IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2025.Doi Number*

# A Comprehensive Evaluation of Generative Models for Privacy-Preserving Synthetic Student Data

**Divine Iloh[1], Senior Member, IEEE, Grace Oku[2], Shaozhi Jiang[3], Senior Member, IEEE, Rashmi Choudhary, Senior Member, IEEE[4], Rohit Kapa[5], Nitesh Chilakala[6]**

[1]University of Arkansas at Little Rock, Little Rock, AR 72204, USA
[2]Illinois State University, Normal, IL 61761, USA
[3]Independent Researcher
[4]Independent Researcher
[5]Prudential Financial, Newark, NJ 07102, USA
[6]University of Houston, Clear Lake, Houston, TX, 77058, USA

Corresponding author: Divine Iloh (e-mail: divineiloh@gmail.com).
This research received no external funding.
Reference [8] is prior work by the first author.

**ABSTRACT** Privacy regulations and institutional policies limit the sharing of educational data, constraining reproducibility in learning analytics. Prior evaluations of synthetic data on benchmarks such as OULAD have examined statistical or adversarial synthesizers in isolation, rarely jointly assessing fidelity, utility, privacy, and explainability. We compare three synthesis paradigms, statistical (Gaussian Copula), adversarial (CTGAN), and diffusion-based (TabDDPM), on two benchmarks (OULAD: 32,593 records; ASSISTments: 8,519) across five evaluation axes: distributional fidelity (SDMetrics), downstream utility (Train on Synthetic, Test on Real), discriminative realism (classifier two-sample test), membership-inference privacy, and feature-importance preservation (SHAP). The pipeline is repeated over five random seeds with bootstrap confidence intervals and Bonferroni-corrected permutation tests ($\alpha' \approx 0.0028$). Four findings emerge: First, TabDDPM delivers the strongest classification utility: on OULAD, a Random Forest achieves TSTR AUC = $0.962 \pm 0.001$, within 0.5 percentage points of the real-data baseline. Second, all synthesizers exhibit near-chance membership-inference risk (worst-case effective AUC $\leq 0.527$). Third, distributional fidelity does not predict task utility; CTGAN scores highest on SDMetrics yet does not yield the smallest utility gap. Fourth, TabDDPM best preserves real-data feature-importance rankings on OULAD (Spearman $\rho = 0.846$, $p < 0.001$). These results provide task-driven guidance for selecting a synthesizer in learning analytics.

**INDEX TERMS** ASSISTments, CTGAN, diffusion models, educational data mining, feature importance, Gaussian Copula, generative adversarial networks, learning analytics, membership inference attacks, OULAD, SHAP, synthetic data, tabular data synthesis.

## I. INTRODUCTION

Learning analytics seeks to improve educational outcomes by analyzing data on student behavior, engagement, and performance [1]. Advances in this field depend on access to representative datasets for benchmarking and reproducible research, yet privacy regulations and institutional policies routinely restrict data sharing [2]–[4]. In the United States, FERPA limits disclosure of personally identifiable information from education records [5]; in the European Union, the GDPR imposes strict requirements on lawful processing and cross-jurisdictional transfer of personal data [6]. Beyond legal mandates, educational records encode sensitive learning trajectories whose inappropriate reuse can produce material harm, and institutions commonly impose additional governance layers (ethical review, data use agreements, restricted-access platforms) that further limit broad reuse.

These constraints create a reproducibility bottleneck: many studies cannot release their underlying data, making it difficult to verify findings or compare methods on common benchmarks. When datasets are shared, access is often gated by multi-step approval workflows that vary by institution.

Synthetic data generation offers a principled pathway to broader sharing while reducing exposure of individual students [7]. A generative model is trained on a private dataset and releases synthetic records that preserve statistical structure without revealing whether any specific individual contributed to the training data. For learning analytics, a useful synthetic dataset must simultaneously satisfy (i) fidelity, preserving distributions and inter-variable relationships; (ii) utility, enabling models trained on synthetic data to generalize to real data on downstream tasks; (iii) privacy, preventing an adversary from inferring membership in the training set [7]; and (iv) explainability, preserving the feature-level drivers that inform actionable educational decisions.

However, the degree to which modern tabular generative models jointly satisfy these requirements for educational data remains insufficiently characterized [8]. Results from healthcare and finance do not transfer automatically, because educational datasets are commonly mixed-type, class-imbalanced, and shaped by engagement dynamics and

institutional course structure. Three representative paradigms have emerged—copula-based statistical models [9], conditional GANs [10], and denoising diffusion models for heterogeneous tabular data [11], yet head-to-head evaluations on educational datasets remain limited. Prior work on OULAD has typically evaluated Gaussian Copula and/or CTGAN in isolation [8], and existing evaluations rarely assess fidelity, utility, realism, privacy, and explainability within a single framework.

This paper addresses these gaps through a comprehensive empirical comparison of three representative synthesizers, namely Gaussian Copula (statistical), CTGAN (adversarial), and TabDDPM (diffusion), on two widely used learning analytics benchmarks: OULAD (32,593 student records) and ASSISTments (8,519 records). All synthesizers are evaluated within a single deterministic pipeline repeated across five random seeds (0–4), with 95% bootstrap confidence intervals (1,000 resamples) and paired permutation tests under Bonferroni correction ($\alpha' = 0.05/18 \approx 0.0028$).

We evaluate each synthesizer along five complementary axes: distributional fidelity (SDMetrics [9]), downstream utility (Train on Synthetic, Test on Real [TSTR] classification AUC and regression MAE, benchmarked against a real-data baseline), discriminative realism (classifier two-sample test [12]), membership-inference privacy (worst-case effective AUC across multiple attacker models [13]), and feature-importance preservation (SHAP-based rank correlation between synthetic and real-data models). Full metric definitions are provided in Section III-D.

Four principal findings emerge from this evaluation:

**F1 (Utility):** TabDDPM delivers the strongest downstream classification transfer on both datasets; on OULAD, the mean TSTR AUC (averaged across RF and LR learners) = 0.957 ± 0.004 across seeds, within 1.0 percentage point of the real-data baseline (0.964 ± 0.001). TabDDPM also yields the lowest regression MAE on both benchmarks.

**F2 (Privacy):** All three synthesizers exhibit near-chance membership-inference risk, with worst-case effective AUC ≤ 0.527 across all dataset–synthesizer configurations evaluated.

**F3 (Fidelity–Utility Decoupling):** Distributional fidelity does not reliably predict task utility. CTGAN attains the highest SDMetrics quality scores on both datasets, yet TabDDPM consistently yields the smallest utility gap to the real baseline, demonstrating that synthesizer selection should be driven by downstream task performance rather than distributional similarity alone.

**F4 (Explainability):** SHAP-based feature-importance analysis reveals that TabDDPM best preserves the real-data feature ranking on OULAD classification (Spearman $\rho = 0.846$, $p < 0.001$), while all synthesizers achieve near-perfect rank preservation on the lower-dimensional ASSISTments feature space ($\rho \geq 0.950$).

Together, these results provide task-driven guidance for practitioners selecting tabular synthesizers under joint constraints on privacy, utility, explainability, and computational resources in learning analytics.

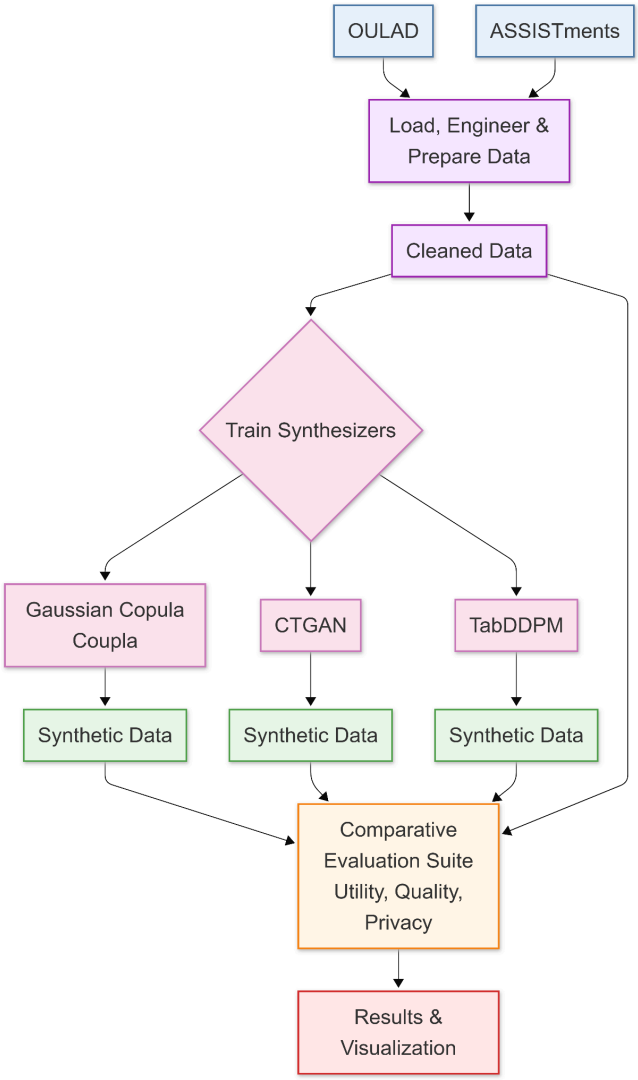


**Figure 1.** End-to-end experimental workflow. Real data from OULAD and ASSISTments are split into training and test partitions. Three synthesizers (Gaussian Copula, CTGAN, TabDDPM) generate synthetic training tables, which are evaluated across five axes: fidelity (SDMetrics), realism (C2ST), utility (TSTR), privacy (MIA), and explainability (SHAP). The pipeline is repeated across five random seeds.

## II.BACKGROUND AND RELATED WORK

### A. Privacy Constraints in Learning Analytics

Sharing student-level behavioral and performance data is governed by layered legal and institutional constraints. FERPA restricts disclosure of personally identifiable information in the United States [5], while the GDPR regulates lawful processing and cross-jurisdictional transfer of personal data in the European Union [6]. Institutions commonly impose additional governance (ethical review, data use agreements, restricted-access platforms) that further limit distribution [1]–[4].

Conventional de-identification (removing direct identifiers) can be insufficient for high-dimensional behavioral data:

Narayanan and Shmatikov [14] demonstrated that re-identification is feasible when an adversary possesses auxiliary information. This risk is particularly salient for educational datasets that combine engagement traces, demographic attributes, and outcome labels.

Two technical approaches address privacy more formally. Differential privacy provides a mathematical framework for bounding information leakage about any individual [15], but strong guarantees can substantially reduce analytical utility, particularly with small cohorts or imbalanced outcomes common in learning analytics. Federated learning reduces data centralization by training models across institutions without pooling raw records [16] but introduces coordination overhead and reproducibility constraints that limit its suitability for open benchmarking.

Synthetic data generation occupies a middle ground: it enables broader sharing without distributing raw records, while privacy properties can be empirically assessed under explicit threat models. The central question is whether a synthesizer can preserve task-relevant structure while demonstrably limiting disclosure risk at acceptable computational cost.

### B. Tabular Data Synthesis Methods

Tabular data synthesis methods commonly fall into three paradigms [7]:

**1) Statistical models.** Copula-based methods estimate per-column marginal distributions and model inter-variable dependence through a copula function [9]. The Gaussian Copula implementation uses closed-form estimation, making it computationally efficient, though the Gaussian dependence assumption may struggle with complex multimodal or non-linear relationships.

**2) Adversarial models.** Generative adversarial networks (GANs) learn a generator alongside a discriminator through adversarial training [17]. CTGAN adapts this framework to tabular data by applying mode-specific normalization and conditional sampling to mixed-type features [10]. Privacy-oriented variants such as PATE-GAN incorporate differential privacy via teacher ensembles [18]. In practice, GAN training can exhibit hyperparameter sensitivity, instability, or mode collapse [17].

**3) Diffusion models.** Denoising diffusion probabilistic models (DDPMs) generate samples by learning to reverse a progressive noising process [19]. TabDDPM adapts this framework to heterogeneous tabular data with separate mechanisms for continuous (Gaussian diffusion) and categorical (multinomial diffusion) variables [11]. Diffusion approaches incur a higher computational cost than copula methods but provide stable optimization (avoiding adversarial min-max instability) and have demonstrated competitive or superior quality across tabular benchmarks [11].

### C. Synthetic Data in Educational Settings

Synthetic data studies specifically addressing educational datasets remain limited [8]. Educational data pose domain-specific challenges: mixed data types, strong class imbalance in outcomes such as dropout, temporal and institutional structure, and patterns shaped by engagement behavior and course design [1], [20].

Prior work has demonstrated that synthetic student data can support downstream prediction tasks with partial retention of performance [8], but most evaluations emphasize distributional similarity and task utility while providing less systematic assessment of privacy risk. Prior learning analytics studies on OULAD have typically evaluated Gaussian Copula and/or CTGAN individually, whereas TabDDPM has been benchmarked primarily on general tabular datasets rather than on educational data. No prior study, to our knowledge, has compared all three paradigms on educational benchmarks within a single protocol that jointly assesses fidelity, realism, utility, privacy, and explainability.

Fairness considerations are also relevant: synthesizers can replicate or amplify biases present in the training data, which may require audits or constraints depending on the intended use [21]. While we do not perform explicit fairness audits in this study, we note this as an important direction for future work (Section V-C).

### D. Evaluation Dimensions for Synthetic Tabular Data

Synthetic data evaluation is inherently multi-dimensional; no single metric suffices for assessing fitness for release [7]. Five complementary axes structure the evaluation landscape: (1) distributional fidelity, comparing real and synthetic distributions via standardized quality scores [9]; (2) discriminative realism, testing whether a classifier can distinguish real from synthetic records (C2ST [12]); (3) downstream utility, measuring whether models trained on synthetic data transfer to real data on target tasks (TSTR [8]); (4) privacy risk, assessing whether an adversary can infer membership in the training set via membership inference attacks [13]; and (5) feature-importance preservation, quantifying whether SHAP-derived feature rankings are conserved across real and synthetic training regimes [22]. This last axis is particularly relevant for educational applications, where understanding which variables predict dropout or performance informs actionable interventions. Existing learning analytics evaluations do not consistently report all five axes across modern synthesizer paradigms with rigorous statistical testing; this gap motivates the unified framework presented in Section III-D.

## III. METHODOLOGY

### A. Datasets and Prediction Targets

We evaluate synthetic data generation across two learning analytics benchmarks with complementary structures, scales, and feature dimensions. Table I summarizes their key characteristics.

**TABLE I: Dataset Characteristics**

| Property | OULAD | ASSISTments |
|---|---|---|
| Source | Open University VLE [23] | Intelligent tutoring system [24] |
| Student records | 32,593 | 8,519 |
| Total columns | 23 | 7 |
| Feature columns | 18 (10 numeric, 8 categorical) | 4 (all numeric) |
| Target columns | 2 (dropout, final_grade) | 2 (high_engagement, student_pct_correct) |
| Identifier columns | 3 (id_student, code_module, code_presentation) | 1 (user_id) |
| Train/test split | 70/30 GroupShuffleSplit on id_student | 70/30 GroupShuffleSplit on user_id |
| Classification target | dropout (binary) | high_engagement (binary) |
| Regression target | final_grade (0–100 scale) | student_pct_correct ([0, 1]) |

**1) Open University Learning Analytics Dataset (OULAD) [23]**. OULAD provides student demographic attributes and aggregated VLE interaction features across multiple Open University courses. Prediction targets are dropout (binary classification) and final_grade (0–100 scale regression). Note that code_module and code_presentation serve dual roles as both grouping keys and categorical features.

We use a 70/30 train/test split via GroupShuffleSplit on id_student, ensuring no student appears in both partitions. Because the split is repeated across five random seeds, exact partition sizes vary slightly (e.g., 22,852 train / 9,741 test).

**2) ASSISTments (2012–2013 school-year data) [24].** We aggregate interaction-level records from an intelligent tutoring system into 8,519 student rows. Feature engineering uses only the first K = 20 interactions per student to compute four behavioral features: unique_skills (distinct skills attempted), hint_rate (mean hint usage), avg_attempts (mean attempts per problem), and avg_response_time (mean first-response time in ms). Two prediction targets are derived: high_engagement (binary; 1 if total interactions ≥ cohort median of 40.0) and student_pct_correct (continuous in [0, 1]; mean correctness over the early window). The same 70/30 GroupShuffleSplit protocol is applied on user_id (e.g., 5,963 train / 2,556 test).

**Data integrity note**. Because features use only the first K = 20 interactions, while the engagement target is derived from total interactions, we compute a leakage diagnostic: the maximum absolute Pearson correlation between features and targets is $|r| = 0.047$, indicating negligible leakage.

**Preprocessing and feature handling.** Identifier columns are excluded from all downstream models. Numeric features are median-imputed and z-score standardized; categorical features are mode-imputed and one-hot encoded (unknown categories ignored at test time). All transformations are fit on the training partition only and applied to held-out and synthetic data through the same fitted preprocessor, preventing information leakage.

### B. Synthesizers

We compare three synthesizers spanning statistical, adversarial, and diffusion-based paradigms, each trained on the real training partition and used to sample a synthetic table matched in row count and schema:

**Gaussian Copula (Statistical)** [9]: models each marginal distribution independently and captures inter-variable dependence through a Gaussian copula correlation structure in a latent space. We use the SDV implementation with closed-form estimation and sample a synthetic table matching the size of the training set.

**CTGAN (Adversarial)** [10]: A conditional GAN designed for mixed-type tabular data, employing mode-specific normalization for continuous variables and conditional sampling to address class imbalance. We train CTGAN for 300 epochs with a batch size of 500 and sample a synthetic table matched to the training set size.

**TabDDPM (Diffusion)** [11]: A denoising diffusion probabilistic model adapted for heterogeneous tabular features, with Gaussian diffusion for continuous variables and multinomial diffusion for categorical variables. We use the Synthcity implementation and train for 1,200 iterations, then sample a synthetic table matched to the training set size.

### C. Experimental Protocol

For each dataset–synthesizer combination, we execute an identical end-to-end pipeline: (1) split the real data into train/test partitions via *GroupShuffleSplit*; (2) fit the synthesizer on the real training partition; (3) sample a synthetic training table with matched row count and schema; (4) evaluate along all five axes (fidelity, realism, utility, privacy, explainability). The full pipeline is repeated across five random seeds (0–4), and results are aggregated as mean ± SD across seeds.

**Uncertainty quantification.** Within each seed, 95% bootstrap confidence intervals (1,000 resamples, percentile method) are computed for utility metrics: test-set resamples for AUC-ROC; per-sample absolute-error resamples for MAE.

**Statistical comparisons.** Pairwise synthesizer and TSTR-vs.-TRTR comparisons use paired permutation tests (10,000 sign-flips on per-sample loss differences) [25], yielding two-sided p-values, mean differences, and Cohen's d. Bonferroni correction across 18 planned comparisons (3 pairs × 2 tasks × 2 learners + 6 TSTR-vs.-TRTR) sets $\alpha' = 0.05/18 \approx 0.0028$. Effect sizes (Cohen's d) are reported for TSTR utility comparisons where per-sample loss vectors

enable paired tests; they are not applicable to single-value-per-seed metrics (SDMetrics, C2ST, MIA, SHAP ρ).

### D. Evaluation Metrics

We evaluate synthetic data quality along five complementary axes:

**1) Distributional fidelity (SDMetrics)**. The SDMetrics Quality Report [9] is computed between real and synthetic training tables. The overall quality score (range [0, 1]) summarizes column-level distributional similarity and inter-column relationship preservation.

**2) Discriminative realism (C2ST).** A Random Forest classifier (300 trees) is trained to distinguish real from synthetic records on a balanced, stratified 70/30 internal split [12]. ID columns and target columns are excluded from the feature set. We report effective AUC = max(AUC, 1−AUC), where values near 0.5 indicate that real and synthetic records are indistinguishable to the classifier.

**3) Downstream utility (TSTR)**. Supervised models are trained on synthetic data and evaluated on the held-out real test split. For classification, we report AUC-ROC from both a Random Forest (300 trees) and Logistic Regression (max_iter = 1,000, solver = "liblinear"); for regression, we report MAE from both a Random Forest (300 trees) and Ridge regression (default regularization). The TSTR score is the average across the two learner types per task. A Train on Real, Test on Real (TRTR) baseline provides the upper-bound reference. Per-sample loss vectors (log-loss for classification; absolute error for regression) are retained for the paired permutation tests described in Section III-C.

**4) Membership-inference privacy (MIA)**. We implement a distance-based membership inference attack. For each record, two features are computed from k-nearest-neighbor distances (k = 5) to the synthetic dataset: the minimum distance (d_min) and the mean distance to the k neighbors (d_mean). Three attacker classifiers are then trained on balanced shadow sets drawn from the real training partition (members) and real test partition (non-members): Logistic Regression, Random Forest (100 trees), and XGBoost. Privacy risk is reported as worst-case effective AUC = max over attackers of max(AUC, 1−AUC), where values near 0.5 indicate near-chance membership prediction.

**Design note (column exclusion).** C2ST excludes both ID and target columns, assessing realism on the feature space alone. MIA excludes only ID columns (targets are included) to reflect a conservative worst-case attacker with access to all non-identifier attributes, including outcome labels.

**5) Feature-importance preservation (SHAP).** For each seed, dataset, task, and data source (TRTR or TSTR), a Random Forest (100 trees) is trained through the same preprocessing pipeline, and SHAP values are computed via TreeExplainer [22] on a test-set subsample (200 samples for OULAD; 500 for ASSISTments). Mean absolute SHAP values per feature (one-hot columns aggregated to parent features) yield importance rankings; TSTR rankings are compared to TRTR via Spearman's ρ, averaged across seeds. This quantifies whether synthesis preserves the explanatory structure that informs intervention design. SHAP beeswarm plots are also generated to visualize the directionality of feature effects beyond rank ordering.

**6) Computational efficiency.** Wall-clock time is recorded for the full end-to-end pipeline (data loading, synthesis, and all evaluations) per seed and reported as mean ± SD across seeds.

## IV. RESULTS

We report results for Gaussian Copula, CTGAN, and TabDDPM under the unified five-axis protocol described in Section III. All metrics are aggregated across five random seeds (0–4) and reported as mean ± SD (n = 5). Where applicable, "effective AUC" denotes max(AUC, 1−AUC) as defined in Section III-D. Table numbering continues from Table I (dataset characteristics) in Section III.

### A. Downstream Utility (TSTR)

**Classification**

Table II presents TSTR classification utility (AUC-ROC) for individual downstream learners and their average (primary TSTR score per Section III-D).

**TABLE II: Downstream Classification Utility (TSTR AUC-ROC, mean ± SD across 5 seeds)**

| Configuration | TRTR Baseline | Gaussian Copula | CTGAN | TabDDPM |
|---|---|---|---|---|
| OULAD – RF | 0.967 ± 0.001 | 0.936 ± 0.002 | 0.948 ± 0.002 | 0.962 ± 0.001 |
| OULAD – LR | 0.962 ± 0.001 | 0.939 ± 0.006 | 0.953 ± 0.002 | 0.952 ± 0.007 |
| OULAD – Mean | 0.964 ± 0.001 | 0.938 ± 0.003 | 0.951 ± 0.002 | 0.957 ± 0.004 |
| ASSISTments – RF | 0.838 ± 0.006 | 0.548 ± 0.037 | 0.504 ± 0.050 | 0.678 ± 0.014 |
| ASSISTments – LR | 0.722 ± 0.005 | 0.516 ± 0.115 | 0.491 ± 0.125 | 0.722 ± 0.006 |
| ASSISTments – Mean | 0.780 ± 0.003 | 0.532 ± 0.053 | 0.498 ± 0.046 | 0.700 ± 0.008 |

*Note: AUC-ROC is mean ± SD across seeds (n = 5). "Mean" is the average of the RF and LR TSTR scores per seed, then averaged across seeds. Higher is better.*

On OULAD, TabDDPM achieves the strongest TSTR classification utility: mean AUC = 0.957 ± 0.004, within 0.7 percentage points (pp) of the TRTR baseline (0.964 ± 0.001). The per-learner breakdown reveals that the TabDDPM Random Forest (0.962 ± 0.001) nearly matches the real-data Random Forest (0.967 ± 0.001), a gap of only 0.5 pp. CTGAN follows at 0.951 ± 0.002, and Gaussian Copula trails at 0.938 ± 0.003. All pairwise differences on OULAD classification are statistically significant after Bonferroni correction ($p < 0.001$). Effect sizes range from negligible (TabDDPM vs. CTGAN, $d \approx 0.14$) to small ($d \approx 0.25$) and reach medium for TabDDPM vs. Gaussian Copula ($d \approx 0.65$–$0.71$).

On ASSISTments, classification utility degrades markedly for Gaussian Copula (0.532 ± 0.053) and CTGAN (0.498 ± 0.046), both near chance against a TRTR baseline of 0.780 ± 0.003, with very high cross-seed LR variance (SD = 0.115 and 0.125, respectively) reflecting unstable learned boundaries. TabDDPM preserves substantially more utility (0.700 ± 0.008), though it still falls 8.0 pp short of the baseline. Notably, TabDDPM's LR TSTR score (0.722 ± 0.006) matches the TRTR LR baseline (0.722 ± 0.005), suggesting that the linear decision boundary is fully preserved while the nonlinear structure captured by RF degrades. The pronounced utility gap likely reflects the smaller training set (≈5,963 vs. ≈22,800), the compressed 4-feature space, and the derived engagement target. With only four features, any synthesis error in even one marginal (e.g., avg_response_time, which exhibits a heavy right tail) can shift decision boundaries substantially, an effect amplified by the small sample size.

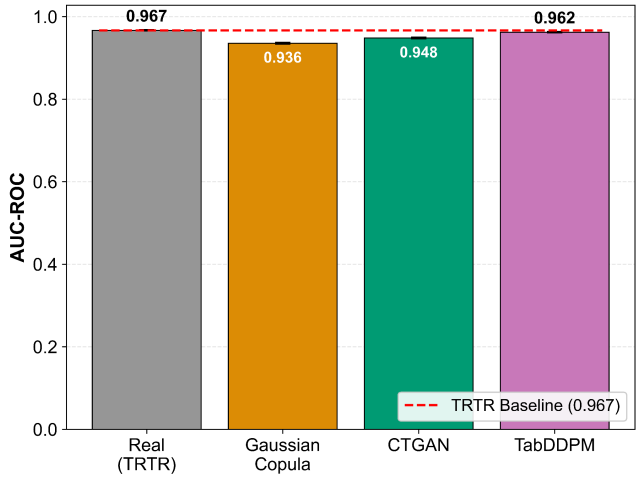


***Figure 2.*** *OULAD TSTR classification utility. Mean AUC-ROC (n = 5 seeds) for RF models. Error bars: ±1 SD.*

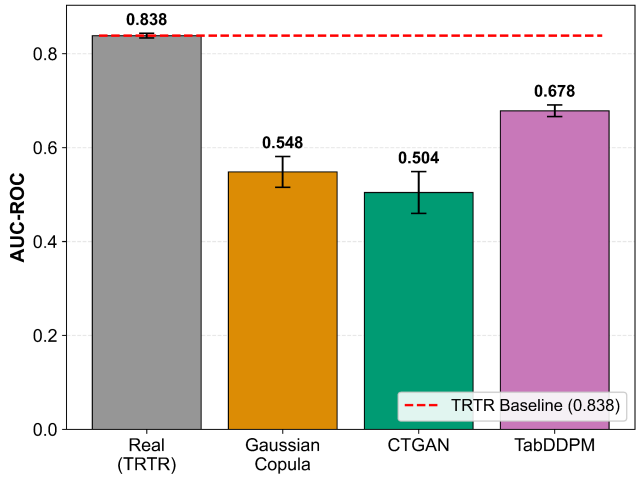


**Figure *3.*** *ASSISTments downstream classification utility (TSTR). Mean AUC-ROC across seeds (n = 5) for Random Forest models. Error bars: ±1 SD.*

**Regression**

Table III presents TSTR regression utility (MAE) per learner and their average.

**TABLE III: Downstream Regression Utility (TSTR MAE, mean ± SD across 5 seeds)**

| Configuration | TRTR Baseline | Gaussian Copula | CTGAN | TabDDPM |
|---|---|---|---|---|
| OULAD – RF | 8.18 ± 0.05 | 19.18 ± 0.65 | 13.13 ± 0.48 | 9.95 ± 0.22 |
| OULAD – Ridge | 15.21 ± 0.13 | 22.19 ± 1.85 | 18.18 ± 0.40 | 18.74 ± 0.96 |
| OULAD – Mean | 11.70 ± 0.09 | 20.68 ± 1.15 | 15.66 ± 0.44 | 14.34 ± 0.58 |
| ASSISTments – RF | 0.174 ± 0.002 | 0.218 ± 0.007 | 0.231 ± 0.039 | 0.198 ± 0.004 |
| ASSISTments – Ridge | 0.202 ± 0.001 | 0.213 ± 0.009 | 0.203 ± 0.004 | 0.201 ± 0.002 |
| ASSISTments – Mean | 0.188 ± 0.001 | 0.215 ± 0.005 | 0.217 ± 0.022 | 0.200 ± 0.003 |

*Note: MAE is mean ± SD across seeds (n = 5). "Mean" is the average of RF and Ridge MAEs per seed. Lower is better.*

On OULAD (TRTR baseline mean MAE = 11.70 ± 0.09), the large per-learner spread (RF = 8.18 vs. Ridge = 15.21) reflects the Random Forest's greater capacity for nonlinear grade prediction. Among TSTR methods, TabDDPM yields the lowest mean MAE (14.34 ± 0.58), followed by CTGAN (15.66 ± 0.44) and Gaussian Copula (20.68 ± 1.15). TabDDPM's RF MAE (9.95 ± 0.22) approaches the baseline RF (8.18 ± 0.05), while all synthesizers produce substantially higher Ridge MAE, indicating that linear learners are more sensitive to synthesis-induced distributional shifts. All pairwise differences are significant ($p < 0.001$), with effect sizes from small (TabDDPM vs. CTGAN, $d \approx 0.30$–$0.41$) to medium (TabDDPM vs. Gaussian Copula, $d \approx 0.55$–$0.62$).

On ASSISTments, the regression task is easier (continuous target in [0, 1] with low variance), and all synthesizers approach the TRTR baseline (0.188 ± 0.001). TabDDPM matches most closely (0.200 ± 0.003), with Gaussian Copula (0.215 ± 0.005) and CTGAN (0.217 ± 0.022) marginally worse. Utility differences on ASSISTments regression are small in absolute terms, and many pairwise comparisons do not reach significance after Bonferroni correction, with negligible effect sizes ($|d| < 0.20$).

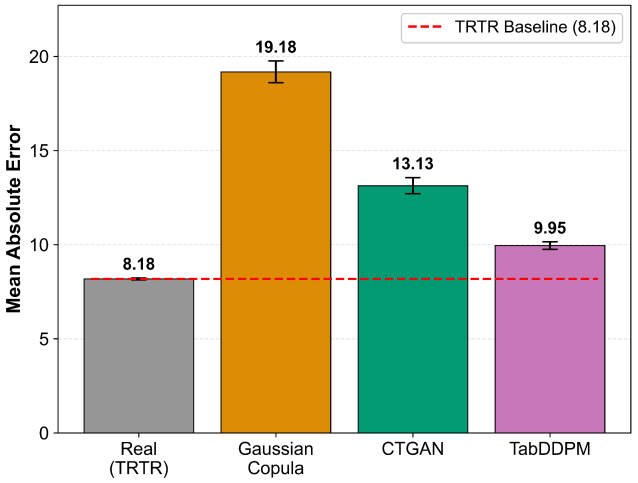


**Figure** *4. OULAD downstream regression utility (TSTR). Mean MAE across seeds (n = 5) for Random Forest models. Error bars: ±1 SD. Lower is better.*

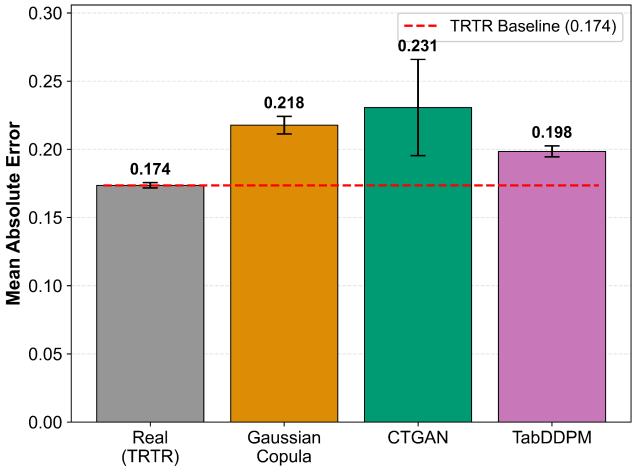


**Figure** *5. ASSISTments downstream regression utility (TSTR). Mean MAE across seeds (n = 5) for Random Forest models. Error bars: ±1 SD. Lower is better.*

### B. Distributional Fidelity (SDMetrics)

Table IV reports SDMetrics overall quality scores for each synthesizer–dataset combination.

**TABLE IV: SDMetrics Quality Report (Overall Score, mean ± SD across 5 seeds)**

| Dataset – Synthesizer | Overall Quality Score |
|---|---|
| OULAD – Gaussian Copula | 0.852 ± 0.003 |
| OULAD – CTGAN | 0.868 ± 0.011 |
| OULAD – TabDDPM | 0.809 ± 0.010 |
| ASSISTments – Gaussian Copula | 0.847 ± 0.004 |
| ASSISTments – CTGAN | 0.924 ± 0.031 |
| ASSISTments – TabDDPM | 0.807 ± 0.003 |

*Note: Higher is better. Range [0, 1].*

CTGAN attains the highest fidelity on both datasets (0.868 ± 0.011 OULAD; 0.924 ± 0.031 ASSISTments), followed by Gaussian Copula (0.852 ± 0.003; 0.847 ± 0.004) and TabDDPM (0.809 ± 0.010; 0.807 ± 0.003). The spread is larger on ASSISTments (CTGAN exceeds TabDDPM by 11.7 pp). Crucially, this fidelity ranking does not align with the utility ranking (Section IV-A), where TabDDPM leads; a decoupling is discussed further in Section IV-F.

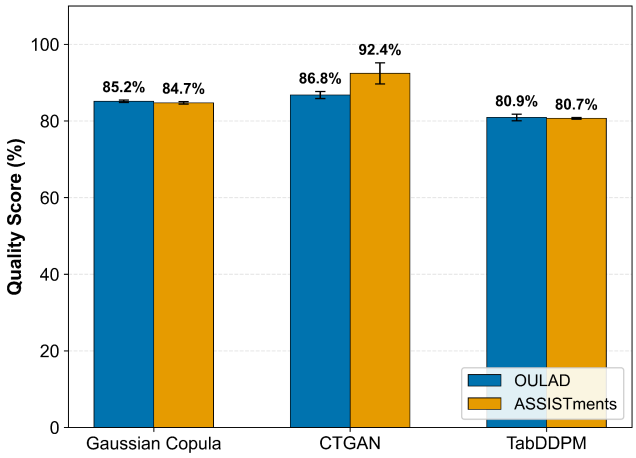


**Figure** *6. SDMetrics overall quality scores for each synthesizer on OULAD and ASSISTments (mean across 5 seeds). Error bars: ±1 SD. Higher is better.*

### C. Discriminative Realism (C2ST)

C2ST results (Table V) are strongly dataset-dependent. On OULAD, all synthesizers are readily distinguishable from real data: Gaussian Copula = 0.999 ± 0.000, CTGAN = 0.962 ± 0.010, TabDDPM = 0.826 ± 0.006. The high values reflect OULAD's complex categorical structure (8 features with many levels), which stresses each synthesizer's categorical mechanism differently: Gaussian Copula maps categories through reversible integer transforms that discard inter-level semantics [9], CTGAN encodes them via conditional vectors with mode-specific normalization [10], and TabDDPM models them through multinomial diffusion that learns per-category transition probabilities [11]. These architectural differences likely explain the wide C2ST spread across synthesizers on OULAD (0.826–0.999) compared to the all-numeric ASSISTments feature space.

On ASSISTments (4 numeric features), TabDDPM achieves better realism (0.671 ± 0.014) than Gaussian Copula (0.950 ± 0.005) or CTGAN (0.959 ± 0.008), though all remain above chance. Notably, high C2ST scores do not preclude high utility: TabDDPM achieves C2ST = 0.826 on OULAD yet delivers TSTR AUC = 0.957 (within 0.7 pp of the baseline), suggesting that C2ST-detectable deviations reside in aspects of the joint distribution irrelevant to supervised targets.

### D. Privacy Risk Under Membership Inference (MIA)

Table V reports both C2ST and MIA results.

**TABLE V: Discriminative Realism (C2ST) and Privacy Risk (MIA), mean ± SD across 5 seeds**

| Dataset – Synthesizer | C2ST Effective AUC | MIA Worst-Case Effective AUC | Attack Advantage |
|---|---|---|---|
| OULAD – Gaussian Copula | 0.999 ± 0.000 | 0.512 ± 0.004 | 0.012 |
| OULAD – CTGAN | 0.962 ± 0.010 | 0.512 ± 0.003 | 0.012 |
| OULAD – TabDDPM | 0.826 ± 0.006 | 0.511 ± 0.004 | 0.011 |
| ASSISTments – Gaussian Copula | 0.950 ± 0.005 | 0.518 ± 0.010 | 0.018 |
| ASSISTments – CTGAN | 0.959 ± 0.008 | 0.516 ± 0.008 | 0.016 |
| ASSISTments – TabDDPM | 0.671 ± 0.014 | 0.527 ± 0.007 | 0.027 |

*Note: lower is better (ideal = 0.50). C2ST effective AUC = max(AUC, 1−AUC). MIA worst-case effective AUC = max over three attacker models (LR, RF, XGBoost) of max(AUC, 1−AUC). Attack advantage = worst-case effective AUC − 0.50.*

Across both datasets and all three synthesizers, worst-case membership-inference effective AUC remains close to 0.50, indicating near-chance-level membership prediction under the evaluated threat model (kNN distance features with three attacker classifiers; see Section III-D). On OULAD, worst-case effective AUCs range from 0.511 (TabDDPM) to 0.512 (Gaussian Copula and CTGAN), with attack advantages of approximately 1 pp. On ASSISTments, values are marginally higher, ranging from 0.516 (CTGAN) to 0.527 (TabDDPM), and the largest observed attack advantage is 2.7 pp.

The per-attacker breakdown provides additional insight (Figures 7–9). On OULAD, no individual attacker exceeds effective AUC = 0.512 for any synthesizer. On ASSISTments, the per-attacker variation is slightly larger (e.g., TabDDPM: LR = 0.514, RF = 0.517, XGBoost = 0.521), but all remain within 3 pp of chance. The absence of a systematically dominant attacker further supports the conclusion that privacy risk under this threat model is negligible for all three synthesizers.

TabDDPM on ASSISTments exhibits the highest attack advantage (0.027), possibly reflecting the smaller dataset (less anonymity mass) and TabDDPM's closer distributional modeling. Nevertheless, 2.7 pp remains well below thresholds associated with meaningful re-identification risk [13].

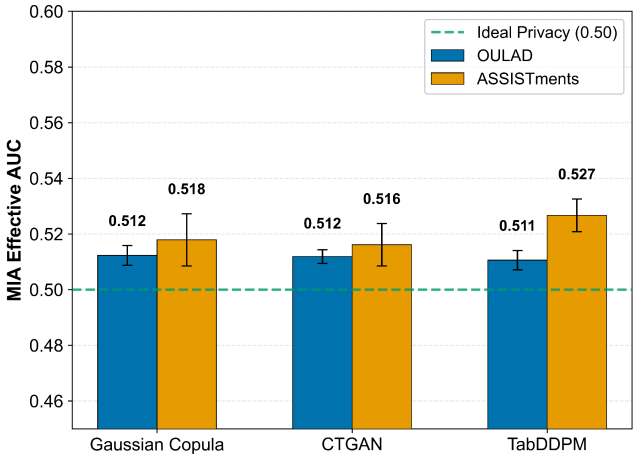


**Figure 7.** *MIA's worst-case effective AUC by dataset. Mean across seeds (n = 5). Error bars: ±1 SD. Dashed green line: ideal privacy (0.50).*

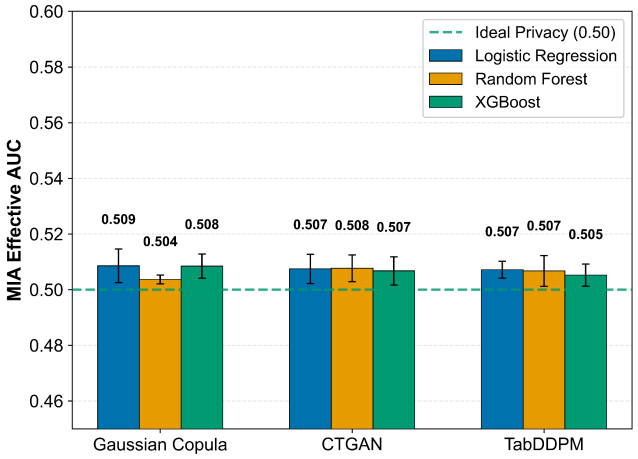


**Figure 8.** *OULAD MIA effective AUC broken down by attacker model (LR, RF, XGBoost). Mean across seeds (n = 5). Error bars: ±1 SD.*

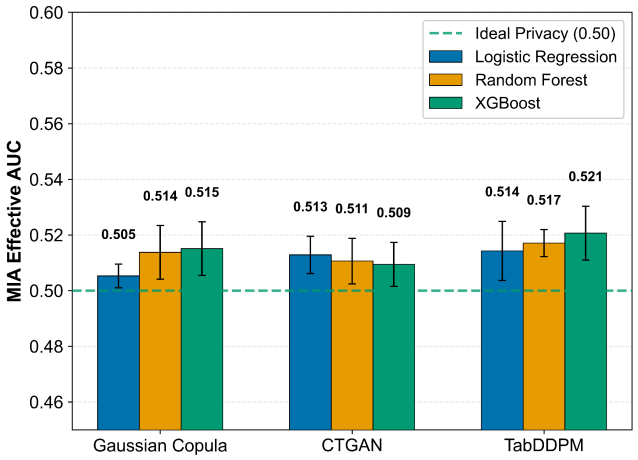


**Figure 9.** *ASSISTments MIA effective AUC broken down by attacker model (LR, RF, XGBoost). Mean across seeds (n = 5). Error bars: ±1 SD.*

### E. Feature-Importance Preservation (SHAP)

Table VI reports Spearman rank correlations (ρ) between SHAP-derived feature-importance rankings from TSTR

(synthetic) models and the TRTR (real-data) baseline, computed via TreeExplainer on Random Forest classifiers/regressors as described in Section III-D.

**TABLE VI: SHAP Feature-Importance Rank Correlations (Spearman ρ, averaged across 5 seeds)**

| Dataset | Task | n_features | Gaussian Copula | CTGAN | TabDDPM |
|---|---|---|---|---|---|
| OULAD | Classification | 60 | 0.770*** | 0.804*** | **0.846***** |
| OULAD | Regression | 60 | 0.707*** | 0.719*** | **0.748***** |
| ASSISTments | Classification | 5 | 0.975* | 0.950* | **1.000***** |
| ASSISTments | Regression | 5 | 0.975* | 0.950* | **1.000***** |

*Note: *** indicates $p < 0.001$; * indicates $p < 0.05$. Bold indicates the highest ρ per row. n_features is the dimensionality of the one-hot-expanded feature space. ASSISTments has only 5 columns (4 features + 1 target-derived), 3 of which have near-zero SHAP importance, yielding near-degenerate rankings. ASSISTments ρ values of 0.950 and 0.975 reflect occasional single-feature reorderings across seeds.*

**OULAD (60-dimensional feature space).** On OULAD classification, TabDDPM best preserves the real-data feature-importance ranking ($\rho = 0.846$, $p < 0.001$), followed by CTGAN ($\rho = 0.804$) and Gaussian Copula ($\rho = 0.770$). This ordering mirrors the downstream classification utility ranking. On OULAD regression, TabDDPM again achieves the highest rank correlation ($\rho = 0.748$), followed by CTGAN ($\rho = 0.719$) and Gaussian Copula ($\rho = 0.707$). All correlations are strong and statistically significant ($p < 0.001$).

Correlations in the 0.71–0.85 range leave room for subtle reordering of mid-tier features, but the top-ranked features (e.g., total_vle_clicks and is_unregistered) are consistently identified as most important regardless of data source, supporting the use of synthetic data for exploratory feature-level analysis.

Figures 14–15 complement these quantitative correlations with SHAP beeswarm plots comparing TRTR and TSTR (TabDDPM) models on OULAD classification; TabDDPM is shown as it achieves the highest rank correlation ($\rho = 0.846$) and thus provides the most informative directional comparison, aggregated across all 5 seeds ($n = 250$ test samples). Beyond rank preservation, the beeswarm visualization (top 10 features) confirms that TabDDPM preserves the directionality of feature effects: higher total_vle_clicks consistently pushes predictions toward non-dropout under both real-data (Figure 14) and synthetic (Figure 15) training, while is_unregistered = 1 drives predictions toward dropout in both models. This directional alignment is critical for educational stakeholders who rely on knowing how features influence outcomes when designing interventions.

**ASSISTments (5-dimensional feature space).** With only 5 features, of which 3 (*hint_rate*, *avg_attempts*, and the target itself) have near-zero SHAP importance, the feature-importance ranking is effectively determined by 2 features (*unique_skills* and *avg_response_time*). TabDDPM is the only synthesizer that consistently achieves $\rho = 1.000$ across all seeds for both classification and regression. Gaussian Copula ($\rho = 0.975$) and CTGAN ($\rho = 0.950$) achieve near-perfect correlations on average but exhibit occasional single-feature reorderings in individual seeds ($\rho = 0.875$ in 1–2 of 5 seeds). These near-perfect correlations are predominantly a ceiling effect of low dimensionality rather than evidence of exceptional synthesis quality, though TabDDPM's perfect consistency does provide a marginal discriminative signal.

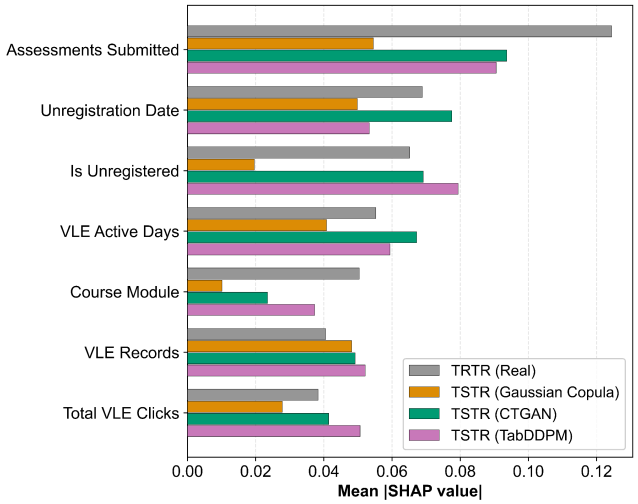


**Figure *10.*** *OULAD SHAP feature importance: Classification (top 7 features by TRTR importance). Mean |SHAP| values aggregated across seeds (n = 5). One-hot encoded categories are summed to their parent feature.*

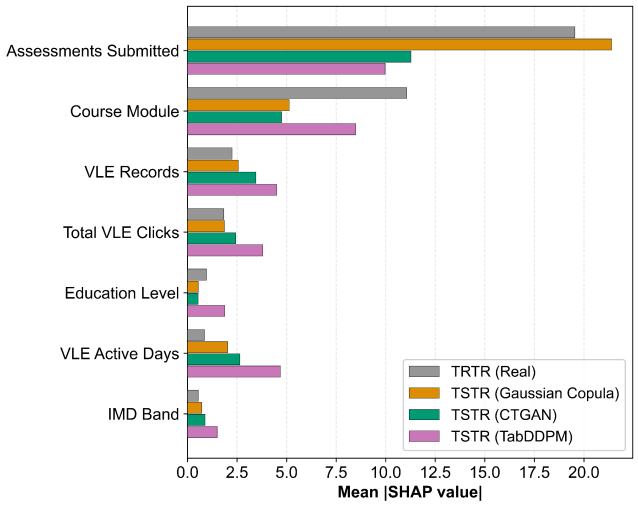


**Figure *11.*** *OULAD SHAP feature importance — Regression (top 7 features by TRTR importance). Mean |SHAP| values aggregated across seeds (n = 5). One-hot encoded categories are summed back to their parent feature.*

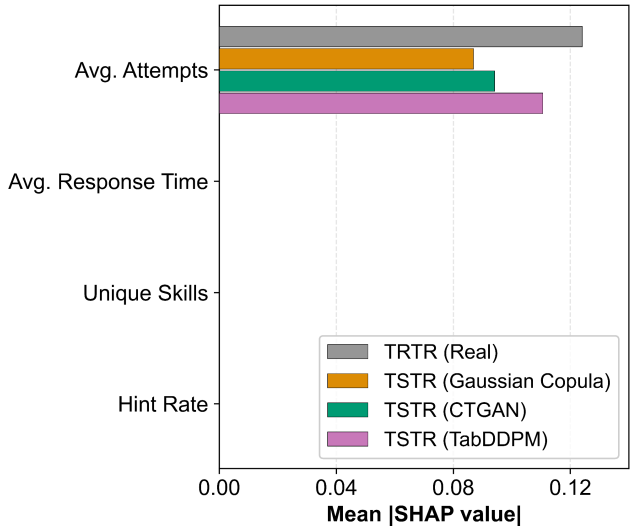


**Figure *12.*** *ASSISTments SHAP feature importance — Classification. Mean |SHAP| values aggregated across seeds (n = 5). All four behavioural features are shown.*

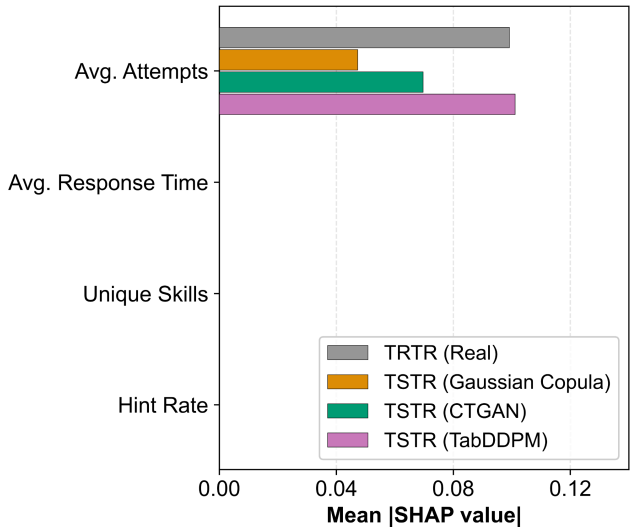


**Figure *13.*** *ASSISTments SHAP feature importance — Regression. Mean |SHAP| values aggregated across seeds (n = 5). All four behavioural features are shown.*

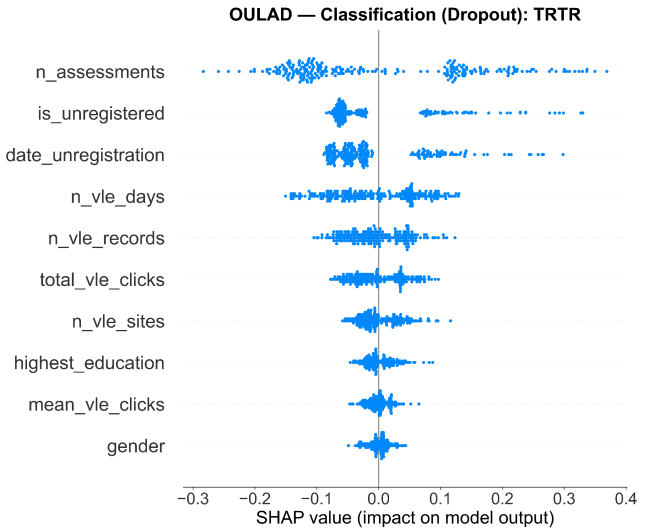


**Figure *14.*** *SHAP beeswarm plot for OULAD classification (TRTR). Aggregated across seeds (n = 5, 250 test samples). Top 10 features by mean |SHAP| value. Each dot represents one test-set student; color indicates feature value (red = high, blue = low); horizontal position indicates SHAP contribution to prediction.*

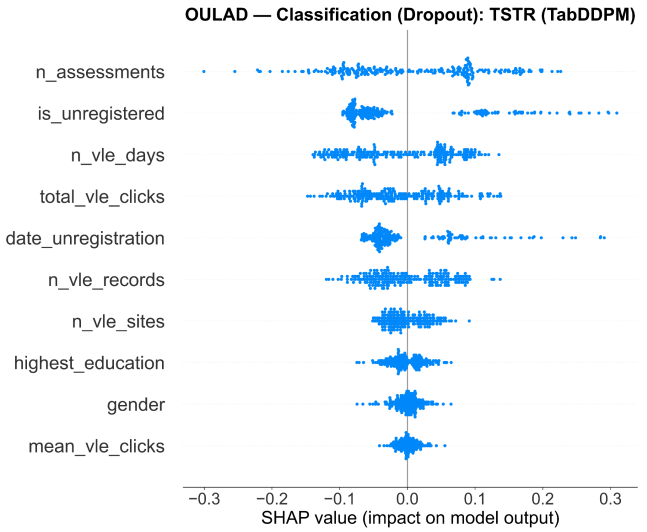


**Figure *15.*** *SHAP beeswarm plot for OULAD classification (TSTR, TabDDPM). Aggregated across seeds (n = 5, 250 test samples). Top 10 features by mean |SHAP| value. Directional alignment with Figure 14 confirms that TabDDPM preserves feature-effect structure beyond rank ordering.*

### F. Multi-Objective Trade-offs

Figures 16–17 present normalized heatmaps summarizing all five evaluation axes for each synthesizer, facilitating multi-objective method selection. Scores are normalized to a 0–100 scale, where higher values indicate better performance on each axis (for metrics where lower is better, specifically C2ST, MIA, and MAE, the normalization inverts the scale so that values closer to the ideal map to higher scores). Note: the heatmap columns cover fidelity, realism, privacy, and utility (classification + regression); SHAP rank correlations are reported separately in Table VI

because ρ values are not directly commensurable with the percentage-scale axes.

The heatmaps visually confirm three key patterns:

**1) Fidelity–utility decoupling.** CTGAN achieves the highest fidelity scores on both datasets, yet not the highest utility. On ASSISTments, this is particularly stark: CTGAN's normalized classification utility appears high (99.7%) because all methods cluster near chance (raw AUC = 0.498). This artifact of normalization can be misleading; practitioners should always consult raw metrics (Tables II–III) alongside normalized heatmaps.

**2) Realism deficit.** The Realism column shows that all synthesizers struggle with C2ST on OULAD (scores ≤ 34.0 on the normalized scale), whereas TabDDPM achieves markedly better realism on the simpler ASSISTments feature space.

**3) Universal privacy safety.** The Privacy column is uniformly high (≥ 96.5%) across all cells, confirming that membership-inference risk is negligible regardless of synthesizer choice.

TabDDPM emerges as the most balanced synthesizer across axes: it achieves the best or near-best scores on utility, realism, privacy, and SHAP preservation, with the only trade-off being lower distributional fidelity relative to CTGAN.

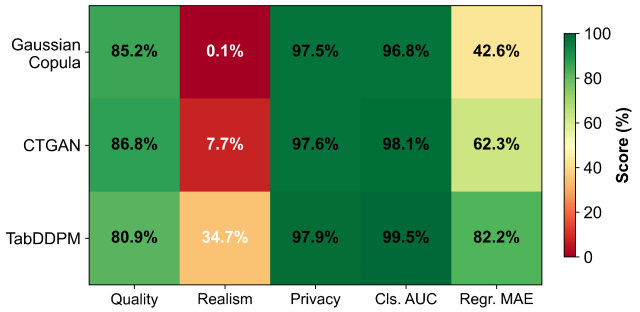


**Figure** *16. OULAD multi-objective heatmap. Normalized scores (0–100) across five evaluation axes, aggregated across seeds (n = 5). Higher (greener) is better.*

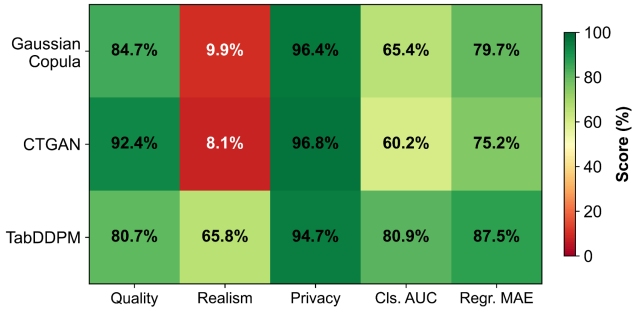


**Figure** *17. ASSISTments multi-objective heatmap. Normalized scores (0–100) across five evaluation axes, aggregated across seeds (n = 5). Higher (greener) is better.*

### G. Computational Efficiency

Table VII reports synthesis-only wall-clock time (training + sampling) per dataset–synthesizer combination.

**TABLE VII: Computational Efficiency (Synthesis Time, mean ± SD across 5 seeds)**

| Dataset – Synthesizer | Synthesis Time (s) | Synthesis Time (min) | Relative to Gaussian Copula |
|---|---|---|---|
| OULAD – Gaussian Copula | 17.8 ± 3.2 | 0.3 ± 0.1 | 1× |
| OULAD – CTGAN | 1,649.2 ± 316.9 | 27.5 ± 5.3 | 93× |
| OULAD – TabDDPM | 1,916.0 ± 416.7 | 31.9 ± 6.9 | 108× |
| ASSISTments – Gaussian Copula | 1.5 ± 0.6 | 0.03 ± 0.01 | 1× |
| ASSISTments – CTGAN | 158.5 ± 32.2 | 2.6 ± 0.5 | 105× |
| ASSISTments – TabDDPM | 358.4 ± 90.8 | 6.0 ± 1.5 | 238× |

Note: Synthesis time includes model training and synthetic data generation. All experiments run on a 12-core CPU with 15.7 GB RAM (no GPU; Windows 10, Python 3.11). Total end-to-end pipeline time (including all evaluation steps) is substantially longer.

Gaussian Copula completes in seconds (93–238× faster than the neural methods). On OULAD, CTGAN, and TabDDPM require ≈28–32 min with differing cost profiles: CTGAN is training-dominated (300 adversarial epochs), while TabDDPM splits cost between training (1,200 iterations) and iterative denoising. On ASSISTments, TabDDPM's relative overhead is more pronounced (238× vs. 108× on OULAD), reflecting fixed per-step diffusion cost. Both neural synthesizers exhibit high timing variance across seeds (CTGAN SD = 316.9s; TabDDPM SD = 416.7s on OULAD), likely reflecting convergence dynamics and sampling variability, respectively.

## V. DISCUSSION

### A. Principal Insights

Five cross-cutting insights emerge from the unified five-axis evaluation.

**1) Diffusion-based synthesis delivers the strongest downstream utility.** TabDDPM achieves the smallest synthetic-to-real gap on both datasets and both tasks (Tables II–III), with all pairwise OULAD differences significant after Bonferroni correction ($p < 0.001$). On ASSISTments, the per-learner decomposition is instructive: TabDDPM's LR TSTR score exactly matches the TRTR LR baseline, indicating that the linear decision boundary is fully preserved while the nonlinear structure captured by RF degrades. This suggests that diffusion-based synthesis preserves conditional class structure more faithfully than marginal-level methods, even when overall utility still falls short of the real-data baseline.

**2) Membership-inference privacy risk is negligible under the evaluated threat model.** All worst-case effective MIA AUCs remain within 2.7 pp of chance (Table

V), with no individual attacker systematically dominant. The slightly elevated advantage for TabDDPM on ASSISTments likely reflects the smaller dataset rather than a fundamental vulnerability. These empirical bounds are reassuring but must not be conflated with formal privacy guarantees (see Section V-C).

**3) Distributional fidelity does not reliably predict downstream utility.** CTGAN leads on SDMetrics and C2ST, while TabDDPM leads on TSTR (Tables II–IV), and high C2ST scores on OULAD coexist with strong classification utility across all synthesizers. This persistent decoupling suggests that marginal-level distributional agreement and joint-distributional realism do not index the conditional structure (class boundaries, regression surfaces) that drives downstream performance. Synthesizer selection should therefore be driven by task-relevant evaluation (TSTR), not by fidelity or realism metrics alone.

**4) SHAP-based feature-importance analysis reveals that TabDDPM best preserves explanatory structure on the higher-dimensional benchmark.** On OULAD, the synthesizer ordering for SHAP rank correlation (Table VI) exactly mirrors the utility ordering (Tables II–III), suggesting that preserving downstream utility also preserves the explanatory structure that drives predictions. On ASSISTments, the 5-feature space creates a ceiling effect that limits discriminative power among synthesizers. For educational applications, this alignment is particularly consequential: stakeholders need to trust that models trained on synthetic data identify the same drivers of student outcomes (e.g., VLE engagement, registration status) as models trained on real data. The beeswarm comparison (Figures 14–15) visually confirms this directional alignment for TabDDPM on OULAD.

**5) Dataset characteristics modulate synthesis difficulty.** The two benchmarks produce markedly different profiles: on OULAD, all synthesizers achieve useful classification transfer but are highly distinguishable under C2ST, while on ASSISTments, classification utility collapses for all but TabDDPM despite an ostensibly simpler feature space. These divergent patterns likely stem from interactions among dataset scale, feature dimensionality, categorical complexity, and target construction. The implication is that synthesizer rankings are not universal: evaluation on the target domain is essential.

### B. Limitations

Several factors constrain the generalizability of these findings:

**Scope of evaluation.** Only tabular student-level representations are evaluated; sequential, temporal, textual, and multimodal educational data require different synthesis approaches.

**Downstream model selection.** Utility is assessed with two learners per task (RF + LR for classification; RF + Ridge for regression). Other architectures (deep networks, gradient-boosted ensembles) may exhibit different sensitivity to synthetic data quality.

**Privacy threat model.** The MIA evaluation covers three attacker classifiers with kNN distance features, but does not exhaust all possible attacks (e.g., shadow-model or auxiliary-data attacks). Reported bounds should be interpreted as specific to this threat model, not as guarantees against all adversaries.

**Hyperparameter sensitivity.** Synthesizer configurations reflect controlled experimental defaults rather than exhaustive tuning. Rankings could shift under alternative settings, though TabDDPM's consistency across five seeds mitigates this concern.

**SHAP analysis scope.** SHAP preservation uses test-set subsamples (200 for OULAD; 500 for ASSISTments), sized inversely to feature dimensionality because TreeExplainer cost scales with the one-hot feature count (60 vs. 5), which may introduce sampling variance. ASSISTments' 5-feature space yields near-degenerate rankings with limited discriminative power.

**Single hardware configuration.** Timing reflects one machine (12-core CPU, no GPU). Absolute runtimes and relative rankings may shift with different hardware.

**Fairness.** Synthesizer-induced bias amplification or attenuation across demographic subgroups was not measured; fairness auditing remains an important open direction (Section V-C).

### C. Future Directions

Five extensions are especially actionable:

**1) Formal privacy mechanisms.** Integrating differential privacy (e.g., DP-SGD for CTGAN; DP noise for copula methods) and sweeping $\varepsilon$ budgets would map utility–privacy frontiers, which is essential for deployments requiring regulatory compliance.

**2) Fairness auditing.** Synthesizers can replicate, attenuate, or amplify demographic biases [21]. Systematic subgroup audits (e.g., by gender, age band, IMD band) would determine whether synthesis introduces or mitigates bias.

**3) Sequential and temporal data.** Extending synthesis to clickstream sequences and temporal interaction logs would address a broader range of learning analytics use cases requiring models of temporal dependencies and variable-length sequences.

**4) Cross-institution and federated synthesis.** Evaluating synthesis under distributed governance constraints would test federated protocols relevant to multi-site educational research consortia.

**5) Expanded explainability methods.** Extending beyond global SHAP rankings to local per-student explanations and alternative attribution methods (permutation importance, LIME) would assess the robustness of preservation findings.

## VI. PRACTICAL GUIDANCE

Synthesizer choice depends on the constraints and objectives of the intended release. We organize decision criteria by primary constraint:

**1) Maximizing downstream task utility.** Select based on TSTR performance on a held-out real test split matched to the intended task. TabDDPM consistently delivers the strongest task transfer across both datasets and both tasks (Tables II–III), and is the recommended default for releases intended to support predictive modeling (e.g., dropout classifiers, grade predictors), provided computational budget permits (see criterion 3).

**2) Preserving feature-importance structure for interpretable analysis.** When synthetic data will inform intervention design (e.g., identifying which engagement metrics drive dropout risk), validate SHAP rank correlation between TSTR and TRTR models before release. TabDDPM provides the strongest preservation on both tasks (Table VI), with directional feature-effect alignment visually confirmed via beeswarm comparison (Figures 14–15), consistent with its utility advantage.

**3) Operating under strict computational budget constraints.** Gaussian Copula completes synthesis in seconds (Table VII) and is appropriate for rapid prototyping or resource-constrained settings, accepting larger utility gaps (Tables II–III). CTGAN requires similar wall-clock time to TabDDPM (≈28–32 min on OULAD) but achieves the highest distributional fidelity (Table IV), making it a middle-ground option when both efficiency and marginal-level fidelity matter.

**4) Balancing multiple objectives.** The normalized heatmaps (Figures 16–17) summarize trade-offs across all five axes. TabDDPM achieves the best or near-best scores across utility, realism, privacy, and SHAP preservation; its only trade-off is lower fidelity relative to CTGAN. Because all three synthesizers exhibit near-chance MIA risk (Table V), privacy is not a differentiating factor, making utility and explainability the decisive criteria.

**Critical caveat.** As demonstrated in Section IV-F, distributional fidelity does not reliably predict downstream utility: CTGAN leads on SDMetrics yet trails TabDDPM on TSTR. Synthesizer selection should therefore be driven by task-relevant evaluation (TSTR), not by fidelity or realism metrics alone. As a practical guideline, releases should define explicit acceptance thresholds (e.g., TSTR AUC within 2 pp of TRTR; MIA effective AUC $\leq 0.55$; SHAP $\rho \geq 0.75$) and follow the evaluation protocol described in Section III-D.

## VII. CONCLUSION

This study compared three tabular synthesis paradigms (Gaussian Copula, CTGAN, and TabDDPM) on two educational benchmarks (OULAD and ASSISTments) under a unified five-axis evaluation protocol covering fidelity, utility, realism, privacy, and explainability, replicated across five random seeds with full statistical inference.

Four principal findings emerge. First, diffusion-based synthesis delivers the strongest downstream utility: TabDDPM consistently achieves the smallest synthetic-to-real gap on both datasets and both tasks, with all pairwise differences on OULAD statistically significant after Bonferroni correction (Tables II–III). Second, membership-inference privacy risk is negligible under the evaluated threat model, with worst-case attack advantages never exceeding 2.7 pp across all synthesizers (Table V); however, these empirical bounds do not constitute formal privacy guarantees. Third, distributional fidelity does not reliably predict downstream utility: CTGAN leads on SDMetrics yet trails TabDDPM on TSTR, underscoring that synthesizer selection should be driven by task-relevant evaluation rather than fidelity metrics alone (Tables II–IV). Fourth, SHAP-based feature-importance preservation mirrors the utility ranking, with TabDDPM best preserving the real-data explanatory structure on OULAD — confirmed both quantitatively (Table VI) and visually via SHAP beeswarm comparison (Figures 14–15) — supporting trustworthy use of synthetic data for interpretable educational analytics.

### Contributions to Learning Analytics and Synthetic Data Research

This work makes three methodological and empirical contributions:

**1) Unified five-axis evaluation framework.** Prior work has typically assessed fidelity and utility in isolation. This study demonstrates that fidelity, utility, realism, privacy, and explainability are complementary and can produce divergent synthesizer rankings, enabling informed multi-objective selection.

**2) First diffusion-model evaluation on educational benchmarks.** TabDDPM [11] has not previously been benchmarked on learning analytics datasets. This study demonstrates that diffusion models deliver the strongest utility and feature-importance preservation on educational data, extending findings from general tabular benchmarks [8].

**3) Empirical privacy benchmarking with cross-seed replication.** The MIA evaluation employs a conservative threat model (kNN distance features; worst-case over three attackers) replicated across five seeds, providing the most rigorous empirical privacy assessment to date for tabular synthesis on educational data [8].

Future work should prioritize formal privacy integration, fairness auditing, and extension to sequential educational data (see Section V-C).

## DATA AVAILABILITY

The complete experimental pipeline and configurations are available in the Synthla-Edu V2 repository: https://github.com/divineiloh/synthla-edu-v2. The datasets (OULAD [23]; ASSISTments [24]) are accessible under their respective licenses.

## REFERENCES

[1] G. Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013.

[2] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *British Journal of Educational Technology*, vol. 45, no. 3, pp. 438–450, 2014.

[3] H. Drachsler and W. Greller, "Privacy and analytics: It's a DELICATE issue: A checklist for trusted learning analytics," in *Proc. 6th Int. Conf. Learning Analytics & Knowledge (LAK'16)*, 2016, pp. 89–98.

[4] P. Prinsloo and S. Slade, "An evaluation of policy frameworks for addressing ethical considerations in learning analytics," in *Proc. 7th Int. Learning Analytics & Knowledge Conf.*, 2017, pp. 470–474.

[5] U.S. Congress, "20 U.S.C. § 1232g: Family educational and privacy rights (FERPA)," *United States Code*, Office of the Law Revision Counsel.

[6] Regulation (EU) 2016/679 (General Data Protection Regulation), *Official Journal* L 119, 4 May 2016 (EUR-Lex).

[7] J. Jordon et al., "Synthetic data: What, why and how?," arXiv preprint arXiv:2205.03257, 2022.

[8] D. Iloh, "Generative private synthetic student data for learning analytics: An empirical study," *IEEE Access*, vol. 12, pp. 154091–154106, 2024.

[9] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *Proc. IEEE Int. Conf. Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410.

[10] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[11] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "TabDDPM: Modelling tabular data with diffusion models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 17564–17579.

[12] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," in *Int. Conf. Learning Representations (ICLR)*, 2017.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symposium on Security and Privacy*, 2017, pp. 3–18.

[14] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[15] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloquium on Automata, Languages and Programming (ICALP)*, 2006, pp. 1–12.

[16] K. Bonawitz et al., "Towards federated learning at scale: System design," in *Proc. Machine Learning and Systems (MLSys)*, 2019, pp. 374–388.

[17] I. Goodfellow et al., "Generative adversarial networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[18] J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *Int. Conf. Learning Representations (ICLR)*, 2019.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[20] R. S. J. d. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*, J. A. Larusson and B. White, Eds. Springer, 2014, pp. 61–75.

[21] Z. Xu et al., "FairGAN: Fairness-aware generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 570–575.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[23] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Scientific Data*, vol. 4, article 170171, 2017.

[24] ASSISTmentsData, "2012–13 School Data with Affect," ASSISTmentsData (Google Sites). [Online]. Available: https://sites.google.com/site/assistmentsdata/datasets/2012-13-school-data-with-affect. Accessed: Feb. 3, 2026.

[25] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, vol. 57. Chapman and Hall/CRC, 2018 (reprint).

**DIVINE ILOH** received the B.S. degree in Public Administration from the University of Nigeria and the M.S. degree in Business Information Systems and Analytics from the University of Arkansas at Little Rock. He is a researcher and data analyst specializing in artificial intelligence, machine learning, educational technology, and data-driven systems.
He is the co-founder of SabiScholar, an AI-driven EdTech platform that delivers personalized learning to students in emerging markets. He has contributed to peer-reviewed research in AI, cybersecurity, and IoT applications, with an emphasis on agentic systems, anomaly detection, and context-aware automation. His work focuses on developing intelligent systems and AI-powered tools that enhance learning, decision-making, and digital resilience. His current research interests include AI for social impact, intelligent learning systems, IoT-based education solutions, and the application of machine learning in real-world environments.
Mr. Iloh is an active member of IEEE and supports interdisciplinary efforts that bridge AI, education, and equitable access to technology.

**GRACE OKU** is a digital experience specialist whose interdisciplinary work integrates technology, early childhood literacy, and mental health promotion. Currently, she leads messaging strategy and stakeholder engagement for Dreambook and Chronicle Creations, developing empathetic, parent-centered communication frameworks and advocating for inclusive design principles in digital literacy tools.
Her professional expertise encompasses user-centered design, web development, workflow optimization, and strategic program evaluation. She specializes in distilling complex academic research and technical information into accessible, actionable communications for diverse audiences, with a particular focus on educational technology and public-facing educational content. This translation work ensures that research insights drive practical implementation and community impact.
Grace's research interests focus on participatory design methodologies, youth mental health interventions, and the development of scalable, community-centered technologies. She is particularly committed to exploring how digital platforms can be designed to support emotional well-being and resilience in children and adolescents from diverse backgrounds.

**SHAOZHI JIANG**, Senior Member, IEEE, is a Software Architect and Technical Leader specializing in Artificial Intelligence, Cyber Security, and Advanced Computing.
With over two decades of experience in cloud-based infrastructure, software supply chains, and telecommunications, his work focuses on developing AI-driven solutions for critical national priorities, including resilient cloud computing and secure infrastructure. He has pioneered innovative platforms, such as the Aegis Framework for Secure Software Supply Chains, which leverages AI/ML for automated vulnerability detection and automated code remediation, directly supporting national security and economic stability. He has served as a Peer Reviewer for the IEEE International Conference on AI and Data Analytics (ICAD) and as an Invited Technical Judge and Mentor for multiple hackathons and AI-powered social good initiatives.
His research interests and primary contributions center on the intersection of privacy-preserving AI, explainability (XAI), and operational resilience. In this paper, he pioneers the integration of SHAP-based feature-importance analysis to quantitatively prove semantic preservation in synthetic models. His work emphasizes deploying trustworthy, secure, and intelligent systems, ensuring that AI translates complex data into actionable insights while remaining transparent, robust, and explicitly audited for fairness in critical sectors.

**RASHMI CHOUDHARY** is a Staff Data Scientist at Zum Services, Redwood City, CA, USA. She received her undergraduate degree in Computer Science in India and her M.S. in Data Science from the University of Illinois Chicago in 2017. She has over seven years of experience in applied machine learning, with six years focused on large-scale behavioral and geospatial data analysis. Her work emphasizes the design and rigorous evaluation of machine learning systems under real-world constraints, including privacy, robustness, and interpretability. Her research interests include synthetic data generation, learning analytics, and trustworthy AI for high-stakes decision-making systems.

**ROHIT KAPA** is a Vice President, Data Scientist at Prudential Financial, where he leads Data Science and AI initiatives for the life insurance business. His work focuses on the development and operational integration of machine learning systems, including emerging applications of large language models and generative AI in regulated financial environments. Prior to joining Prudential, he worked in distributed computing and big data engineering at Tata Consultancy Services. His research interests include applied machine learning, large language model systems, decision optimization, and scalable data platforms. He holds an M.S. in Business Analytics from the University of Connecticut and a B.Tech. in Computer Science and Engineering.

**NITESH CHILAKALA** received his M.S. in Management Information Systems from the University of Houston – Clear Lake. Currently serving as a Senior Manager in Data and AI, his research focuses on AI/ML data pipeline architecture and responsible AI governance — exploring Apache NiFi as a governed ingestion backbone for machine learning workflows, with proof-of-concept work on embedding data lineage, policy enforcement, and auditability directly into NiFi flow configurations. In his role leading cloud platform modernization, he has driven the migration of enterprise NiFi deployments from EC2 and on-premises environments to containerized Kubernetes architectures on Amazon EKS, enabling scalable, low-latency data flows required by modern AI systems, and developed automated tooling for NiFi Parameter Context migration. He actively contributes optimization best practices to the Apache NiFi open-source community. With over 13 years of experience in cloud platform engineering and data architecture, he has delivered measurable impact across large-scale streaming and data platforms, achieving a 40% improvement in data processing efficiency and 50% gain in system reliability. He holds certifications as an AWS Solutions Architect Associate and SAFe Agile Practitioner.