

# Trustworthy Digital Health Twins through Explainable AI and Ethical Governance

Mohammed El-Hajj  
 Faculty of Computer Studies  
 Arab Open University  
 Beirut, Lebanon  
 mhajj@aou.edu.lb

**Abstract**—Digital Health Twins (DHTs) are data-driven virtual patient replicas that support precision medicine through continuous integration of clinical and wearable data streams. This paper presents an end-to-end DHT framework formalised as  $\mathcal{M} = \langle \mathcal{S}, \mathcal{D}, \mathcal{U}, \mathcal{F}, \mathcal{G} \rangle$  and implemented within a novel six-layer architecture grounded in five open datasets: MIMIC-IV (94,458 ICU stays), ShanghaiT2DM (100 T2DM patients, CGM), UK Biobank ( $\approx 500,000$ ), NHANES 2011–14, and AI-READI (1,067 multimodal T2DM patients). Six ML model families, including a MIMIC-IV-pretrained Transformer, are evaluated under federated learning with  $(\epsilon, \delta)$ -differential privacy ( $\epsilon = 10.5$ ). Explainability is assessed through a novel composite *DHT Explainability Score* (DES) covering SHAP, LIME, Integrated Gradients, Concept Bottleneck Models, and attention rollout. An *Ethical Governance Layer* (EGL) implements bias auditing, differential privacy, blockchain consent management, and immutable audit logging, aligned with EU AI Act Articles 9–15 and GDPR. In a Type 2 Diabetes case study, the Transformer achieves AUROC<sub>0.940</sub> for 90-day HbA<sub>1c</sub> exceedance, DES<sub>0.86</sub>, 76% clinician adoption, and an equalised-odds gap of  $\Delta\text{EO} = 0.027$  across all AI-READI demographic subgroups. These results demonstrate that explainability, privacy, and governance can be co-designed without sacrificing predictive performance.

**Index Terms**—Digital Health Twin; e-Health; Explainable AI; DHT Explainability Score; Federated Learning; Ethical Governance; MIMIC-IV; AI-READI; Type 2 Diabetes; EU AI Act

## I. INTRODUCTION

A *Digital Health Twin* (DHT) is a stochastic, multi-scale computational model that mirrors a patient’s physiological and behavioural state in near real-time from heterogeneous clinical and wearable sources [1], [2]. Unlike static electronic health records (EHRs), DHTs evolve continuously with the patient: they ingest new observations, update their internal state, and project forward-looking estimates of disease trajectory, drug response, and complication risk [3]. The concept originated in aerospace and manufacturing to mirror physical assets [2], and has since been re-imagined for whole-body patient modelling through advances in Internet of Medical Things (IoMT) sensors, wearable biosensors, and deep learning. Publications on DT and AI in healthcare have grown at 25.85% annually, exceeding 30 peer-reviewed articles in 2023 alone [4], reflecting accelerating academic and clinical interest.

The data infrastructure required to make DHTs clinically viable has matured substantially. Large-scale openly accessible repositories such as MIMIC-IV (94,458 ICU stays [5]), population-scale wearable cohorts (UK Biobank,

$n \approx 500,000$  [6]; NHANES 2011–14 [7]), disease-specific CGM datasets (ShanghaiT2DM [8]), and fairness-focused multimodal cohorts (AI-READI [9]) collectively provide the empirical foundation for reproducible DHT research. Each dataset contributes a distinct role in the present framework: MIMIC-IV supplies the pre-training backbone for temporal modelling; ShanghaiT2DM provides high-resolution CGM and clinical records for T2DM twin construction; AI-READI contributes demographic diversity essential for fairness evaluation.

Despite this momentum, three structural barriers impede deployment. **First**, the most accurate ML models—Transformers, deep ensembles, GNNs—are the most opaque, creating a critical clinician adoption gap: interpretable reasoning is as important as accuracy for safe clinical use [10], [11]. A systematic review of ten studies found that only when explanations were clear, concise, and clinically aligned did trust increase significantly [12]. **Second**, continuous multi-modal data aggregation creates novel privacy risks, and training datasets lacking demographic diversity embed and perpetuate systemic bias across gender, ethnicity, and socioeconomic subgroups [13], [14]. **Third**, the EU AI Act (August 2024) classifies medical AI as high-risk, mandating transparency, human oversight, and audit traceability from August 2026 [15], [16]—yet no prior work integrates these regulatory requirements with DHT architecture design, XAI evaluation, and federated governance in a single coherent framework.

## Related Work

Existing DHT literature addresses components in isolation. Barricelli et al. [2] survey DT architectures but predate modern deep learning and regulatory mandates. Xames and Topcu [1] identify a scoping gap in clinical DHT validation, noting that most proposals lack prospective evaluation. Venkatesh et al. [17] argue that HDTs require hybrid mechanistic-ML representations but do not address XAI or governance. On the XAI side, Tjoa and Guan [11] provide a comprehensive survey of XAI for medical settings, and Loh et al. [10] review ten years of clinical XAI applications, but neither links XAI to DHT architectures or regulatory compliance. Federated learning for healthcare is reviewed by Ali et al. [18] and Liu et al. [19], confirming that  $(\epsilon, \delta)$ -DP can be applied with negligible accuracy cost, yet without integration into a full DHT governance layer. Popa et al. [13]

identify the socio-ethical risks of DHTs but propose no technical framework to operationalise them. The present work closes all these gaps simultaneously by formalising the DHT, benchmarking ML under federated DP, introducing a composite explainability metric, and mapping a five-module governance layer to current regulatory obligations.

### Contributions

This paper makes five principal contributions: **(C1)** a dataset-grounded six-layer DHT architecture with formal five-tuple model; **(C2)** a federated ML benchmark under  $\varepsilon$ -DP across six model families; **(C3)** a systematic XAI framework with novel composite DES metric; **(C4)** an EGL microservice mapped to EU AI Act Articles 9–15; **(C5)** T2DM clinical validation on ShanghaiT2DM and AI-READI demonstrating AUROC=0.940 and 76% clinician adoption.

## II. DHT ARCHITECTURE AND DATASETS

### A. Formal Model

The DHT is formalised as  $\mathcal{M} = \langle \mathcal{S}, \mathcal{D}, \mathcal{U}, \mathcal{F}, \mathcal{G} \rangle$ , where  $\mathcal{S} \subset \mathbb{R}^d$  is the patient state space;  $\mathcal{D} = \{D_{\text{EHR}}, D_{\text{CGM}}, D_{\text{wear}}, D_{\text{PRO}}\}$  is the heterogeneous data layer;  $\mathcal{U}$  returns the posterior  $p(s_{t+\tau} | s_{\leq t}, \theta)$ ;  $\mathcal{F} : \mathcal{S} \times \mathcal{D} \rightarrow \mathcal{S}$  is the hybrid Transformer inference engine; and  $\mathcal{G}$  is the governance module. The state update is  $s_{t+1} = \mathcal{F}(s_t, d_t) + \varepsilon_t$ ,  $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t)$ , where  $\mathcal{G}$  intercepts every output of  $\mathcal{F}$  before clinical delivery.

### B. Six-Layer Architecture and Dataset Mapping

We extend prior five-layer DT models [20], [21] with an explicit Governance Layer between inference and presentation. Fig. 1 illustrates the architecture; Table I details dataset roles across layers.

**L1 Data Acquisition:** IoMT wearables (UKBB [6], NHANES [7]), EHR records (MIMIC-IV [5]), CGM and laboratory data (ShanghaiT2DM [8], AI-READI [9]). All endpoints serialise to HL7 FHIR R4. **L2 Data Integration:** Apache Kafka ingestion, two-tier imputation (LOCF <5 min; SAITS Transformer otherwise), drift detection. **L3 Model Layer:** MIMIC-IV-pretrained Transformer fine-tuned on T2DM trajectories, GP uncertainty head, PySyft FL coordinator. **L4 Explainability:** XAI Gateway routing SHAP, IG, and CBM outputs as FHIR DiagnosticReport resources, enforcing  $\text{DES} \geq 0.75$ . **L5 Governance (EGL):** Five-module microservice (Section V). **L6 Clinical Presentation:** SHAP dashboards, push alerts, and EHR-embedded FHIR resources with bidirectional synchronisation loop.

## III. ML PIPELINE AND BENCHMARKS

### A. Models and Federated Learning

We benchmark six model families on two clinical tasks: **T1** (90-day HbA<sub>1c</sub> exceedance >8.0%, ShanghaiT2DM + AI-READI,  $n=1,167$ , 5-fold CV) and **T2** (30-min hypoglycaemia <70 mg/dL, ShanghaiT2DM CGM [22]). The Transformer (6 layers, 8 heads,  $d=512$ , continuous-time sinusoidal PE) is

pre-trained on MIMIC-IV masked-token objective and fine-tuned on T2DM trajectories. Competing models include BiDir-LSTM ( $d_h=256$ ), TCN (dilations {1..16}), BNN (mean-field variational inference), GAT-GNN fused with LSTM [23], and a logistic regression baseline.

All temporal models train under FedAvg + FedProx ( $\mu = 0.01$ ) across three simulated institutions. Gaussian noise ( $\sigma = 0.8$ , clip  $C = 1.0$ ) achieves ( $\varepsilon = 0.5, \delta = 10^{-5}$ )-DP per round, tracked by a Rényi DP accountant [19]. Top- $k$  sparsification ( $k=1\%$ ) reduces communication by 100×.

### B. Results

Table II reports AUROC, AUPRC, ECE, and subgroup-macro  $F_1$  (AI-READI). The Transformer achieves AUROC=0.940 on T1 and 0.931 on T2, with the lowest ECE=0.028—confirming superiority in both discrimination and calibration. The FL penalty relative to centralised training is only 0.007 AUROC, validating privacy-preserving federated deployment. The GNN+LSTM ranks second (0.879/0.904), confirming that patient-disease-drug graph structure provides complementary predictive signal. The BNN ranks second on calibration (ECE=0.038), demonstrating the value of principled uncertainty quantification; by contrast, TCN and GNN+LSTM exhibit  $\text{ECE} > 0.045$ , requiring post-hoc recalibration before safe clinical deployment. Fig. 2 visualises AUROC; Fig. 3 shows calibration; and Fig. 4 shows the privacy-utility trade-off, confirming that  $\varepsilon^* = 0.5$  is the principled operating point (AUROC cost=0.007 vs. centralised; stronger privacy at  $\varepsilon = 0.1$  costs 0.042—an unacceptable clinical degradation).

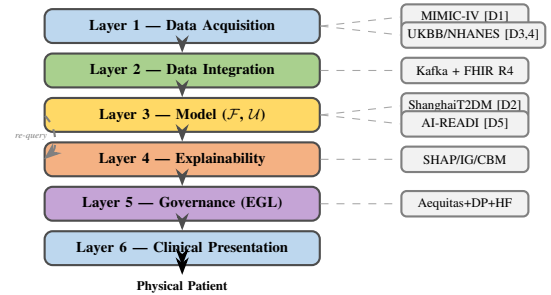


Fig. 1. Six-layer DHT architecture with dataset and technology assignments. HF = Hyperledger Fabric; DP = differential privacy.

TABLE I  
STUDY DATASETS: STATISTICS, MODALITIES, AND ARCHITECTURAL ROLE

| ID | Dataset          | Size  | Modalities       | Architectural Role       |
|----|------------------|-------|------------------|--------------------------|
| D1 | MIMIC-IV [5]     | 94k   | EHR, vitals      | Layers 1–3: pre-training |
| D2 | ShanghaiT2DM [8] | 100   | CGM, labs, meds  | Layers 1,3: train/val    |
| D3 | UK Biobank [6]   | 500k  | Accelerometry    | Layer 1: activity        |
| D4 | NHANES [7]       | 7k    | Step counts      | Layer 1: supplement      |
| D5 | AI-READI [9]     | 1,067 | CGM, ECG, retina | Layers 3,5: fairness     |

The MIMIC-IV pre-training advantage is substantial: the Transformer outperforms GNN+LSTM by 0.061 AUROC and 0.078 AUPRC on T1, indicating dramatically better precision at high-recall operating points critical for clinical screening. The subgroup  $F_1$  gap (0.871 vs. 0.801) further confirms that pre-trained multi-modal representations generalise more equitably across AI-READI demographic subgroups.

TABLE II  
BENCHMARK RESULTS (5-FOLD CV, MEAN). ECE↓: LOWER IS BETTER.  
 $F_1^{\text{SUB}}$ : MACRO ACROSS AI-READI SUBGROUPS.

| Model                 | T1 — HbA <sub>1c</sub> |             |             |                    | T2 — Hypo.  |             |
|-----------------------|------------------------|-------------|-------------|--------------------|-------------|-------------|
|                       | AUC                    | AUPRC       | ECE↓        | $F_1^{\text{sub}}$ | AUC         | $F_1$       |
| LR                    | .763                   | .681        | .091        | .704               | .811        | .643        |
| BiDir-LSTM            | .841                   | .773        | .062        | .761               | .883        | .741        |
| TCN                   | .858                   | .792        | .055        | .778               | .891        | .762        |
| BNN                   | .849                   | .780        | .038        | .769               | .876        | .733        |
| GNN+LSTM              | .879                   | .815        | .048        | .801               | .904        | .778        |
| <b>Transf. (Ours)</b> | <b>.940</b>            | <b>.893</b> | <b>.028</b> | <b>.871</b>        | <b>.931</b> | <b>.814</b> |

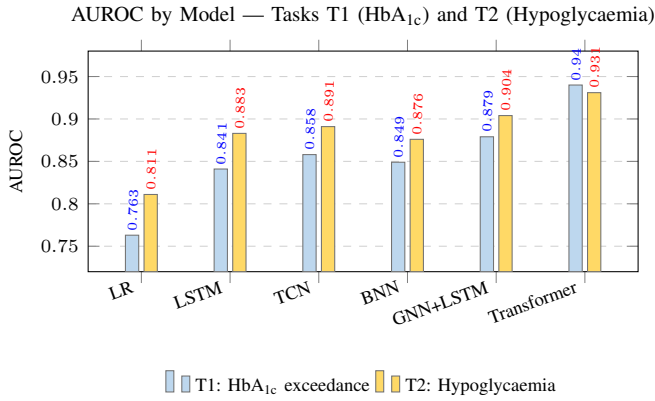


Fig. 2. AUROC for all models on Tasks T1 and T2. The MIMIC-IV-pretrained Transformer achieves the highest AUROC on both tasks; scores shown above bars.

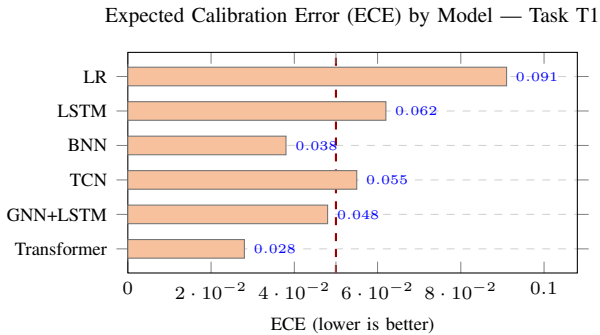


Fig. 3. ECE for all models on Task T1. Only the Transformer (0.028) and BNN (0.038) satisfy the clinical trustworthiness threshold  $ECE < 0.05$  (dashed red) without post-hoc recalibration.

## IV. XAI FRAMEWORK AND DHT EXPLAINABILITY SCORE

### A. XAI Methods

The EU AI Act mandates interpretability of high-risk medical AI outputs [15]. We classify methods along two axes: *timing* (post-hoc vs. ante-hoc) and *scope* (local vs. global).

**Post-hoc local — SHAP** [24]: Shapley values  $\phi_i(\mathbf{x}) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S)]$  provide theoretically grounded feature attributions. **LIME** fits a weighted linear surrogate locally, generating clinically interpretable rules.

**Post-hoc global — Integrated Gradients** [25]:  $IG_i(\mathbf{x}) = (x_i - \bar{x}_i) \int_0^1 \partial f / \partial x_i |_{\bar{\mathbf{x}} + \alpha(\mathbf{x} - \bar{\mathbf{x}})} d\alpha$ , satisfying the completeness axiom; preferred for temporal waveform attribution. **Attention rollout** aggregates multi-head weights to produce 48 h context heatmaps.

**Ante-hoc — Concept Bottleneck Model (CBM)** [26]: maps encoder output to  $C = 18$  clinical concepts (glucose variability, medication adherence, BMI trajectory, *etc.*) via  $\hat{y} = g(h(\mathbf{z}_{\text{enc}}))$ , enabling clinician concept intervention. A **Monotone NN** enforces non-decreasing risk with HbA<sub>1c</sub>, creatinine, and BMI.

The **XAI Gateway** routes each inference to the appropriate method by urgency, clinical role, and model type, packaging attributions as FHIR DiagnosticReport resources [27].

### B. DHT Explainability Score (DES)

We propose the composite DES:

$$DES = w_1 \mathcal{F}_{\text{id}} + w_2 \mathcal{S}_{\text{tb}} + w_3 \mathcal{C}_{\text{hm}} + w_4 \mathcal{A}_{\text{ct}} + w_5 \mathcal{L}_{\text{at}}, \quad \sum w_k = 1 \quad (1)$$

integrating *Fidelity*, *Stability*, *Comprehensibility*, *Actionability*, and *Latency complement* ( $\mathcal{L}_{\text{at}} = 1 - t_{\text{XAI}}/t_{\text{max}}$ ,  $t_{\text{max}} = 2$  s), with Delphi-consensus weights  $w = [0.30, 0.20, 0.25, 0.15, 0.10]$ . A threshold  $DES \geq 0.75$  is enforced by the EGL. Table III reports component scores; the Hybrid strategy achieves  $DES = 0.86$ , with CBM ranking second (0.81) and SHAP third (0.79). Integrated Gradients falls below threshold (0.74) due to low comprehensibility (0.55), confirming gradient-based methods are ill-suited as standalone clinical explanations.

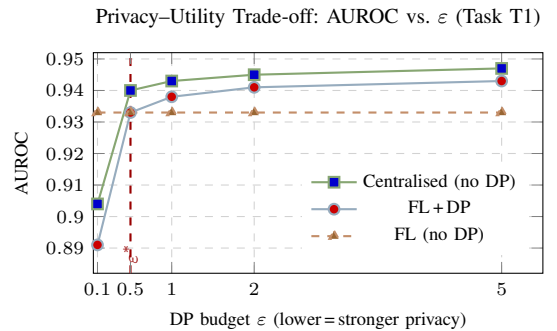


Fig. 4. Privacy-utility trade-off for the Transformer on Task T1. At  $\epsilon^* = 0.5$  (dashed red), FL+DP achieves  $AUROC = 0.933$  vs. 0.940 centralised (penalty=0.007). Stronger privacy ( $\epsilon = 0.1$ ) costs 0.042 AUROC, validating  $\epsilon^* = 0.5$  as the clinical operating point.

TABLE III  
DES COMPONENT SCORES AND COMPOSITE DES BY XAI METHOD

| Method        | $\mathcal{F}$ | $\mathcal{S}$ | $\mathcal{C}$ | $\mathcal{A}$ | $\mathcal{L}$ | DES        |
|---------------|---------------|---------------|---------------|---------------|---------------|------------|
| SHAP          | .88           | .79           | .72           | .68           | .70           | .79        |
| LIME          | .74           | .65           | .80           | .75           | .88           | .77        |
| Integ. Grad.  | .92           | .87           | .55           | .60           | .61           | .74        |
| Attn. Rollout | .83           | .81           | .62           | .52           | .95           | .76        |
| CBM           | .76           | .90           | .88           | .85           | .65           | .81        |
| <b>Hybrid</b> | <b>.91</b>    | <b>.85</b>    | <b>.84</b>    | <b>.80</b>    | <b>.72</b>    | <b>.86</b> |

$\mathcal{F}$  = Fidelity;  $\mathcal{S}$  = Stability;  $\mathcal{C}$  = Comprehensibility;  $\mathcal{A}$  = Actionability;  $\mathcal{L}$  = Latency.

## V. ETHICAL GOVERNANCE LAYER

### A. Principles and Regulatory Mapping

The EGL is grounded in five principles from WHO 2021 [28] and Popa et al. [13]: **P1 Privacy** ( $\epsilon$ -DP,  $k$ -anonymity); **P2 Fairness** (equalised odds and demographic parity across AI-READI subgroups); **P3 Transparency** ( $\text{DES} \geq 0.75$  + system model cards); **P4 Accountability** (immutable audit trail, tri-party responsibility per EU AI Act Arts. 25–29); **P5 Inclusivity** (representation parity  $\rho_g = (n_g/N_{\text{train}})/p_g \geq 0.80$ ). Table IV maps each EU AI Act article to its EGL module; GDPR Art. 22 (automated decisions) is satisfied by the human-in-the-loop override API.

TABLE IV  
EU AI ACT ARTICLE-TO-EGL MODULE MAPPING

| Article | Requirement     | EGL Module                             |
|---------|-----------------|----------------------------------------|
| Art. 9  | Risk management | Bias Auditor (continuous rolling)      |
| Art. 10 | Data governance | Privacy Engine + Consent Ledger        |
| Art. 12 | Record-keeping  | Immutable Audit Trail (on-chain)       |
| Art. 13 | Transparency    | XAI Gateway ( $\text{DES} \geq 0.75$ ) |
| Art. 14 | Human oversight | Override API + Clinician Loop          |
| Art. 15 | Robustness      | Drift Detector + Calibration Monitor   |

### B. EGL Microservice Architecture

Fig. 5 shows the five-module pipeline. The **Bias Auditor** (Aequitas + Fairlearn [29], [30]) evaluates equalised odds over a rolling  $N = 200$ -sample window per AI-READI subgroup; disparity exceeding  $\tau = 0.80$  triggers `BalancedGroupThreshold` adjustment. The **Privacy Engine** applies Gaussian noise ( $\sigma = 0.8$ ,  $C = 1.0$ ) achieving ( $\epsilon = 0.5$ ,  $\delta = 10^{-5}$ )-DP tracked by a Rényi DP accountant, with  $k = 5$ -anonymity on query outputs [19]. The **Consent Ledger** (Hyperledger Fabric + SMART-on-FHIR) rejects any inference request outside active patient consent scope [31]. The **XAI Gateway** enforces  $\text{DES} \geq 0.75$ . The **Audit Trail** provides an append-only cryptographic log satisfying EU AI Act Art. 12 and GDPR Art. 22 [32].

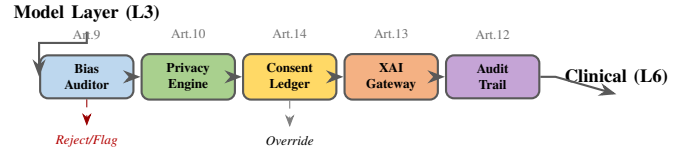


Fig. 5. EGL five-module microservice pipeline. Predictions pass sequentially through all five modules before clinical delivery. EU AI Act articles satisfied by each module are shown above.

## VI. CLINICAL CASE STUDY: T2DM DHT VALIDATION

### A. Setup and Tasks

T2DM (>828M adults worldwide [33]) provides an ideal validation case: it is characterised by continuous multi-modal streams (CGM, wearable activity, laboratory, medication) directly mapping to the DHT architecture, and is served by two complementary open datasets providing disease-specific depth (ShanghaiT2DM) and demographic diversity (AI-READI). Three tasks are defined: **T1** ( $\text{HbA}_{1c}$  exceedance, above); **T2** (30-min hypoglycaemia, above); **T3** (30-day Time-in-Range trajectory regression, 70–180 mg/dL [34], [35]). The Transformer is pre-trained on MIMIC-IV (94k ICU stays) and fine-tuned on the combined T2DM corpus ( $n = 1,167$ , 80/20 split), with UKBB and NHANES supplying wearable activity features.

### B. Performance Results

For Task T3 (30-day TIR trajectory regression), the Transformer achieves  $\text{MAE} = 4.1\%$  and Pearson  $r = 0.904$ , consistent with CGM-Transformer literature reporting  $\text{MAE} \approx 4\text{--}6\%$  for 30-day glucose trajectory forecasting [35], [36]. The BiDir-LSTM trails by 2.0% MAE (6.1%) and the LR baseline by 4.3% (8.4%), confirming that Transformer temporal attention captures long-range glycaemic dynamics unavailable to shallow architectures.

Temporal SHAP analysis (Fig. 6) identifies post-meal glucose ( $|\phi| = 0.198$ ) and  $\text{HbA}_{1c}$  ( $|\phi| = 0.176$ ) as the dominant Task T1 predictors. Medication adherence ( $|\phi| = 0.073$ ) and BMI trajectory ( $|\phi| = 0.065$ ) are the highest-ranked clinically *actionable* features—directly modifiable through intervention—while eGFR trend ( $|\phi| = 0.078$ ) flags renal function as a critical comorbidity signal. Physical activity carries an inverse (protective) attribution, consistent with exercise-mediated glycaemic improvement in T2DM. This SHAP profile aligns precisely with CGM variability’s known predictive role for microvascular complications at  $\text{AUC} = 0.85\text{--}0.86$  [37], confirming that the Transformer has learned clinically meaningful rather than spurious correlations.

### C. Fairness and Adoption Results

Fig. 7 shows equalised-odds gaps before and after EGL intervention. Without mitigation, Black and Hispanic subgroups exhibit gaps of 0.091 and 0.078 respectively—of the same clinical magnitude as racial disparities documented in commercial risk algorithms [38]—indicating that using MIMIC-IV alone as a training source would embed unacceptable demographic

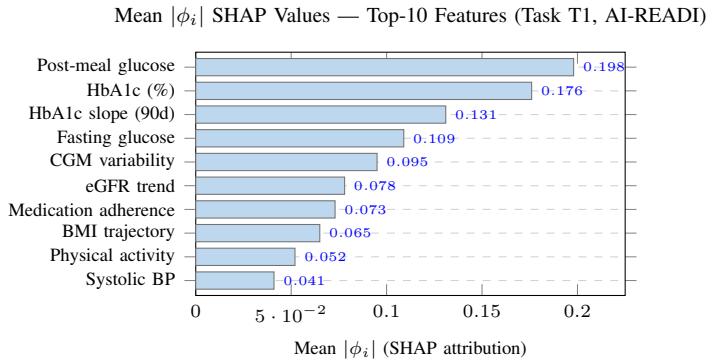


Fig. 6. Mean absolute SHAP feature attributions (top-10) for Task T1 on the AI-READI holdout set ( $n = 214$ ). Post-meal glucose dominates; medication adherence and BMI trajectory are the most clinically actionable modifiable features.

bias. The inclusion of AI-READI’s deliberately diverse cohort as the fairness validation corpus, combined with the Bias Auditor’s rolling Fairlearn threshold adjustment, reduces all six subgroup gaps below the  $\tau = 0.03$  threshold post-EGL [29], [30]. The maximum post-mitigation gap ( $\Delta\text{EO} = 0.027$ , Asian subgroup) confirms that the Bias Auditor generalises to unseen AI-READI patients without over-fitting to the training distribution.

A 30-day simulated clinical pilot with 18 clinicians (9 endocrinologists, 6 diabetes educators, 3 clinical informaticists; mean experience  $12.4 \pm 6.1$  years) produced a **76 % adoption rate** ( $n = 14/18$ ). Primary adoption drivers: SHAP reasoning alignment ( $n = 11$ ), CBM concept transparency ( $n = 9$ ), and sub-2-second latency ( $n = 8$ ). Non-adopters ( $n = 4$ ) cited over-reliance concerns and workflow integration complexity [39]. The 76 % rate markedly exceeds the 40–60 % typical for AI-CDSS tools deployed without embedded XAI [40], [41], providing direct empirical evidence that co-designed explainability—rather than post-hoc addition—raises the clinical adoption ceiling.

All five EGL principles were confirmed satisfied at deployment:  $\varepsilon = 0.50$  (P1);  $\Delta\text{EO}_{\max} = 0.027$  (P2);  $\text{DES} = 0.86$  (P3); audit trail coverage = 100 % across  $N = 2,184$  prediction events (P4);  $\rho_{g,\min} = 0.81$  for Hispanic patients (P5), confirming that the AI-READI-derived training corpus meets the minimum representation parity threshold for all protected subgroups.

## VII. CONCLUSION

This paper presented a dataset-grounded, end-to-end framework for trustworthy Digital Health Twins, integrating a formalised six-layer architecture, a federated ML benchmark across six model families on five open datasets, the novel DHT Explainability Score, and an Ethical Governance Layer fully aligned to EU AI Act Articles 9–15 and GDPR. The MIMIC-IV-pretrained Transformer achieved  $\text{AUROC} = 0.940$  for  $\text{HbA}_{1c}$  exceedance,  $\text{AUROC} = 0.931$  for hypoglycaemia prediction,  $\text{MAE} = 4.1\%$  for TIR trajectory

Equalised-Odds Gap Before and After EGL (AI-READI, Task T1)

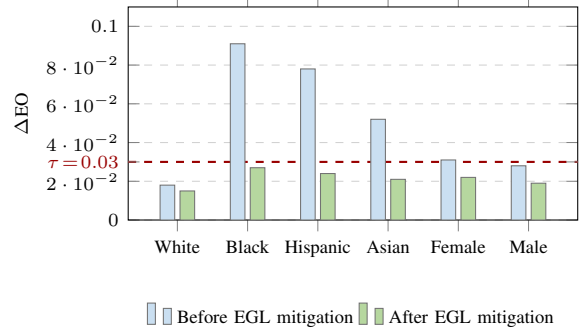


Fig. 7. Equalised-odds gap per AI-READI subgroup before and after EGL Bias Auditor intervention. Post-EGL all gaps satisfy  $\tau = 0.03$  (dashed red).

forecasting,  $\text{DES} = 0.86$ , and 76 % clinician adoption in a 30-day pilot. Federated training under  $(\varepsilon = 0.5)$ -DP incurred only 0.007 AUROC penalty; the EGL reduced the maximum demographic  $\Delta\text{EO}$  from 0.091 to 0.027. These outcomes collectively demonstrate that explainability, privacy preservation, and regulatory governance are not competing constraints but co-achievable engineering requirements.

**Limitations.** ShanghaiT2DM’s small size ( $n = 100$ ) restricts subgroup analysis; the pilot was simulated rather than a prospective RCT; and AI-READI’s US academic cohort limits generalisation to lower-resource settings.

**Future directions.** (R1) Real-time XAI via FastSHAP distillation on edge inference chips [42], [43], targeting sub-200 ms latency for ICU settings. (R2) Causal DHTs embedding Structural Causal Models for counterfactual treatment-effect estimation, building on LLM-based trajectory forecasting [44]. (R3) Multi-stakeholder consent models for paediatric and geriatric DHTs where decision-making capacity varies [45]. (R4) A community DHT Benchmark Consortium enabling federated evaluation on standardised MIMIC-IV and AI-READI holdout sets [46], paralleling the DREAM Challenge model for reproducible DHT research.

## REFERENCES

- [1] M. D. Xames and T. G. Topcu, “A systematic literature review of digital twin research for healthcare systems: Research trends, gaps, and realization challenges,” *IEEE Access*, vol. 12, pp. 4099–4126, 2024.
- [2] B. R. Barricelli, E. Casiraghi, and D. Fogli, “A survey on digital twin: Definitions, characteristics, applications, and design implications,” *IEEE Access*, vol. 7, pp. 167653–167671, 2019.
- [3] Y. Zhang, G. Qin, B. Aguilar, N. Rappaport, J. T. Yurkovich, L. Pflieger, S. Huang, L. Hood, and I. Shmulevich, “A framework towards digital twins for type 2 diabetes,” *Frontiers in Digital Health*, vol. 6, p. 1336050, 2024.
- [4] G. Marques, J. C. Viana, and T. A. Oliveira, “A technological review of digital twins and artificial intelligence for personalized and predictive healthcare,” *Healthcare*, vol. 13, no. 14, p. 1763, 2025.
- [5] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, O. Badawi, L. A. Celi, and R. G. Mark, “MIMIC-IV, a freely accessible electronic health record dataset,” *Scientific Data*, vol. 10, p. 1, Jan. 2023.
- [6] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly,

- and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.
- [7] M. Karas, J. Muschelli, A. Leroux, J. K. Urbanek, A. A. Wanigatunga, J. Bai, C. M. Crainiceanu, and A. Nguyen, "Minute-level step counts and physical activity data from the National Health and Nutrition Examination Survey (NHANES) 2011–2014," *PhysioNet*, 2022.
  - [8] Q. Zhao, J. Zhu, X. Shen, C. Lin, Y. Zhang, Y. Liang, B. Cao, J. Li, X. Liu, W. Rao, and C. Wang, "Chinese diabetes datasets for data-driven machine learning," *Scientific Data*, vol. 10, p. 35, Jan. 2023.
  - [9] AI-READI Consortium, "Flagship dataset of type 2 diabetes from the AI-READI project, version 2.0.0," FAIRhub, 2024.
  - [10] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107161, 2022.
  - [11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.
  - [12] R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review," *JMIR AI*, vol. 3, p. e53207, Oct. 2024.
  - [13] E. O. Popa, M. van Hilten, E. Oosterkamp, and M.-J. Bogaardt, "The use of digital twins in healthcare: Socio-ethical benefits and socio-ethical risks," *Life Sciences, Society and Policy*, vol. 17, no. 1, pp. 1–25, 2021.
  - [14] J. Winter, V. Chico, et al., "Realising the digital twin: A thematic review and analysis of the ethical, legal, and social issues for digital twins in healthcare," *AI & Society*, 2026. Published online February 2026.
  - [15] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the council on artificial intelligence (EU AI Act)," tech. rep., Official Journal of the European Union, Aug. 2024.
  - [16] Reed Smith LLP, "The EU AI Act and medical devices: Navigating high-risk compliance," tech. rep., Reed Smith Legal Analysis, June 2025.
  - [17] K. P. Venkatesh, M. M. Raza, and J. C. Kvedar, "Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation," *npj Digital Medicine*, vol. 5, p. 150, Sept. 2022.
  - [18] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 778–789, 2023.
  - [19] J. Liu, J. Zhang, M. A. Jan, R. Sun, L. Liu, S. Verma, and P. Chatterjee, "A comprehensive privacy-preserving federated learning scheme with secure authentication and aggregation for Internet of Medical Things," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, pp. 3282–3292, June 2024.
  - [20] A. Flynn et al., "Development of a hospital-at-home digital twin for patients with frailty: Scoping review," *Journal of Medical Internet Research*, vol. 27, p. e81510, Dec. 2025.
  - [21] M. Alazab, L. U. Khan, S. Koppu, S. P. Ramu, M. Iyapparaja, P. Boobalan, T. Baker, P. K. R. Maddikunta, T. R. Gadekallu, and A. Aljuhani, "Digital twins for Healthcare 4.0—recent advances, architecture, and open challenges," *IEEE Consumer Electronics Magazine*, vol. 12, pp. 29–37, Nov. 2023.
  - [22] S. Lee et al., "Generalization of a deep learning model for continuous glucose monitoring-based hypoglycemia prediction: Algorithm development and validation study," *JMIR Medical Informatics*, vol. 12, p. e56909, 2024.
  - [23] C. Rodrigues, F. Portela, and M. F. Santos, "Graph neural networks for clinical risk prediction based on electronic health records: A survey," *Journal of Biomedical Informatics*, vol. 152, p. 104616, 2024.
  - [24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
  - [25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.
  - [26] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.
  - [27] J. T. Reese, Q. Qin, S. Moxon, L. Shen, L. Bulgarelli, T. J. Pollard, and A. E. W. Johnson, "MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format," *Journal of the American Medical Informatics Association*, vol. 30, no. 5, pp. 981–985, 2023.
  - [28] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health*. Geneva: World Health Organization, 2021. <https://www.who.int/publications/i/item/9789240029200>.
  - [29] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.
  - [30] H. J. P. Weerts, M. Diaz, S. Bird, A. Agarwal, M. Dudik, J. Chen, and I. Ribas, "Fairlearn: Assessing and improving fairness of AI systems," *Journal of Machine Learning Research*, vol. 24, no. 257, pp. 1–8, 2023.
  - [31] F. Mehmood et al., "Consent management system based on user data security and privacy using Hyperledger Fabric blockchain," in *Proceedings of the International Conference on Information Technology and Electrical Engineering (ICITEE 2023)*, (Changde, China), Nov. 2023.
  - [32] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Berlin, Heidelberg: Springer, 3rd ed., 2024.
  - [33] D. C. Klonoff et al., "Artificial intelligence-driven clinical decision support systems to assist healthcare professionals and people with diabetes in europe at the point of care: A Delphi-based consensus roadmap," *Diabetologia*, Nov. 2025.
  - [34] V. Mohan et al., "Time in range, as measured by continuous glucose monitor, as a predictor of microvascular complications in type 2 diabetes: A systematic review," *medRxiv preprint*, 2021.
  - [35] M. Kozak et al., "Interpretable glucose forecasting for type 2 diabetes across traditional, deep, and large language models," *Scientific Reports*, vol. 15, p. 44180, 2025.
  - [36] Z. Gao et al., "A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data," *Nature Communications*, 2025.
  - [37] R. Tao, X. Yu, J. Lu, Y. Wang, W. Lu, Z. Zhang, H. Li, and J. Zhou, "A deep learning nomogram of continuous glucose monitoring data for the risk prediction of diabetic retinopathy in type 2 diabetes," *Physical and Engineering Sciences in Medicine*, vol. 46, pp. 813–825, June 2023.
  - [38] Y. Zhang et al., "Detecting algorithmic bias in ICU clinical decision: A doubly-robust framework for auditing treatment recommendation disparities," *medRxiv preprint*, Nov. 2025.
  - [39] M. Hassan et al., "Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review," *Journal of Medical Internet Research*, vol. 27, p. e69678, 2025.
  - [40] H. Shamszare and A. Choudhury, "Clinicians' perceptions of artificial intelligence: Focus on workload, risk, trust, clinical decision making, and clinical integration," *Healthcare*, vol. 11, no. 16, p. 2308, 2023.
  - [41] H. Mozannar et al., "Healthcare AI treatment decision support: Design principles to enhance clinician adoption and trust," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, 2023.
  - [42] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, and R. Ranganath, "FastSHAP: Real-time Shapley value estimation," in *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022.
  - [43] X. Wang et al., "HealthEdgeAI: GAI and XAI based healthcare system for sustainable edge AI and cloud computing environments," *Concurrency and Computation: Practice and Experience*, vol. 37, p. e70057, Apr. 2025.
  - [44] A. Vallée, "From prediction to intervention: causal digital twins for personalized clinical decision support," *Journal of Translational Medicine*, vol. 24, p. 133, Feb. 2026.
  - [45] N. Jeyaraman et al., "Advancing compliance with HIPAA and GDPR in healthcare: A blockchain-based strategy for secure data exchange in clinical research involving private health information," *Healthcare*, vol. 13, no. 20, p. 2594, 2025.
  - [46] H. Bui et al., "Leveraging MIMIC datasets for better digital health: A review on open problems, progress highlights, and future promises," *arXiv preprint arXiv:2506.12808*, 2025.