Temporal Drift and Uncertainty Analysis in Deep Learning-Based Glucose Forecasting

Yash Kumar Panjwani, Jyotirmoy Das, Satya Prakash, and Asmit Raj

Lovely Professional University, Phagwara, Punjab, India
ypanjwani1110@gmail.com, jyotirmoyofficial11@gmail.com,
sp68028222@gmail.com, asmitraj81@gmail.com

Abstract. Accurate short-term forecasting of glucose is essential for managing Type 1 diabetes using continuous glucose monitoring (CGM) systems. However, the deep learning models deployed in real-world conditions frequently experience performance degradation over time due to temporal drift in physiological states and behavioral patterns. This study investigates how it impacts the temporal drift on forecasting accuracy, uncertainty calibration, and explainability. Using the OhioT1DM dataset from Kaggle, CGM data is partitioned into four chronological windows (T1-T4) to mimic long-term deployment. The Long Short-Term Memory (LSTM) network predicts 30-minute-ahead glucose levels. The epistemic uncertainty is estimated through Monte Carlo Dropout (MC-Dropout) and the evaluated using Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and uncertainty-error correlation. SHAP is applied to quantify the temporal changes in feature contributions. Then Performance declines progressively from T2 to T4, confirming temporal drift in glucose dynamics. In our research, PICP remains below the ideal 0.95 threshold across all windows, indicating under-calibrated uncertainty intervals. SHAP analysis reveals the substantial explainability drift, including the decreasing influence of raw glucose values and increasing dependence on short-term glucose deltas. Whereas Temporal drift significantly affects the predictive accuracy, uncertainty calibration, and feature relevance in deep learning-based glucose forecasting. The paper findings demonstrate the need for more drift-aware monitoring, periodic recalibration, and adaptive modeling strategies for reliable deployment in clinical CGM environments.

Keywords: Glucose forecasting · Temporal drift · Deep learning · Uncertainty estimation · SHAP explainability· Uncertainty estimation.

1 Introduction

Continuous glucose monitoring (CGM) enables high-frequency acquisition of glucose dynamics and has become central to automated insulin delivery and real-time decision support. Short-term glucose forecasting, particularly at 30-minute horizons, is essential for mitigating adverse glycemic events. Deep learning models, most notably Long Short-Term Memory (LSTM) networks have shown strong capability in modeling nonlinear and temporally correlated CGM patterns. Most previous forecasting studies treat glucose patterns as if they remain constant over time, even though this is rarely true in real-life situations.

Glucose trajectories are influenced by evolving behavioral and physiological factors, including meal timing, insulin dosing strategies, circadian rhythm variation, stress, sensor noise, and long-term metabolic changes. These factors introduce temporal drift, in which the data distribution gradually shifts over time. Under such drift, a model trained on early CGM segments may become misaligned with later segments, leading to degradation in forecasting accuracy.Recent work in medical time-series modeling has shown that temporal drift is a key failure mode for deployed clinical machine learning systems, yet glucose forecasting literature continues to rely on random train-test splits that obscure drift effects.

Another critical aspect during clinical implementation is uncertainty of quantification. Predictive models should not only produce the point estimates but also deliver calibrated uncertainty intervals that reflect epistemic risk, especially when patterns drift or are out of distribution. Monte Carlo Dropout (MC-Dropout) provides a feasible Bayesian approach for approximation the estimate of epistemic uncertainty in deep networks through stochastic forward passes. Nonetheless, its reliability through calibration under temporal drift—specifically, prediction interval coverage and uncertainty-error correlation—has not been systematically assessed in glucose forecasting contexts.

Model interpretability is equally imperative. The ability to understand shifts in feature relevance over time may provide valuable insight into physiological variability as well as dependencies of the model on input variables. Although SHAP (SHapley Additive exPlanations) provides theoretically defensible explanations of feature attributions, previous glucose forecasting studies have not examined how SHAP values differ across time-segmented datasets. As drift occurs, shift in feature importance is possible which can be described or reflected in terms of explainablity drift, as model behavior evolves.

To address these gaps, this paper provided a comprehensive assessment of temporal drift, uncertainty drift, and explainabilty drift for deep learning based glucose prediction systems. Using CGM data divided into four time-ordered windows (T1 - T4), we evaluated the time-based changes in an LSTM model's performance, the reliability of uncertainty from MC-Dropout, and the contributions of features derived from SHAP. Our findings indicate that drift has a meaningful impact on predictive performance, uncertainty reliability, and feature importance, which signifies a need for adequate drift monitoring, and other adaptive modeling strategies, for long-term automated glucose prediction systems.

2 Related Work

Research on glucose forecasting has progressed through advances in deep learning, uncertainty estimation, and temporal drift analysis. This section reviews related works across three major areas: glucose prediction models, temporal distribution shift, and uncertainty and explainability methods.

2.1 Glucose Forecasting Models

Early machine learning applications for glucose prediction relied on shallow neural networks and classical approaches. Zecchin et al. [1] demonstrated that feed-forward neural networks could capture basic glucose patterns but lacked robustness due to limited physiological inputs. Plis et al. [2] compared classical regression and early ML models, emphasizing the need for more expressive temporal architectures.

Deep learning significantly improved glucose forecasting performance. Li et al. [3] introduced GluNet, a CNN-LSTM hybrid capable of modeling multi-scale CGM dynamics. Zhu et al. [4] proposed a multi-scale residual CNN for short-term glucose prediction, achieving improved performance by capturing hierarchical temporal representations. Sun et al. [5] evaluated multiple deep architectures, concluding that LSTM-based models perform best for 30-minute-ahead forecasting.

Contextual feature integration has also been investigated. Alqudah et al. [6] showed that glucose prediction accuracy improves when longer historical windows are provided to LSTM models. Mirshekarian et al. [7] in their work incorporated meal, insulin, and activity indicators, reducing forecasting error in real-world conditions. Personalized approaches such as those by Montaser et al.[8] show that models specific to subjects tend to perform better than population-level ones. Mohiuddin et al. [9], recent survey, noted that most of the glucose forecasting studies still rely on random data splits, often missing the temporal drift issues that might appear during real-world deployment.

2.2 Temporal Drift in Medical Time Series

Temporal drift, or distribution shift, occurs whenever data characteristics change over time.Quinonero-Candela et al. [10] formally defined major types of drift and their impact on predictive modeling. Widmer and Kubat [11] introduced early adaptive learning methods for non-stationary environments.Surveys by Gama et al.[12] and Lu et al.[13] describe several drift methods along with adaptive modeling strategies.

Healthcare systems are particularly sensitive and vulnerable to drift. Liu et al. [14] showed that temporal distribution shifts degrade clinical prediction models, emphasizing the need for time-aware evaluation. Purushotham et al. [15] demonstrated that deep models trained on historical EHR data decline in accuracy when applied to later time segments. Ricci et al. [16] further reported that drift impact both prediction accuracy and feature importance stability in patient time series.

Despite these findings, glucose forecasting research rarely addresses drift. Existing works mostly focus on assuming stable glucose dynamics, even though CGM signals naturally shift due to behavioral, physiological, and device-related changes. This gap highlights the need for a systematic analysis of performance drift, uncertainty drift, and explainability drift in glucose forecasting models.

2.3 Uncertainty Estimation and Explainability

Reliability of medical systems, which are heavily based on AI and ML requires calibration to estimate the uncertainty. Gal and Ghahramani [17] in their paper they had introduced Monte Carlo Dropout (MC-Dropout), to prove that dropout can approximate Bayesian inference and can quantify epistemic uncertainty. Lakshminarayanan et al. [18] proposed that deep ensembles used for robust uncertainty estimation can surpass the single-model approaches at the drawbacks of increased computational cost. Kuleshov et al. [19] demonstrated that the deep neural networks sometimes give poorly calibrated predictive distributions, motivating post-hoc calibration strategies. Levi et al. [20] applied calibration techniques to time series data for a medical use case and observed under-coverage of uncertainty intervals under drift. Conformal prediction is reviewed by Papadopoulos [21], which offers a distribution-free guarantee that features as a benchmark and a systematic method for uncertainty evaluation.

Explainability in medical scenarios is equally crucial for medical simulation. Lundberg and Lee [22] introduced SHAP, a Shapley-value-based framework for interpreting complex models that is based on Shap. Moitra et al. [23] had extended the explainability function to clinical time series and noted that feature attributions may vary across temporal segments. Zhou et al. [24] proposed that monitoring explainability drift in deployed ML systems, proving that feature reliance can shift even when predictive accuracy remains stable. Another recent work by Zhang et al. [25] proposed feature importance monitoring of feature importance which is tailored for real-world medical prediction workflows, emphasizing the role of explainability drift in long-term model maintenance.

Collectively, prior works suggest that the existing glucose forecasting approach frameworks rarely integrate temporal drift evaluation, uncertainty quantification, and explainability analysis. This motivates our study, which examines predictive drift, uncertainty drift, and SHAP-based feature drift across four chronological time windows using a deep learning-based glucose forecasting model.

3 Methodology

This section describes the dataset, data preprocessing, feature engineering, temporal segmentation strategy, model building using LSTM, MC dropout uncertainty estimation, and explainability analysis using SHAP which is used to evaluate temporal drift in deep learning–based glucose forecasting.

3.1 Dataset and Preprocessing

In our experiment, we used the **OhioT1DM** datasets from Kaggle for our continuous glucose monitoring (CGM) analysis. This dataset has features like timestamps, glucose readings after every 5 minutes, insulin doses, carbohydrate doses, basal rate and other contextual features. Each patient contribute their own time-series data.

The data preprocessing involved data cleaning, timestamp formatting, and normalization. We converted insulin and meal intake to binary values so that it shows whether a recent intake or dose occurred. We also encoded time using cyclic features to keep the daily patterns in data.

$$hour\_sin = \sin\left(2\pi \frac{hour}{24}\right) \tag{1}$$

$$hour\_cos = \cos\left(2\pi \frac{hour}{24}\right) \tag{2}$$

Categorical insulin types were encoded using one-hot encoding. Continuous features were standardized using statistics computed only from the early temporal segments to prevent information leakage.

3.2 Feature Engineering and Temporal Segmentation

We created some of the features to track the glucose levels. We looked at the 5 minutes and 10 minutes differences using the first and second order calculations

$$\Delta g_5(t) = g(t) - g(t-5) \tag{3}$$

$$\Delta g_{10}(t) = g(t) - g(t-10) \tag{4}$$

g(t) and g(t) just measure how much the glucose level changes over the last 5 or 10 minutes. They basically help us catch quick ups and downs in glucose that the raw CGM values might miss

For a slower, overall trend, we also computed a 30-minute slope:

$$slope_{30}(t) = \frac{g(t) - g(t-30)}{30} \tag{5}$$

$slope_{30}(t)$ shows how the glucose level has been trending over the last 30 minutes. It basically tells us how fast the glucose is rising or falling during that time.

We extracted rolling mean and rolling standard deviation of a glucose level using six previous 5 minutes glucose readings. These features combines the local trends and variability that helps the model to understand the recent glucose behavior.

$$\mu_{30}(t) = \frac{1}{6} \sum_{i=1}^{6} g(t-5i) \tag{6}$$

$\mu_{30}(t)$ basically just means the average glucose over the last 30 minutes. We get it by using the six previous 5-minute readings. It gives a smoother idea of where the glucose has been recently, so the model can understand what's going on at time t.

$$\sigma_{30}(t) = \sqrt{\frac{1}{6}\sum_{i=1}^{6}(g(t-5i)-\mu_{30}(t))^2} \tag{7}$$

Similarly, $\sigma_{30}(t)$ tells us how much the glucose level has been changing over the last 30 minutes. If $\sigma_{30}(t)$ is high, it means the glucose has been pretty unstable, which helps the model guess when a sudden rise or drop might happen.

Insulin-on-board (IOB) was estimated with a simple exponential decay model:

$$IOB(t) = \sum_{\tau=0}^{T} d(t-\tau)\exp\left(-\frac{\tau}{\lambda}\right). \tag{8}$$

This formula just tells us how much insulin is still active at time t. Here, $d(t-\tau)$ is the insulin taken τ minutes earlier, and the exponential part shows how its effect slowly fades. The value λ basically controls how fast the insulin wears off. Our prediction target was the glucose level 30 minutes into the future:

$$y(t) = g(t + 30\,min) \tag{9}$$

Here, $y(t)$ is the value we want to predict — the glucose level 30 minutes from the current time t. This is the 30-minute-ahead prediction we use in all our experiments

We split the dataset chronologically into four equal partitions to study temporal drift across different time segments.

$$\{T1, T2, T3, T4\}.$$

T1–T2 were used for training, T3 for temporal validation, and T4 for temporal testing, reflecting real deployment scenarios.

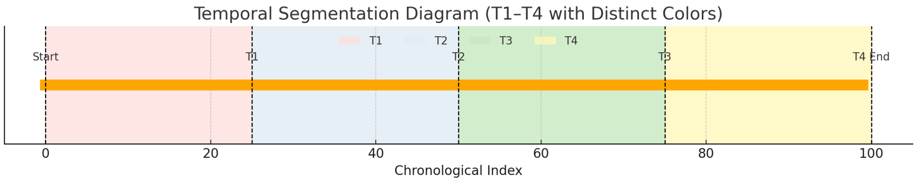

Fig. 1. Temporal segmentation of the OhioT1DM dataset into four chronological windows (T1–T4) used for evaluating temporal drift.

3.3 Sequence Modeling and LSTM Forecasting

Sequences were constructed using a lookback window of L steps:

$$X_t = [x(t-L+1), \ldots, x(t)],$$

with $y(t)$ as the target. Sequence generation was performed independently per subject.

We created a forecasting model LSTM which consists of a spatial dropout1D layer, a 128-unit LSTM with recurrent dropout, a dense hidden layer (ReLU), and a linear output layer for regression. We used Mean Absolute Error (MAE) loss function during training.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{10}$$

The term $\mathcal{L}$ denotes the Mean Absolute Error (MAE) loss used for training the LSTM forecasting model. Here, N represents the total number of training samples, y_i is the true glucose value at the i-th time step, and $\hat{y}_i$ is the model's predicted value. The MAE measures the average absolute deviation between predictions and ground-truth readings, providing a stable and interpretable error metric for physiological time-series forecasting. Model optimization was performed using the Adam optimizer with early stopping to prevent overfitting.

3.4 Uncertainty and Explainability Analysis

Epistemic uncertainty was estimated using Monte Carlo Dropout (MC-Dropout) by performing T stochastic forward passes:

$$\hat{y}_t^{(k)} = f(X_t; W, dropout = p), \quad k = 1, \ldots, T$$

Predictive mean and variance were computed as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{11}$$

$$\mu_t = \frac{1}{T}\sum_{k=1}^{T}\hat{y}_t^{(k)} \tag{12}$$

$$\sigma_t^2 = \frac{1}{T}\sum_{k=1}^{T}\left(\hat{y}_t^{(k)} - \mu_t\right)^2 \tag{13}$$

A 95% prediction interval was formed as:

$$PI_t = [\mu_t - 1.96\sigma_t,\ \mu_t + 1.96\sigma_t]$$

Model explainability was assessed using SHAP values. For an input x, the prediction is decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{M}\phi_i \tag{14}$$

where ϕ_i is the contribution of feature i. Comparing SHAP distributions across T1–T4 allows quantifying explainability drift.

4 Results and Discussion

This section provides the performance of LSTM under four time segments (T1-T4), followed by uncertainty estimation using monte carlo dropout and a temporal Explainability evaluation using SHAP. Together, these analyses quantify temporal drift in accuracy, uncertainty calibration, and feature contributions.

4.1 Predictive Performance Across Temporal Windows

Table 1 states the Mean absolute error (MAE) and Root mean square error (RMSE) accross all the time segments. As we can see that T2 achieves lowest MAE (15.554 mg/dL) and RMSE (20.717 mg/dL) score which indicates best performance under T2. But T3 and T4 achieves greater Error scores which clearly indicates a degradation of performance as a later portion of the data.

Table 1. Forecasting performance across temporal windows.

Window	MAE (mg/dL)	RMSE (mg/dL)
T1	16.912	24.041
T2	15.554	20.717
T3	17.543	24.856
T4	18.483	25.575

From T2 to T4, the error values keep going up, which shows that the model is not doing as well in the later time windows as it did earlier. This drop in performance known as temporal drift. In long-term glucose data, the patterns often change because the person's routine also changes — things like food timing, sleep shifts, stress levels, or general physiological behavior. When these changes happen, the model starts seeing patterns it never learned during training, and naturally, its predictions become less accurate.

4.2 Uncertainty Behavior Under Temporal Drift

Monte Carlo dropout (MC dropout) was calculated to measure the uncertainty in glucose prediction model. For every input sample, the model made 50 different predictions and based on those we calculated predictive mean, epistemic uncertainty and calibration metrics such as Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and the correlation between absolute error and predicted uncertainty.

Across all time stamps, PICP remained below the ideal 0.95 threshold, indicating *under-coverage* of uncertainty intervals. While MPIW remained relatively stable, the true error increased over time, showing that MC-Dropout does not automatically adapt to temporal drift.

The uncertainty error correlation remained positive across all windows, confirming that epistemic uncertainty still serves as a useful risk indicator. However,

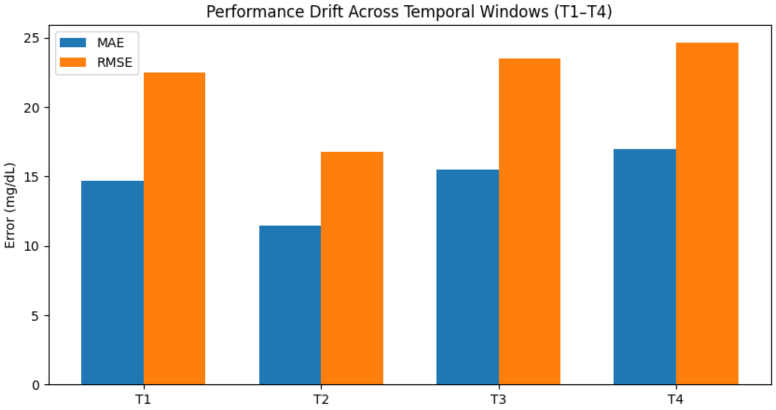

Fig. 2. Prediction performance (MAE and RMSE) across temporal windows T1–T4, illustrating accuracy degradation due to temporal drift.

Window	**PICP (95% target)**	**MPIW**	**Corr($\|err\|$, σ)**
T1	0.761	42.50	0.205
T2	0.816	41.29	0.154
T3	0.731	41.47	0.225
T4	0.690	41.21	0.171

Table 2. Uncertainty metrics across temporal windows.

its magnitude decreased from early to late windows, suggesting weakened reliability of uncertainty estimates as drift intensifies.

These results show that standard MC-Dropout captures general uncertainty structure but fails to calibrate intervals appropriately under drift. A recalibration mechanism, such as conformal prediction, may be necessary for real deployment.

4.3 Explainability Drift via SHAP

SHAP was used to compute global feature attributions for each temporal window. Figure 3 summarizes the importance of features between T1–T4.

Several patterns indicate clear *explainability drift*:

- The contribution of raw glucose values decreased from T1 to T4, suggesting reduced predictive power due to change in physiological dynamics.
- Rate-of-change features (glucose deltas and slopes) increased in importance in later windows, which indicates the model relied more heavily on short-term dynamics when long-term patterns became less stable.
- The variability measures like the rolling of standard deviation, fluctuated significantly, and reflecting changing volatility in glucose pattern.

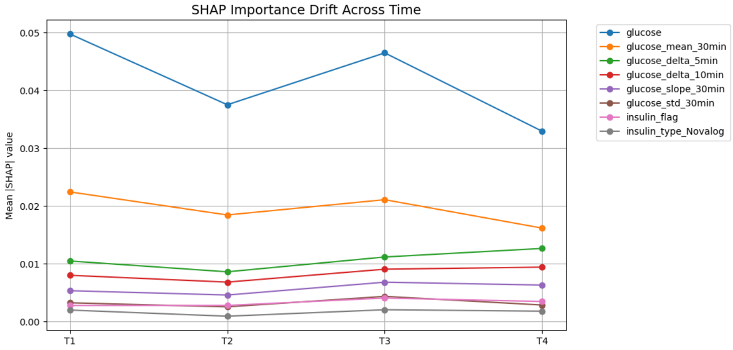

Fig. 3. SHAP feature global importance across temporal windows.

– The Behavioural features like meal, insulin etc exhibit small but noticeable shifts across windows (T1-T4) and also it consistent with real world lifestyle changes

These trends corroborate the drift observed in predictive performance. As glucose behavior evolves, the model changes its reliance on different features, revealing the shifts in physiological regimes and user behavior. Explainability drift therefore serves as an important companion signal for diagnosing temporal drift in medical forecasting systems.

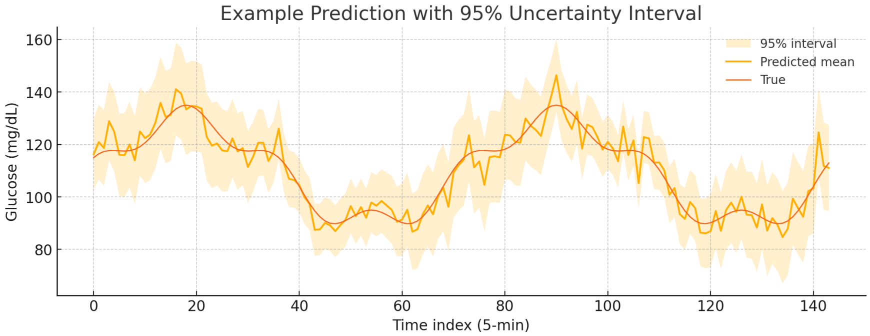

Fig. 4. Example glucose prediction showing predicted mean and the corresponding 95% uncertainty interval generated by MC-Dropout. Wider intervals correlate with regions of greater model uncertainty.

4.4 Discussion

The combined results shows that glucose forecasting model that is trained on early time segments face significant challenges on the future time segments due to several lifestyle changes. Below, we shown the temporal drift in three ways:

1. **Accuracy Drift** - The model performance from T2 to T4 has been decreased which shows the relationship between input features and future glucose levels that decrease over time.
2. **Uncertainty Drift** - During temporal drift, the uncertainty estimates also starts to drift. We noticed that the prediction intervals often fail to cover the true values, and the link between the model's uncertainty and its actual error becomes weaker. This basically shows that MC-Dropout is not as well calibrated once the distribution shifts.
3. **Explainability Drift** - SHAP analysis reveals that feature importance also changes across all the time stamps. It showed glucose behavior changed and model adapted by relying on different signals.

Together, these findings show the necessity of drift-aware model monitoring, periodic recalibration, and potentially adaptive forecasting frameworks for real-world continuous glucose monitoring. The proposed analysis framework showcases that how accuracy, uncertainty, and explainability jointly reveal hidden temporal failures that conventional random-split evaluations overlook.

5 Conclusion

This work investigated the influence of temporal drift on deep learning-based glucose forecasting by conducting a detailed evaluation of prediction accuracy, calibration uncertainty, and the dynamics of explainability.Utilizing the LSTM model, trained on earlier timeframes and tested on subsequent chronological windows, we established that forecasting performance declines as glucose patterns change over time.The uncertainty revealed by Monte Carlo Dropout consistently showed under-coverage of prediction intervals across all segments, indicating that epistemic estimates become increasingly unreliable in the face of drift.Furthermore, SHAP analysis indicated the importance of distribution shifts across temporal segments, reflecting changes in physiological behavior, glucose variability, and the model's dependence on both short-term and long-term predictors.In summary, these findings indicate that temporal drift affects not just accuracy, but also the model's internal decision-making processes and confidence in its decisions.

Together, the results emphasize the need to have drift-aware evaluation protocols during deploying glucose forecasting models in real-world clinical environment. Conventional random data splits failed to expose long-term degradation, whereas temporal evaluation reveals meaningful shifts during model behavior that must be accounted in the time of during deployment and monitoring.

References

1. Zecchin, C., et al.: Neural network models for glucose prediction. IEEE Trans. Biomed. Eng. **59**(9), 2470–2477 (2012)
2. Plis, K., et al.: Machine learning methods for blood glucose prediction. Diabetes Technol. Ther. **16**(6), 381–390 (2014)
3. Li, K., et al.: GluNet: A deep learning framework for continuous glucose monitoring. IEEE J. Biomed. Health Inform. **23**(1), 188–198 (2019)
4. Zhu, T., et al.: Blood glucose prediction via multi-scale deep residual networks. IEEE Access **7**, 22627–22640 (2019)
5. Sun, Q., et al.: Short-term blood glucose prediction with deep learning. Sensors **20**(18), 1–14 (2020)
6. Alqudah, A., et al.: LSTM-based continuous blood glucose prediction. Comput. Biol. Med. **136**, 104675 (2021)
7. Mirshekarian, S., et al.: MV-LSTM: Multi-variable LSTM-based blood glucose prediction. In: EMBC, pp. 7064–7067. IEEE (2019)
8. Montaser, E., et al.: Personalized deep learning for glucose forecasting. IEEE Access **10**, 18090–18100 (2022)
9. Mohiuddin, M., et al.: A review of glucose prediction algorithms for diabetes management. J. Healthc. Inform. Res. **5**, 1–20 (2021)
10. Quinonero-Candela, J., et al.: Dataset Shift in Machine Learning. MIT Press, Cambridge (2009)
11. Widmer, G., Kubat, M.: Learning in the presence of concept drift. Mach. Learn. **23**(1), 69–101 (1996)
12. Gama, J., et al.: A survey on concept drift adaptation. ACM Comput. Surv. **46**(4), 1–37 (2014)
13. Lu, J., et al.: Learning under concept drift: A review. IEEE Trans. Knowl. Data Eng. **31**(12), 2346–2363 (2019)
14. Liu, B., et al.: Temporal distribution shifts in medical AI. Nat. Mach. Intell. **3**, 199–205 (2021)
15. Purushotham, S., et al.: Benchmarking deep learning models on EHR data. Sci. Rep. **8**, 1–13 (2018)
16. Ricci, F., et al.: Temporal instability of predictive models in healthcare. Patterns **2**(9), 100361 (2021)
17. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty. In: ICML, pp. 1050–1059 (2016)
18. Lakshminarayanan, B., et al.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, pp. 6405–6416 (2017)
19. Kuleshov, V., et al.: Accurate uncertainties for deep learning. In: ICML, pp. 2796–2804 (2018)
20. Levi, D., et al.: Uncertainty calibration for medical time series. In: NeurIPS ML4H Workshop (2020)
21. Papadopoulos, H.: Conformal Prediction for Reliable Machine Learning. Springer, Cham (2023)
22. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: NeurIPS, pp. 4765–4774 (2017)
23. Moitra, D., et al.: Explainability in medical time series deep learning models. Artif. Intell. Med. **130**, 102318 (2022)
24. Zhou, B., et al.: Feature importance drift monitoring in machine learning systems. In: KDD Workshop on Monitoring and Explainability (2020)
25. Zhang, Y., et al.: Monitoring feature drift in deployed medical prediction models. IEEE J. Biomed. Health Inform. **28**(2), 700–712 (2024)