International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert/2026/v13/i7/v13i701
Volume 13, Issue 7, July 2026, Pp.1-7

Research Paper

Personalized Forecasting and Anomaly Detection for IoT Health Monitoring

[1*] T. Durga, [2] P. Venkata Krishna

[1*] *Department of Computer Science and Engineering, Sri Padmavati Mahila University,Tirupati, India*
Email: durga.csit05@gmail.com
[2] *Professor, Department of Computer Science and Engineering, Sri Padmavati Mahila University, Tirupati, India*
Email: parimalavk@gmail.com

**Corresponding Author: durga.csit05@gmail.com*

Received: 02/04/2026 Revised: 18/05/2026 Accepted: 25/07/2026 Published: 31/07/2026

Abstract: - Reliable analytics for Internet of Things (IoT) health monitoring require temporal evaluation, correctly aligned anomaly scores and meaningful baselines. This study evaluates statistical and machine-learning methods using two fully synthetic data families. We first reproduce an existing comparison of univariate ARIMA and a multilayer perceptron (MLP) forecaster, and of Isolation Forest and an MLP autoencoder, on ten one-day simulated patients. Although the reported results reproduce, the anomaly experiment uses in-sample scoring and future observations, and averages window errors against point labels. Holding the fitted autoencoder unchanged and correcting only score alignment increases glucose F1 from 0.164 to 0.882. In a separate causal evaluation on 30 simulated glucose trajectories, mean F1 is 0.833 for Isolation Forest and 0.826 for the corrected autoencoder. A training-mean baseline reduces glucose RMSE from ARIMA's 18.89 to 13.32 mg/dL on the original fixed-origin task. Further experiments use seven-day trajectories with daily structure, temporally correlated fluctuations and injected events. On 30 fresh trajectories, a patient-specific daily template with optional recent-residual correction achieves 30-minute RMSE of 3.37 mg/dL, compared with 4.37 for previous-day prediction and 5.75 for Kalman forecasting. Seasonal residual alerts with a separate sensor-quality check perform well on the tested perturbations; adding CUSUM does not improve recall for these scenarios. These findings support stronger evaluation and a simple personalized baseline, but do not establish clinical effectiveness or general superiority across model classes.

Keywords- IoT health monitoring; glucose forecasting; anomaly detection; causal evaluation; patient-specific baseline; synthetic time series.

---

1 Introduction

Wearable and portable monitoring systems motivate the analysis of repeated physiological measurements outside conventional clinical encounters [1]. An IoT architecture can provide a route from sensors to patient and clinician interfaces, but an architectural description does not establish that its forecasts or alerts are reliable. Such claims require evidence about data provenance, information available at prediction time, calibration and performance on unseen observations.

The central problem addressed here is the evaluation of forecasting and anomaly detection within a proposed chronic-disease monitoring framework. We distinguish unusual measurements from clinical events and sensor faults. A statistically rare reading is not automatically a harmful event, and success in detecting a deliberately injected spike is not evidence of improved disease management. We therefore study measurable computational tasks and explicitly restrict all reported performance to simulation.

An initial analytical pipeline combined Isolation Forest with a forecaster described as VARIMA. Inspection of the executable comparison identifies independent univariate ARIMA (2, 1, and 2) fits instead. The analysis also reveals that the autoencoder scores a centred window while using a label for a single timestamp. These implementation details can change conclusions more substantially than replacing one algorithm with a more complex model.

The contributions are threefold. First, we reproduce the supplied experiment and isolate the effect of score-to-label alignment. Second, we introduce chronological, past-only evaluation and straightforward forecasting and detection baselines. Third, we compare Kalman-based and seasonal approaches on temporal simulations with explicit event and false-alert metrics. The seasonal follow-up is exploratory and does not constitute a novel algorithm or a validated clinical system.

2 Related methods and system scope

Isolation Forest isolates observations through random feature partitions, using shorter isolation paths as evidence of unusualness [2]. Its score is distinct from the decision threshold. A threshold selected with a known injected anomaly proportion answers a different question from one fixed before future data arrive. Reconstruction-based detection likewise depends on which reconstruction error is associated with each labelled timestamp; a window-level target and a point-level target are not interchangeable.

State-space filtering provides a principled way to estimate an evolving latent level and trend from noisy observations [3]. Cumulative sum (CUSUM) monitoring accumulates deviations and can be evaluated as a change detector [4]. We examine these methods as candidates rather than presume that their temporal structure guarantees better performance. Repeating daily behaviour also motivates a patient-specific time-of-day baseline, whose assumptions must be checked against less regular recordings.

The intended IoT architecture comprises data acquisition, timestamped transmission, storage, analytics and an interface for reviewing measurements and alerts. This study implements the analytics and synthetic-data stages only. Physical sensing, wireless transport, cloud deployment, encryption, interoperability, clinician interfaces and hardware performance are architectural objectives, not experimentally demonstrated components. The analytics separate sensor-quality flags, numerical forecasts and statistical-deviation alerts. No automated diagnosis or treatment recommendation is implemented.

Real longitudinal resources provide a route to later evaluation. OhioT1DM describes eight weeks of glucose-related recordings for 12 people with type 1 diabetes [5], while DiaTrend describes longitudinal glucose-monitor and insulin-pump data from 54 participants [6]. Neither dataset is used in the present experiments. Their inclusion here identifies potential validation resources, not evidence supporting the numerical results below.

3 Data and experimental design

3.1 One-day Gaussian simulations

The supplied generator creates ten trajectories, each with 1,440 one-minute observations of glucose, systolic and diastolic blood pressure, heart rate, SpO2 and cholesterol. It reconstructs a process described in an earlier manuscript; it is not the original data-generation implementation. Across these ten simulated patients, parameter configurations are identical and random-noise seeds differ. The six variables are generated independently conditional on deterministic daily and meal components, without physiological cross-variable coupling.

$$y_{p,j}(t) = \mu_j + A_j \sin[2\pi(t - \varphi_j)/1440] + B_j(t) + \varepsilon_{p,j}(t) \quad (1)$$

Here ε is independent Gaussian noise with standard deviation σj. B is zero except for glucose, where it is the sum of Gaussian-shaped bumps centred at 08:00, 13:00 and 19:00, each with amplitude 18 mg/dL and width 45 minutes. The process is not a random walk. SpO2 is clipped to [0, 100]. The cholesterol series is a computational surrogate and does not establish feasible minute-by-minute cholesterol measurement.

Table I. Parameters of the reconstructed one-day generator

Variable	Unit	Mean	Noise SD	Daily amplitude	Phase min
Glucose	mg/dL	100	12	4	0
Systolic BP	mmHg	120	9	6	240
Diastolic BP	mmHg	80	7	4	240
Heart rate	bpm	75	9	12	300
SpO2	%	97	1.2	0.8	0
Cholesterol	mg/dL	180	6	1	0

The spike injector selects 5% of timestamps per variable and patient without replacement, then ads randomly signed perturbations of 4–7 noise standard deviations. Labels identify selected injection indices. Because SpO2 is clipped again after injection, one of its 720 positive labels corresponds to no observable change and 26 correspond to changes below 1.2 percentage points. These labels represent injection provenance rather than independent clinical adjudication.

3.2 Temporal glucose simulations

A second, glucose-only generator produces seven days sampled every five minutes, or 2,016 observations per trajectory. Patient-specific baseline, daily amplitude and phase are drawn uniformly from 90–120 mg/dL, 4–10 mg/dL and −90 to 90 minutes. Meal-bump amplitude and width are drawn from 12–24 mg/dL and 35–60 minutes. The same nominal meal times repeat daily. Temporally correlated fluctuations follow $u(t) = 1.25u(t-1) - 0.4u(t-2) + \eta(t)$, with η drawn from N(0, 1), and independent measurement noise has standard deviation 2 mg/dL. No parameters are estimated from clinical recordings.

The development experiment uses seeds 4001–4010, 5001–5010 and 6001–6010. A seasonal follow-up uses fresh seeds 9001–9010, 10001–10010 and 11001–11010. Each set therefore contains 30 independently simulated trajectories. Fresh seeds test Monte Carlo variability within the same generator; they are not external validation. The seasonal design was selected after examining development results. Component-removal analyses were subsequently examined on the same fresh set without selecting hyperparameters from its test labels

3.3 Evaluation stages

Table II. Distinct evaluation protocols

Stage	Data and split	Purpose
Reproduction	10 one-day trajectories; forecast 80/20 split; anomaly fitting and scoring on all data	Verify the supplied results
Score ablation	Same glucose networks and data as reproduction	Isolate window versus point scoring
Causal evaluation	30 one-day glucose trajectories; chronological 60/20/20 split	Evaluate unseen observations with past-only inputs
Temporal development	30 seven-day trajectories; train days 1–3, tune day 4, calibrate day 5, test days 6–7	Test Kalman and CUSUM candidates
Seasonal follow-up	30 fresh seven-day trajectories; same day allocation	Evaluate daily patterns and component ablations

For the causal one-day evaluation, the first half of validation selects Kalman parameters and the second half calibrates detection thresholds. Models fit a known-clean synthetic training prefix; only the test suffix is replaced with

injected readings for anomaly evaluation. Forecasting uses the corresponding uncorrupted series. Three repeats use generation base seeds 42, 1042 and 2042, and injection seeds 123, 1123 and 2123. Patient identifiers 1–10 are added to each generation base seed. The clean-training assumption is favourable and is not guaranteed in deployment.

4 Analytical methods

4.1 Reproduced forecasting and detection

The original forecasting comparison fits ARIMA(2,1,2) independently to each patient-variable series using its first 80%, then predicts all remaining 288 minutes from one fixed origin. The MLP forecaster standardizes the training prefix and uses 15 lagged values, hidden layers of 32 and 16 ReLU units, Adam optimization, learning rate 0.005, L2 penalty 0.001 and up to 1,500 iterations. Predictions are recursive. Each standardized prediction is clipped to the training minimum minus three or training maximum plus three. This implementation tests neither a multivariate VARIMA model nor a general class of neural forecasters.

The original Isolation Forest uses 200 trees and contamination 0.05, with fitting and prediction on the same series. The original autoencoder uses 15-value centred windows, hidden layers of 10, 2 and 10 tanh units, learning rate 0.01, L2 penalty 0.0001 and up to 1,500 iterations. Its scaler and network fit all evaluation windows. Reconstruction error is averaged over the window and thresholded at its 95th percentile. Both model types use random state 42. Seven future observations enter a centred window at its current timestamp.

For the isolated ablation, we retain the fitted autoencoder, input windows and 5% decision quota, changing only the anomaly score to the squared reconstruction error at the centre position. This remains an in-sample, acausal diagnostic. It isolates score alignment without representing it as a corrected prospective experiment.

4.2 Causal anomaly detection and simple baselines

The held-out autoencoder uses trailing windows of 15 observations and only the endpoint reconstruction error. Standardization and training use complete training windows; the first 14 padded windows are excluded from fitting. All remaining architecture and optimization settings match the original autoencoder. Isolation Forest fits the clean training prefix with 200 trees and contamination set to auto; its negative score_samples output is subsequently calibrated on validation data. The robust baseline scores absolute deviations from the training median, divided by 1.4826 times the median absolute deviation (MAD), with a scale floor of 10^{-6}.

$$a(t) = [x(t) - r(t)]^2; \quad b(t) = |y(t) - c| \,/\, s \qquad (2)$$

In (2), x and r are the standardized endpoint and its reconstruction; c and s are the training median and scaled MAD. Each method's threshold is the 99th percentile of its clean calibration scores. Thresholds remain fixed during testing and are not chosen from test anomaly prevalence. Hyperparameters, preprocessing and learned distributions use only their designated training or validation segments. A calibration quantile does not guarantee the same false-positive rate on a later segment.

Forecasting baselines include the training mean, last observed value and the average of the preceding 60 minutes. The temporal simulations additionally include the value at the same time on the previous day. For rolling evaluation, all methods receive observations through the same origin and forecast targets strictly within the test segment. Rolling and original fixed-origin errors are reported separately.

4.3 Kalman filtering and cumulative deviations

We implement a damped local level–trend state-space model with transition matrix F = [[1, 0.98], [0, 0.98]] and observation vector H = [1, 0]. Observation-noise variance is estimated from the training differences as $\max\{[1.4826\ \mathrm{MAD}(\Delta y)]^2/2,\ 10^{-6}\}$. Process-noise covariance is diagonal with entries qlevelR and qslopeR. A fixed steady-state gain is obtained from the algebraic Riccati equation, with a scalar local-level solution when slope noise is zero. The state starts at the first observation with zero slope. Each innovation is calculated before assimilating the current observation.

Validation selects among (qlevel, qslope) = (0.0001, 0), (0.001, 0), (0.01, 0), (0.1, 0), (0.001, 0.00001) and (0.01, 0.0001), minimizing one-step validation MSE for rolling experiments. The added fixed-origin baseline selects parameters using a 30-minute horizon within the original training prefix. State updates during testing use only observations that have arrived; parameters are not refitted on future data.

$$C^+(t) = \max\{0,\ C^+(t-1) + z(t) - 0.5\};$$

$$C^-(t) = \max\{0,\ C^-(t-1) - z(t) - 0.5\} \qquad (3)$$

For temporal detection, signed innovations are centred and scaled using tuning data. CUSUM issues an alert when either accumulation exceeds its calibrated threshold. It resets after an alert and ignores updates during a 30-minute refractory interval shared with competing alert methods. Candidate thresholds are 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50, 100 and 10^6. The least restrictive threshold satisfying at most one observed alert per calibration day is chosen. This is an empirical calibration target, not a test-time guarantee.

4.4 Patient-specific daily template

For each seven-day trajectory, the first three days are averaged by time of day to obtain a 288-bin template. A circular five-bin mean smooths the template over a 25-minute window. All template values derive from completed training days; the template remains fixed through testing. Let s(b) denote its value for time-of-day bin b, and $e(t) = y(t) - s[b(t)]$ the observed residual.

$$m(t) = (1 - \alpha)m(t-1) + \alpha e(t);$$

$$\hat{y}(t+h|t) = s[b(t+h)] + \rho^h m(t) \qquad (4)$$

The recent-residual state starts at zero. For each horizon, tuning evaluates α in {0, 0.05, 0.2, 0.5, 1} and ρ in {0.5, 0.8, 0.95, 1}, selecting the pair with lowest day-4 MSE. Setting α to zero gives the uncorrected daily template. h is measured in five-minute samples. Future calendar time is known, but no future measured value enters the forecast.

Seasonal detection scores deviations from the fixed daily template, using the median and MAD scale of day-4 residuals. Absolute-residual thresholds are selected from calibration-score quantiles 0.90, 0.95, 0.975, 0.99, 0.995, 0.999 and 1.0,

with an additional value just above the maximum, to meet the same calibration alert budget. A separate quality branch flags three consecutive zero increments in the unrounded signal. Because exact repeats are extremely unlikely in the continuous-valued generator, this branch is a controlled sensor-fault diagnostic and requires resolution-aware redesign for real devices.

We also evaluate seasonal CUSUM, a hybrid of seasonal point scores and CUSUM with the quality branch, and component-removal variants. The hybrid applies a shared multiplier selected from {1, 1.25, 1.5, 2, 3, 5, 10, 1000} to its component thresholds to satisfy the calibration alert budget. All variants share the refractory interval. The simpler seasonal-residual-plus-quality variant removes CUSUM without introducing a new tuning search.

5 Performance measures

Forecast accuracy is reported as RMSE and MAE in each variable's own unit, averaged across trajectories. Raw errors across glucose, blood pressure, pulse, SpO2 and cholesterol are not combined into a primary score. Point detection reports precision, recall, F1, average precision and the false-positive fraction among normal test timestamps. Average precision summarizes score ranking; F1 reflects the threshold fixed before testing.

Temporal tests include a no-event condition and four perturbation types. Each non-null condition inserts four nonoverlapping events per trajectory, for 120 events per type across 30 trajectories. Event starts are near test offsets 50, 180, 310 and 440 samples, with an independent ±8-sample jitter. Signs alternate. Event generation is independent of detector scores.

Table III. Controlled temporal perturbations

Type	Duration	Perturbation
Spike	One sample	±20 mg/dL at one timestamp
Sustained shift	90 minutes	±15 mg/dL throughout the event
Gradual shift	180 minutes	Linear ramp from 0 to ±20 mg/dL
Stuck sensor	90 minutes	Repeat the observation immediately before onset

An event is detected only when an alert falls between its onset and end. Alerts outside event windows, including post-event recovery alerts, count as false and are normalized by observed normal time in days. Delay is measured from onset to the first alert and is conditional on detection; missed events are represented by recall, not zero delay. Reported delays average each trajectory's mean detected-event delay. No-event trajectories independently measure background alert burden.

Paired percentile bootstrap intervals resample complete simulated trajectories 5,000 times using seed 20260912. Differences are computed between methods on corresponding trajectories. These exploratory intervals quantify variability within a simulation family; they are unadjusted for multiple comparisons and do not imply clinical statistical significance. All comparisons use the same observations within their own evaluation stage.

6 Results

6.1 Reproduction and missing baselines

The original anomaly results reproduce with mean F1 of 0.857407 for Isolation Forest and 0.164352 for the MLP autoencoder across six parameters and ten trajectories. Original forecasting results also match the reported precision, with no missing forecast metrics in the reproduction. However, reproducibility does not resolve the in-sample design, future inputs or score-alignment problem. Precision, recall and F1 coincide in the quota-based experiment because the number of predicted positives matches the number of injected positives.

Table IV. Reproduced original results with distinct forecasting And anomaly protocols

Variable	ARIMA RMSE	MLP RMSE	Forest F1	Autoencoder F1
Glucose (mg/dL)	18.89	41.90	0.943	0.164
Systolic BP (mmHg)	9.29	30.40	0.935	0.163
Diastolic BP (mmHg)	7.35	26.49	0.946	0.167
Heart rate (bpm)	10.57	31.62	0.881	0.172
SpO2 (%)	1.27	5.23	0.476	0.154
Cholesterol (mg/dL)	5.98	24.16	0.964	0.167

On the identical original fixed-origin glucose task, the training mean achieves RMSE 13.32 mg/dL, lower than ARIMA's 18.89 and the MLP's 41.90. The 29.5% reduction relative to ARIMA shows that the earlier model comparison omitted a strong simple baseline. Kalman fixed-origin RMSE is 22.45, providing no improvement for this task. These comparisons do not establish that a mean model will dominate on temporally structured clinical data.

Table V. Glucose forecasting on the original 288-minute Fixed-origin test

Method	RMSE mg/dL	MAE mg/dL
ARIMA	18.89	16.00
Kalman fixed origin	22.45	19.57
Previous 60-minute mean	17.01	14.14
Last observed value	20.55	17.75
MLP forecaster	41.90	34.65
Training mean	13.32	10.74

6.2 Score alignment and prospective detection

Holding the original glucose autoencoder unchanged and replacing mean window error with centre-point error increases F1 from 0.163889 to 0.881944. The change isolates the consequences of comparing a distributed window score with a point label. It remains a diagnostic on the fitted data; it is not evidence of out-of-sample accuracy.

Table VI. Held-out causal glucose detection across 30 one-day trajectories

Method	Precision	Recall	F1	Avg precision	Normal FPR
Isolation Forest	0.753	0.960	0.833	0.937	0.018
Causal autoencoder	0.717	0.989	0.826	0.989	0.022
Kalman residual	0.746	0.996	0.849	0.993	0.019
Robust baseline	0.780	0.971	0.858	0.983	0.016

In the separate causal experiment, the corrected autoencoder and Isolation Forest are close: mean F1 is 0.826 and 0.833, respectively. Their paired F1 difference is −0.007, with a 95% bootstrap interval of [−0.042, 0.031]. This does not demonstrate equivalence, but provides no clear separation in this simulation. The robust baseline achieves F1 0.858; its difference from Isolation Forest is 0.025 [0.005, 0.047]. The corrected experiment changes training, calibration and window direction together, so its improvement should not be attributed solely to one adjustment.

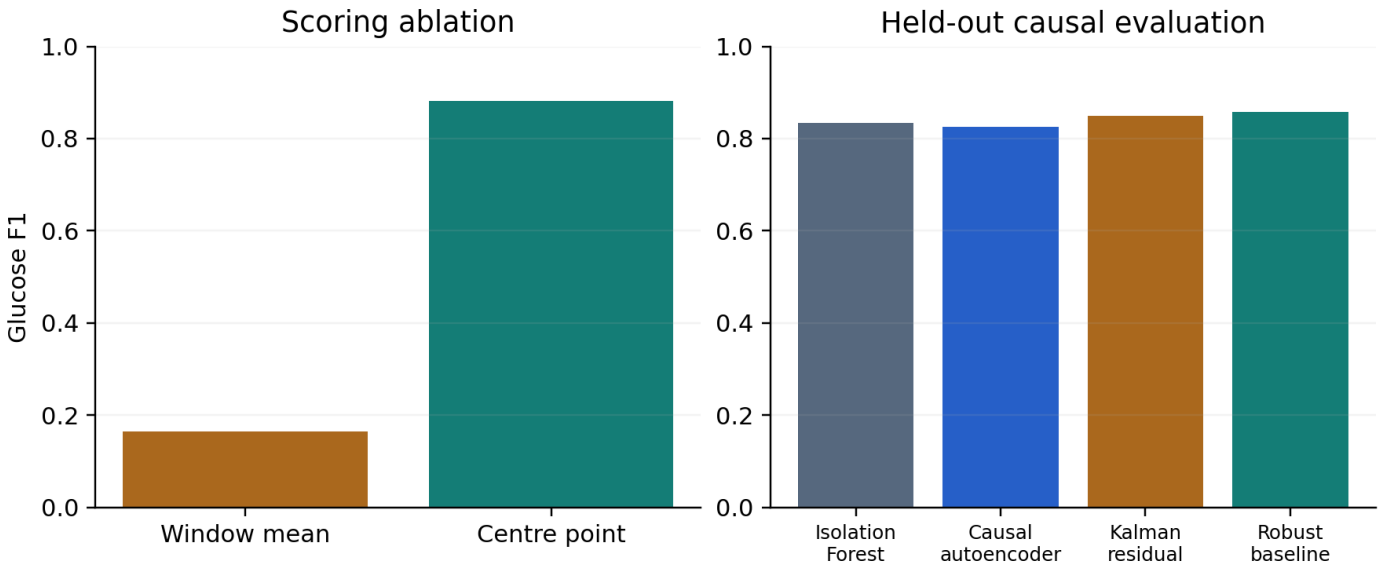

Fig 1. Glucose anomaly scoring. Left: isolated centre-point scoring ablation, retaining the original in-sample and a causal design. Right: a separate held-out causal experiment on 30 synthetic trajectories. The two panels are not the same evaluation protocol.

6.3 Temporal development and seasonal forecasting

Kalman filtering provides useful smoothing on the original noisy glucose family, but its temporal assumptions do not guarantee superiority. In the seven-day development experiment, 30-minute RMSE is 5.750 mg/dL for Kalman and 4.424 for the previous-day baseline. Kalman CUSUM detects 55.8% of gradual shifts, compared with 78.3% for Isolation Forest; its false alerts per normal day are also higher in that condition, 2.578 versus 1.844. These results motivate evaluating daily structure explicitly rather than treating Kalman/CUSUM as a mandatory replacement.

Table VII. Forecast RMSE on 30 fresh seven-day glucose trajectories

Method	15 min	30 min	60 min
Last observed value	4.019	5.343	7.625
Previous-day value	4.373	4.373	4.376
Kalman	4.420	5.750	7.950
Daily template	3.368	3.368	3.369
Template with correction	3.170	3.370	3.369

On fresh trajectories, the template with optional residual correction achieves RMSE 3.170, 3.370 and 3.369 mg/dL at 15, 30 and 60 minutes. Relative to previous-day prediction, the paired RMSE differences are −1.203 [−1.281, −1.130], −1.003 [−1.085, −0.923] and −1.008 [−1.090, −0.925] mg/dL. The correction helps primarily at 15 minutes; at longer horizons it is essentially tied with the fixed template. Because the generator repeats daily and meal patterns, these results demonstrate suitability to that structure rather than general superiority.

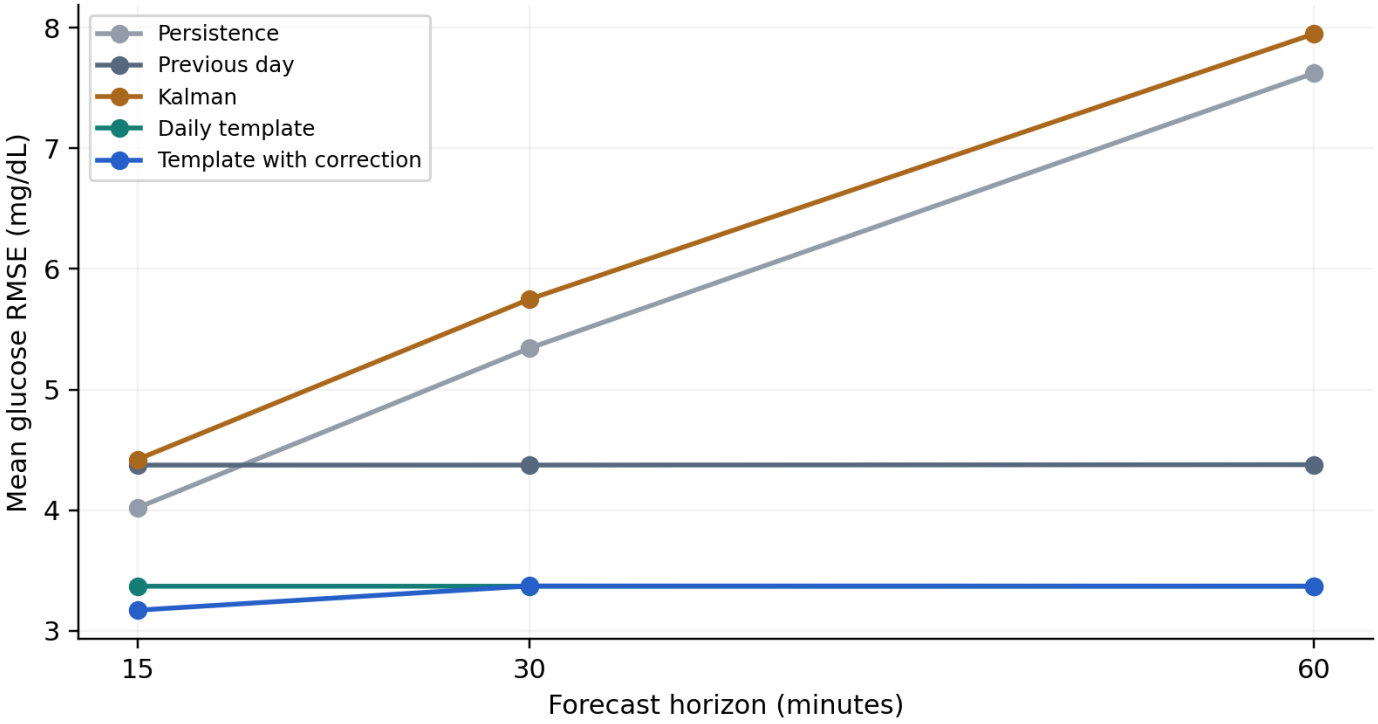

Fig 2. Forecast RMSE on fresh synthetic trajectories. All models use the same rolling origins within each horizon. The repeated daily structure favours seasonal methods; these errors cannot be compared directly with the original fixed-origin results in Table V.

6.4 Events and sensor-quality ablations

Table VIII. Event recall percentages and no-event alert burden on fresh simulations

Method	Spike %	Sustained %	Gradual %	Stuck %	False alerts per day
Isolation Forest	41.7	52.5	87.5	0.0	1.367
Kalman CUSUM	10.8	50.8	47.5	0.0	0.783
Seasonal residual	97.5	100.0	100.0	12.5	1.167
Seasonal CUSUM	3.3	100.0	100.0	36.7	1.583
Seasonal hybrid + quality	97.5	100.0	100.0	100.0	1.600
Seasonal residual + quality	97.5	100.0	100.0	100.0	1.167

Seasonal residual alerts with the quality branch detect 97.5% of spikes and all sustained shifts, gradual shifts and stuck-sensor events in this controlled test. The corresponding Isolation Forest recalls are 41.7%, 52.5%, 87.5% and 0%. The stuck-sensor result derives mainly from the explicit repeat-value rule. It must not be interpreted as detection of clinical deterioration. For the seasonal residual-plus-quality method, mean conditional delays are 0.0 minutes for detected spikes, 0.79 for sustained shifts, 71.0 for gradual shifts and 10.04 for stuck sensors.

The same method produces 1.167 false alerts per day in the no-event condition versus 1.367 for Isolation Forest. The paired difference is −0.200 alerts/day with a 95% bootstrap interval of [−0.733, 0.383], so a reliable reduction in background alert burden is not established. False alerts per normal day in the spike, sustained, gradual and stuck conditions are 1.141, 1.143, 1.133 and 1.143 for the seasonal method, compared with 1.359, 1.429, 1.467 and 1.429 for Isolation Forest. These rates are observed outcomes, not an exactly matched test-time operating point.

Adding CUSUM to the seasonal point detector and quality branch does not increase recall for these event types, while no-event false alerts rise from 1.167 to 1.600 per day. Removing the quality branch from the full hybrid reduces stuck-sensor recall to 35.0%. The ablations therefore favour the simpler point-residual detector with a dedicated quality check for this prototype. They do not exclude potential CUSUM benefits for smaller shifts or other temporal processes.

7 Discussion and limitations

The results distinguish reproducibility from validity. Reproducing a numerical result establishes consistency with an implementation, but does not establish that the implementation answers the intended research question. Here, the input window, score definition and evaluation split strongly affect detector rankings. The original experiment consequently cannot support a broad neural-versus-classical conclusion. Both model families must be evaluated using aligned targets and the information available at the decision time.

The positive seasonal results should be interpreted in light

of their assumptions. Averaging several previous days reduces noise in a recurring pattern, while a recent-residual correction can help at short horizons. A fixed pattern may nevertheless fail when meal timing, medication, sleep, activity or baseline physiology changes. The current generator does not model these clinical mechanisms, irregular schedules, missingness, comorbidity or cross-sensor dependence. Explicit tests under such changes are needed before recommending an operating model.

All trajectories and event labels are synthetic. Large spikes and repeated values are comparatively easy perturbations, and injection magnitude is not a clinical severity definition. The 30 trajectories per experiment are independent simulated draws, not 30 recruited participants. Three days of template training and one day of calibration are short; thresholds can overfit normal calibration behaviour. Observed test alert rates exceed the one-alert-per-day calibration target, emphasizing that calibration is not a prospective guarantee.

The seasonal follow-up is exploratory: its model family was chosen after development results, and component-removal findings were inspected on the same fresh simulation family. Bootstrap intervals capture only within-generator variability and are not adjusted for the number of model and scenario comparisons. Larger independent cohorts, prespecified hypotheses and a final untouched evaluation set would be needed for confirmatory claims.

The intended IoT setting also remains unvalidated. No physiological sensor was tested, and no end-to-end communication latency, energy use, memory constraint, data-security property, regulatory compliance or user experience was measured. Prediction intervals and adaptive conformal calibration were not implemented in these experiments. The methods output forecasts, statistical deviations and sensor-quality flags; they do not establish clinically actionable alarm thresholds or treatment benefits.

A next study should use real longitudinal glucose recordings, define target events with appropriate domain input, retain simple and corrected neural baselines, and separate within-patient future prediction from generalization to new patients. Model selection should use validation data only, with longer calibration, explicit missing-data rules and reporting of event sensitivity at clinically meaningful alert burdens. OhioT1DM and DiaTrend are candidate resources for that next stage [5,6], subject to their access conditions and suitability for the selected outcome.

8 Conclusion

Evaluation design changes the interpretation of IoT health-monitoring analytics. Correcting point-score alignment raises the original glucose autoencoder F1 from 0.164 to 0.882; held-out causal evaluation places it close to Isolation Forest. A training-mean baseline outperforms ARIMA on the original glucose task. On temporal simulations, a patient-specific daily template, optional recent-residual correction and separate sensor-quality monitoring provide a stronger candidate than Kalman/CUSUM. These findings support a focused prototype and transparent evaluation, but depend on simulated daily regularity. Real longitudinal validation is required before claiming clinical readiness or general model superiority.

References

[1] Z. Deng, L. Guo, X. Chen, and W. Wu, “Smart Wearable Systems for Health Monitoring,” Sensors, vol. 23, no. 5, p. 2479, Feb. 2023, doi: 10.3390/s23052479.

[2] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422, Dec. 2008, doi: 10.1109/icdm.2008.17.

[3] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” Journal of Basic Engineering, vol. 82, no. 1, pp. 35–45, Mar. 1960, doi: 10.1115/1.3662552.

[4] E. S. PAGE, “CONTINUOUS INSPECTION SCHEMES,” Biometrika, vol. 41, no. 1-2, pp. 100–115, 1954, doi: 10.1093/biomet/41.1-2.100.

[5] C. Marling and R. Bunescu, “The OhioT1DM dataset for blood glucose level prediction: Update 2020,” in *Proc. 5th Int. Workshop Knowledge Discovery in Healthcare Data (KDH@ECAI)*, vol. 2675, 2020, pp. 71–74. [Online]. Available: https://ceur-ws.org/Vol-2675/paper11.pdf

[6] T. Prioleau, A. Bartolome, R. Comi, and C. Stanger, “DiaTrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions,” Scientific Data, vol. 10, no. 1, Aug. 2023, doi: 10.1038/s41597-023-02469-5.