, targeted adjustments. N-BEATS, with its deep architecture and built-in trend/seasonality modeling, generated smooth and well-aligned trajectories that captured both short- and long-term patterns effectively. This made it especially suitable for variable-length and dynamic clinical sequences.

One important strength across all models was their ability to make selective changes. Features like BMI, Creatinine, and Platelet Count—often less immediately tied to short-term survival—were mostly left untouched. Instead, the models focused their adjustments on more relevant features like Heart Rate, Troponin T, and Bicarbonate. This targeted behavior adds to the interpretability and clinical credibility of the generated counterfactuals.

Altogether, the results point to real potential for using counterfactual modeling in clinical forecasting. These models don't just help explain what the model saw—they can also be used to simulate "what-if" scenarios, which could be useful for decision support, treatment planning, or even early warning systems in critical care.

Still, there are limitations to address. Our evaluation so far has been mainly qualitative, and future work should include more rigorous quantitative metrics for plausibility, safety, and causal validity. Clinical deployment would also require expert review and integration with medical knowledge. And while we assume that changes to exogenous inputs cause changes in outcomes, this assumption should be carefully examined, especially when working with observational data that may include confounders.

In summary, our findings show that counterfactual generation in time series is both feasible and valuable. By comparing a variety of modeling strategies, we offer insights into how different models handle the trade-offs between flexibility, coherence, and interpretability—laying the groundwork for more transparent and clinically grounded machine learning

in healthcare.

6 Conclusion

References

[AALC20] Emre Ates, Burak Aksar, Vitus Leung, and Kaan Coşkun. Counterfactual explanations for machine learning on multivariate time series data. *CoRR*, 2020.

[ASFWHY+25] Nur Farah Afifah Ahmad Sukri, Wan Mohd Amir Fazamin Wan Hamzah, Mohd Kamir Yusof, Ismahafezi Ismail, Harmy Mohamed Yusoff, and Azliza Yacob. A systematic literature review on machine learning in healthcare prediction. *International Journal of Online & Biomedical Engineering*, 21(6), 2025.

[BAI25] Mamoune Benaida, Ibtissam Abnane, and Ali Idri. Deep learning based one step and multi-steps ahead forecasting blood glucose level. *Expert Syst. J. Knowl. Eng.*, 42(1), 2025.

[BP10] Barry Borlaug and James Paulus. Heart failure with preserved ejection fraction: Pathophysiology, diagnosis, and treatment. *European heart journal*, 32:670–679, 2010.

[CHN+21] Ran Cui, Chirath Hettiarachchi, Christopher Nolan, Elena Daskalaki, and Hanna Suominen. Personalised short-term glucose prediction via recurrent self-attention network. In *34th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2021, Aveiro, Portugal, June 7-9, 2021*, pages 154–159. IEEE, 2021.

[Dia] Last Accessed: Jul. 5, 2025. Available: https://www.who.int/health-topics/diabetes.

[DMML+14] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator: New features. *Journal of diabetes science and technology*, 8:26–34, 2014.

[EMA+24a] Nuha Elsayed, Rozalina McCoy, Grazia Aleppo, Kirthikaa Balapattabi, Elizabeth Beverly, Kathaleen Early, Dennis Bruemmer, Osagie Ebekozien, Justin Echouffo-Tcheugui, Laya Ekhlaspour, Jason Gaglia, Rajesh Garg, Kamlesh Khunti, Rayhan Lal, Ildiko Lingvay, Glenn Matfin, Naushira Pandya, Elizabeth Pekas, Scott Pilla, and Raveendhara Bannuru. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2025. *Diabetes Care*, 48:S27–S49, 2024.

[EMA+24b] Nuha Elsayed, Rozalina McCoy, Grazia Aleppo, Kirthikaa Balapattabi, Elizabeth Beverly, Kathaleen Early, Dennis Bruemmer, Justin Echouffo-Tcheugui, Laya Ekhlaspour, Rajesh Garg, Kamlesh Khunti, Rayhan Lal, Ildiko Lingvay, Glenn Matfin, Naushira Pandya, Elizabeth Pekas, Scott Pilla, Sarit Polsky, Alissa Segal, and Raveendhara Bannuru. 6. glycemic goals and hypoglycemia: Standards of care in diabetes—2025. *Diabetes Care*, 48:S128–S145, 2024.

[FAJ+18] Ian Fox, Lynn Ang, Mamta Jaiswal, Rodica Pop-Busui, and Jenna Wiens. Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1387–1395. ACM, 2018.

[GBV+24] Raffaele Giancotti, Pietro Bosoni, Patrizia Vizza, Giuseppe Tradigo, Agostino Gnasso, Pietro Hiram Guzzi, Riccardo Bellazzi, Concetta Irace, and Pierangelo Veltri. Forecasting glucose values for patients with type 1 diabetes using heart rate data. *Comput. Methods Programs Biomed.*, 257:108438, 2024.

[Gui24] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

[HMH+25] Yue Hu, Fanghui Ma, Mengjie Hu, Binbing Shi, Defeng Pan, and Jingjing Ren. Development and validation of a machine learning model to predict the risk of readmission within one year in hfpef patients: Short title: Prediction of hfpef readmission. *International Journal of Medical Informatics*, 194:105703, 2025.

[HSL+23] Chenlu Hong, Linjuan Sun, Guangwen Liu, Boyuan Guan, Chengfu Li, and Yanan Luo. Response of global health towards the challenges presented by population aging. *China CDC Weekly*, 5(39):884, 2023.

[HSYZ23] Jianing Hao, Qing Shi, Yilin Ye, and Wei Zeng. TimeTuner : Diagnosing time representations for time-series forecasting with counterfactual explanations. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–11, 2023.

[JHS+22] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.

[JPS+16] Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[KBKB23] Pınar Özen Kavas, Mehmet Recep Bozkurt, İbrahim Kocayiğit, and Cahit Bilgin. Machine learning-based medical decision support system for diagnosing hfpef and hfref using ppg. *Biomedical Signal Processing and Control*, 79:104164, 2023.

[KCS+20] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A Pickett, and Varun Dutt. Ai in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, 3:4, 2020.

[KM25] Deepjyoti Kalita and Khalid B. Mirza. Multivariate glucose forecasting using deep multihead attention layers inside neural basis expansion networks. *IEEE J. Biomed. Health Informatics*, 29(5):3654–3663, 2025.

[KS20] Zahra Karevan and Johan AK Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

[LBF13] Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

[LCP+25] Francisco J. Lara-Abelenda, David Chushig-Muzo, Pablo Peiro-Corbacho, Ana M. Wägner, Conceição Granja, and Cristina Soguero-Ruíz. Personalized glucose forecasting for people with type 1 diabetes using large language models. *Comput. Methods Programs Biomed.*, 265:108737, 2025.

[LH+24] Pei Hung Liao, Chung Lieh Hung, et al. Diagnostic yield and model prediction using wearable patch device in hfpef. In *Innovation in Applied Nursing Informatics*, pages 25–30. IOS Press, 2024.

[MB20] Cindy Marling and Razvan Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. *CEUR workshop proceedings*, 2675:71–74, 2020.

[MGL+20] Awais Malik, Gauravpal S Gill, Fahad K Lodhi, Lakshmi S Tummala, Steven N Singh, Charity J Morgan, Richard M Allman, Gregg C Fonarow, and Ali Ahmed. Prior heart failure hospitalization and outcomes in patients with heart failure with preserved and reduced ejection fraction. *The American journal of medicine*, 133(1):84–94, 2020.

[MKT+24] Kirsty McDowell, Toru Kondo, Atefeh Talebi, Ken Teh, Erasmus Bachus, Rudolf A De Boer, Ross T Campbell, Brian Claggett, Ashkay S Desai, Kieran F Docherty, et al. Prognostic models for mortality and morbidity in heart failure with preserved ejection fraction. *JAMA cardiology*, 9(5):457–465, 2024.

[MMA+21] Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, et al. 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European heart journal*, 42(36):3599–3726, 2021.

[MPS19] Nishita Mehta, Anil Pandit, and Sharvari Shukla. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *Journal of biomedical informatics*, 100:103311, 2019.

[MSG14] Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13, 2014.

[OHH+06] Theophilus E Owan, David O Hodge, Regina M Herges, Steven J Jacobsen, Veronique L Roger, and Margaret M Redfield. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *New England Journal of Medicine*, 355(3):251–259, 2006.

[RCR23] Ignacio Rodríguez-Rodríguez, María Campo-Valera, and José-Víctor Rodríguez. Forecasting glycaemia for type 1 diabetes mellitus patients by means of iomt devices. *Internet Things*, 24:100945, 2023.

[RNZ17] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, pages 323–350, 2017.

[SBA+24] Ikgyu Shin, Nilay Bhatt, Alaa Alashi, Keervani Kandala, and Karthik Murugiah. Predicting 30-day and 1-year mortality in heart failure with preserved ejection fraction (hfpef). *medRxiv : the preprint server for health sciences*, 2024.

[SK25a] Yiheng Shen and Samantha Kleinberg. Personalized blood glucose forecasting from limited CGM data using incrementally retrained LSTM. *IEEE Trans. Biomed. Eng.*, 72(4):1266–1277, 2025.

[SK25b] Yuyang Sun and Panagiotis Kosmas. Integrating bayesian approaches and expert knowledge for forecasting continuous glucose monitoring values in type 2 diabetes mellitus. *IEEE J. Biomed. Health Informatics*, 29(2):1419–1432, 2025.

[VDH21] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756*, 2021.

[WMSP23] Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou. Counterfactual explanations for time series forecasting. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, pages 1391–1396. IEEE, 2023.

[WSMP24] Zhendong Wang, Isak Samsten, Ioanna Miliou, and Panagiotis Papapetrou. COMET: constrained counterfactual explanations for patient glucose multivariate forecasting. In *37th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2024, Guadalajara, Mexico, June 26-28, 2024*, pages 502–507. IEEE, 2024.

[Xie18] Jinyu Xie. Simglucose v0.2.1. https://github.com/jxx123/simglucose, 2018. Accessed on: 16-09-2024.

[ZGW+21] Liye Zhou, Zhifei Guo, Bijue Wang, Yongqing Wu, Zhi Li, Hongmei Yao, Ruiling Fang, Haitao Yang, Hongyan Cao, and Yuehua Cui. Risk prediction in patients with heart failure with preserved ejection fraction using gene expression data and machine learning. *Frontiers in genetics*, 12:652315, 2021.

[ZLH+18] Taiyu Zhu, Kezhi Li, Pau Herrero, Jianwei Chen, and Pantelis Georgiou. A deep learning algorithm for personalized blood glucose prediction. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018), Stockholm, Sweden, July 13, 2018*, CEUR Workshop Proceedings, pages 64–78. CEUR-WS.org, 2018.

A Appendix

A.1 Detailed Classification Metrics for MIMIC Dataset

The following tables provide comprehensive classification metrics (precision, recall, F1-score, and overall accuracy) for GRU and N-BEATS models across all patient clusters and gender subgroups in the MIMIC dataset.

Feature types	Specifics	Occurrences
Targets		
	Death within 30 days	136
	Death within 1 year	150
Vital signs and laboratory values		
	BMI	1845
	Heart Rate	1845
	SpO2	1845
	Diastolic BP	1837
	Systolic BP	1837
	Temperature	1829
	Creatinine	1825
	Sodium	1825
	Bicarbonate	1823
	Hemoglobin	1814
	WBC Count	1814
	Platelet Count	1812
	Troponin	958
	NT-proBNP	76
Comorbidities		
	Diabetes	798
	RD	768
	Coronary Artery Disease	761
	COPD	658
	Hypertension	599
	PVD	577
	Atrial Fibrillation	424
	AMI	230
	CEVD	218
	Prior Admission	188
	Pulmonary Hypertension	170
	Diabetes + Complications	156
	Dementia	75
	Cancer	65
	Metastatic Cancer	50
	Rheumatoid Disease	50
	Mild LD	46
	PUD	40
	Moderate/Severe LD	30
	HP/PAPL	27
Gender		
	Male	773
	Female	1072

Table 2: All features of the MIMIC data divided by feature types: Target, Vital signs and laboratory values, Comorbidities, and Gender.

Table 4: Classification Metrics for GRU and N-BEATS Models Cluster 0

Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.91	0.86	0.88
		Died	0.00	0.00	0.00
		Overall Accuracy		0.7882	
	1 year	Survived	0.90	0.91	0.90
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8235	
N-BEATS	30 days	Survived	0.92	0.97	0.94
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8922	
	1 year	Survived	0.90	0.96	0.93
		Died	0.06	0.02	0.03
		Overall Accuracy		0.8706	

Feature	Pair 71, 48		Pair 27, 82		Pair 27, 11		Pair 68, 31		Pair 74, 75	
	Act	Norm	Act	Norm	Act	Norm	Act	Norm	Act	Norm
Basal Insulin	2.1575	0.5678	3.1525	0.6280	2.0002	0.2891	2.7798	0.3388	2.9895	0.5143
Bolus Insulin	19.4057	0.5012	21.2238	0.8798	19.8393	0.4116	10.8953	0.2978	15.0757	0.3838
Carbohydrates	100.5660	0.4907	80.1914	0.5888	125.0327	0.4353	31.7565	0.4943	143.2101	0.5476
Exercise Intensity	1.1705	0.5956	1.5558	0.9469	1.1187	0.5106	0.9448	0.2775	0.8671	0.7661

Table 5: Euclidean distances between exogenous features for the five closest target pairs in the OhioT1DM dataset, showing both actual and normalized values.

Table 6: Classification Metrics for GRU and N-BEATS Models Cluster 0

Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.91	0.86	0.88
		Died	0.00	0.00	0.00
		Overall Accuracy		0.7882	
	1 year	Survived	0.90	0.91	0.90
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8235	
N-BEATS	30 days	Survived	0.92	0.97	0.94
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8922	
	1 year	Survived	0.90	0.96	0.93
		Died	0.06	0.02	0.03
		Overall Accuracy		0.8706	

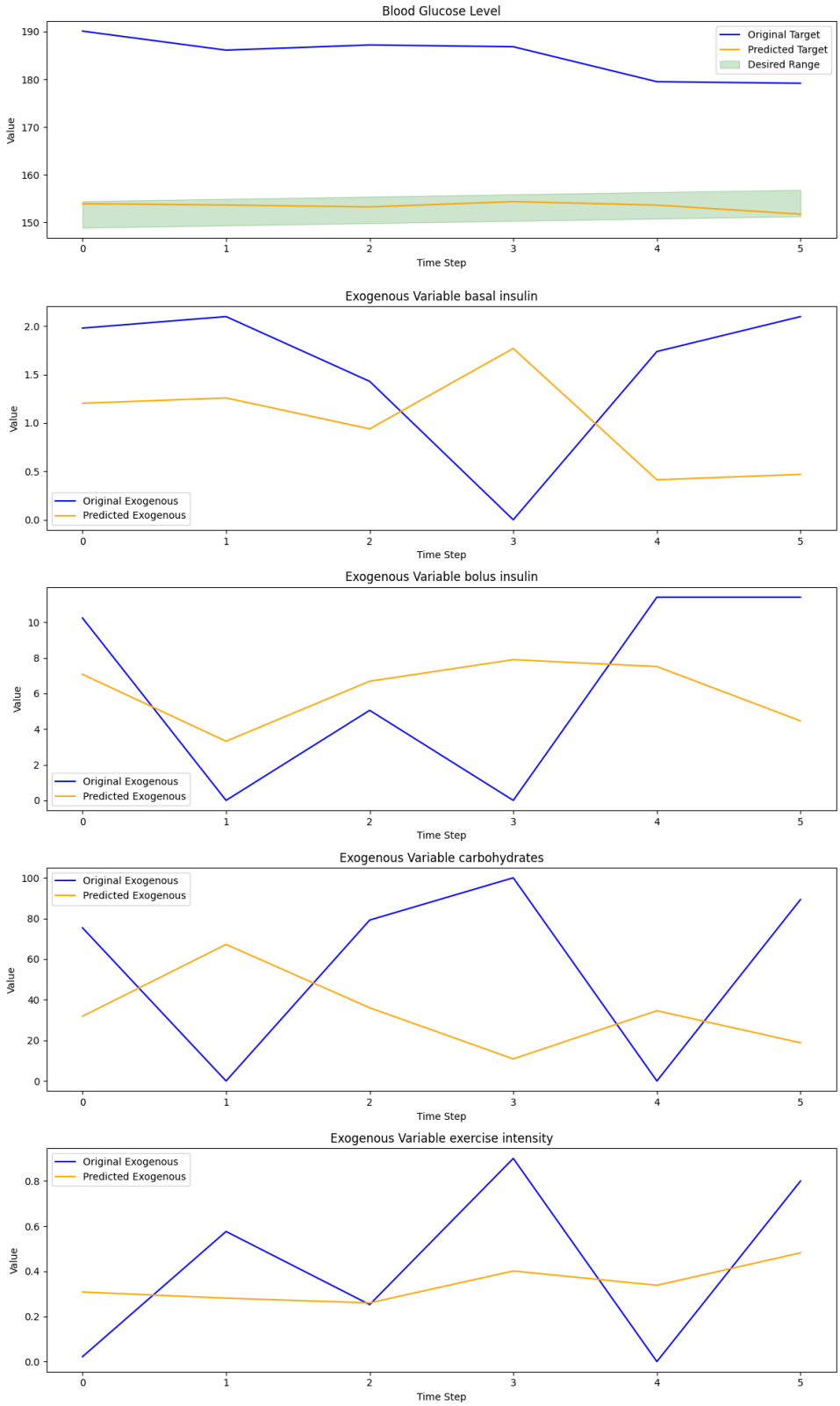


Figure 10: Example of generated counterfactuals for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples. Bottom: Comparison of exogenous variables (original in blue, counterfactual in yellow, bounds in green).

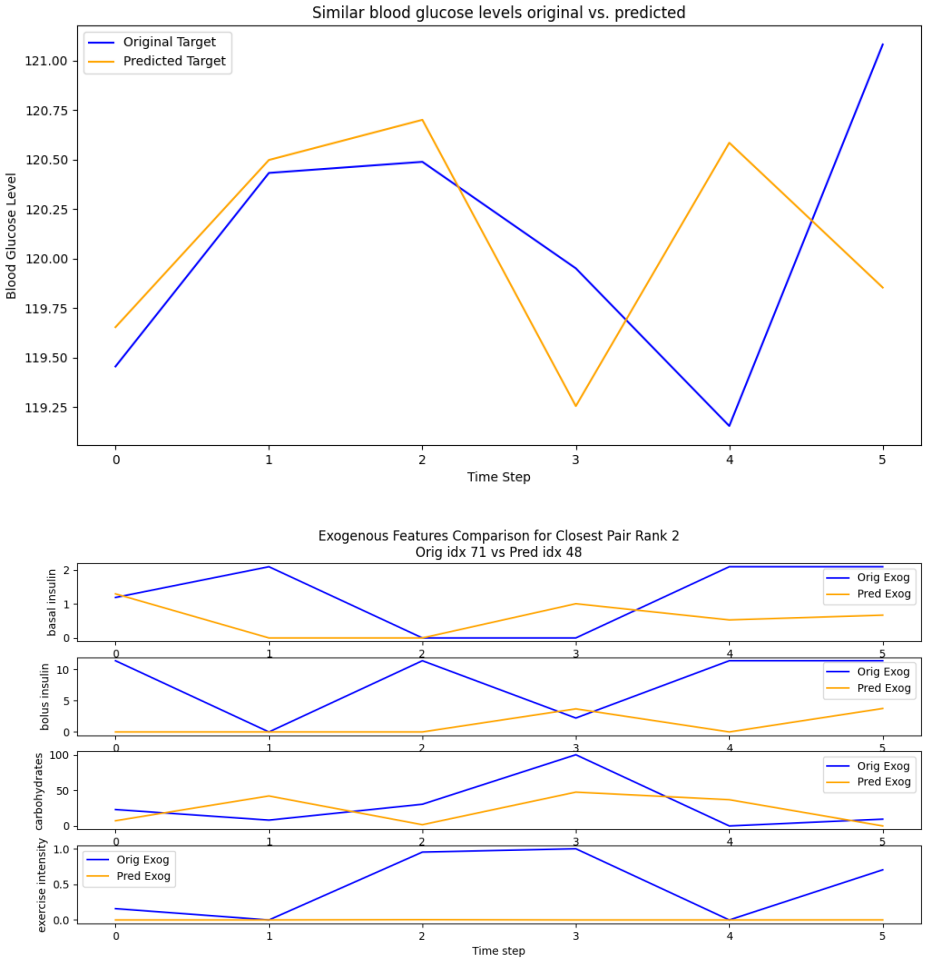


Figure 11: Comparison of two samples from the OhioT1DM dataset with similar blood glucose levels. Top: Blood glucose targets. Bottom: Exogenous variables for both samples (original in blue, comparison sample in yellow).

Table 7: Classification Metrics for GRU and N-BEATS Models Cluster 1

Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.80	0.22	0.35
		Died	0.07	0.50	0.12
		Overall Accuracy		0.2486	
	1 year	Survived	1.00	0.05	0.10
		Died	0.10	1.00	0.19
		Overall Accuracy		0.1475	
N-BEATS	30 days	Survived	0.90	0.93	0.91
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8424	
	1 year	Survived	0.91	0.97	0.94
		Died	0.26	0.08	0.13
		Overall Accuracy		0.8871	

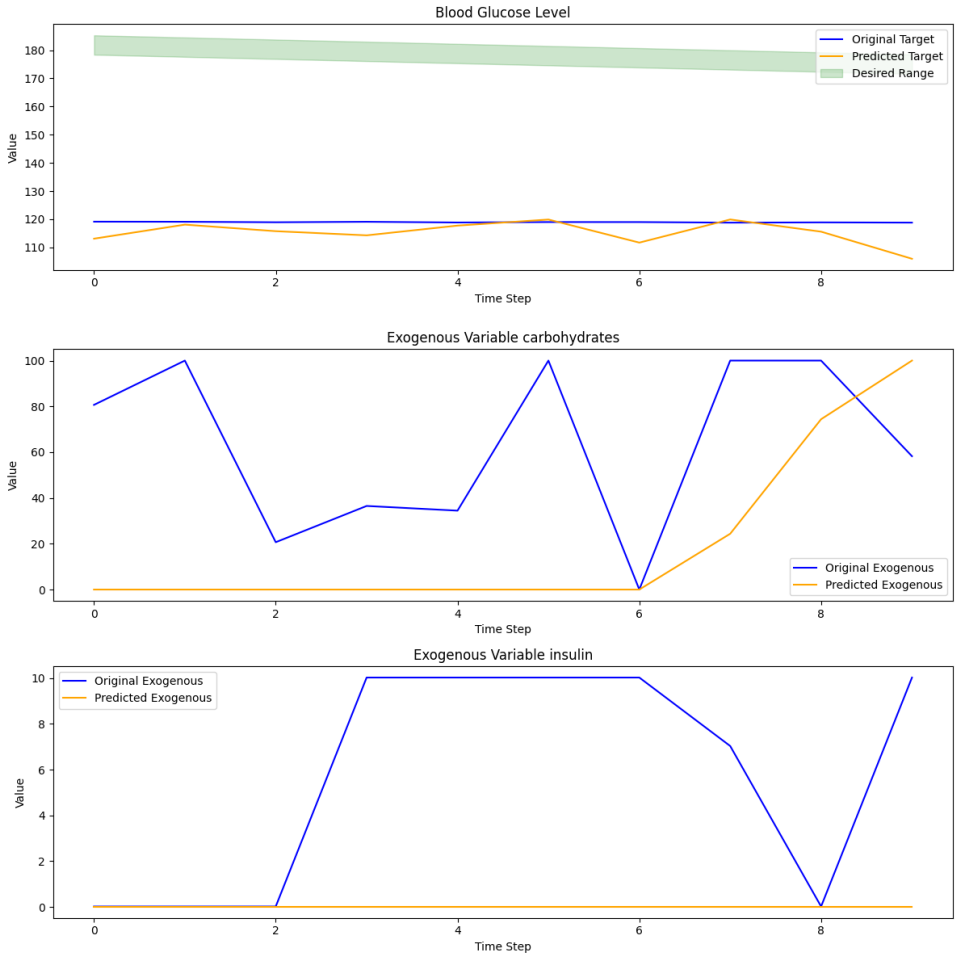


Figure 12: Example of the counterfactuals generated for the SimGlucose dataset, with the blood glucose level within the desired bounds. The different exogenous variables, with the original values in blue and the predicted in yellow.

Table 8: Classification Metrics for GRU and N-BEATS Models Cluster 2

Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.90	0.53	0.67
		Died	0.12	0.50	0.19
		Overall Accuracy		0.5291	
	1 year	Survived	0.91	0.61	0.73
		Died	0.14	0.50	0.21
		Overall Accuracy		0.5979	
N-BEATS	30 days	Survived	0.89	0.99	0.94
		Died	0.18	0.02	0.04
		Overall Accuracy		0.8807	
	1 year	Survived	0.90	0.88	0.89
		Died	0.15	0.18	0.16
		Overall Accuracy		0.7982	

Figure 13: Example of the counterfactuals generated for the MIMIC dataset using SARIMAX, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

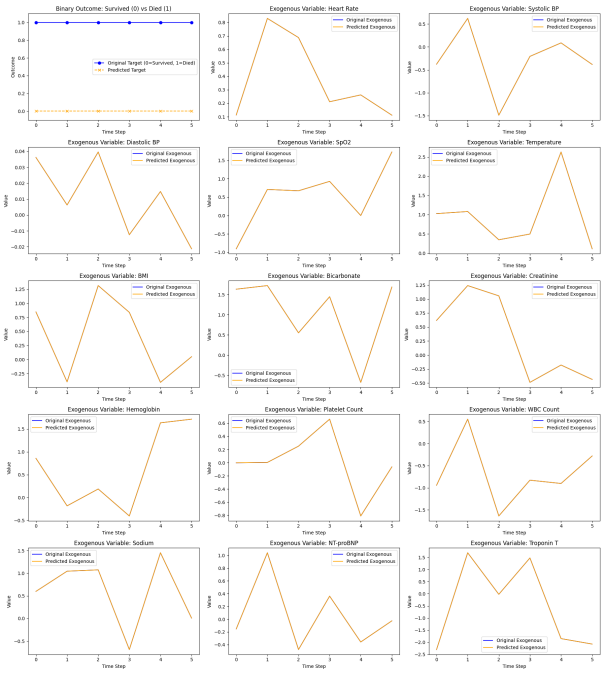


Figure 14: Example of the counterfactuals generated for the MIMIC dataset using OLS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

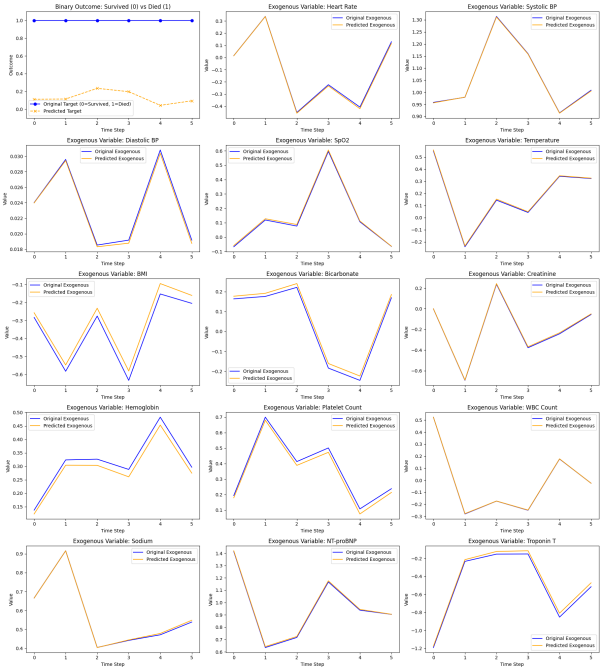


Figure 15: Example of the counterfactuals generated for the MIMIC dataset using GRU, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

Figure 16: Example of the counterfactuals generated for the MIMIC dataset using NBEATS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

Table 9: Classification Metrics for GRU and N-BEATS Models Cluster 3

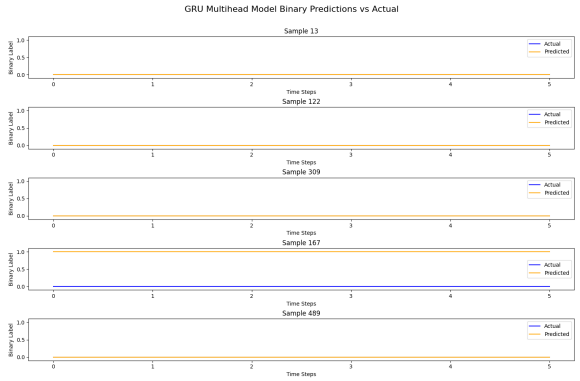
Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	
	1 year	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	
N-BEATS	30 days	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9484	
	1 year	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	

Table 10: Classification Metrics for GRU and N-BEATS Models Female

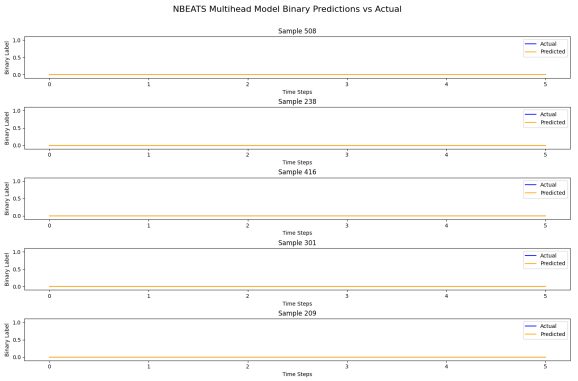
Model	**Class**	**Precision**	**Recall**	**F1-Score**	
GRU	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9414	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9369	
N-BEATS	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9398	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9362	

Table 11: Classification Metrics for GRU and N-BEATS Models Male

Model	Class	Precision	Recall	F1-Score	
GRU	30 days	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9299	
	1 year	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9299	
N-BEATS	30 days	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9298	
	1 year	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9296	

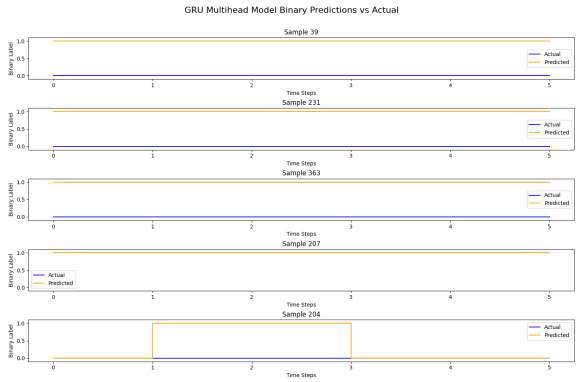


(a) Results of the multivariate forecasting using GRU.

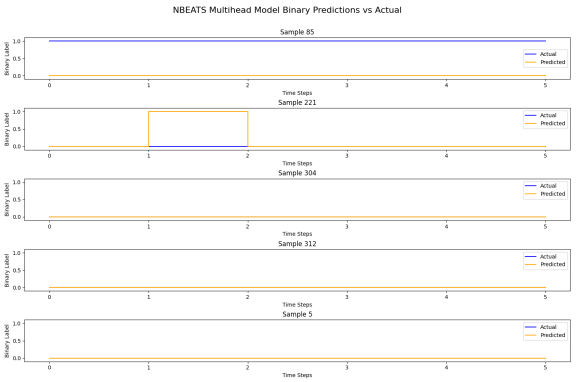


(b) Results of the multivariate forecasting using NBEATS.

Figure 20: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

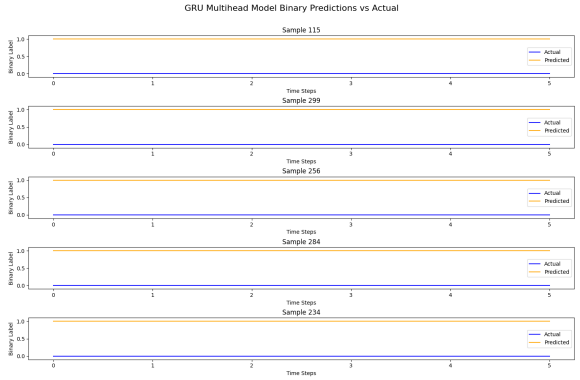


(a) Results of the multivariate forecasting using GRU.

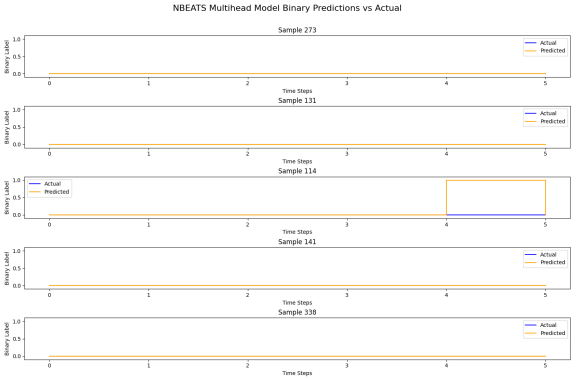


(b) Results of the multivariate forecasting using NBEATS.

Figure 21: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

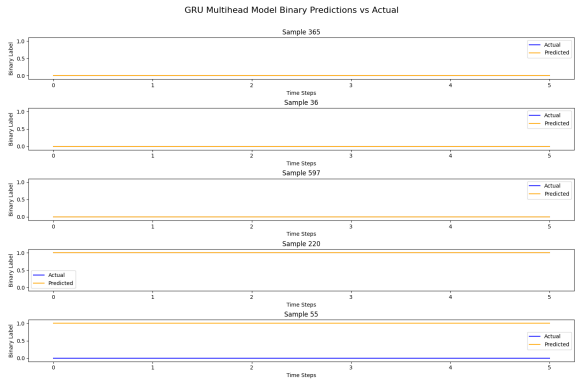


(a) Results of the multivariate forecasting using GRU.

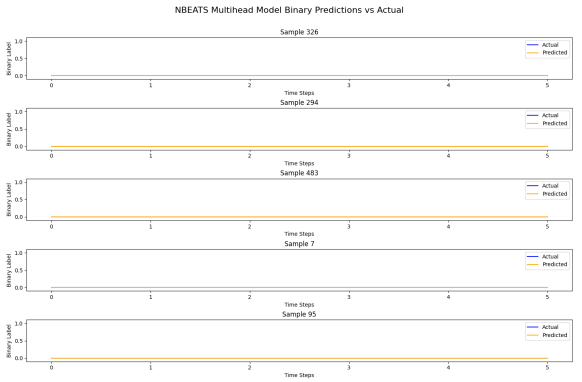


(b) Results of the multivariate forecasting using NBEATS.

Figure 22: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

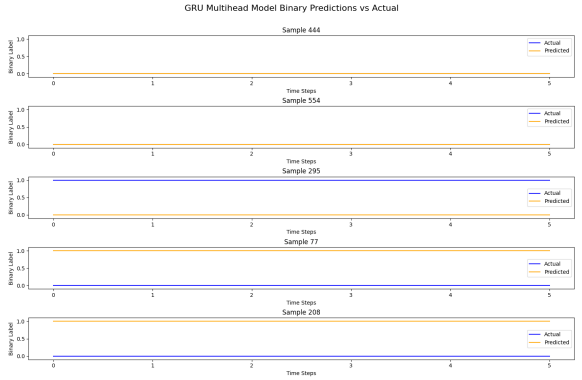


(a) Results of the multivariate forecasting using GRU.

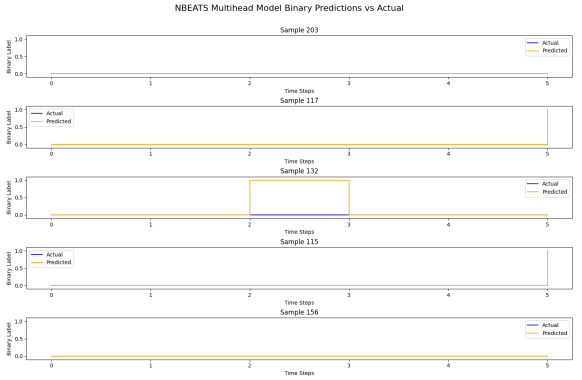


(b) Results of the multivariate forecasting using NBEATS.

Figure 23: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

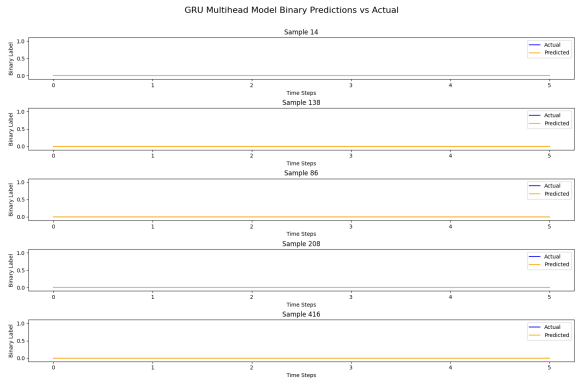


(a) Results of the multivariate forecasting using GRU.

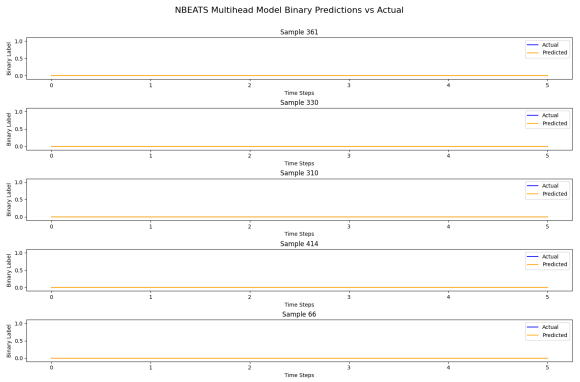


(b) Results of the multivariate forecasting using NBEATS.

Figure 24: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

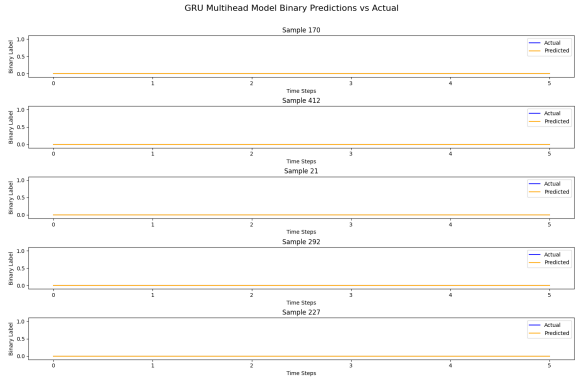


(a) Results of the multivariate forecasting using GRU.

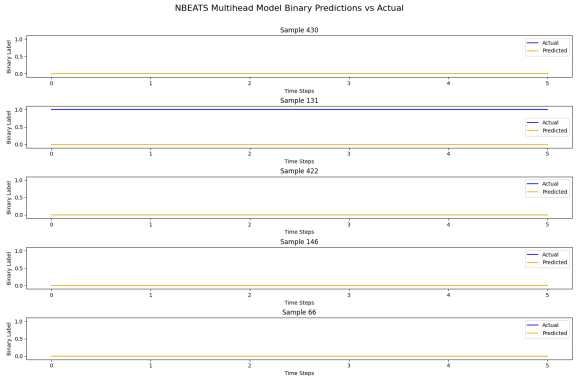


(b) Results of the multivariate forecasting using NBEATS.

Figure 25: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

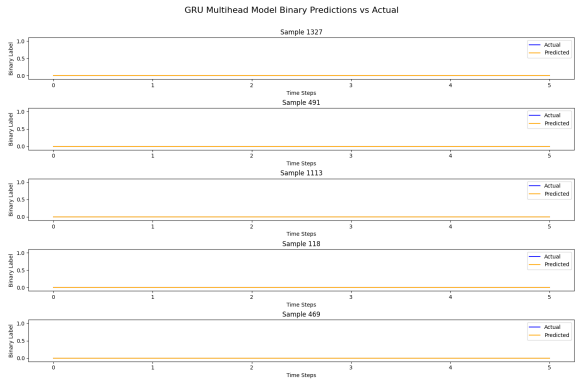


(a) Results of the multivariate forecasting using GRU.

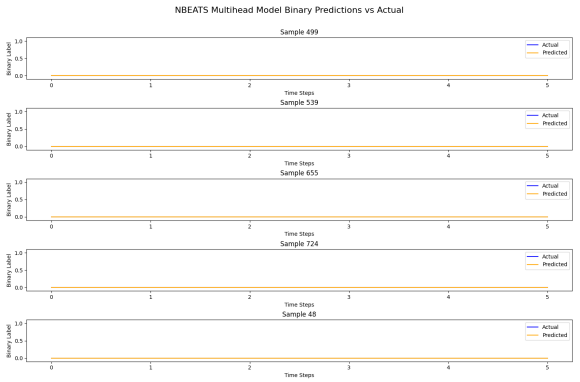


(b) Results of the multivariate forecasting using NBEATS.

Figure 26: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

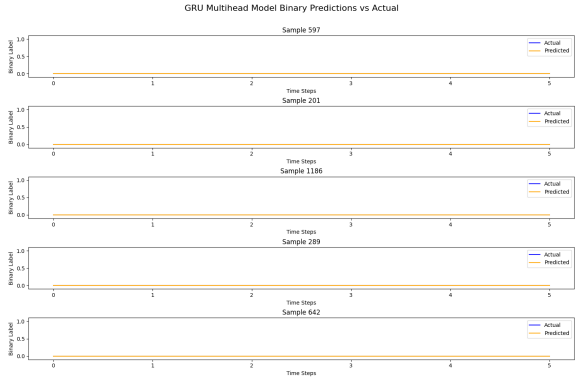


(a) Results of the multivariate forecasting using GRU.

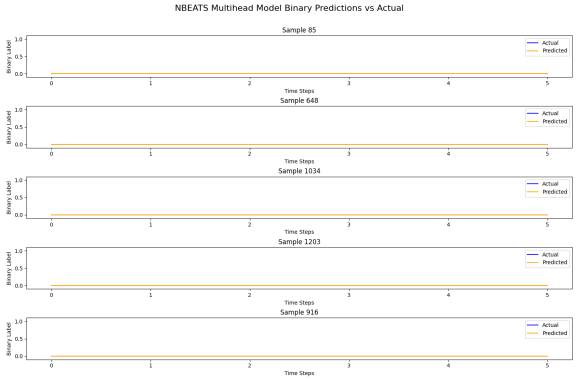


(b) Results of the multivariate forecasting using NBEATS.

Figure 27: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

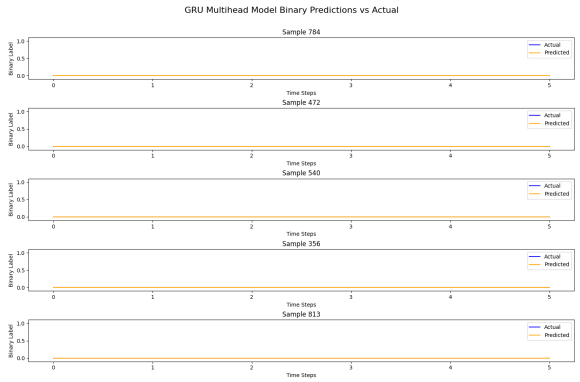


(a) Results of the multivariate forecasting using GRU.

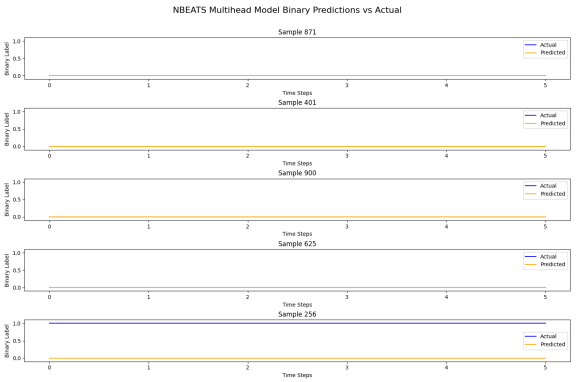


(b) Results of the multivariate forecasting using NBEATS.

Figure 28: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

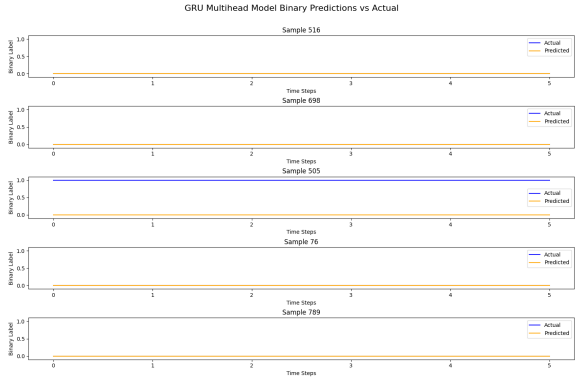


(a) Results of the multivariate forecasting using GRU.

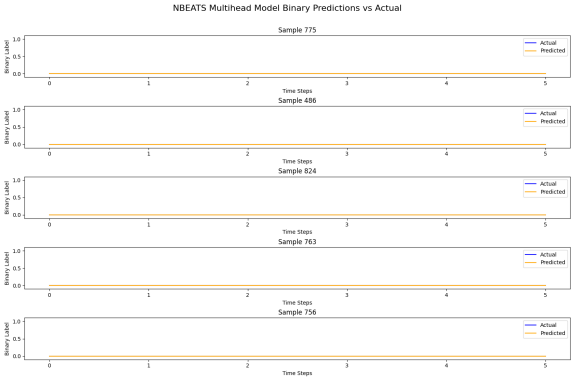


(b) Results of the multivariate forecasting using NBEATS.

Figure 29: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using NBEATS.

Figure 30: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.